



**POLSKA AKADEMIA NAUK
MIĘDZYKARODOWE CENTRUM BIOCYBERNETYKI
INTERNATIONAL CENTRE OF BIOCYBERNETICS
POLISH ACADEMY OF SCIENCES**

4 Ks. Trojdena Str., 02-109 Warsaw, Poland, phone: (00 4822) 5925973, fax: (00 4822) 5925998
Director: Prof. Piotr Ładyżyński Ph.D., D.Sc., e-mail: mcb@ibib.waw.pl

LECTURE NOTES OF THE ICB SEMINAR

The 10th International Seminar

STATISTICS AND CLINICAL PRACTICE

Warsaw, May 2016

**Edited by: L. Bobrowski
Z. Valenta
C. Enăchescu**

Warsaw 2016

The event is co-financed from the financial sources of the Polish Academy of Sciences

CONTENTS	3
PREFACE	5

ORAL PRESENTATIONS

M. Genin, C. Preda, A. Duhamel, C. Gower-Rousseau - Isotonic spatial scan statistics: application to the epidemiology of Crohn's disease in northern France.....	9
C. Preda, V. Vandewalle - Clustering categorical functional data.....	14
B. Alexe - Using a bivariate scan statistics for detecting disease clusters	19
C. Enachescu, D. Enachescu - IRIS biometrics: adaptive resonance networks for indexing and retrieving data.....	25
Z. Valenta, P. Ošťádal, D. Vondráková, M. Průcha, A. Kruger, M. Janotka - Survival and 30-days cerebral performance in patients with successful cardiopulmonary resuscitation following cardiac arrest: statistical inference in the presence of incomplete data	32
Ł. Mierzejewski, W. Niemiro, W. Rejchel, M. Zalewska - Infections caused by CLOSTRIDIUM DIFFICILE: feature selection via ORDINAL REGRESSION.....	37
M. Ćwiklińska-Jurkowska, A. Wolińska-Welcz - Dimensionality reduction and visualization of genomic data.....	46
A. Oniško - Modeling uncertain medical knowledge with Bayesian networks: engineering and applications.....	57
L. Bobrowski - Biclustering as extraction of collinear patterns.....	62
P. Munro - A framework for combining unsupervised and supervised learning procedures	68

POSTER SESSIONS

I. Chmiel, M. Górkiewicz - Combining two correlated variables into one factor: an application to obesity measurements.....	75
M. Ćwiklińska-Jurkowska - Ensembles of variables selection and combined classifiers for medical differentiation on the basis of genes expression data set.....	81
M. Górkiewicz, J.N Peña-Sánchez, I. Chmiel - Item response theory methods can support validity of 4cornersat scale for career satisfaction of physicians	89
A. Korzyńska, L. Roszkowiak, J. Zak, D. Pijanowska - Software framework for validation of segmentation results (VoS).....	95
M. Krętowska - Oblique survival trees and competing risks	98
T. Łukaszuk, A. Gyenesei, Leon Bobrowski - Application of the Relaxed Linear Separability method for the analysis of genomic data	101
P. Malinowski, W. Dąbrowski, B. Karolinczak - Comparison of support vector regression and classical transformed linear model in constructed wetland treatment context	106
K. Sałapa, T. Darocha, J. Majkowski, T. Sanak, P. Podsiadło, S. Kosiński, R. Drwiła - Assessing agreement between two measurement methods of core body temperature.....	110

The special session addressed to medical doctors
STATISTICAL STANDARDS IN MEDICAL RESEARCH PROJECTS
AND PUBLICATIONS

*The session organized by the Polish National Group
of the International Society for Clinical Biostatistics (ISCB)*

J.N. Peña-Sánchez, M. Górkiewicz - Scale of career satisfaction in medicine	117
M. Polak - Meta-analysis in biomedical research	117
U. Cwalina, D. Jankowska, A.J. Milewska, D. Citko, R. Milewski - The selection of a representative sample in medical research	118
K. Szafraniec - Multivariable analysis. The essentials.....	122
A. Wolińska-Welcz - Statistical inference in multivariate analysis of medical data	122

PREFACE

The tenth International Seminar "*Statistics and Clinical Practice*" was held between 15 -18 May 2016 in the Institute of Biocybernetics and Biomedical Engineering, Polish Academy of Sciences, Warsaw, Poland.

The Seminars "*Statistics and Clinical Practice*" are organised since 1994 in Warsaw in the framework of the International Centre of Biocybernetics (*ICB*). Up to now ten seminars were carried out in:

1994, 1996, 1998, 2000, 2002, 2005, 2008, 2011, 2014 and 2016 year.

The below topics were addressed during the tenth Seminar:

- prognostic models selection
- exploratory analysis of big biomedical data sets
- data mining methods in bioinformatics
- inclusion of genomic data in biostatistical modeling
- new models in survival analysis
- clinical epidemiology
- study designs for individualized medicine

The ICB Seminars "*Statistics and Clinical Practice*" have become a tradition as a meeting space for statisticians from various countries, dealing with problems related to medicine. Both scientific and didactic goals form a part of the Conference framework. The Seminar provides an opportunity for professional discussions among biostatisticians, while medical doctors are offered an opportunity to discuss statistical problems with leading experts in the field. The didactic session for medical doctors were organized by the Polish National Group of the International Society for Clinical Biostatistics (*ISBC*).

Leon Bobrowski

Warsaw, May 2016

ORAL PRESENTATIONS

ISOTONIC SPATIAL SCAN STATISTICS: APPLICATION TO THE EPIDEMIOLOGY OF CROHN'S DISEASE IN NORTHERN FRANCE

M. Genin¹, C. Preda², A. Duhamel¹, C. Gower-Rousseau³

¹ *Univ. Lille, EA 2694 – Santé publique: épidémiologie et qualité des soins, Lille, France*

² *Inria Lille-Nord-Europe – Equipe-Projet MODAL, Villeneuve d'Ascq, France*

³ *Lille University Hospital, Lille, France*

Abstract

We address the issue of cluster detection by means of spatial scan statistics methods. These quite recent methods provide high power for detecting significant spatial clusters. Firstly, we explain the principle of spatial scan statistics and the underlying mathematical model. Secondly, we perform an application of these methods in the search of spatial clusters of Crohn's disease in northern France, during the period from 1990 to 2011, using the data provided by EPIMAD registry. The application of spatial scan statistics allowed to highlight 8 significant spatial clusters: 4 clusters of high-incidence and 4 clusters of low-incidence.

Keywords: Crohn's disease, Cluster detection, Spatial Scan Statistics

I. INTRODUCTION

The detection of clusters of events is an area of statistics that is particularly extensive in recent decades. First, the scientific community has worked to develop methods in the one-dimensional framework (e.g. time) and, subsequently extended these methods to the multidimensional case, including two-dimensional (space). Among all the events cluster detection methods, three types of methods can be distinguished. The first concerns global tests that detect a global clustering, without locating any clusters. The second type corresponds to the focused tests that are used when an a priori knowledge allows to define a source point (date or spatial location) and test the aggregation of events around it. The last type includes the cluster detection tests that allow localization, without a priori knowledge, of clusters and test their statistical significance. It is in this latter type of methods that spatial scan statistics (SSS) are embedded. These methods, originally proposed by [1, 2], proved to be very powerful [3, 4] as part of the objective detection of events spatial clusters and test their statistical significance. In other words, these methods allow the detection of spatial areas where the probability of an event is abnormally high (or low). In medical research, SSS were applied to many fields such as oncology [5-7], cardiology [8, 9], infectious diseases [10, 11] or gastroenterology [12-14].

In this paper, we explain the principle of the SSS and present the application of these methods in the detection of spatial clusters of Crohn's Disease (CD). The paper is organized as follows. In Section 2, we present the principle of SSS and the underlying mathematical model. In section 3, we present the results of the application of SSS to CD epidemiology in northern France, through the data from the EPIMAD Registry. In Section 4, we propose a discussion of the results. Section 5 will conclude the paper.

II. METHODOLOGY

Spatial scan statistics

Consider a studied region modeled by a random process $X = \{X_d, d \in D\}$ indexed by a fixed discrete spatial set $D = \{d_1, d_2, \dots, d_n\} \subset \mathbb{R}^2$. For instance, each site $d_i, 1 \leq i \leq n$, may correspond to the centroid of a municipality and X_{d_i} corresponds to the observed number of events. In case of count data, the random variables X_{d_i} are often considered as Poisson distributed with parameter $p\mu(d_i)$ where p is the probability of apparition of an event over the whole studied area and $\mu(d_i)$ is the underlying population related to the location d_i . Thus, for each $A \subset D$ we can define $\mu(A) = \sum_{d_i \in A} \mu(d_i)$ and $X_A = \sum_{d_i \in A} X_{d_i}$, X_A being Poisson distributed with parameter $p\mu(A)$.

The aim of the the SSS is to test the null hypothesis H_0 of absence of cluster against an alternative hypothesis H_1 supporting the existence of at least one area $Z \subset D$ in which the probability of apparition of an event, p , is higher than in the rest of the studied area, q . The two hypotheses are summarized in the following equation:

$$\begin{cases} H_0 : p = q, X_A \sim P(p\mu(A)), \forall A \subset D \\ H_1 : p > q, X_A \sim P(p\mu(A \cap Z) + q(A \cap Z^c)), \forall A \subset D. \end{cases}$$

The SSS are defined by a two-step process: cluster detection and statistical inference. During the detection step, the method uses a circular scanning window Z of variable size which moves across the studied area, using as centre the centroid of each spatial unit. At each position, the radius of the circular window varies from 0 to a maximum so that the window never contains more than 50% of the total underlying population. Thus, the detection step yields to an important collection \mathcal{Z} of potential clusters. To each potential cluster is associated a likelihood ratio (LR) defined by

$$L(Z) = \frac{\sup_{\{p>q\}} L(Z, p, q)}{\sup_p L(p)},$$

where $L(Z, p, q)$ and $L(p)$ are respectively the likelihood functions under H_1 and H_0 . The potential cluster that maximizes the LR, \tilde{Z} , is called the Most Likely Cluster (MLC). The LR related to \tilde{Z} is the test statistic λ used to test H_0 against H_1 :

$$\lambda = \sup_{\{Z \in \mathcal{Z}\}} L(Z) = L(\tilde{Z}).$$

The detection step yields to the identification of the MLC and the test based on λ allows to test the significance of the MLC. However, the distribution of λ under H_0 has no analytical form. Thus, a Monte-Carlo hypothesis testing procedure is used [15].

Isotonic version of spatial scan statistics

Instead of considering the risk as constant in a cluster, the isotonic spatial scan statistics (ISSS) consider the risk as a decreasing function from the center of the cluster [16]. During the detection step, a likelihood function is calculated and it models the potential cluster using an isotonic regression function with successively decreasing risk with increasing

distance from the cluster center. Actually, for a given window of radius d , the risk is modeled as a non-increasing function $r(d)$ of the distance of the centroid with multiple locations where the function takes a step down.

The risk function is fitted by means of an isotonic regression and there is no a priori assumption about the number of steps. It is linked to the isotonic regression function chosen, among all possible non-increasing functions, thanks to maximum likelihood method.

Each significant isotonic cluster is assigned a global relative risk (RR) which is defined as the risk in the cluster compared to the rest of the studied area. Moreover, each step of a cluster is assigned a RR which is defined as the risk in the cluster step compared to the rest of studied area outside the cluster.

Application to the epidemiology of CD

This study was conducted in northern France, which has a surface of 24,862 km² including both rural industrial and urban regions. It has an average total population of 5,858,921 inhabitants, which corresponds to approximately 9.4% of the French population. Northern France is divided into four large administrative units: Nord, Pas-de-Calais, Somme and Seine-Maritime. These large administrative units can be further divided into 273 cantons (a French small administrative area). Canton centroid, as defined by the geographical center (longitude and latitude), was used for our statistical analysis.

Epidemiological data were extracted from the EPIMAD registry of inflammatory bowel disease in northern France. The methodology of EPIMAD registry has been previously described in details [17]. For the purpose of this study, we considered all CD incident cases from 1990 to 2011 (n=8,970). Each CD case was attributed to a canton using the zip code of residence of the patient.

ISSS adjusted for age and gender have been used to highlight high and/or low-incidence clusters of CD.

III. RESULTS

From 1990 to 2011, the mean crude annual CD incidence rate was $7.1/10^5$ inhabitants in the region covered by the EPIMAD registry.

The age and gender adjusted ISSS highlighted eight significant clusters of CD incidence, which the spatial locations and characteristics are described in Fig. 1. Among these clusters, four cluster of high-incidence of CD and four clusters of low-incidence of CD can be distinguished. In the clusters of high-incidence, the global RR varies from 1.27 to 1.46. The RR associated to the first step of each high-incidence cluster varies from 1.64 to 1.89. Among the low-incidence cluster, the global RR varies from 0.69 to 0.77 and the RR associated to the first step of each cluster varies from 0.14 to 0.65.

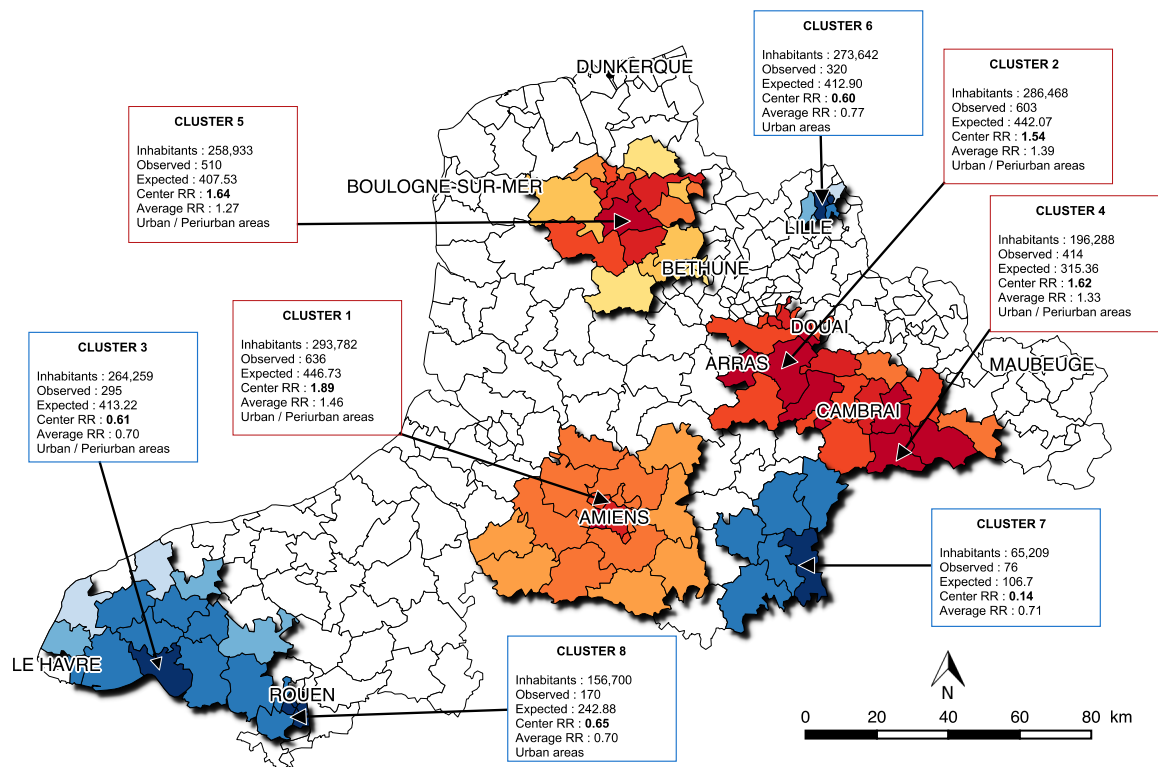


Fig. 1. Relative risks of Crohn’s disease (CD) in northern France during the period from 1990 to 2011. Spatial clusters detected by isotonic spatial scan statistics adjusted for age and gender.

IV. DISCUSSION

The application of ISSS to CD epidemiology yields to the highlighting of a strong geographical heterogeneity of CD incidence by means of the identification of 8 clusters using ISSS. Among these latter, 4 clusters of high-incidence have been identified, which are characterized by urban and peri-urban area. Simultaneous use of powerful methods and high number of CD cases from EPIMAD registry ensure the robustness of these ecological results. Moreover, these latter pave the way to the evaluation of environmental causes of the disease by crossing the cluster data with pollution data.

V. CONCLUSION

The SSS are powerful methods for cluster detection without pre-selection bias. Moreover, the isotonic version of SSS is more useful in an epidemiological point of view when considering the risk as a decreasing function from the cluster center. When seeking for the causes of a disease and dealing with large clusters, the ISSS provide an “epicenter” of the cluster that allows the physicians to focus their etiological researches.

Acknowledgements

The authors would like to thank the EPIMAD group for providing the high quality data related to CD incidence in northern France.

References

1. Kulldorff, M. and N. Nagarwalla, *Spatial disease clusters: detection and inference*. Statistics in Medicine, 1995. **14**(8): p. 799-810.

2. Kulldorff, M., *A spatial scan statistic*. Communications in Statistics - Theory and Methods, 1997. **26**(6): p. 1481-1496.
3. Kulldorff, M., T. Tango, and P.J. Park, *Power comparisons for disease clustering tests*. Computational Statistics & Data Analysis, 2003. **42**(4): p. 665-684.
4. Genin, M., *Discrete scan statistics and generalized likelihood ratio test*. Romanian Journal of Pure and Applied Mathematics, 2015. **60**(1): p. 83-92.
5. Amin, R., et al., *Epidemiologic mapping of Florida childhood cancer clusters*. Pediatr Blood Cancer, 2010. **54**(4): p. 511-8.
6. DeChello, L.M. and T.J. Sheehan, *Spatial analysis of colorectal cancer incidence and proportion of late-stage in Massachusetts residents: 1995-1998*. Int J Health Geogr, 2007. **6**: p. 20.
7. Klassen, A.C., M. Kulldorff, and F. Curriero, *Geographical clustering of prostate cancer grade and stage at diagnosis, before and after adjustment for risk factors*. Int J Health Geogr, 2005. **4**(1): p. 1.
8. Kuehl, K.S. and C.A. Loffredo, *A cluster of hypoplastic left heart malformation in Baltimore, Maryland*. Pediatr Cardiol, 2006. **27**(1): p. 25-31.
9. Pedigo, A., T. Aldrich, and A. Odoi, *Neighborhood disparities in stroke and myocardial infarction mortality: a GIS and spatial scan statistics approach*. BMC Public Health, 2011. **11**: p. 644.
10. Weisent, J., et al., *Detection of high risk campylobacteriosis clusters at three geographic levels*. Geospatial Health, 2011. **6**(1): p. 65-76.
11. Wu, S., et al., *Incidence analyses and space-time cluster detection of hepatitis C in fujian province of china from 2006 to 2010*. PLoS ONE, 2012. **7**(7): p. e40872.
12. Aamodt, G., et al., *Geographic distribution and ecological studies of inflammatory bowel disease in southeastern Norway in 1990-1993*. Inflamm Bowel Dis, 2008. **14**(7): p. 984-91.
13. Aamodt, G., S.O. Samuelsen, and A. Skrondal, *A simulation study of three methods for detecting disease clusters*. Int J Health Geogr, 2006. **5**: p. 15.
14. Genin, M., et al., *Space-time clusters of Crohn's disease in northern France*. Journal of Public Health, 2013. **21**(6): p. 497-504.
15. Dwass, M., *Modified randomization tests for nonparametric hypotheses*. The Annals of Mathematical Statistics, 1957. **28**(1): p. 181-187.
16. Kulldorff, M., *An isotonic spatial scan statistic for geographical disease surveillance*. Journal of the National Institute of Public Health, 1999. **48**(2): p. 94-101.
17. Gower-Rousseau, C., et al., *Incidence of inflammatory bowel disease in northern France (1988-1990)*. Gut, 1994. **35**(10): p. 1433-8.

CLUSTERING CATEGORICAL FUNCTIONAL DATA

C. Preda¹, V. Vandewalle²

¹*ISMMA Romanian Academy, Bucharest, Romania*

²*University of Lille 2, Lille, France*

Abstract

Categorical functional data represented by paths of a stochastic jump process with continuous time are considered for clustering. For Markov models we propose an EM algorithm to estimate a mixture of Markov processes. A simulation study as well as a real application on medical discharge letters will be presented.

Keywords: Categorical functional data, clustering, EM algorithm

I. INTRODUCTION

Let $S = \{s_1, \dots, s_m\}$, $m > 1$, be a set of m states and $X = \{X_t : t > 0\}$ be a S -valued family of categorical random variables. A path of X is a sequence of states s_{ij} and times points t_i of transitions from one state to another one : $\{(s_{i1}, t_1), (s_{i2}, t_2), \dots\}$, with $s_{ij} \in S$ and $t_i > 0$. We call the sample paths of the process X *categorical functional data*. The Figure 1 presents graphically an example of categorical functional data.

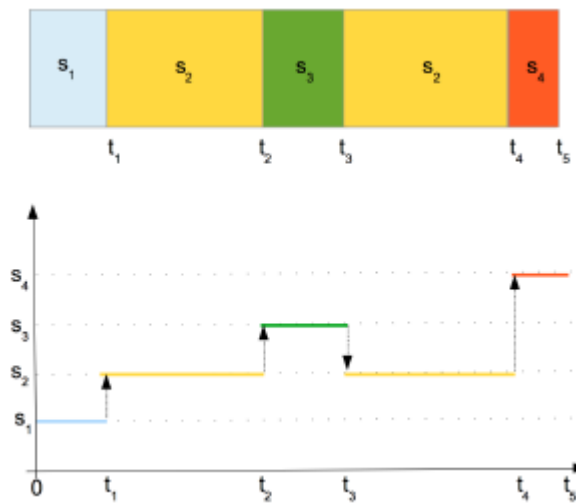


Fig1. Categorical functional data representation.

In this work we present a model-based methodology of clustering categorical functional data. Instead of the classical setting considering a fixed length of the paths of X , i.e. the process is observed over a fixed length of time T , $T > 0$, we consider that the process X has an absorbing state and thus, we allow sample paths of different lengths. In the Markovian framework, based on the likelihood function, we derive an EM algorithm for clustering categorical functional data. A simulation study and an application on clustering medical discharge letters according to their status of dictating, type-writing and delivery to the end-user (patient or medicine) are presented.

II. METHODOLOGY

It is supposed that the n paths come from K different Markov processes. Given that the path i comes from cluster k the probability density function of the path i is written $p(x_i; \theta_k)$ where θ_k are the parameters describing the process in cluster k .

Let first introduce some notations:

- $S = \{1, 2, \dots, m\}$ the state space, m being an absorbing state.
- n is the number of observed paths of X
- d_i is the length of the i -th path (in terms of state changes)
- e_{ijh} equals to 1 if the j -th state from the i -th path is h , 0 otherwise
- $e_{ij} = (e_{ij1}, e_{ij2}, \dots, e_{ijd_i})$ the binary coding of the j -th state for the path i
- s_{ij} is the time spent in the j -th state for the path i
- $x_i = (e_{i0}, s_{i1}, e_{i2}, s_{i2}, \dots, e_{id_i})$ is the data from the path i
- $x = (x_1, x_2, \dots, x_n)$ the whole dataset

By the chain rule we have

$$p(x_i; \theta_k) = p(e_{i0}; \theta_k) \prod_{j=1}^{d_i} p(s_{ij}, e_{ij} | e_{i0}, s_{i1}, e_{i1}, \dots, s_{i(j-1)}, e_{i(j-1)}; \theta_k)$$

where θ_k are the parameters describing the process in cluster k .

Four assumptions are made:

Assumption 1 : The distribution of (s_{ij}, e_{ij}) is independent of the past given $e_{i(j-1)}$.

Assumption 2 : The distribution of s_{ij} is independent of e_{ij} given $e_{i(j-1)}$.

Assumption 3 : The distribution of s_{ij} given $e_{i(j-1)}$ is an exponential one.

Assumption 4 : The distribution of the initial state does not depend on the cluster.

Consequently, $p(x_i; \theta_k)$ can be written as

$$p(x_i; \theta_k) = p(e_{i0}; \theta_k) \prod_{j=1}^{d_i} p(s_{ij}, e_{ij} | e_{i0}, s_{i1}, e_{i1}, \dots, s_{i(j-1)}, e_{i(j-1)}; \theta_k)$$

The assumptions 1, 2 and 3 allow to deal with a Markov process and assumption 4 is made to not involve directly the initial state in the clustering process. The assumptions 2 and 3 can be easily weakened, in this case we would work in a semi-Markov framework. Assumption 4 can also be easily weakened by making the initial state depend on the cluster, but perhaps it would give a too large weight on it in the clustering process.

The parameters θ_k are decomposed in two parts, on the one hand the parameters governing the transitions probabilities α_k and on the other hand the parameters λ_k governing the sojourn times which follow an exponential distributions. Thus,

- $\alpha_{khh'}$ is the transition probability from state h to h' in the cluster k
- α_k is the matrix $(\alpha_{khh'}, 1 \leq h, h' \leq m)$
- λ_{kh} is the parameter of the sojourn time in state h for the cluster k
- $\lambda_k = (\lambda_{k1}, \lambda_{k2}, \dots, \lambda_{k(m-1)})$
- $\theta_k = (\alpha_k, \lambda_k)$

Under assumptions 1- 4, simple calculations yield to:

$$p(\mathbf{x}_i; \theta_k) = p(\mathbf{e}_{i0}) \prod_{j=1}^{d_i} \prod_{h=1}^{m-1} (\lambda_{kh} e^{-\lambda_{kh} s_{ij}})^{e_{i(j-1)h}} \prod_{h'=1}^m \alpha_{khh'}^{e_{i(j-1)h} e_{ijh'}}$$

Since the class membership is unknown, the probability density function for the path i can be written as a mixture in the following way. Let $\pi = (\pi_1, \dots, \pi_K)$ be the prior weights for the K clusters and denote by $\theta = (\pi, \theta_1, \dots, \theta_K)$. Then,

$$p(\mathbf{x}_i; \theta) = \sum_{k=1}^K \pi_k p(\mathbf{x}_i; \theta_k).$$

The EM algorithm.

The aim is now to estimate the parameters of the model. This can be performed by maximum likelihood. Making the assumption that the paths are independent and identically distributed, the expression of the log-likelihood, denoted by $l(\theta; \mathbf{x})$ is

$$\ell(\theta; \mathbf{x}) = \sum_{i=1}^n \log \left(\sum_{k=1}^K \pi_k p(\mathbf{x}_i; \theta_k) \right).$$

The direct maximisation of $l(\theta; \mathbf{x})$ with respect to θ is hard to perform due to the logarithm of the sum. We propose to use the EM algorithm in order to perform the parameters estimation and thus, giving for each path the posterior probabilities to belong to some cluster.

III. RESULTS

We apply the clustering algorithm to the medical discharge letters. Data come from the Saint Philibert Hospital (Northern France). We have a set of 443 325 letters for which we know, at all moment its status (states):

1. the doctor is dictating the letter
2. the letter is "waiting" to be type-writing by an assistant (queue)
3. the letter is type-writing by the assistant
4. the letter is "waiting" for doctor validation (queue)

5. the letter is in validation process by the doctor
6. the letter is "waiting" to be affected to an assistant (queue)
7. the letter is treated by the assistant
8. the letter is sent to the patient (end).

A summary description of the data is presented in Table 2 and Figure 3.

Number of jumps (length of the path) :

Length	2	3	4	5	6	7	8
Frequence	336181	1118	2752	8157	23688	8541	62888

Number of transitions from one state to another state :

from \ to	1	2	3	4	5	6	7	8
1	0	93042	201	0	0	0	0	335306
2	0	0	90453	2849	32	0	0	317
3	0	0	0	100452	113	974	1	73
4	0	0	0	0	92351	6629	191	6694
5	0	0	0	0	0	76523	887	15353
6	0	0	0	0	0	0	81180	3184
7	0	0	0	0	0	0	0	82398

Table 2. Transitions between states

Quantile	Time (s)
0%	1.00
10%	47.00
20%	57.00
30%	69.00
40%	90.00
50%	147.00
60%	383.00
70%	2643.00
80%	231211.40
90%	637633.20
100%	89717350.00

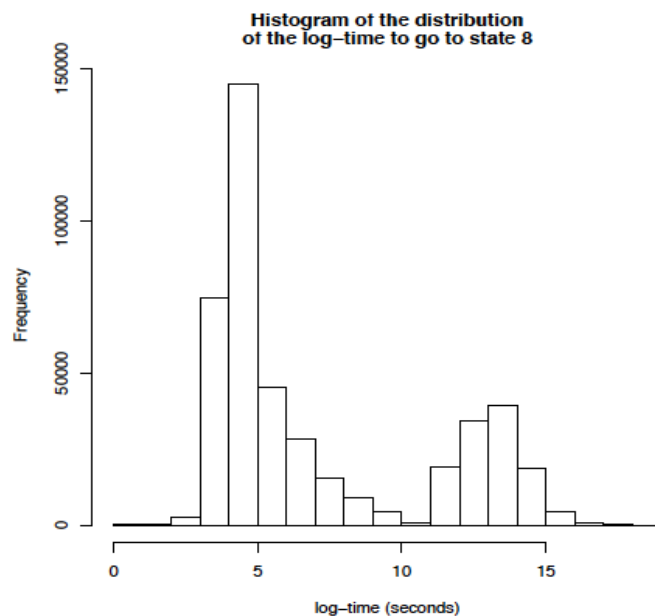


Fig. 3. Distribution of the lengths of paths (time to go in state 8)

We performed the EM clustering algorithm with $K = 4$ clusters. The cluster distributions and the average time for each state within clusters are presented in Table 4 and Figure 5.

Cluster	1	2	3	4
Size	37783	23159	358498	23844

Table 4. Clusters distribution

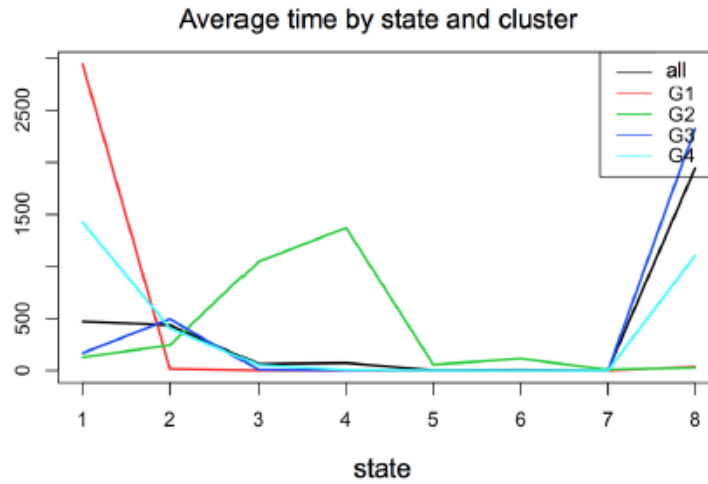


Fig. 5. Cluster profile: state-time distribution

Clearly, the cluster 1 contains letters for which dictating is the state which take the most of time until de delivery. The cluster 2 is a cluster where the state 3 and 4 are important states (assistants type-writing and validation from the doctor). Clusters 3 and 4 are characterized by an important time spent in state 2.

Acknowledgements

The present study was supported by a grant from ANR, France.
This work was performed within the Project CLINMINE sponsored by the ANR, France.

USING A BIVARIATE SCAN STATISTICS FOR DETECTING DISEASE CLUSTERS

B. Alexe

Institute of Mathematical Statistics and Applied Mathematics, Bucharest, Romania
Faculty of Mathematics and Computer Science, University of Bucharest, Romania

Abstract

In this paper we address the problem of detecting disease clusters using a bivariate scan statistics. We define it as the maximum number of events in a scan window and give an algorithm for determining its empirical distribution. We apply it for detecting clusters in children leukemia by assuming an underlying uniform bivariate Poisson process and simulate the distribution of the bivariate scan statistics based on the population and the number of ill cases. We compare the described method to other methods such as flexible spatial scan statistic and circular scan statistic.

Keywords: scan statistics, cluster detection.

I. INTRODUCTION

Scan statistics [1-6] addresses the problem of unusual event detection, trying to answer in a statistical manner whether is natural or not to observe large or small number of events grouped in a region. We define the scan statistic as the maximum number of events in a window of fixed dimensions and propose an algorithm for obtaining its empirical distribution for the case where events follow a bivariate uniform Poisson distribution. The distribution of the scan statistics is used in many application domains such as epidemiology, computer vision, reliability theory. We use it here for detecting disease clusters applied to children leukemia. We compare our results to other cluster detection methods such as the flexible spatial scan statistic of [8] and the scan statistic using a circular window of [4].

II. METHODOLOGY

In this section we define the scan statistics, present an algorithm for obtaining its empirical distribution and compare the simulation results to [1] and [6].

Definition 1. Let X_1, X_2, \dots, X_N be random points in the interval $[0, T]$. We define the one dimensional scan statistic S_w as the maximum number of points which are found in an interval of length w in $[0, T]$. The interval of length w is called the scan window.

Of great interest in the literature is the case where the random variables X_1, X_2, \dots, X_N are a trajectory of N points from a Poisson process $\{X_t, t \geq 0\}$. We define next the bivariate scan statistic when the random variables X_1, X_2, \dots, X_N are a selection of a uniform bivariate Poisson process. We give first the definition of the bivariate Poisson process [7].

Definition 2. A process consisting of randomly occurring points in the plane is said to constitute a uniform bivariate Poisson process having rate $\lambda, \lambda > 0$, if

- i) the number of points occurring in any given region of area A is Poisson distributed with mean λA ;
- ii) the number of points occurring in disjoint regions are independent.

Definition 3. Let $I = [0, T] \times [0, L]$ and let $u, v > 0$ two positive numbers such that $0 < u < T < \infty, 0 < v < L < \infty$, u and v define the size of the bivariate scan window. Assume that in the interval I there are N points X_1, X_2, \dots, X_N which constitutes the trajectory of uniform bivariate Poisson process $\{X_t, t \geq 0\}$ with intensity λ . We denote with $v_{t,s} = v_{t,s}(u, v)$ the number of points which fall in the rectangular scan window $[t, t + u] \times [s, s + v]$. We define the bivariate scan statistic S as

$$S = S(u, v, \lambda, T, L) = \max_{0 \leq t \leq T-u, 0 \leq s \leq L-v} v_{t,s} \quad (1.1)$$

the maximum number of points $v_{t,s}(u, v)$ over all rectangular scan windows $[t, t + u] \times [s, s + v]$ from I .

We are interested in computing the distribution of the scan statistic denoted by:

$$P(S((u, v), \lambda, T, L) \geq k) = P(u, v, \lambda, T, L, k). \quad (1.2)$$

The probability distribution (1.2) is hard to compute. Therefore a simulation procedure is one way to estimate it. In [6] is introduced a method for estimating the probability distribution (1.2) using the simulation of conditional scan statistic and the relationship between scan statistic and conditional scan statistic. For estimating the probability distribution given in (1.2) we use the following simulation algorithm.

Algorithm 1

Input: T, L, m, u, v, λ ;

Step 1. For $j = 1, m$ do

1.1 simulate a bivariate Poisson process $\{X_1, X_2, \dots, X_k\}$ on $[0, T] \times [0, L]$;

1.3 determine $S = S((u, v, \lambda, T, L, k))$ and denote $K_j = S$;

Step 2. Determine the empirical distribution of the sample K_1, \dots, K_m as follows:

2.1 Determine the order statistics $K_{(1)} < K_{(2)} < \dots < K_{(r)}, r < m$

2.2 Determine the frequencies $f_i, 1 \leq i \leq r, f_i =$ number of sampling values K 's equal to

$$K_{(i)}, \sum_{i=1}^r f_i = m.$$

2.3 Determine the relative frequencies (sampling probabilities) $\pi_i = \frac{f_i}{m}$.

Output: $P(S((u, v), \lambda, T, L) \geq k) = \sum_{i, K_i \geq k} f_i$.

Algorithm 1 takes as inputs the dimension T and L of the interval I , the number m of the simulations employed, the dimension u and v of the scan statistic window and the intensity λ of the bivariate Poisson process. The bivariate Poisson process is simulated at step 1.1. Step

1.2 computes the bivariate scan statistic S . Based on the m simulations of the scan statistic S , Step 2 builds-up a frequency distribution, i.e. a histogram of the scan statistics. If m is large enough, then the sampling distribution of K_1, K_2, \dots, K_m converges to (1.2) (according to the consistency property of the estimates π_i).

Step 1.1 in Algorithm 1 is implemented by Algorithm 2 which simulates [7] an uniform bivariate Poisson process with intensity λ on the interval $[0, T] \times [0, L]$.

Algorithm 2

Input: T, L, λ ;

Step 1. Simulate $0 < T_1 < T_2 < \dots < T_k$ a uniform one dimensional Poisson process with rate $\lambda, T_k \leq T$, i.e. k is random variable.

1.1 Initialize $t = 0, k = 0$;

1.2 Repeat

1.2.1 Generate $E \sim \text{Exp}(1)$;

1.2.2 Take $k = k+1, t = t + \frac{E}{\lambda}, T_k = \lfloor t \rfloor$;

until $t \geq T$.

Step 2. Generate L_1, L_2, \dots, L_k uniform on $[0, L]$;

Output $(T_1, L_1), (T_2, L_2), \dots, (T_k, L_k)$.

The step 1.2 in Algorithm 1, determining the value of the bivariate scan statistic $S = S((u, v, \lambda, T, L, k))$ is implemented by Algorithm 3, described below.

Algorithm 3

Input: $(T_1, L_1), (T_2, L_2), \dots, (T_N, L_N), T, L, u, v$.

Initialize $S = 0$.

Step 1: for $i = 1$ to N

1.1 Put a scan window of size $u \times v$ in point (T_i, L_i) that defines the rectangular $R_i = [T_i, T_i + u] \times [L_i, L_i + v]$.

1.2 Count the number of points n_i that are contained in R_i .

1.3 $S = \max(S, n_i)$.

Output: S .

In practice, the calculation employed in step 1.1 can be reduced by running a scanning window mechanism which reuses some computations. Also, due to computations issues, the points $(T_1, L_1), (T_2, L_2), \dots, (T_k, L_k)$ are generated with integer coordinates. This can be achieved by scaling the initial generated points and also scaling the window sizes u and v .

$\lambda = 0.01, T = L = 10, u = v = 1$

K	P(S<=k)	[6]	[1]
1	0.9818	0.9826	0.9959
2	1.0000	0.9998	0.9999

$\lambda = 0.01, T = 200, L = 100, u = v = 1,$

K	P(S<=k)	[6]	[1]
2	0.9720	0.9713	0.9808
3	1.0000	0.9998	0.9999

$\lambda = 0.5, T = 10, L = 10, u = v = 1,$

K	P(S<=k)	[6]	[1]
2	0.9859	0.9854	0.9905
3	0.9998	0.9996	0.9997
4	1.0000	0.9999	0.9999

$\lambda = 0.5, T = 10, L = 10, u = v = 1$

K	P(S<=k)	[6]	[1]
4	0.7865	0.7938	0.8343
5	0.9692	0.9707	0.9759
6	0.9968	0.9970	0.9974
7	0.9999	0.9997	0.9997

Table 1. Simulation results of the scan statistic $S = S(u, v, \lambda, T, L)$ based on $m = 10000$ simulations and different values for parameters u, v, λ, T, L .

III. RESULTS

We present in Table 1 simulations results using Algorithm 1 for assessing the empirical distribution of the scan statistics $S(u, v, \lambda, T, L, k)$ based on $m = 10000$ simulations, comparing our results in column 2 with the results obtained by [6] and [1]. Table 1 shows similar results of our method when compared to [6] and [1]. We used it next for detection of disease clusters.

We apply our method for identifying the potential disease clusters in the region of Nord Pas de Calais, north of France. In this region, from 2001 and 2003, there were registered 497 cases of children leukemia among a population of 573500 people. The region is divided in two departments, containing a total of 156 cantons. Among these, only 123 cantons exhibit ill cases. As some of these cantons have the same administrative center in the end it results 96 cantons with different administrative centers with ill cases. The canton with the highest disease incidence is Lievin Sud, with 9 ill cases registered at a population of 1600 people. The question is whether this canton is a potential disease cluster, with the number of ill cases higher than normal values. We answer this question using our method. We assume an underlying uniform bivariate Poisson process of the ill cases on the entire region with intensity λ . Denoting with P the population size and q the number of ill cases, we obtain $\lambda = \frac{q}{P} = 0,000866$. We model the population as a square of length $T = L = \sqrt{P} = 757.3$. As we want to decide if the number of ill cases registered in region Lievin Sud, 9 ill cases over 1600 population, is abnormal we take the scanning window with dimension $u = v = \sqrt{q} = 40$ and compute the distribution of the scan statistic $S(40,40,0.000866,757.3,757.3)$. In particular we are interested what is the probability that this statistic is higher or equal than 9. Using Algorithm 1 we obtain:

$$P(S(40,40,0.000866,757.3,757.3) \geq 9) = 0.178$$

which shows that the 9 ill cases from Lievin Sud do not constitute an abnormal event. We decide that Lievin Sud is not a disease cluster.

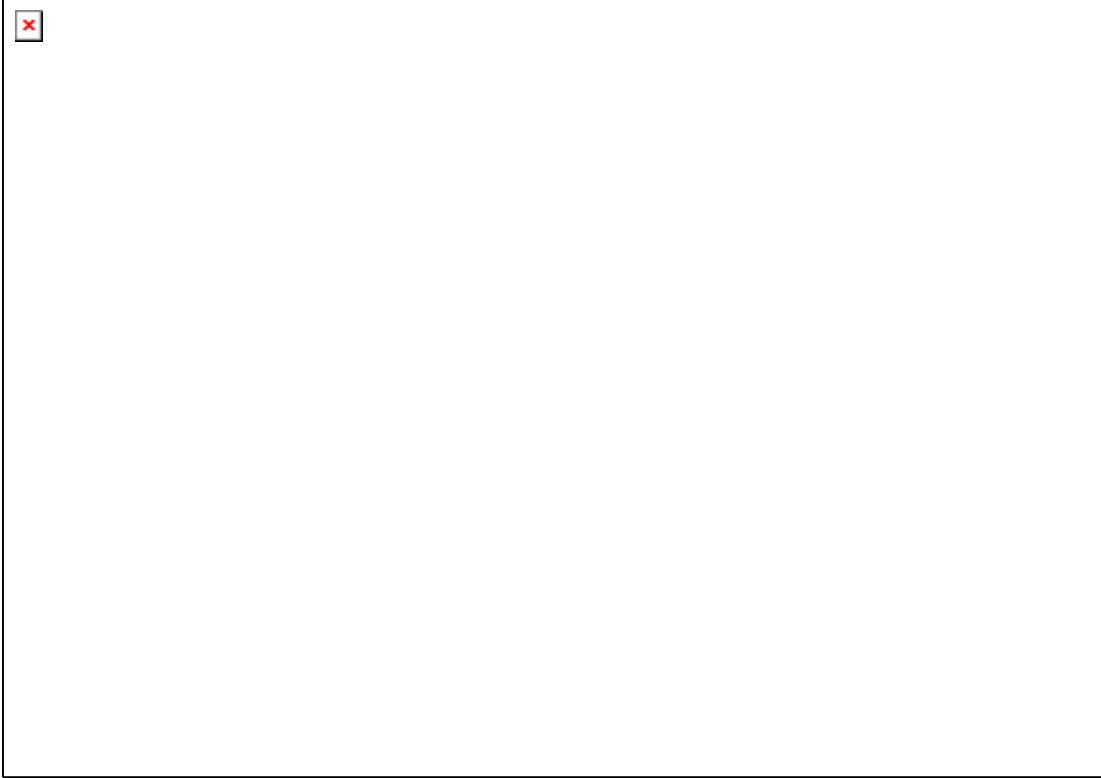


Figure 1. The map of Nord Pas de Calais devised in cantons. The blue colored cantons have a higher disease incidence that the red colored cantons.

IV. COMPARISON TO OTHER METHODS

We compare our method to the circular scan statistics method of [4] and to the flexible shaped spatial scan statistics of [8]. The method of [4] detects clusters using a circular scan statistic window centered in an administrative center of a canton and using variable radii for the scan window. All cantons whose administrative center is included in the circular windows are added. This procedure results in small deviations from the circular shape of a potential cluster. While the method of [4] can detect only clusters with shape almost circular, the method of [8] overcomes this drawback by allowing a flexible shape scan window. A scan window is build by starting with an administrative center and including in the window the corresponding region. At each step, the method allows adding adjacent regions to the ones already included up to a maximum number of C regions. Consequently, this method could potential identify disease clusters with irregular shape, such as the clusters along rivers. Both methods detect potential disease cluster regions as all the regions in a scan window Z based on the likelihood ratio:

$$\sup_{Z \in \mathcal{Z}} \left(\frac{n(\mathbf{Z})}{\xi(\mathbf{Z})} \right)^{n(\mathbf{Z})} \left(\frac{n(\mathbf{Z}^C)}{\xi(\mathbf{Z}^C)} \right)^{n(\mathbf{Z}^C)} I \left(\frac{n(\mathbf{Z})}{\xi(\mathbf{Z})} > \frac{n(\mathbf{Z}^C)}{\xi(\mathbf{Z}^C)} \right), \quad (1.3)$$

where $n(\mathbf{Z})$ is the number of ill cases in window Z , $\xi(\mathbf{Z})$ is the expected number of ill cases under the null hypothesis H_0 of equal means for all windows Z , \mathbf{Z}^C denotes the set of regions not included in Z , I is the indicator function. The distribution of the statistics in (1.3) under the null hypothesis H_0 can be computed using Monte Carlo simulations. The

	Flexible scan window [8]	Circular scan window [4]
C=15	Log likelihood ratio = 9.96 p-value = 0.016 third cluster found	Log likelihood ratio = 9.96 p-value = 0.002 second cluster found
C=20	Log likelihood ratio = 9.96 p-value = 0.051 third cluster found	Log likelihood ratio = 9.96 p-value = 0.002 second cluster found

Table 2. Results for methods [4] and [8].

window Z^* with the highest value given in (1.3) defines the most probable cluster. We obtained results for both methods using the Flex – Scan software [8]. It is important to mention that this software uses the Poisson modeling of the underlying process of ill cases. To obtain the clusters we need the coordinates of the administrative center of each canton, the adjacent matrix for all cantons, the number of ill cases and the total population of each canton. Table 2 shows the log likelihood ratios, the corresponding p-values and the rank of the log likelihood ratio among all other clusters found for the Lievin region when taking the maximum C numbers of regions in a cluster to be 15 or 20. Both methods [4] and [8] suggest that Lievin is a disease cluster, with the method [4] having a very small p-value of 0.002 indicating a high confidence for this decision.

V. CONCLUSION

In this paper we have used a bivariate scan statistics for cluster detection in children leukemia. We compared this method with the circular [4] and flexible scan window [8] methods for deciding whether the region Lievin, the region with the highest disease incidence, is a potential disease cluster.

References

1. Alm S. E.: On the distribution of the scan statistic of a two-dimensional Poisson process. *Adv. in Appl. Probab.* 1997, 29 1–16.
2. Glaz J., Balakrishnan N.: *Scan Statistics and Applications*. 1999, Birkhäuser, Boston.
3. Glaz J., Naus J., Wallenstein S: *Scan Statistics*. 2001, Springer, New York.
4. Kulldorff M.: A spatial scan statistic. *Comm. Statist. Theory Methods*, 1997, 26 1481–1496.
5. Kulldorff M.: Spatial scan statistics: Models, calculations, and applications. In: *Scan Statistics and Applications* (J. Glaz and N. Balakrishnan, eds.), 1999, Birkhäuser, Boston.
6. Haiman G., Preda C.: A new method for estimating the distribution of scan statistics for a two-dimensional Poisson process. *Journal of Methodology and Computing in Applied Probability*, Volume 4, Issue 4, Pages 393-407.
7. Ross S.: *Simulation*, 1997, Academic Press, San Diego, London.
8. Tango T., Takahashi K.: A flexible shaped spatial scan statistic for detecting clusters, *International Journal of Health Geographics*, 2005.

IRIS BIOMETRICS: ADAPTIVE RESONANCE NETWORKS FOR INDEXING AND RETRIEVING DATA

C. Enachescu¹, D. Enachescu^{1,2}

¹*Institute for Mathematical Statistics and Applied Mathematics, Bucharest, Romania*

²*University of Bucharest, Faculty of Mathematics and Computer Science, Bucharest, Romania*

Abstract

Biometric exploits discriminable physiological characteristics to identify a legitimate individual. Among the present biometric traits, iris is found to be the most reliable and accurate because iris is distinct and intrinsic organ, which is externally visible and yet secured one.

This work presents a new method for iris indexing and recognition. Given a query iris image, the goal of indexing is to identify and retrieve a small subset of candidate irises from the database in order to determine a possible match. This can significantly improve the response time of iris recognition systems operating in the identification mode.

Our contribution consists in applying an Adaptive Resonance Architectures neural network based on a Mahalanobis distance (ARTMAH) to find small clusters of irises and adapt this network to a supervised learning strategies in order to identify the genuine iris.

The performance of the proposed algorithms is validated and compared with other algorithms using *ND-CrossSensor-Iris-2013* database. Experiments show a substantial decrease of the false acceptance and false rejection rates (FAR and FRR) in the recognition of iris images.

Keywords: Adaptive Resonance Theory networks, cluster analysis, iris recognition, Mahalanobis distance, supervised learning

I. INTRODUCTION

Amongst a variety of biometric traits, iris based recognition systems are more valued due to its distinct pattern. Iris patterns are believed to be unique due to its rich, distinctive and complex pattern of crypts, furrows, arching, collarets and pigment spots. It is very precise and most stable personal identification biometric. The two iris patterns are not similar or identical even if those of identical twins, even between the same individuals left and right eye [1].

Iris recognition process is quite complex and is divided into three main steps: (1) pre-processing, (2) feature extraction and (3) recognition. The original contribution of the present paper is the learning mechanism corresponding to the recognition step. The learning mechanism is a modified version of an Adaptive Resonance Architectures (ART) network that is adaptive enough in order to accept the iris templates given in a random order for training. This modified unsupervised neural network will create and store the digital identities of the enrolled users, identities that are further used to test the recognition and identification process. ART networks are able of stable categorization of an arbitrary sequence of unlabeled input patterns in real time. They are capable of continuous training with non-stationary inputs. ART networks also solve the stability-plasticity dilemma; namely, they let the network to adapt preventing the current inputs from destroying past training.

This paper is organized into the following sections. Section II presents a detailed discussion on iris classification using ARTMAH classifier. Comparative analysis of experimental results is reported in Section III. Section IV concludes the paper.

II. METHODOLOGY





Multiple approaches of Machine Learning (ML) techniques, both supervised and unsupervised, have been made in biometrics over the last several years. Regarding supervised learning methods, there are different types of neural network classifier reported such as: probabilistic neural network (PNN), multilayer perceptron (MLP) neural network, radial basis function (RBF) neural network. Regarding unsupervised learning methods, the most frequently reported results are for k -means (which, for instance, in [2] is used as an iris image classifier) and for Self-Organizing Maps (SOM) – [3]. Narrowing down the list of related works to the ones utilizing ART networks, is seen that these neural networks are applied in biometrics mainly in handwritten signature verification approaches, in speaker recognition systems, and in palm vein recognition. In [4] was proposed an iris biometric system based on a Fuzzy ART, tested on CASIA iris image database. The authors performed several tests, using different values for the vigilance parameter, and reported the FAR and FRR values.

Back propagation network is very powerful in the sense that it can simulate any continuous function given a certain number of hidden neurons and a certain forms of activation functions. But, training a back propagation network is quite time consuming. It takes thousands of epochs for the network to reach the equilibrium and it is not guaranteed that it can always land at the global minimum. Once a back propagation is trained, the number of hidden neurons and the weights are fixed. The network cannot learn from new patterns unless the network is re-trained. Thus we consider the back propagation networks don't have *plasticity*. Assuming that the number of hidden neurons can be kept constant, the plasticity problem can be solved by retraining the network on the new patterns using on-line learning rule. However, it will cause the network to forget about old knowledge rapidly. We say that such algorithm is not *stable*. The contradiction between plasticity and stability phenomenon is called *plasticity/stability dilemma* [5].

Adaptive Resonance Theory (ART) is a type of neural network designed by Grossberg in 1976 [6] to solve *plasticity /stability dilemma*. The most important ART networks are: ART-2, used to cluster analog data, ARTMAP, a supervised learning mechanism for binary data, and Fuzzy ART, a supervised learning algorithm for analog data.

The typical ART network is a recurrent unsupervised neural network.

The basic architecture of the ART network consists of a layer of linear perceptrons representing prototype vectors whose outputs are acted on by a winner-take-all network. This architecture differs from a competitive net in that the linear prototype units are allocated dynamically, as needed, in response to novel input vectors. Once a prototype unit is allocated, appropriate lateral-inhibitory and self-excitatory connections are introduced so that the allocated unit may compete with preexisting perceptrons. Alternatively, one may assume a prewired architecture as in Figure 1 with a large number of inactive (zero weight) units. Here, a unit becomes active if the training algorithm decides to assign it as a cluster prototype unit, and its weights are adapted accordingly.

The goal of the network training is to find a set of templates, which best represent the underlying structure of the samples. The general idea behind ART training is as follows. Every training iteration consists of taking a training example \mathbf{x} , examining existing prototypes (weight vectors ) that are sufficiently similar to \mathbf{x} . If a prototype  is found to "match" \mathbf{x} (according to a "vigilance" test based on a preset matching threshold), example \mathbf{x} is added to the cluster represented by  and  is modified to make it better match \mathbf{x} . If no prototype

matches \mathbf{x} , then \mathbf{x} becomes the prototype for a new cluster. The details of the ART clustering procedure are considered next.

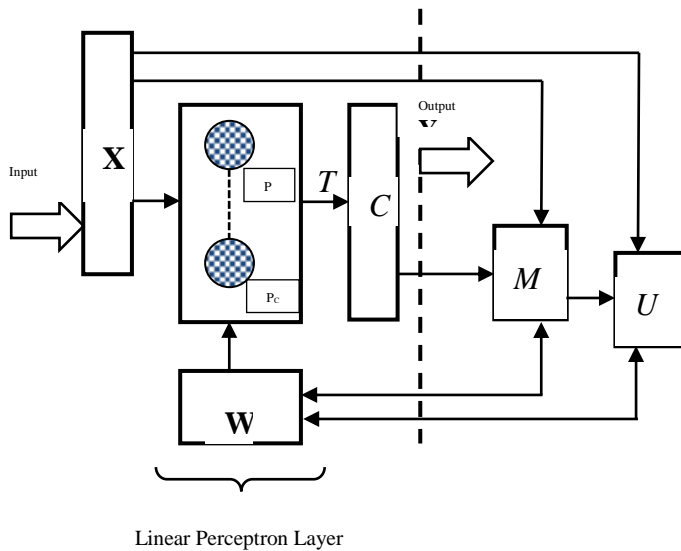


FIGURE 1 The architecture of ARTMAH.

Each learned cluster, say, cluster j , is represented by the weight vector of the j -th prototype unit. Every time an input vector \mathbf{x} is presented to the net, each existing prototype unit compute an *activation function* $T(\mathbf{x}, \mathbf{w}_j)$ and output it to the winner-take-all *compet function* for determining the winner unit. The compet function computes a "winner" unit i . Subject to further verification, the weight vector of the winner unit \mathbf{w}_i now represents a potential prototype for the input vector.

The verification comes in the form of passing the *vigilance test*, i.e. $\frac{\|\mathbf{x} - \mathbf{w}_i\|}{\|\mathbf{x}\|} \leq \rho$. This is called the *match function* and is used to qualify how good is the likeness of \mathbf{x} to \mathbf{w}_i . The function is used in conjunction with the *vigilance parameter* $0 < \rho \leq 1$. The vigilance is the most important network parameter that determine its resolution: smaller value allow for large deviations from cluster centers and hence lead to a small set of clusters; larger vigilance value normally yields larger number of output nodes and good precision.

If the vigilance test is passed by the winner unit i for a given input \mathbf{x} (here, the network is said to be in *resonance*), then \mathbf{x} joint cluster i , and this unit's weight vector \mathbf{w}_i is updated according to the *update function* $U(\mathbf{x}, \mathbf{w}_j)$. If the unit i don't pass the vigilance test, the unit is deactivated (its output is clamped to zero until a new input arrives) and the test is repeated with the unit with the next highest output. If this scenario persists even after all existing prototype units are exhausted, then a new unit representing a new cluster j is allocated and its weight vector is initialized with \mathbf{x} .

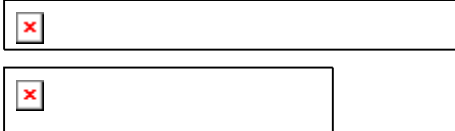
Note that the learning dynamics described above constitute a search through the prototype vectors looking at the closest, next closest, etc., according to the compet and match functions. It also should be noted that this search only occurs before stability is reached for a given training set. After that, each prototype vector is matched on the first attempt, and no search is needed.

Regardless of the setting of ρ , the ART network is stable for a finite training set; i.e., the final clusters will not change if additional training is performed with one or more patterns drawn from the original training set. A key feature of the ART network is its continuous

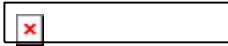
learning ability. This feature, coupled with the preceding stability result, allows the ART net to follow nonstationary input distributions.

ART implement an unsupervised clustering method. ARTMAH (see [7]) on the other hand performs incremental supervised learning of labeled patterns. Comparing with the ART learning algorithm, ARTMAH has one extra loop which checks if the label of the pattern matches with the label of template. If labels matches, the algorithm proceeds as in the ART case. If label doesn't match, the vigilance is boosted to activation value of current candidate plus a small positive number, and current winning node is suppressed.

In the case of ARTMAH the activation and match functions are:



The candidate is the one which has the minimum activation value,



After resonance happened, in addition to the centers of the ellipsoids, the covariance matrices S_j for each cluster have to be updated as well. The recurrent equation for updating means and covariance matrices are:



The classification is decided by the Mahalanobis distance, or activation function. The label of the template with which a pattern has the minimum Mahalanobis distance is the prediction.

ARTMAH, is used below to create the digital identities of an iris biometric verification system. The number of neurons resulted after training the ARTMAH network corresponds to the total number of enrolled users.

III RESULTS

In order to compare the performance of the ARTMAH network with other iris biometric systems we follow the experiments methodology presented in [8]. The experiments are made with ICs (Iris Codes) obtained from iris images captured with the LG4000 sensor. These iris images are found within *ND-CrossSensor-Iris-2013* database 9. . The unwrapped iris segments are extracted with CFIS2 (Circular Fuzzy Iris Segmentation) and the iris codes are obtained by applying the Log-Gabor encoder. The training (*TrainIC*) and testing (*TestIC*) iris codes are firstly mean-shifted in the preprocessing phase:

$$Mean = \frac{1}{nc} \sum_{i=1}^{nc} IC_i,$$

$$TrainIC_i = IC_i + Mean, i = 1 : nc,$$

$$TestIC_j = IC_j + Mean, j = 1 : mc$$

where *mc* is the total number of test iris templates, and *nc* is the total number of training templates. For enrollment, and for testing, each user honestly claimed its true identity, allowing the simulation of a logically consistent iris biometric system.

In the following, the performance of the proposed ARTMAH network is investigated in terms of two empirical measures, specific for biometric iris recognition systems, namely consistency and comfort. Both, consistency (1) and comfort (2), are acquired through satisfying adequate preliminary conditions, P_C , for enrollment. Therefore, in a consistent theory of iris recognition, denoting by S_{c^I} an imposter score, i.e. the Mahalanobis distance of the imposter to the prototype of the cluster it was classified, and by S_{c^G} a genuine score, if the preliminary conditions of the enrollment procedure, P_C , are consistent, then is implied that the maximum imposter score is smaller than the minimum genuine score obtained with the same classifier:

$$P_C \rightarrow [\max(S_{c^I}) < \min(S_{c^G})], \quad (1)$$

In a comfortable theory of iris recognition, if the preliminary conditions of the enrollment procedure are satisfied, then between the maximum imposter score and the minimum genuine score may be fitted a comfortable (generous) safety band.

$$P_C \rightarrow [\max(S_{c^I}) \ll \min(S_{c^G})], \quad (2)$$

It should be noted that the two conditions require that the intersection of supports of the genuine-score distribution and impostors-score distribution is empty. This implies that the confusion matrix, a general empirical measure used in classification problems, is a diagonal matrix, hence the classification error is zero.

Organizing the Data

Eight tests are performed in order to evaluate the robustness of the above introduced ARTMAH network. The tests are grouped in two sets, according to the training setup. For the first set of experiments, the training templates are sent in order, while for the second one the order is randomized. By ‘order’ should be understood that, for each individual that is currently in the enrolling phase, all its corresponding iris codes are presented, one by one, to the digital identity learning mechanism, before starting to enroll another individual. By ‘random’ should be understood that the templates of all the individuals that will be enrolled are not user-ID dependent when are presented to the ARTMAH network.

There are used two training samples: a 100-individuals sample, containing 300 templates, 3 iris template per individual, and a 200-individuals one, with the same structure. The iris templates are selected randomly from the *ND-CrossSensor-Iris-2013* database (*LG 4000* sensor). For tests, each individual from the training set contributes with other 10 different iris codes. A structured view on the number of individuals and templates involved in each test, on the ordered or on the randomly trained ARTMAH network, is available in TABLE I.

TABLE I. THE NUMBER OF INDIVIDUALS AND TEMPLATES, RESPECTIVELY, INVOLVED IN THE PERFORMED EXPERIMENTS.

	1 st test		2 nd test	
	Train	Test	Train	Test
No. individuals	100	100	200	200
No. ICs	300	1000	600	2000

In the training phase of the ARTMAH network, several values for the vigilance parameter, ρ , have been used, until it managed to create the correct number of clusters, i.e. 100 for the first

experiment and 200 for the second one. In both cases, the value of the vigilance parameter that allowed forming the 100 and the 200 clusters was 0.7.

The obtained genuine and imposter score distributions are illustrated below, and will show that our iris biometric system is both consistent and comfortable, maintaining a clear separation between its users, avoiding accidental or intended impersonations among them.

The experiments performed on the multi-enrollment iris biometric system with 100 enrolled users proved that it is both consistent and comfortable, having null FAR and FRR and a large safety band between the minimum genuine score and the maximum imposter score, scores obtained for the templates from the test dataset. This iris recognition system is able to recognize the identities of all his enrolled users, regardless of whether the network was trained randomly or organized.

In this second set of tests (200-users), the maximum imposter scores and the minimum genuine scores continued to have values that are distant enough, allowing the system to perform according to the consistent and comfortable theory. It can be seen that, as observed for the 100-users multi-enrollment iris biometric system, the imposter scores have the tendency of conglomerating under the 0.7 threshold, allowing null FAR and FRR. Also, there is no difference between the results obtained for the tests performed on the system that used an ARTMAH network trained with iris templates in user's ID order, and the tests performed on the one that used an ARTMAH network trained with randomly organized iris templates.

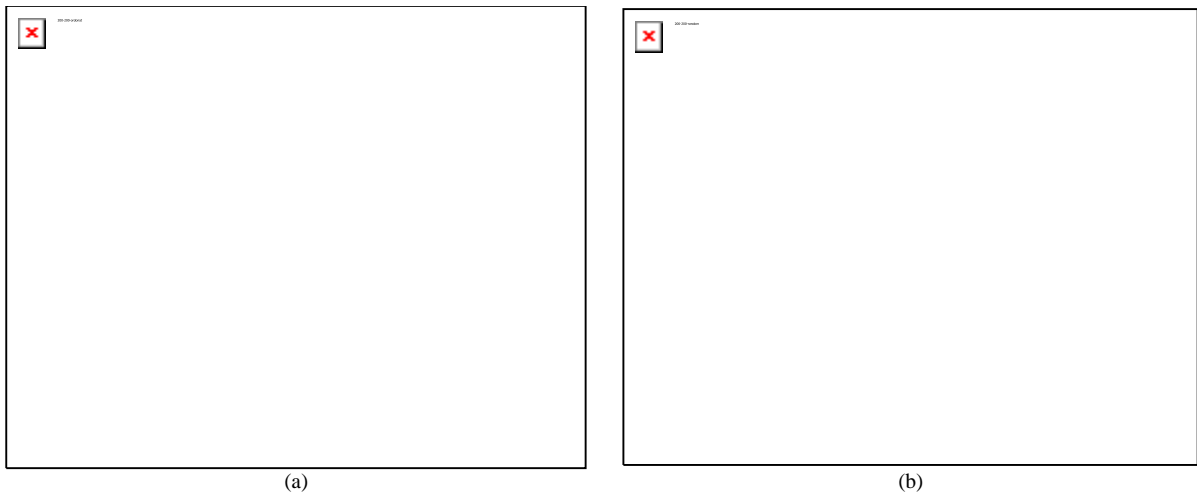


FIGURE 2 The y-scale logarithmic representation of genuine and imposter score distributions of a multi-enrollment iris biometric system with 200 enrolled users. The genuine and imposter score comparisons are obtained on the test dataset. (a) ordered training (b) random training.

It can be noticed, see Figure 2, that the genuine scores from interval 0.86-0.94 have a different behavior, indicating that the iris templates from the test dataset might have been subject to noisy acquisition. The results obtained for the four tests allow the conclusion that, no matter if the digital identities are established randomly or ordered – the proposed iris biometric verification system is consistent and able to offer both stability and comfort. Also, these two different manners of presenting the iris codes to the network in the training phase show that the proposed architecture of the neural network is independent of the order of the training patterns.

IV. CONCLUSIONS

This paper presents an iris biometric system based on an ARTMAH neural network that achieves good results in terms of both consistency and comfort. The genuine and imposter

score distributions showed that the system successfully recognizes its users, obtaining null FAR and FRR. The system proved to be robust, by performing very well in all the tested scenarios, allowing a generous safety band between the minimum genuine and maximum imposter scores.

References

1. L.F. Araghi, H. Shahhosseini, F. Setoudeh, Iris recognition using neural network, In Proceedings of The International Multiconference of Engineers and Computer Scientists, Vol. 1, pp. 338-340, 2010.
2. H. Proenca, L.A. Alexandre, Iris segmentation methodology for non-cooperative recognition, In Vision, Image and Signal Processing, IEE Proceedings, Vol. 153, No. 2, pp. 199-205, IET, 2006, April.
3. S. Campos, R. Salas, H. Allende, C. Castro, Multimodal algorithm for iris recognition with local topological descriptors, In Progress in Pattern Recognition, Image Analysis, Computer Vision, and Applications, pp. 766-773, Springer Berlin Heidelberg, 2009.
4. A. Taherian, M.A. Sh, Noise Resistant Identification of Human Iris Patterns Using Fuzzy ARTMAP Neural Network, International Journal of Security and Its Applications, Vol. 7, Issue 1, pp. 105-118, 2013.
5. R. O. Duda, P. E. Hart and D. G. Strok. *Pattern Classification*. John Wiley & Sons, Inc, 2001.
6. S. Grossberg. Adaptive Pattern Classification and Universal Recoding, I: Parallel Development and coding of neural feature Detectors. *Biological Cybernetics*, Vol. 23: 121-134, 1976.
7. M. I. Vuskovic, H. Xu and S. Du. Simplified ARTMAP Network Based on Mahalanobis distance. *Proceeding of the 2002 International Conference on Mathematics and Engineering Techniques in Medicine and Biological Science*, Las Vegas, Nevada, June 27-27, 2002.
8. C.M. Noaica, D. Enachescu, An Iris Recognition System based on a modified Adaptive Resonance Network, Procc. of DACS 2015, Analele Universitatii din Bucuresti, seria Informatica, 2015.
9. ND-CrossSensor-Iris-Database-2013, http://www3.nd.edu/~cvrl/CVRL/Data_Sets.html.

SURVIVAL AND 30-DAYS CEREBRAL PERFORMANCE IN PATIENTS WITH SUCCESSFUL CARDIOPULMONARY RESUSCITATION FOLLOWING CARDIAC ARREST: STATISTICAL INFERENCE IN THE PRESENCE OF INCOMPLETE DATA

Z. Valenta¹, P. Ošťádal², D. Vondráková², M. Průcha³, A. Kruger², M. Janotka²

¹*Dept. of Medical Informatics & Biostatistics, Institute of Computer Science,
Czech Academy of Sciences, Czech Republic*

²*Heart Center, Department of Cardiology, Na Homolce Hospital, Prague, Czech Republic*

³*Dept. of Clinical Biochemistry, Hematology and Immunology, Na Homolce Hospital,
Prague, Czech Republic*

Abstract

We analyzed data of 111 patients with successful cardiopulmonary resuscitation following out-of-hospital cardiac arrest. All patients were treated for 24 hours with mild therapeutic hypothermia (33°C) using endovascular temperature-controlling system. Clinical results were assessed according to the cerebral performance category (CPC) at 30 days. Markers of oxidative damage to DNA, RNA and phospholipids were measured at admission, 12 and 30 hours of follow-up. While the information on patients' survival and CPC at 30 days follow-up was complete, markers of oxidative damage and few other predictors could not all be measured as planned. We use multiple imputation (MI) of missing values following partial regression reconstruction of missing values in markers of oxidative damage at baseline. We use Bayesian proportional odds logistic regression (BPOLR) in the context of MI and extend the R package 'mi' to allow for calculating confidence limits for the OR. We further use Weibull modeling of right- and interval-censored data in the presence of incomplete data. Similarly as with BPOLR we use rules for combining variance from within and between MI samples to draw inference from Weibull regression in the context of MI.

Keywords: Bayesian POLR, Cardiac arrest, Cardiopulmonary resuscitation, Incomplete data, Weibull survival

I. INTRODUCTION

Primary purpose of our study involving 111 patients with successful cardiopulmonary resuscitation following out-of-hospital cardiac arrest was to characterize the association between the values of selected markers of oxidative damage measured at baseline and the cerebral performance category (CPC) determined after 30 days of follow-up. All patients were managed according to the international guidelines for post-resuscitation care and underwent mild therapeutic hypothermia (33°C) for 24 hours using endovascular temperature-controlling system [1]. Markers of oxidative damage to DNA, RNA and phospholipids (Isoprostan) were measured at admission and 12 and 30 hours after admission to the intensive care unit (ICU). Baseline values of these markers were, however, not completely available and the records exhibited large patterns of incompleteness.

The principal outcome variables assessing 30-days cerebral performance (good neurological result, moderate neurological dysfunction, serious neurological damage, persisting coma or death during follow-up) and patients' survival were completely observed. However, the missing patterns observed in the levels of markers of oxidative stress at baseline would not appear suitable for direct employment of multiple imputation. We were, however, able to

recover some missing baseline marker values from those observed later during the follow-up.

We used regression reconstruction (RR) as an approximate method of evaluating missing marker values at baseline. From a clinical perspective it was not considered optimal to directly substitute the values missing at baseline with those observed later during the follow-up and that notion was further confirmed by statistical testing. We therefore used simple linear regression as an approximate method for reconstructing missing baseline marker values from those measured 12 or 30 hours into the follow-up. Residual missingness patterns following RR appeared suitable for employing multiple imputation (MI) [2, 3] and the MAR assumption would not appear to be violated.

Patterns in the original data are shown in Figure 1. Apart from those observed at baseline we also display associated patterns observed in the three markers at 12 and 30 hours into the follow-up. The upper panel of Figure 2 displays reduction in missingness patterns of baseline marker values following RR, the lower panel shows average completed data after MI.

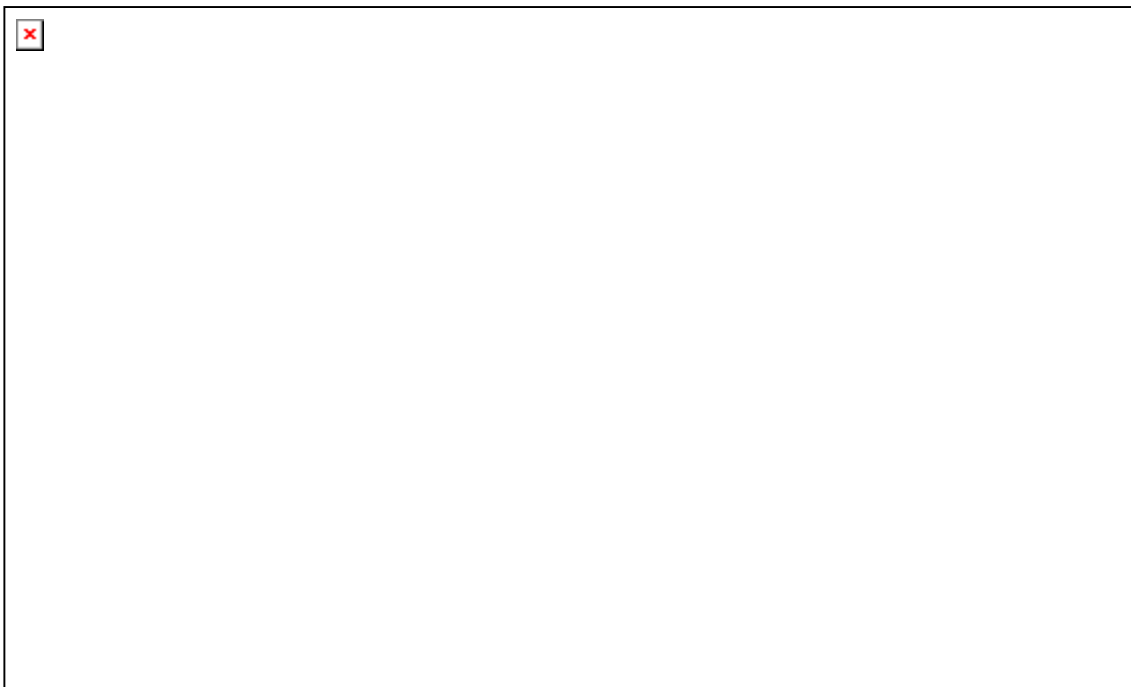


Figure 1. Missigness patterns in the original data.

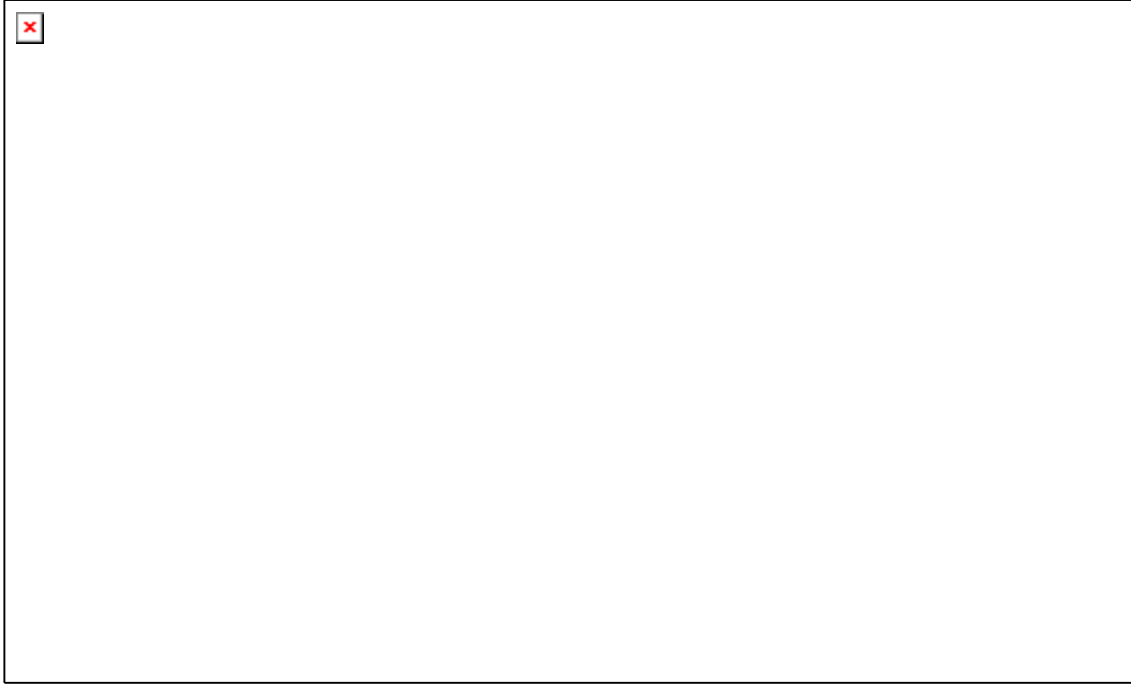


Figure 2. Missigness patterns after MI following RR.

II. METHODOLOGY

From the clinical standpoint an adverse outcome of primary interest was defined as death of a patient or the occurrence of serious neurological damage during 30-days of follow-up. The employed statistical methodology was carefully selected to match the clinical perspective. Patient data allowed for performing both the cohort (longitudinal) as well as ordinal cross-sectional analysis. While the fatal events were directly observed, occurrence of serious neurological damage was determined at 30 days of follow-up. The latter events, however, could have occurred at any time between the patient's study entry and the end of follow-up. Longitudinal data thus expressed a mixture of right- and interval-censored data suitable for Weibull survival modeling.

A compatible methodological approach to cross-sectional analysis of our data was represented by the Proportional Odds Logistic Regression (POLR). Both aforementioned approaches had to be adapted to the MI scenario. We used R's packages 'mi' and 'arm' [4] and the rules for combining variance from within and between MI samples to extend 'mi' package's capability and obtain confidence limits (CL) for the odds ratios (OR) based on the Bayesian POLR model. Similarly, in the longitudinal context we used rules for combining variance from within and between MI samples to draw inference from Weibull regression in the context of MI.

III. RESULTS

A summary of the results obtained from both cross-sectional and longitudinal analysis in the context of MI are shown in Tables 1 and 2.

By employing Bayesian POLR model in the context of MI after RR we modelled the odds of favourable neurological result vs. serious neurological damage or death within 30 days of follow-up, in parallel with the odds of favourable result or serious neurological damage vs. death of a patient.

In contrast, a cohort (longitudinal) data analysis was estimating the hazard associated with adverse outcomes (serious neurological damage or death) thus rendering a reciprocal

interpretation. Results from the Weibull survival model for right- and interval-censored data obtained in the MI context after RR are shown in Table 2.

Table 1: Bayesian POLR model results following RR & MI data completion

Covariates (baseline values)	Odds Ratio (OR)	95% Confidence Limits for the OR	p-value
log(RNA) marker	1.482	(0.852, 2.577)	0.1607
log(DNA) marker	1.863	(0.674, 5.149)	0.2113
log(Isoprostan) marker	0.258	(0.115, 0.579)	0.0015
log(ROSC)	0.069	(0.018, 0.265)	0.0002
log(CRP)	0.428	(0.244, 0.752)	0.0046
log(Lactate)	0.232	(0.096, 0.561)	0.0019
Gender	3.329	(1.091, 10.153)	0.0348

In both instances, the inference is based on 5 MI-completed datasets which was a default option for the 'mi' package. This allowed us to verify the variance estimation procedure in the context of BPOLR programming using the 'arm' package. A non-Bayesian averaging of parameter estimates over 5 completed datasets was applied in case of Weibull modeling and again the rules for combining the variance from within- and between-MI samples were applied to draw inference shown below.

Table 2: Weibull modeling results for right- and interval-censored data following RR & MI data completion

Covariates (baseline values)	Hazard Ratio (HR)	95% Confidence Limits for the HR	p-value
log(RNA)	0.759	(0.528, 1.092)	0.1374
log(DNA)	0.801	(0.428, 1.498)	0.4872
log(Isoprostan)	1.429	(0.935, 2.184)	0.0992
log(ROSC)	4.804	(2.071, 11.145)	0.0003
log(CRP)	1.645	(1.179, 2.294)	0.0034
log(Lactate)	2.934	(1.682, 5.117)	0.0001
Gender	0.361	(0.166, 0.784)	0.0101

IV. DISCUSSION

Both the cross-sectional modeling of ordinal outcome and the longitudinal data analysis performed in the context of incomplete data rendered similar results. Multiple imputation following regression reconstruction of missing data allowed us examining all three genetic markers simultaneously as part of one statistical model. This would not be possible when analyzing complete case subsets of our data. Such approach would be also be highly susceptible to rendering biased results. Extending the capabilities of the 'mi' package by writing the additions with the use of the 'arm' package in R allowed us to obtain additional inference from the BPOLR model and obtain similar inference from the Weibull survival modeling in the context of MI.

V. CONCLUSION

Multiple imputation following regression reconstruction helped in assessing simultaneous impact of the three target genetic markers of oxidative damage on the survival and the occurrence of serious neurological damage to the brain within 30 days of patients' follow-up. This would not be possible if we were just using the complete subsets of our data while such approach would be highly susceptible to rendering biased results.

Acknowledgements

The study was supported by institutional grant MH CZ – DRO, Nemocnice Na Homolce – NNH, No. 00023884, and by institutional support RVO:67985807 of the Institute of Computer Science, Czech Academy of Sciences.

References

1. Nolan JP, Soar J, Cariou A, Cronberg T, Moulaert VR, Deakin CD, Bottiger BW, Friberg H, Sunde K, Sandroni C. European Resuscitation Council and European Society of Intensive Care Medicine Guidelines for Post-resuscitation Care 2015: Section 5 of the European Resuscitation Council Guidelines for Resuscitation 2015. *Resuscitation*. 2015 Oct;95:202-22.
2. A. Gelman and J. Hill. *Data Analysis Using Regression and Multilevel/Hierarchical Models. Analytical Methods for Social Research*, volume 1. Cambridge University Press, 2006.
3. R.J.A. Little and D.B. Rubin. *Statistical Analysis with Missing Data*, volume 1 of Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, Inc., 1987.
4. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.

INFECTIONS CAUSED BY CLOSTRIDIUM DIFFICILE: FEATURE SELECTION VIA ORDINAL REGRESSION

Ł. Mierzejewski¹, W. Niemiński², W. Rejchel³, M. Zalewska⁴

¹*Medical University of Warsaw, Poland*

²*Faculty of Mathematics, Informatics and Mechanics, University of Warsaw, Poland*

³*Faculty of Mathematics and Computer Science, Nicolaus Copernicus University, Toruń, Poland*

⁴*Department of Prevention of Environmental Hazards and Allergology, Medical University of Warsaw, Poland*

Abstract

Clostridium difficile is a type of bacteria frequently causing hospital infections. The aim of our study is to find main risk factors, which have significant influence on these infections. We analyse data collected in one of Polish hospitals. The data include many features which are potential risk factors, but only few of them are truly relevant. Thus we consider the statistical problem of feature selection. The response variable, the level of infection, is measured in an ordinal scale. Most of standard methods of feature selection are not suitable to deal with this type of data. Therefore, apart from some analyses using standard statistical tools, we apply a relatively new method, named “ordinal regression”. We explain theoretical foundations of this method, with some recent results. We consider and apply a version of adaptive LASSO for ordinal regression. This algorithm is based on hinge loss, absolute deviations penalty and convex piecewise linear minimization.

Keywords: Hospital infections, Risk factors, Ordinal scale, Feature selection, LASSO, Adaptive LASSO, Least Absolute Deviations, Hinge Loss, Convex optimization.

I. INTRODUCTION

Clostridium difficile (Cd) is a type anaerobic bacteria producing toxins. This bacterium constitutes 3% of the physiological intestinal flora of the adult. It forms the spores immune to the warmth which are able to survive for months or for years. Cd is being regarded as the main pathogen causing hospital intestinal illnesses starting from mild diarrhoea to threatening the life the *pseudomembranous colitis*. Prevalence among hospitalized patients is 20%-40%. Cd is surprisingly omnipresent:

- Present in sand, dung of camels, horses and donkeys, droppings of dogs, cats, birds.
- Also present on human genitals, in digestive tract and faeces.

In the USA, Cd causes about 3 million cases of diarrhoea and *enteritis* (annually from 5 thousand up to 20 thousand demises). Numbers of cases of infection and deaths are increasing. Cd is spreading via dirty hands, toilets, contact with the hospital equipment etc. Cd infections frequently occur after antibiotic therapy.

The aim of our statistical analyses was to discover features which significantly increase the risk of Cd infections in a hospital. The data used in our study are described in more detail in the next section. The database includes explaining variables and the response variable: Cd infection. There are two specific aspects of these data, which make the analyses difficult.

- There are many explaining variables in the data but only a few of them are really relevant (i.e. significantly influencing Cd infections). Thus the main aim is to choose

the relevant features.

- The response variable is measured in an ordinal scale (i.e. it is possible to compare the severity of infection, but it is not appropriate to assign numeric values to the levels of severity).

The problem of feature selection, or model choice, has long been studied in statistical literature and recently attracts even more attention. There are numerous statistical tools and procedures to tackle this problem. One of the most successful approaches is based on the idea of penalized estimation, e.g. different versions of LASSO estimators. However, most of the existing methods of feature selection are developed either for numeric response variables or for categorical type data (qualitative variables). In the case of *ordinal* type response, theory of feature selection is less developed and few algorithms exist. The character of our Cd data motivates the use of specialized and new methods.

II DATA

In our analyses we used a database which was created in a chosen polish hospital, based on a questionnaire about hospital infections. The survey concerned 445 patients and included 45 quantitative and qualitative features. These features were regarded as potential factors for Cd infection, i.e. as explaining variables. The response variable was (presence and degree of) Cd infection. In the database the name of this variable was “zakazenie”, which is polish term for “infection”. This variable was measured on an ordinal scale with 5 values (0 - no infection, values 1-4 correspond to concentration of toxins (in the increasing order).

In the table below there is a list of the original names of variables in the database (abbreviations of polish terms; for example: “antybiot” = “antybiotyterapia” = “antibiotic therapy”, “wc” = “availability of a separate toilet in a hospital room” etc.).

Data: 444 cases (patients), 45 variables (features):

id	patogen	zakazenie
grupa	klinika	chir.int.oit
odc	plec	wiek
data.przyj	data.zak	dni.od.przyj.do.zakazenia
tryb.przyj	hosp.lub.zak.6msc	pielucha
wc	prysznic	ilu.pacjentow
antybiot	cukrzyca	niewyd.nerek
niewyd.watrobymidozyw.lub.wyniszcz		endosk
IPP	antag.H2	zyw.pozajel
immunosupres	prep.krwi	autotransf
transplant	hemodial	kraz.poz aust
radioter	chemia	zgleb.zoladk
nawrot.raz	nawrot.drugi	kolejny.nawr
operacje	reoperacje	wspol.zak
zak.szp.objaw	zak.pozaszp	nklas

Severity of infection is described by variable “zakazenie” (meaning “infection”). This variable has 5 possible values: symbol “0” for the absence of infection and symbols “1”, “2”, “3”, “4” encoding the degree of concentration of toxins. The frequency is given in the following table:

0	1	2	3	4
213	15	19	15	182

Let us emphasize that the levels (degrees) are ordered but do not correspond to concrete numerical values – they are only conventionally assigned descriptions. Put differently, the response variable is measured in an *ordinal* scale. Statistical inference for such data is difficult. The standard procedure is to regard ordinal data either as numeric (i.e. treat conventional levels as measurements in some fictitious units) or as categorical (i.e. disregard the ordering of levels). In the case of our Cd data, we can either treat 0, 1, 2, 3, 4 simply as numbers (and apply e.g. classical linear regression) or treat 0, 1, 2, 3, 4 as unordered labels (and apply e.g. discriminant analysis). Neither of these approaches is satisfactory. The first approach suggests that (for example) level “4” is twice as big as level “2” which is nonsense. The second approach suggests that difference between “0” and “5” is equally important as that between “1” and ‘2”. Of course we can also reduce the levels to a binary variabe (infection absent/present):

0	<input checked="" type="checkbox"/>
213	231

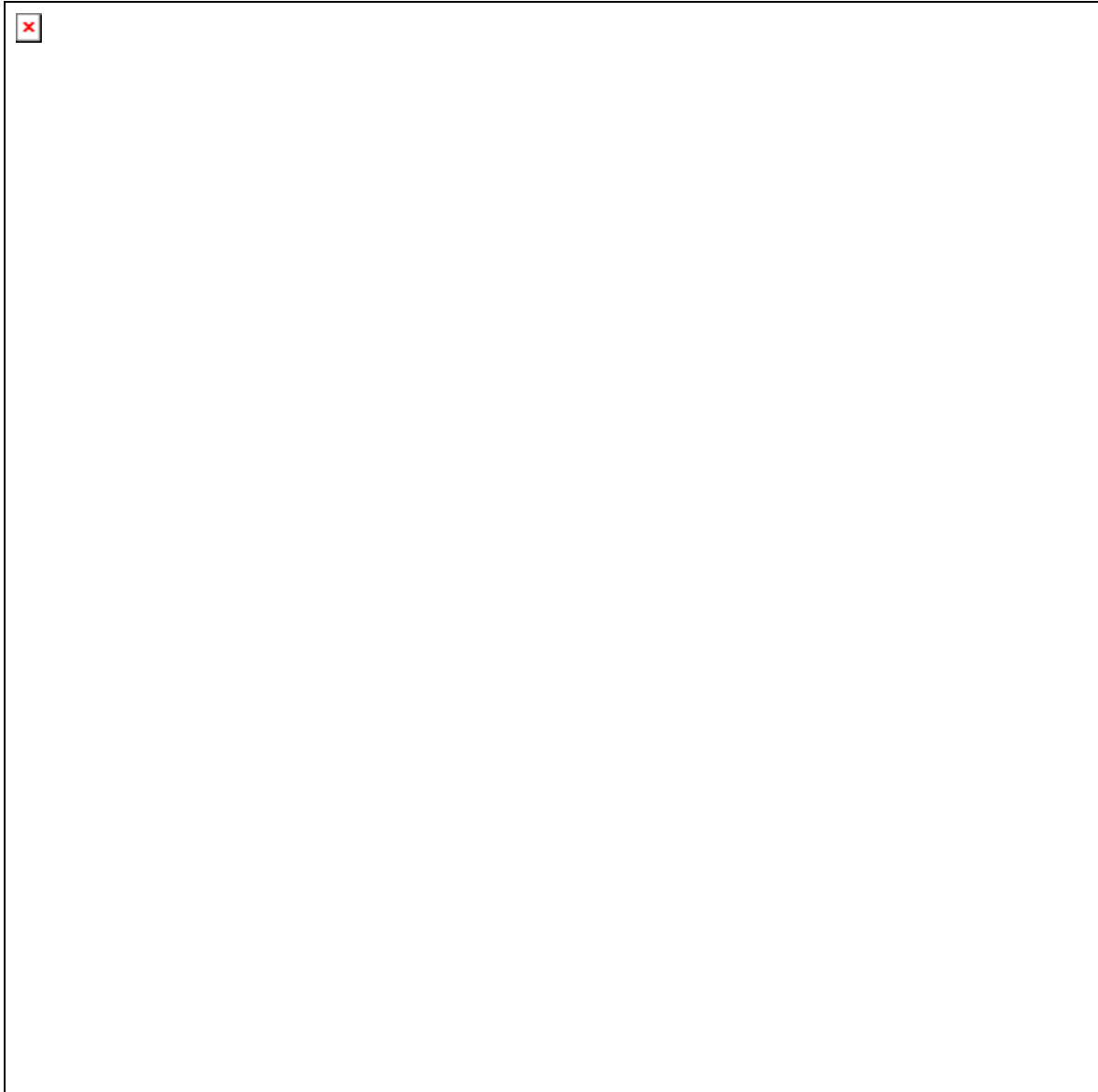
Methods for analysing binary data are much better developed and there is off-the-shelf software easily available (e.g. logistic regression). However, if we reduce 5 levels to 2, some information must be lost!

III. GOAL AND METHODS OF ANALYSIS

The aim of the studu is to find variables that significantly influence Cd infection (risk factors). We applied several statistical methods. Apart from some classical and standard methods, we also applied specialized algorithms for ordinal regression. The following methods have been used:

- Ordinary least squares regression (Degree of Cd infection considered as a *numeric* variable).
- Logistic regression (Degree of Cd infection reduced to a *binary* variable).
- Ordinal variable regression (Degree of Cd infection considered *ordinal* variable). Two approaches to ordinal type data have been applied:
 - “Cumulative risk model” due to Agresti (2002), implemented as R function `polr(MASS)`.
 - “Ordinal regression” introduced by Bobrowski (2007).

The figure below shows the ROC curves corresponding to several logistic functions obtained by the method of Agresti.



The analyses based on logistic regression and Agresti's method have found 6 most relevant features: "antybiot", "klinika", "antag.H2", "endosk", "chemia", "niewyd.nerek" (*antibiotic therapy, clinic, blockers of H2 receptors, endoscopy, chemotherapy, kidney failure*).

In the methodological part of this paper we focus on the last, nonstandard method, which was investigated by the authors in several papers (Niemiro and Rejchel 2009, Rejchel 2009, Rejchel 2014).

IV. ORDINAL REGRESSION

In this section we explain the foundations of the algorithm introduced by Bobrowski (2007) under the name of "rank regression". but Since this term has different connotations, we will use the name "ordinal regression", which is more appropriate in our opinion. Consider the statistical model with the following variables:

- y_i :- response variable measured in *ordinal* scale: $y_i > y_j$ means that object x_i is "better" or "bigger" than x_j with respect to y . We do not assign numeric values to y_i . Examples of such variable are y_i "severity of infection", y_i "degree of

improvement" and similar.

- \mathbf{x}_i - p -dimensional vector of predictors - either numeric or qualitative (e.g. coded in "0-1" scale).

Linear ordinal regression (scoring) function is $f(\mathbf{x}_i)$, where β is the regression parameter.

- The idea is to construct such a function $f(\mathbf{x}_i)$, that the ordering of its values is consistent with the ordering of the y_i -values.
 - Ideally, $f(\mathbf{x}_i)$ should imply that $f(\mathbf{x}_i)$ (function of y_i -ranks objects perfectly).
 - More realistically, we look for such a parameter β ; that $f(\mathbf{x}_i)$ implies that $f(\mathbf{x}_i)$ with "high probability" (function of y_i -ranks objects as correctly as possible with a margin).

Let us now describe a method of fitting ordinal regression. We assume that the data (learning sample) are of the following form:

- y_i - response variable observed for i -objects (e.g. patients) y_i .
- \mathbf{x}_i - vectors of predictors observed for these i -objects.

1. Hinge loss criterion

The ordinal regression function is obtained by minimization of the following criterion with the "hinge loss":

$$\sum_{i=1}^n \max(0, \beta y_i - \mathbf{w}^T \mathbf{x}_i) \quad (1)$$

This loss is 0 if $\mathbf{w}^T \mathbf{x}_i \geq \beta y_i$ implies $\mathbf{w}^T \mathbf{x}_i \geq \beta y_i + \text{margin}$, (linear function of y_i -ranks i s correctly with a margin). The details about this criterion are explained in Bobrowski (2007), Niemiro & Rejchel (2009), or Rejchel (2012). Let us mention that from this criterion is very convenient computationally. Since $\max(0, \beta y_i - \mathbf{w}^T \mathbf{x}_i)$ is a convex piecewise linear function, its minimum can be computed via basis exchange techniques (Bobrowski & Niemiro 1984). An implementation of this method as R function was developed by the authors. The graph below illustrates the *hinge loss* used in the definition of $\max(0, \beta y_i - \mathbf{w}^T \mathbf{x}_i)$.



2. Model selection via regularization

We expect that many features are irrelevant. Formally: some of regression coefficients β_j should be 0. To find zero coefficients, we add a “regularizer” to the loss function, i.e. modify the minimization problem as follows:

$$\min_{\beta} \sum_{i=1}^n (y_i - \beta^T x_i)^2 + \lambda \sum_{j=1}^p |\beta_j| \tag{2}$$

where λ is some weight (penalty for nonzero coefficient β_j). The idea is similar to LASSO estimators in ordinary least square regression, first proposed by Tibshirani (1996). The minimizer of the penalized loss, i.e. the solution to the above optimization problem will be denoted $\hat{\beta}$. Typically $\hat{\beta}$ has some coordinates zero.

3. Population and sample parameters

To understand the sense of the theoretical results to follow, we have to clarify the distinction between parameters which describe a population of objects and their sample counterparts. We need the following notation. The basic criterion function $\sum_{i=1}^n (y_i - \beta^T x_i)^2$ is computed for a *sample* of n objects: $\{(x_i, y_i)\}_{i=1}^n$. Note that the definition of $\sum_{i=1}^n (y_i - \beta^T x_i)^2$ can be rewritten in a slightly different form:

$$\sum_{i=1}^n (y_i - \beta^T x_i)^2 = \sum_{i=1}^n (y_i - \beta^T \bar{x}_i)^2 + n \beta^T \bar{y} \bar{x} - n \beta^T \bar{y} \bar{x} \tag{3}$$

The theoretical counterpart, describing the *population* of objects is the following:

$$\sum_{i=1}^n (y_i - \beta^T x_i)^2 \tag{4}$$

where x_1 and x_2 correspond to two objects independently selected from the population, (considered as random vectors). Therefore:

$$\begin{matrix} x_1 \\ x_2 \end{matrix} \quad (5)$$

4. Prediction and testing sample

Ordinal regression serves to *predict* ordering of new objects (testing sample):

- Consider objects x_1 and x_2 .
- Assume that x_1 and x_2 are observed (measured) but the ordering between x_1 and x_2 is unknown and should be predicted.
- If x_1 then we predict that x_2 .

Of course, prediction makes sense if the new objects are not included in the learning sample. The quality of prediction is measured by the probability of making the correct decision about the ordering:

$$x_1 \quad (6)$$

An estimator of x_1 should be based on a testing sample x_2 , collected independently of the learning sample x_1 :

$$x_1 \quad (7)$$

5. Oracle properties of the estimator

Assume that the population regression parameter β has zeros. As in the previous section, $\hat{\beta}$ denotes the minimizer of the penalized loss function. Define the following sets of coordinates:

- S is the set of “relevant explanatory variables”.
- \hat{S} is the set of variables estimated as relevant.

Let S and \hat{S} .

There are two properties of a “good” estimator/selector, often considered in the literature:

1. $\hat{S} \subseteq S$,
2. $\hat{\beta}_{\hat{S}} \rightarrow \beta_S$, where β_S is the standard asymptotic distribution of an “oracle” estimator based on prior knowledge of S .

The first property has rather obvious meaning. It is called “model consistency”. The second requirement is that the estimator should be asymptotically as good as the fictitious “oracle” estimator.

6. Some theoretical results

Rejchel (2014) has shown that the desired oracle properties hold if

$$x_1 \quad (8)$$

But x_1 is unknown, so the above condition cannot be applied. The question is “how to choose weights?”. An answer given by Rejchel (2014) is based on the idea of adaptive LASSO,

proposed by Zou (2006). We first use a preliminary estimator $\hat{\beta}_0$ and put $\hat{\beta}_0$. These weights are used in the second stage to compute the final estimate $\hat{\beta}_1$. [Rejchel, 2012] Assume $\hat{\beta}_0$ is a \sqrt{n} -consistent, $\hat{\beta}_0$ and $\hat{\beta}_1$. Then

$$\hat{\beta}_1 = \hat{\beta}_0 + \frac{1}{n} \sum_{i=1}^n \frac{y_i - \hat{\beta}_0^T x_i}{\|x_i\|^2} x_i \quad (9)$$

satisfies

1. $\hat{\beta}_1$ is a \sqrt{n} -consistent estimator of β .
2. $\hat{\beta}_1$ is a \sqrt{n} -consistent estimator of β .

(Under some technical “regularity” assumptions, which are omitted. The details and rigorous statement can be found in the cited paper.)

V. RESULTS OBTAINED FOR ORDINAL REGRESSION

The data consisting of n cases (one case was omitted, because it contained missing values) was divided into a learning (training) sample of size n_1 , a validating sample of size n_2 and testing sample of size n_3 . After a preliminary analyses we reduced the number of variables to p . The validating sample was used to choose the best parameter λ in the penalty term. The candidate values of λ were in the range $[\lambda_{\min}, \lambda_{\max}]$. Apart from the adaptive penalized estimation described in the previous section, regularized estimators with ℓ_1 and ℓ_2 type penalty were computed. On the testing sample, the quality of ordering prediction was assessed. The whole procedure (dividing the data into a training sample, validating sample, computing the regression and evaluating its predictive quality) was repeated 30 times. Below we report the averaged final results.

Estimator	Predictive quality (std. dev.)	Number of chosen features
without penalty term	0.695 (0.050)	21
ℓ_1 penalty	0.697 (0.039)	18.5
ℓ_2 penalty	0.706 (0.042)	21
adaptive estimator	0.692 (0.042)	17

The results obtained by this method are consistent with earlier, more classical analyses. In particular, 6 relevant features discovered by Agresti’s method have been always included in the sets of variables found by ordinal regression.

References

1. A. Agresti, *Analysis of ordinal categorical data*. Wiley & Sons (2010).
2. L. Bobrowski, Linear ranked regression – designing principles, *CORES’05, IV International Conference on Computer Recognition Systems, Advances in Soft Computing* (2007).
3. L. Bobrowski and W. Niemi, A method of synthesis of linear discriminant function in the case of nonseparability, *Pattern Recognition* (1984).
4. W. Niemi and W. Rejchel, Rank correlation estimators and their limiting distributions, *Statistical Papers* (2009).

5. S.M. Poutanen, A.E. Simor. Clostridium difficile-associated diarrhea in adults. *JAMC2004*; 171, 51-8.
6. W. Rejchel, Ranking - convex risk minimization. *World Academy of Science, Engineering and Technology* 56 (2009).
7. W. Rejchel, On Ranking and Generalization Bounds, *Journal of Machine Learning Research* (2012).
8. W. Rejchel, Model selection consistency of U-processes with convex loss and weighted Lasso penalty, submitted (2014).
9. R Development Core Team, *R: A Language and Environment for Statistical Computing*, 2011.
10. R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58, 267–288 (1996).

DIMENSIONALITY REDUCTION AND VISUALIZATION OF GENOMIC DATA

M. Ćwiklińska-Jurkowska¹, A. Wolińska-Welcz²

¹*Collegium Medicum, Bydgoszcz, Nicolaus Copernicus University, Toruń, Poland*

²*Maria Curie-Skłodowska University, Lublin, Poland*

Abstract

Microarrays data set is a collection of a few dozen or at most a few hundred multi-dimensional observations containing thousands of variables that measure the level of genes expression. Among these thousands of genes, only a small part is really active. In classification of microarrays we encounter a high-dimension problem. Thus, an important step in the data analysis is a proper selection of a subset of genes in such a way that significant medical information will not be lost.

Grade Correspondence Analysis (GCA) is a multidimensional method that makes use of multidimensional dependencies both between variables and cases. Therefore, the aim of the work is an assessment of usefulness of GCA for dimensionality reduction in comparison with commonly applied methods.

Two huge genomic data sets were analyzed. Whole sets were randomly divided into two subsets with almost equal number of patients. For each discriminating problem, the first subset was applied for a selection of the most discriminating genes. The second subset was used to assess generalization properties of the selected genes. Results obtained are really encouraging. The selected genes overlap to a large extent with the results of the other, more commonly used dimension reduction techniques. Additionally, classification errors assessed on the basis of selected subsets of genes are comparable.

Keywords: Grade Correspondence Analysis, genes selection, microarrays

I. INTRODUCTION

An important medical application of genomic data sets is separation of the sick from the healthy. A choice of a representative subset of genes eliminates information noise, and contributes to the development of medical knowledge. The results can be used in personalized medicine.

An overwhelming number of the known methods of selecting the most active genes are based on the one-dimensional analysis [1], though different methods for multiple tests are developed and selection procedures include corrections for multiple testing.

The Grade Correspondence Analysis (GCA) –a multidimensional method which takes into account dependencies between both variables and cases - was considered. The grade method gives the opportunity of data mapping and visualization. After reduction of dimensionality performed with the grade methodology, visualization of the discriminated patients together with the connection with important genes might be done.

II. METHODOLOGY

The grade (i.e. copula based) approach to multivariate data analysis and its application to explore different sets of clinical data was presented during the last Seminars on Statistics and Clinical Practice. The grade methodology gives an invaluable tool in medical research. Grade methods also turned out to be reliable for high-dimensional data and exploration of huge datasets just as we thought.

Principles of grade methods and models were thoroughly discussed in [2-3]. The main grade methods are: Grade Correspondence Analysis (GCA) and Grade Correspondence Cluster Analysis (GCCA). Both basic grade procedures GCA and GCCA may be applied to investigate and visualize multivariate additive (or approximately additive) datasets with m cases and k variables. GCA permutes rows and columns in a two-way table to achieve a table with the maximal value of Spearman rho (ρ) or of Kendall tau (τ) to emphasize the strongest and the most regular monotone dependence. GCCA performs cluster analysis of a two-way data table starting from GCA so that rows and columns are already ordered according to the observed trend of relationship discovered by maximization of ρ or τ . The numbers of clusters to be formed (for rows and columns or only for rows or only for columns) are *a priori* chosen by the searcher, while their sizes (numbers of elements inside) are found in GCCA by maximization ρ or τ in the table with aggregated clusters.

Datasets

Analysis was performed on two genomic data sets: Prostate data set [4] and Colon data set [5]. Each genomic data set additionally includes information of two groups of the patients' state. In the examined sets (Prostate and Colon data sets), tumor versus normal classification is considered. In the Prostate data set 52 tissues out of 102 are tumor tissues (i.e. 52 prostate tumor samples and 50 non-tumor, normal prostate samples). Expression levels of 6033 genes are reported. In the Colon data set 40 tissues out of 62 are colon tumor tissues and 22 are normal. Expression levels of 2000 genes are reported.

Whole sets (6033 variables x 102 patients and 2000 variables x 62 patients, respectively) were randomly divided into two subsets of approximately equal number of patients. Thus, P1 subset of the Prostate data set has 6033 genes and 51 patients (where 23 are tumor cases) and P2 subset consists of 6033 variables with 51 cases, with 29 tumor subjects. C1 subset of the Colon data set contains expression levels of 2000 genes and 32 patients (with 20 colon tissues) and C2 subset consists of 2000 variables and 30 cases, with 20 colon tissues. For each discriminating problem, the first subset P1 or C1, respectively, was applied for a selection of the most discriminating genes. The second independent set P2 or C2, respectively, was used to assess generalization properties of selected subsets of genes. P2 was divided into $k=10$ cross-validation subsamples and the 10 cross-validation error was calculated. C2 is similarly divided into $k=5$ cross-validation subsamples, because it contains smaller number of cases, namely 30 patients. Cross-validation into k folds was used to assess generalization errors.

Selection of variables methods

Many authors applied different selection of variables methods to solve the microarrays classification problem [6]. Grade methods and thirteen other methods for dimension reduction were performed (Tab.1).

Table 1

The methods for dimension reduction for microarrays.

<i>Selection method name</i>	<i>Description</i>
1. GCA	Grade Correspondence Analysis
2. PAM	Prediction Analysis of Microarrays
3. WilcoxonRanSum	Wilcoxon test
4. SAM	Significance Analysis of Microarrays
5. SAMnPar	Significance Analysis of Microarrays, Wilcoxon
6. PermutAdjPmaxT	permutation adjusted t-test
7. PermutAdjPmaxTWilcox	permutation adjusted t-test
8. BenHochberg	T-test with Benjamini-Hochberg correction for multiple tests
9. PredStength	Prediction strength
10. BetweenClassScatter	Between-class scatter
11. BetweenWithinRatio	Between-within ratio
12. InformGain	Information gain
13. TwoingRule	Twoing rule
14. Gini	Gini index

III. RESULTS

The grade exploration has been done to analyze and visualize genomic datasets. Both basic grade procedures GCA and GCCA have been applied to investigate datasets P1 and C1. After GCA ordering the learning set P1 turned out to be quite regular (grade parameters: $\rho^*_{\max}=0.36$, $\tau=0.24$), while the learning set C1 – completely irregular and difficult for analysis ($\rho^*_{\max}=0.06$, $\tau=0.04$). Next an advanced ordering grade procedure has been applied - only for genes of two connected segments of patients (sick and healthy). Then grade decomposition with division set P1 into 100 ordered and possible homogeneous clusters of genes and C1 into 50 ordered clusters have been done. The selected 13 diagnostic genes (from P1) and 35 genes (from C1) to discriminate between the sick and the healthy have been chosen from the first and the last clusters obtained after an application of GCCA procedure only for variables.

The post GCA maps for selected genes in learning and in testing sets (with their grade parameters) are presented in Fig.1-2 on the coloured appendix Results of Multivariate Grade Data Analysis in Genes Selection for both investigated *multivariate* genomic datasets.

IV. DISCUSSION

Comparisons of genes selected by GCA with other examined reduction dimensionality methods

For the Prostate dataset the agreement of the chosen genes sets for 100, 48 and 13 variables was examined for all methods of dimensionality reduction from Table 1.

The intersection of genes sets chosen by GCA and other examined selection procedures were investigated. Concordance of 100 genes chosen by GCA and other selection procedures is presented in Fig.3 The highest concordance of about 40% is obtained with Prediction Analysis of Microarrays (PAM) or BetweenClassScatter method. In subset of 48 genes chosen by GCA the concordance about 40% is also obtained (Fig.4).

Next, smaller number of 13 genes are obtained according to GCA analysis: 5016, 2694, 1989, 2746, 4255, 2443, 2425, 4448, 1839, 5662, 5639, 1607, 1897 (Fig.5). All these 13 genes are in subset of 48 genes and 12 coincide with variables chosen to subset of 100 genes (variables: 1897, 1607, 1839, 4448, 2425, 5662, 2443, 4255, 2746, 2694, 1989, 5016).

For the smallest number of 13 genes chosen by GCA Venn diagrams are elaborated, which show the number of overlapping genes chosen by 5 methods. Fig. 6 on the left side shows number of overlapping genes obtained by GCA with InfGain, between-within ratio (BW), TwoingRule, and Gini. Similarly, right side of this figure visualize concordance of 13 genes chosen by GCA with PAM, Significance Analysis of Microarrays (SAM), adjusted t-test (PermutT) and Benjamini-Hochberg procedure (BenjHoch).

For 13 genes obtained by GCA, 6 variables are overlapping with TwoingRule. These are genes identified in the set by numbers: 1839, 1989, 2694, 2746, 4255, 5016. Similarly, we obtained four genes overlapped with PAM and those genes are the following: 1839, 2425, 2746, 5016. Three genes with the identifiers 1839, 2425, 5016 are concordant with PermutT. Two genes with identifiers 1839 and 2425 are the same as obtained with BenjHoch.

The applied supervised classification was SVM with $c=1$. For a subset of 15 genes, selected on the basis of GCA, the errors obtained by CV10 are compared with the same classifier, but the selection of genes will be used by popular PAM method of selection. Similarly, an adequate comparison for SVM $c=2$ was made (Fig.7). CV10 estimation was performed on 51 microarrays, remaining after the exclusion from the whole set other 51 microarrays, used previously to the stage of the genes selection. After the division of CV, nine subsets in the cross-validate contains 5 microarrays and one subset is consisted of 6 microarrays. Comparison of the number of misclassified microarrays from the testing cross-validated sample is shown in Fig.7. The results are presented in an aggregated form for the genes chosen by GCA and PAM methods. For the Prostate set, classification results by SVM with the regularization parameters $c=1$ and $c=2$, yielded results comparable to the PAM method for the number of genes 7-12, but the results were better for smaller number of genes. The first four selected genes according to GCA are: 5016, 2425, 4448 and 1839. Further analysis may take advantage of the genes found with the lowest level of errors in the supervised classification. The databases of the functions of genes (genes ontology data bases) may be used for biomedical interpretation.

For the Colon dataset, 100 genes from GCA are considered in comparison to other genes number reduction methods and the results obtained are presented in Fig.8. PAM selection is now concordant in about 50%. A similar agreement is obtained for BetweenClassScatter selection, and even is higher (or at least equal) for number of genes exceeding 50.

Another set obtained by GCA methodology has been also examined - the smaller set of 50 variables. However, the set is not a subset of 100 genes, because only 47 genes are the same. Almost 50% accordance is obtained at the end of the curve (50 genes) for PAM procedure (solid line on Fig.9). The highest concordance for PAM is obtained between 20 and 30 genes – almost 60%. Also SAM method is concordant with GCA selection in more or less 50%.

For chosen 34 genes concordance of about 60% is obtained, depending on the increasing power of subsequent subsets. Above 60% concordance between GCA with PAM is achieved for 16 genes (Fig.10).

V. CONCLUSION

The selected genes subsets overlap to a large extent with the results of the other, more common used dimension reduction techniques. Most frequently, the PAM method from whole examined set of variables selection method is the most concordant with GCA.

Additionally, classification errors assessed on selected subsets of genes are comparable.

The advantage of GCA applied as genes selection method over the known genes selection procedure is that after the dimension reduction, the visualization of discriminated patients together with the connection with important genes might be done. The visualization may be given for both learning and testing subsets, where the appropriateness of selection might be assessed by testing sample. The line for clusters of patients may be added to the plot and based on this visualization, the division misclassifications into tumor and normal cases might be done, so specificity and sensitivity is possible to take into account in medical decision-making.

Acknowledgements

The Authors would like to thank Professor Elżbieta Pleszczyńska - a distinguished Polish statistician and the other founders of the grade methodology.

References

1. Boulesteix A.-L., Strobl C., Augustin T., Daumer M.: Evaluating Microarray-based Classifiers: An Overview. *Cancer Inform.* 2008; 6, 77–97.
2. Szczesny W., Kowalczyk T., Wolińska-Welcz A., Wiech M., Dunicz-Sokolowska A., Grabowska G., Pleszczyńska E.: *Models and Methods of Grade Data Analysis: Recent Developments*, IPI PAN, Warszawa, 2012.
3. Kowalczyk T., Pleszczyńska E., Ruland F.[Eds.], *Grade Models and Methods for Data Analysis. With Applications for the Analysis of Data Populations*. Berlin, Springer Verlag, 2004.
4. Singh D., et al., Gene expression cor635 relates of clinical prostate cancer behavior, *Cancer Cell* 1 2002, 2, 203 –209.
5. Alon U.: Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays. *PNAS*. 1999, 96, 6745-6750.
6. van Sanden S., Lin D., Burzykowski T.: Performance of Gene Selection and Classification Methods in a Microarray Setting: A Simulation Study. *Communications in Statistics - Simulation and Computation* Volume 37, Issue 2, 2008.

Results of multivariate grade data analysis in genes selection



Fig.1. The post-GCA maps for 51 patients (learning group-above, testing group-below) and 13 genes chosen from 6033. The values of grade density are determined by colours according to the blue/purple scale given at the right side of the map. Dark arrow-shaped markers are attached to the sick. Parameters of monotone dependence: $\rho^*_{\max}=0.45$, $\tau=0.31$, the regularity index $\tau_{\max}/\tau_{\text{abs}}=0.68$ (for upper map) and $\rho^*_{\max}=0.43$, $\tau=0.29$, the regularity index $\tau_{\max}/\tau_{\text{abs}}=0.62$ (for bottom map)

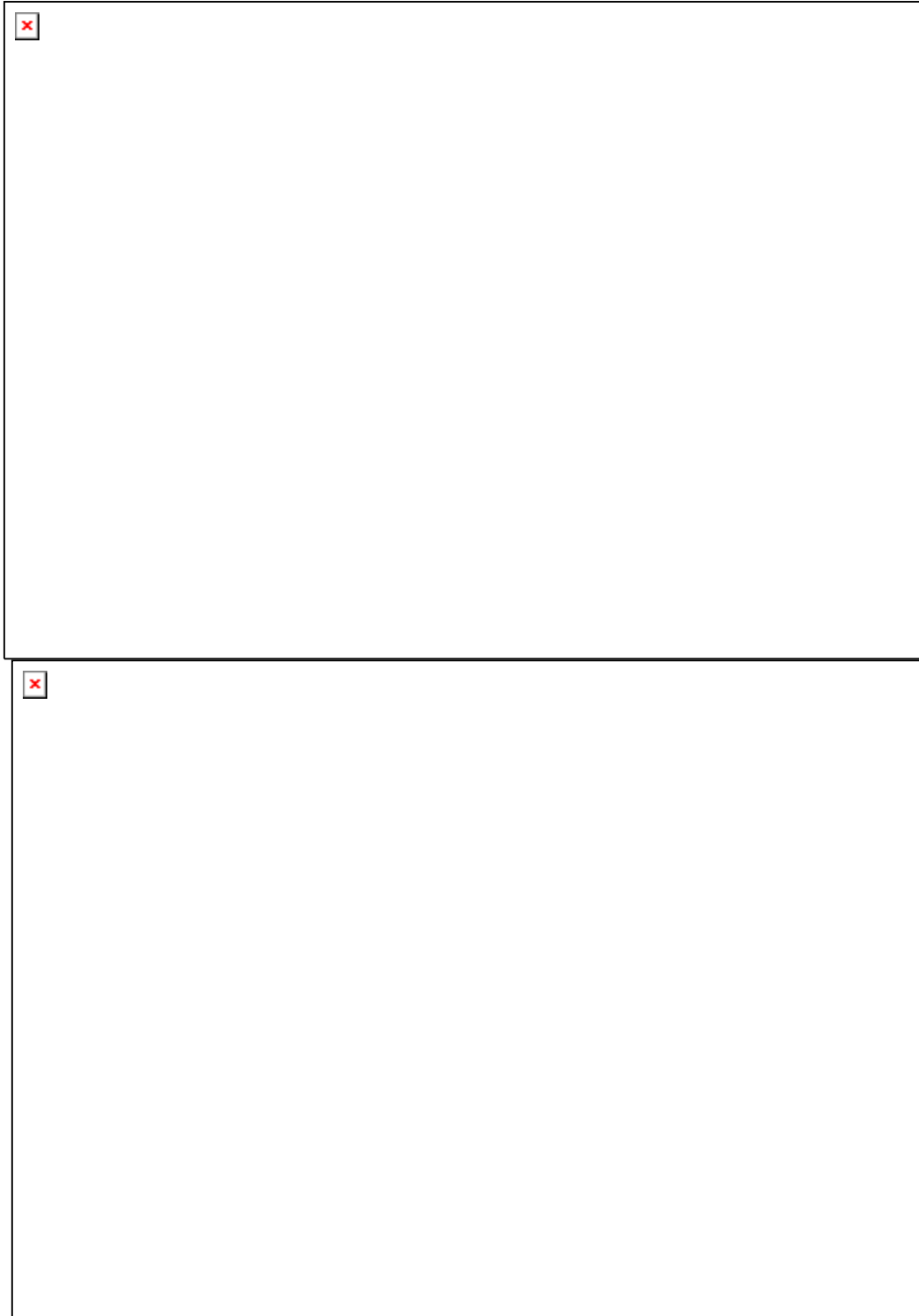


Fig.2. The post-GCA maps for 32 (learning group-above), 30 (testing group-below) patients and 35 genes chosen from 2000. The values of grade density are determined by colours according to the orange/green scale given at the right side of the map. Dark arrow-shaped markers are attached to the sick. Parameters of monotone dependence: $\rho^*_{\max}=0.15$, $\tau=0.10$, the regularity index=0.57 (for upper map) and $\rho^*_{\max}=0.13$, $\tau=0.09$, the regularity index =0.54 (for bottom map).

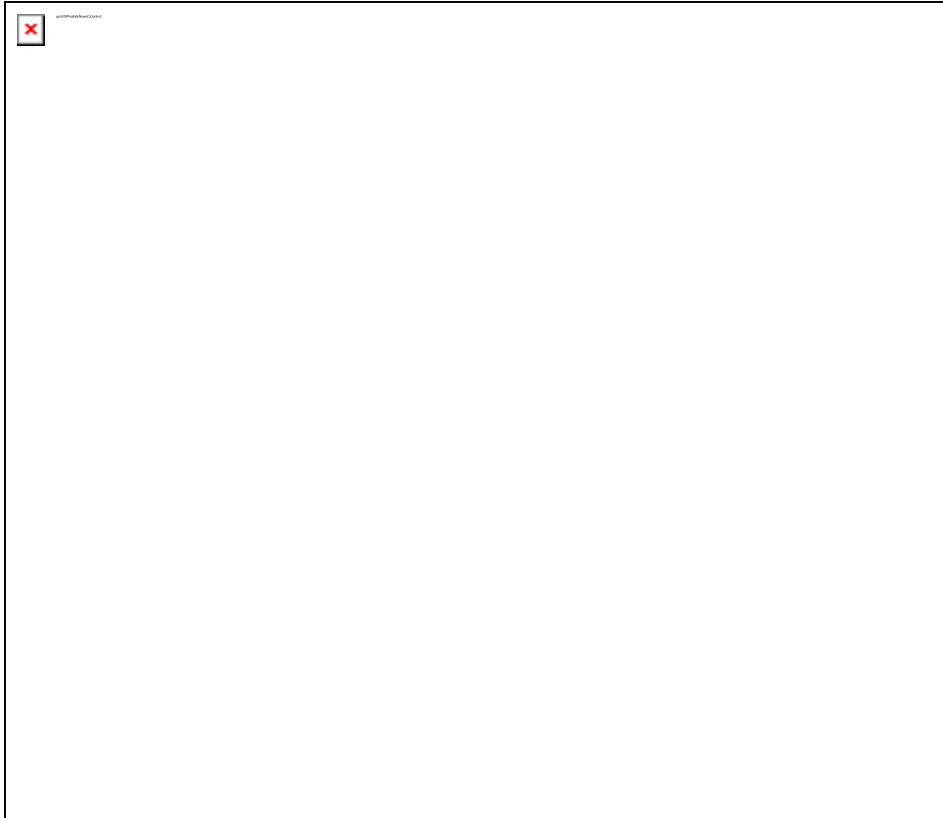


Fig.3. Number of common genes for 100 variables from P1 set chosen by GCA and other selection procedures.



Fig.4. Number of common genes for 48 variables from P1 set chosen by GCA and other selection procedures.

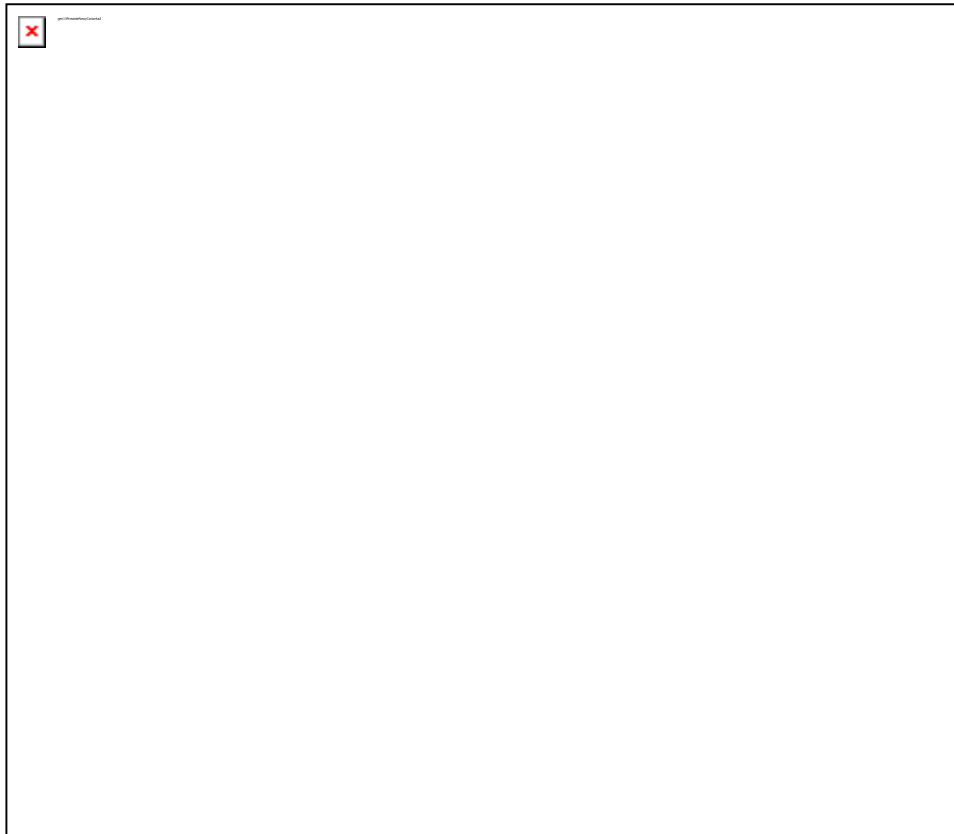


Fig.5. Number of common genes for 13 variables from P1 set chosen by GCA and other selection procedures.

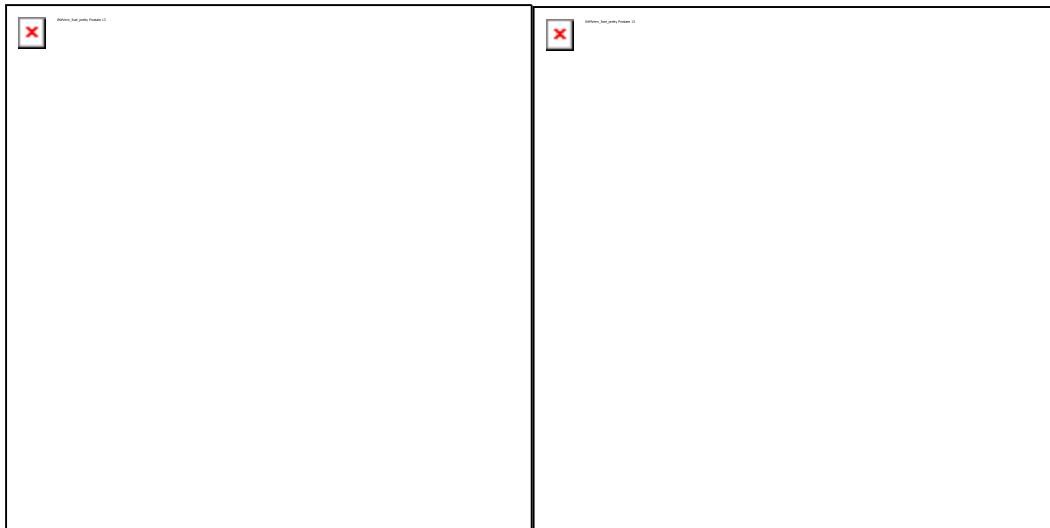


Fig.6. Venn diagram for comparison of 13 genes from P1 set selected by GCA with the highest ranked genes by InfGain, between-within ratio (BW), TwoingRule, and Gini (left side) and PAM, SAM, permutation adjusted t-test and Benjamini Hochberg procedure (right side).



Fig.7. Assessed number of misclassified microarrays by SVM (parameters $c=1$ and $c=2$, respectively) for genes selected by PAM and GCA.



Fig.8. Number of common genes for 100 variables from set C1 chosen by GCA and other selection procedures.

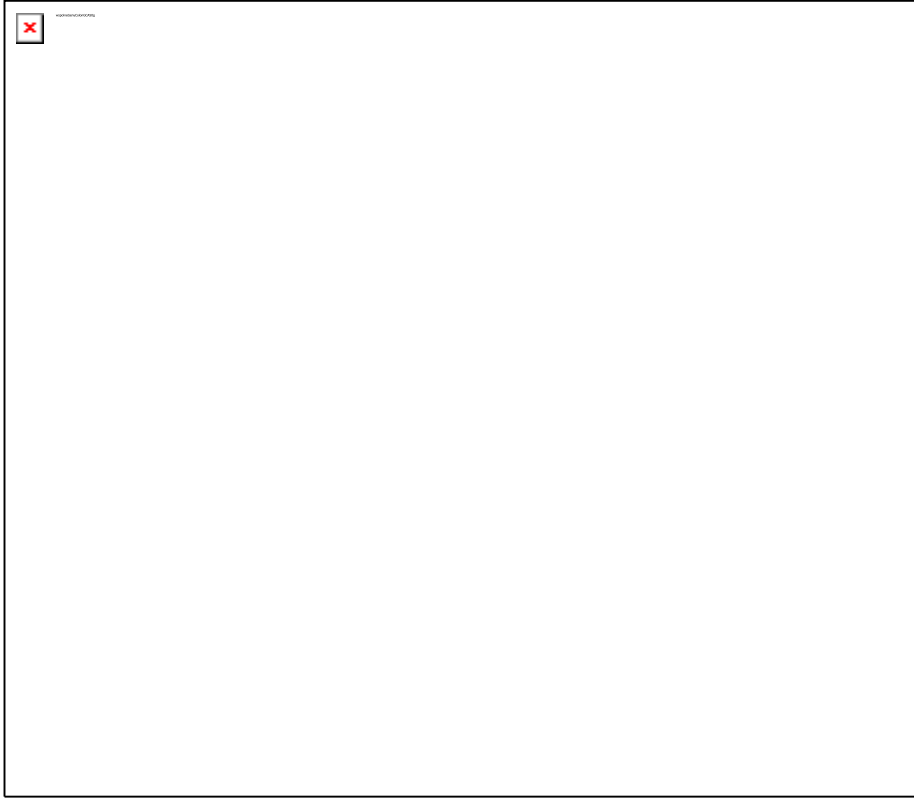


Fig.9. Number of common genes for 50 variables from set C1 chosen by GCA and other selection procedures.

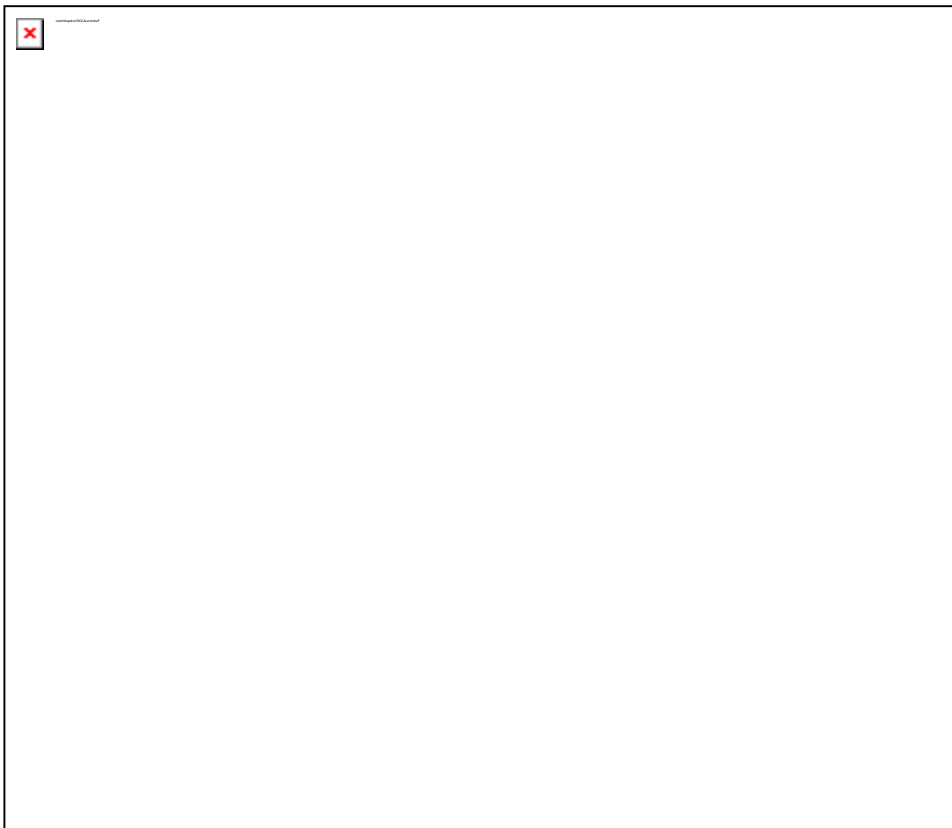


Fig.10. Number of common genes for 35 variables from set C1 chosen by GCA and other selection procedures.

MODELING UNCERTAIN MEDICAL KNOWLEDGE WITH BAYESIAN NETWORKS: ENGINEERING AND APPLICATIONS

A. Oniśko^{1,2}

¹*Faculty of Computer Science, Białystok University of Technology, Białystok, Poland*

²*Magee-Womens Hospital, University of Pittsburgh Medical Center, Pittsburgh, USA*

Abstract

In the last two decades Bayesian networks have proven to be powerful tools for modeling complex uncertain problems, such as those encountered in medical domains. Knowledge engineering for constructing Bayesian networks includes combining data coming from different sources, for example, subjective expert opinion, clinical data, screening data, histopathologically verified data, or the data coming from the questionnaires. This paper discusses knowledge engineering for constructing Bayesian network models along with the examples of these models for diagnostic and prognostic problems in medicine.

Keywords: Bayesian networks, knowledge engineering, medical diagnosis, medical prognosis

I. INTRODUCTION

The support of medical diagnosis and prognosis by computer-based tools has a long history with the first approaches proposed in the 1960s and 1970s (e.g., [1,2]). The medical support systems developed during last decades were based on various approaches. Probabilistic graphical models, such as Bayesian networks, have proven to be powerful tools for modeling complex uncertain knowledge. Bayesian network modeling has found its applications in both medical diagnosis and prognosis. There are quite a few examples of Bayesian network models developed to solve medical problems (e.g., [3,4,5,6]). Knowledge engineering for building Bayesian networks includes knowledge elicitation from domain experts, transforming medical data into the framework of acyclic directed graph as well as, combining data coming from different sources such as subjective expert opinion with objective data. This paper discusses knowledge engineering for constructing Bayesian network models along with the examples of these models for diagnostic and prognostic problems in medicine.

II. METHODOLOGY

Bayesian networks [7] are acyclic directed graphs that allow for modeling probabilistic dependencies and independencies among variables. The graphical part of a Bayesian network reflects the structure of a modeled problem, while local interactions among neighboring variables are quantified by conditional probability distributions. The structure of the directed graph represents a factorization of the joint probability distribution. For example, a Bayesian network encoding n variables: X_1, X_2, \dots, X_n , has the following factorization:

$$P(X_1, X_2, \dots, X_n) = \prod_{(i=1,2,\dots,n)} (X_i / Pa(X_i))$$

where $Pa(X_i)$ indicates parent variables of X_i . Reasoning with Bayesian networks consists of a calculation of a posteriori probability for a target node given the information entered into observed nodes. This probability can be farther interpreted as a probability of developing a disease given observed evidence. Bayesian networks can reflect expert's understanding of the

domain, enhance interaction with a human expert at the model building stage, and are readily extendible with new information.

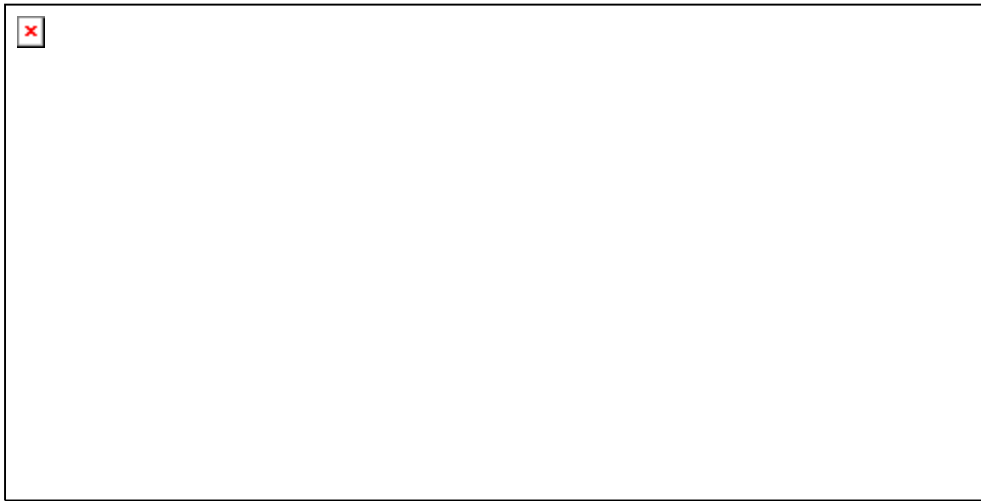


Figure 2: Example of a Bayesian network model.

Figure 1 captures an example of a Bayesian network that models a risk factor of cervical cancer and two tests used in cervical cancer screening: HPV and Pap tests. Each of the arcs in the graph represents a probabilistic relationship between the connected variables. Figure 1 contains also the probability tables for two nodes: a prior probability distribution for the node *Age* and a conditional probability distribution for the node *Pap test*. Assuming that *Cervical cancer* is a target node, we can calculate a posterior probability of developing cervical cancer given observed evidence that in this case is the information about a specific patient introduced into the nodes *Age*, *HPV test*, and *Pap test*. The resulting posterior probability can be further interpreted as a risk of developing a cervical cancer.

There exists a temporal extension of Bayesian networks, dynamic Bayesian network, offering a framework for explicit modeling of temporal relationships and is useful as both prognostic and diagnostic tool.

III. RESULTS

Knowledge engineering for building Bayesian networks includes knowledge elicitation from domain experts, transforming and incorporating the data coming from different sources into the framework of acyclic directed graph and quantifying it by means of conditional probability distributions. The data sources can include, for example, expert subjective opinion, clinical data, screening data, histopathologically verified data, or the data coming from the questionnaires. The knowledge engineer, who is responsible for building the model, has to be aware what is the data from and how to interpret the data.

Building a Bayesian network model is an iterative process that consists of four main elements (see Figure 2). The process begins with selecting the variables of a model. Then the qualitative and quantitative parts of a Bayesian network model are constructed. The fourth crucial element of the process includes model verification and evaluation.

Qualitative and quantitative part of a Bayesian network can be built based on the expert knowledge, i.e., a graphical structure of the model along with its numerical parameters can be assessed based on the expert opinion. This task requires much effort on the knowledge engineer part that is responsible for acquiring the knowledge from the expert. Another method to build the network is based on a hybrid approach, where the structure of the model is assessed by the expert and the numerical parameters are learned from the data. Yet another

approach to build a Bayesian network is to learn it automatically from the available data. There exist the algorithms to learn Bayesian network models from the data. In any of the approaches knowledge engineer plays an important role. Knowledge engineer should not only be trained in techniques that facilitate the process of knowledge acquisition but also should develop basic knowledge of the domain to establish a common language with experts.

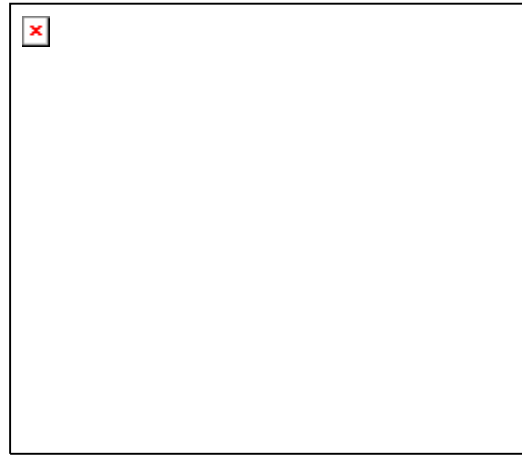


Figure 3: Iterative character of model building.

Bayesian networks are often used in modeling diagnostic problems. We have built several diagnostic models. For example, HEPAR II, a model for diagnosis of liver disorders or the BPH model for diagnosis of benign prostatic hyperplasia [8]. To build these models we have used a hybrid approach, i.e., a graphical part of the model was elicited based on expert opinion while the conditional probability distributions were learned from collected clinical data. Another diagnostic Bayesian network model was built to determine whether adenocarcinoma present in a biopsy or curettage is of endometrial or endocervical origin [9]. To build this model we have used immunohistochemical profile data. Yet another diagnostic model, AutismNET, was constructed based on expert knowledge, i.e., the graphical structure and numerical parameters were elicited from domain experts. The AutismNET model supports early diagnosis of autism and is dedicated to parents [10].

We have developed the Pittsburgh Cervical Cancer Screening Model for risk assessment. The model is a dynamic Bayesian network and was built based on the cervical cancer screening data collected over the period of 11 years. This prognostic model allows for individualized management of patients and computes patient-specific risk based on the patients characteristics, history data, and screening test results [11]. Another prognostic model was built based on screening, histopathological, and clinical data [12]. The model allows us to predict a risk of atypical endometrial hyperplasia and endometrial carcinoma. The aim of the model was to limit the number of performed endometrial biopsies based on clinical and screening data.

An important part of Bayesian network modeling is also discovering probabilistic relationships from data. Given data coming from the questionnaires we have built a Bayesian network that models risk factors and effects of dental caries in three year old children. The model includes over 30 variables and identifies risk factors of dental caries [13]. For example, the model has confirmed that cleaning teeth is a leading factor in preventing dental caries in three year old children.

IV. CONCLUSION

Bayesian networks are powerful tools for modeling uncertain and complex medical knowledge. These models offer a framework for explicit modeling of probabilistic relationships and are useful as both prognostic and diagnostic tools. Knowledge engineering for building Bayesian network models includes transforming and incorporating the knowledge from experts and existing data into the framework of probabilistic graphical model. Knowledge engineer is responsible for this elaborate and time consuming process. Furthermore, a knowledge engineer should be familiar not only with the techniques that facilitate the process of knowledge acquisition from the domain expert or the data, but should also develop basic knowledge of the domain to understand it and to establish a common language with experts.

Acknowledgements

All Bayesian network models that I referred to in this paper were created and tested using SMILE, an inference engine, and GeNIe, a development environment for reasoning in graphical probabilistic models, both developed at the Decision Systems Laboratory and available at <http://www.bayesfusion.com>.

References

1. R. S. Ledley and L. B. Lusted. Reasoning foundations of medical diagnosis. *Science*, 130(3366):9–21, July 1959.
2. G. Anthony Gorry. Computer-assisted clinical decision-making. *Methods of Information in Medicine*, 12:45–51, 1973.
3. Gregory F. Cooper. NESTOR: A Computer-based Medical Diagnostic Aid that Integrates Causal and Probabilistic Knowledge. PhD thesis, Stanford University, Computer Science Department, 1984.
4. I.A. Beinlich, H.J. Suermondt, R.M. Chavez, and G.F. Cooper. The ALARM monitoring system: A case study with two probabilistic inference techniques for belief networks. In *Proceedings of the Second European Conference on Artificial Intelligence in Medical Care*, pages 247–256, London, 1989.
5. F. J. Díez, J. Mira, E. Iturralde, and S. Zubillaga. DIAVAL, a Bayesian expert system for echocardiography. *Artificial Intelligence in Medicine*, 10:59–73, 1997.
6. Peter J. F. Lucas, Linda van der Gaag, and Ameen Abu-Hanna. Bayesian networks in biomedicine and health-care. *Artificial Intelligence in Medicine*, 30:201–214, 2004.
7. Pearl J.: *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers, Inc., San Mateo, CA, 1988.
8. Agnieszka Oniśko. Medical Diagnosis, In Patrick Naim, Olivier Pourret, and Bruce Marcot (eds), *Bayesian Networks: A Practical Guide to Applications*, Wiley & Sons, pages: 15-32, 2008.
9. Mirka W. Jones, Agnieszka Oniśko, David J. Dabbs, Esther Elishaev, Rohit Bhargava. Immunohistochemistry and HPV in situ hybridization in distinction between endocervical and endometrial adenocarcinoma: A comparative tissue microarray study of 76 tumors, *International Journal of Gynecological Cancer*, 23(2):380-4, 2013.
10. Justyna Szczygieł, Agnieszka Oniśko, Jolanta Świdorska, Elżbieta Krysiwicz, Jerzy Sienkiewicz. Probabilistic graphical model supporting early diagnosis of autism spectrum disorder, *Advances in Computer Science Research*, No. 11, pages: 151-164, 2014.

11. R. Marshall Austin, Agnieszka Oniśko, Marek J. Druzdzel. The Pittsburgh Cervical Cancer Screening Model. A Risk Assessment Tool. *Arch Pathol Lab Med.* 134:744–750, 2010.
12. Jing Yu, Agnieszka Oniśko, and R. Marshall Austin. Bethesda System Reporting of Benign-Appearing Endometrial Cells in Women 40 and Older: Analysis of Predictive Value from a Large Academic Women's Hospital Database, United States and Canadian Academy of Pathology's 105th Annual Meeting, March 12-18, 2016, Seattle, WA, USA.
13. Wojciech Łaguna. Probabilistic models in discovering risk factors of dental caries in three-year-old children. Master thesis (in Polish). Faculty of Computer Science, Białystok University of Technology, July 2014.

BICLUSTERING AS EXTRACTION OF COLLINEAR PATTERNS

L. Bobrowski^{1,2}

¹*Faculty of Computer Science, Bialystok University of Technology, Poland*

²*Institute of Biocybernetics and Biomedical Engineering, PAS, Warsaw, Poland*

Abstract.

Data mining techniques based on minimization of the convex and piecewise linear (CPL) criterion functions can be used among others for extraction of collinear (flat) patterns from large, multidimensional data sets. New method of biclustering can be also developed by using this technique. Properties of such collinear biclustering are analyzed in the presented paper.

Keywords: data mining, flat patterns, CPL criterion functions, biclustering

I. INTRODUCTION

Clustering techniques are used in data mining tasks to extract *patterns* from large, multidimensional data set [1]. An extracted pattern is expected to have a form of subset (*cluster*) of feature vectors characterized by a certain type of regularity. *Biclustering* procedures should allow to extract not only clusters of feature vectors but also subsets of features specific for a particular pattern [2]. Biclustering techniques are developed intensively at present for the purpose of genomic data analysis.

We are considering biclustering aimed at extraction of collinear patterns in selected feature subspaces. The collinear (*flat*) pattern can be observed if a large number of feature vectors from the given data set is located on a such hyperplane. Collinear patterns can be extracted by omission of some feature vectors (objects) from the data set combined with neglecting certain features x_i from the feature space [3].

II. DATA SET AND DUAL HYPERPLANES IN PARAMETER SPACE

Let us the data set $C[n]$ contains m feature vectors $\mathbf{x}_j[n] = [x_{j,1}, \dots, x_{j,m}]^T$ belonging to a given n -dimensional feature space $F[n]$ ($\mathbf{x}_j[n] \in F[n]$):

$$C[n] = \{\mathbf{x}_j[n]\}, \text{ where } j = 1, \dots, m \quad (1)$$

Components x_{ji} of the feature vector $\mathbf{x}_j[n]$ can be treated as the numerical results of n standardized examinations of a given object (patient) O_j ($x_{ji} \in \{0, 1\}$ or $x_{ji} \in R$).

Each of m feature vector $\mathbf{x}_j[n]$ from the set $C[n]$ (1) defines the below dual hyperplane h_j in the parameter space R^n ($\mathbf{w}[n] \in R^n$):

$$(\forall \mathbf{x}_j[n] \in C[n]) \quad h_j = \{\mathbf{w}[n]: \mathbf{x}_j[n]^T \mathbf{w}[n] = 1\} \quad (2)$$

Each of n unit vector $\mathbf{e}_i[n] = [0, \dots, 1, \dots, 0]^T$ in the n -dimensional feature space $F[n]$ ($\mathbf{e}_i[n] \in F[n]$) defines the below dual hyperplane h_i^0 in the parameter space R^n :

$$(\forall i \in \{1, \dots, n\}) \quad h_i^0 = \{\mathbf{w}[n]: \mathbf{e}_i[n]^T \mathbf{w}[n] = 0\} = \{\mathbf{w}[n]: w_i = 0\} \quad (3)$$

The set S_k of n_k ($1 \leq n_k \leq n$) feature vectors $\mathbf{x}_{j(i)}[n]$ ($j(i) \in J_k$) and $n - n_k$ unit vectors $\mathbf{e}_{i(i')}[n]$ ($i(i') \in I_k$) allows to define the below matrix $\mathbf{B}_k[n]$:

$$\mathbf{B}_k[n] = [\mathbf{x}_{j(1)}[n], \dots, \mathbf{x}_{j(n_k)}[n], \mathbf{e}_{i(n_k+1)}[n], \dots, \mathbf{e}_{i(n)}[n]]^T \quad (4)$$

If the matrix $\mathbf{B}_k[n]$ (4) is nonsingular, then it is called the k -th *basis*. The basis $\mathbf{B}_k[n]$ allows to compute the k -th *vertex* $\mathbf{w}_k[n]$ through the below equation []:

$$\mathbf{w}_k[n] = (\mathbf{B}_k[n]^T)^{-1} \mathbf{1}[n] \quad (5)$$

where the vector $\mathbf{1}[n] = [1, \dots, 1, 0, \dots, 0]^T$ has the components equal to 1 or to 0 adequately to the feature vector $\mathbf{x}_{j(i)}[n]$ ($j(i) \in J_k$) or to the unit vector $\mathbf{e}_{i(i')}[n]$ ($i(i') \in I_k$).

The vertex $\mathbf{w}_k[n]$ (5) is the intersection point of the hyperplanes h_j (2) and h_i^0 (3) defined by elements $\mathbf{x}_{j(i)}[n]$ ($j(i) \in J_k$) and $\mathbf{e}_{i(i')}[n]$ ($i(i') \in I_k$) of the set S_k (4) [].

Remark 1: Each of n_k dual hyperplane $h_{j(i)}$ (2) defined by the basic feature vector $\mathbf{x}_{j(i)}[n]$ ($j(i) \in J_k$) (4) passes through the vertex $\mathbf{w}_k[n]$ (5):

$$(\forall j(i) \in J_k (4)) \quad \mathbf{w}_k[n]^T \mathbf{x}_{j(i)}[n] = 1 \quad (6)$$

Remark 2: Each of $n - n_k$ dual hyperplane h_i^0 (3) defined by the basic unit vector $\mathbf{e}_{i(i')}[n]$ ($i(i') \in I_k$) (4) passes through the vertex $\mathbf{w}_k[n]$ (5):

$$(\forall i(i') \in I_k (4)) \quad \mathbf{w}_k[n]^T \mathbf{e}_{i(i')}[n] = 0 \quad (7)$$

Each component $w_{k,i(l)}$ of the vertex $\mathbf{w}_k[n] = [w_{k,1}, \dots, w_{k,n}]^T$ (5) linked to the unit vector $\mathbf{e}_{i(l)}[n]$ in the basis $\mathbf{B}_k[n]$ (4) is equal to zero ($w_{k,i(l)} = 0$) as it results from (7).

Definition 1: The *rank* r_k ($1 \leq r_k \leq n$) of the vertex $\mathbf{w}_k[n] = [w_{k,1}, \dots, w_{k,n}]^T$ (5) is defined as the number of such components $w_{k,i}$ which are different from zero ($w_{k,i} \neq 0$).

Definition 2: The vertex $\mathbf{w}_k[n_k]$ (5) is degenerated if and only if more than r_k dual hyperplanes h_j (2) passes through this vertex in the parameter space R^{n_k} .

Definition 3: The *degree of degeneration* d_k of the vertex $\mathbf{w}_k[n_k]$ (5) is defined in the below manner:

$$d_k = m_k - r_k \quad (8)$$

where m_k is the number of such feature vectors $\mathbf{x}_j[n]$ from the data set $C[n]$ (1), which define the hyperplanes h_j (2) passing through this vertex ($\mathbf{w}_k[n]^T \mathbf{x}_j[n] = 1$).

III. CONVEX AND PIECEWISE LINEAR (CPL) CRITERION FUNCTION

The *CPL* penalty functions $\varphi_j(\mathbf{w}[n])$ are defined on the feature vectors $\mathbf{x}_j[n]$ ($\mathbf{x}_j[n] \in F[n]$) from the data set C (1) [3]:

$$\begin{aligned} (\forall \mathbf{x}_j[n] \in C (1)) \quad \varphi_j(\mathbf{w}[n]) &= |1 - \mathbf{w}[n]^T \mathbf{x}_j[n]| = \begin{cases} 1 - \mathbf{w}[n]^T \mathbf{x}_j[n] & \text{if } \mathbf{w}[n]^T \mathbf{x}_j[n] \leq 1 \\ \mathbf{w}[n]^T \mathbf{x}_j[n] - 1 & \text{if } \mathbf{w}[n]^T \mathbf{x}_j[n] > 1 \end{cases} \end{aligned} \quad (9)$$

The k -th criterion function $\Phi_k(\mathbf{w}[n])$ is determined as the weighted sum of the penalty functions $\varphi_j(\mathbf{w}[n])$ defined on the feature vectors $\mathbf{x}_j[n]$ from the data subset $C_k[n]$ (1):

$$\Phi_k(\mathbf{w}[n]) = \sum_{i \in J_k} \beta_i \varphi_i(\mathbf{w}[n]) \quad (10)$$

where $J_k = \{j: \mathbf{x}_j[n] \in C_k[n] \subset C[n] \text{ (1)}\}$ and the positive parameters β_i ($\beta_i > 0$) in the below function $\Phi_k(\mathbf{w}[n])$ can be treated as the *prices* of particular feature vectors $\mathbf{x}_j[n]$. The standard choice of the parameters β_j values is one ($\beta_j = 1.0$).

It can be proved that the minimal value of the convex and piecewise linear criterion function $\Phi_k(\mathbf{w}[n])$ (10) can be found in one of the vertices $\mathbf{w}_k^*[n]$ (5):

$$(\exists \mathbf{w}_k^*[n]) (\forall \mathbf{w}[n]) \Phi_k(\mathbf{w}[n]) \geq \Phi_k(\mathbf{w}_k^*[n]) = \Phi_k^* \geq 0 \quad (11)$$

The basis exchange algorithms which are similar to the linear programming allow to find efficiently the minimal value $\Phi_k(\mathbf{w}_k^*[n])$ of the criterion functions $\Phi_k(\mathbf{w}[n])$ (10) even in the case of large, multidimensional data subsets $C_k[n]$ [4].

The hyperplane $H(\mathbf{w}[n], \theta)$ in the feature space $F[n]$ is defined as follows:

$$H(\mathbf{w}[n], \theta) = \{\mathbf{x}[n]: \mathbf{w}[n]^T \mathbf{x}[n] = \theta\} \quad (12)$$

where $\mathbf{x}[n]$ is the feature vector ($\mathbf{x}[n] \in F[n]$), $\mathbf{w}[n]$ is the *weight vector* ($\mathbf{w}[n] \in R^n$) and θ is the *threshold* ($\theta \in R^1$).

Theorem 1: The minimal value $\Phi_k(\mathbf{w}_k^*[n])$ (11) of the criterion function $\Phi_k(\mathbf{w}[n])$ defined (10) on elements $\mathbf{x}_j[n]$ of the subset $C_k[n]$ ($C_k[n] \subset C[n]$ (1)) is equal to the zero ($\Phi_k(\mathbf{w}_k^*[n]) = 0$), if and only if all the feature vectors $\mathbf{x}_j[n]$ from this subset are situated on some hyperplane $H(\mathbf{w}[n], \theta)$ (12) with $\theta \neq 0$.

Proof: Let us suppose that all the feature vectors $\mathbf{x}_j[n]$ from the subset $C_k[n]$ are situated on the hyperplane $H(\mathbf{w}'[n], \theta')$ (12) with $\theta' \neq 0$:

$$(\forall \mathbf{x}_j[n] \in C_k[n]) \quad \mathbf{w}'[n]^T \mathbf{x}_j[n] = \theta' \quad (13)$$

From this

$$(\forall \mathbf{x}_j[n] \in C_k[n]) \quad (\mathbf{w}'[n] / \theta')^T \mathbf{x}_j[n] = 1 \quad (14)$$

The above equations mean that functions $\varphi_j(\mathbf{w}'[n] / \theta')$ (9) are equal to zero in the point $(\mathbf{w}'[n] / \theta')$:

$$(\forall \mathbf{x}_j[n] \in C_k[n]) \quad \varphi_j(\mathbf{w}'[n] / \theta') = 0 \quad \text{so (10)} \quad (15)$$

$$\Phi_k(\mathbf{w}'[n] / \theta') = 0 \quad (16)$$

On the other hand, if the criterion function $\Phi_k(\mathbf{w}'[n])$ (10) is equal to the zero in some point $\mathbf{w}'[n]$, then each of the penalty functions $\varphi_j(\mathbf{w}'[n])$ (9) has to be equal to zero:

$$(\forall \mathbf{x}_j[n] \in C_k[n]) \quad \varphi_j(\mathbf{w}'[n]) = 0 \quad (17)$$

or

$$(\forall \mathbf{x}_j[n] \in C_k[n]) \quad \mathbf{w}'[n]^T \mathbf{x}_j[n] = 1 \quad (18)$$

The above equations mean that each feature vector $\mathbf{x}_j[n]$ from the subset $C_k[n]$ is located on the hyperplane $H(\mathbf{w}'[n], 1)$ (12). \square

Remark 3: If all the feature vectors $\mathbf{x}_j[n]$ from the subset $C_k[n]$ ($C_k[n] \subset C[n]$ (1)) are located on the hyperplane $H(\mathbf{w}'[n], \theta')$ (12) with $\theta' \neq 0$, then the minimal value $\Phi_k(\mathbf{w}_k^*[n])$ (12) ($\Phi_k(\mathbf{w}_k^*[n]) = 0$) is located in the optimal vertex $\mathbf{w}_k^*[n] = \mathbf{w}'[n] / \theta'$.

The above *Remark* can be justified on the basis of the proof of the *Theorem 1*.

IV. VERTEXICAL PLANES AND LINES IN FEATURE SPACE

The k -th *vertexical plane* in the feature space $F[n]$ is defined by using the basic feature vectors $\mathbf{x}_{j(i)}[n]$ belonging to the basis $\mathbf{B}_k[n]$ (4) which is linked (5) to the vertex $\mathbf{w}_k[n]$ [3]:

$$P_k(\mathbf{x}_{j(1)}[n], \dots, \mathbf{x}_{j(n_k)}[n]) = \{\mathbf{x}[n]: \mathbf{x}[n] = \alpha_1 \mathbf{x}_{j(1)}[n] + \dots + \alpha_{n_k} \mathbf{x}_{j(n_k)}[n]\} \quad (19)$$

where the n_k parameters α_i ($\alpha_i \in \mathcal{R}^1$) fulfill the below condition:

$$\alpha_1 + \dots + \alpha_{n_k} = 1 \quad (20)$$

Remark 4: The dimension of the plane $P_k(\mathbf{x}_{j(1)}[n], \dots, \mathbf{x}_{j(n_k)}[n])$ (19) is equal to $n_k - 1$.

Remark 5: The vertexical plane $P_k(\mathbf{x}_{i(1)}[n], \dots, \mathbf{x}_{i(n)}[n])$ (18) with the n basic vectors $\mathbf{x}_{i(i)}[n]$ is such hyperplane $H(\mathbf{w}_k[n], 1)$ in the n -dimensional feature space $F[n]$ which can be defined by the equation (12) with the k -th vertex $\mathbf{w}_k[n]$ (5) and the threshold $\theta = 1$.

Remark 6: None of the *vertexical planes* $P_k(\mathbf{x}_{j(1)}[n], \dots, \mathbf{x}_{j(n_k)}[n])$ (19) passes through the point zero $\mathbf{0}[n]$ (*origin*).

Theorem 2: The feature vector $\mathbf{x}_j[n]$ ($\mathbf{x}_j[n] \in F[n]$) defines such dual hyperplane h_i (2) which passes through the vertex $\mathbf{w}_k[n]$ (5) supporting the vertexical plane $P_k(\mathbf{x}_{j(1)}[n], \dots, \mathbf{x}_{j(n)}[n])$ (19) if and only if the vector $\mathbf{x}_j[n]$ is situated on this plane.

The proof of a similar *Theorem* can be found in the paper [3]. The *Theorem 2* gives the link between the degeneracy (*Definition 2*) of the vertex $\mathbf{w}_k[n]$ (5) and the location of feature vectors $\mathbf{x}_j[n]$ on the vertexical plane $P_k(\mathbf{x}_{j(1)}[n], \dots, \mathbf{x}_{j(n)}[n])$ (19).

V. EXTRACTION OF COLLINEAR BICLUSTERS

The minimal values $\Phi_k(\mathbf{w}_k^*[n])$ (11) of the criterion functions $\Phi_k(\mathbf{w}[n])$ defined (10) on elements $\mathbf{x}_j[n]$ of the subsets $C_k[n]$ ($C_k[n] \subset C[n]$ (1)) have the below property [3]:

The monotonicity property: (21)

The removal of such a feature vector $\mathbf{x}_j[n]$ from the data subset $C_k[n]$ which is characterized by the positive value $\varphi_j(\mathbf{w}_k^*[n])$ of the penalty function $\varphi_j(\mathbf{w}[n])$ (9) in the optimal vertex $\mathbf{w}_k^*[n]$ (11) causes a decrease of the minimal value Φ_k^* (11) to $\Phi_{k'}^*$:

$$\Phi_k^* - \Phi_{k'}^* \geq \varphi_j(\mathbf{w}_k^*[n]) > 0 \quad (22)$$

where the symbol $\Phi_{k'}^*$ stands for the minimal value (11) of the criterion function $\Phi_{k'}(\mathbf{w}[n])$ (10) defined on the elements $\mathbf{x}_j[n]$ of the reduced set $C_k[n] / \mathbf{x}_j[n]$.

A gradual removal of $m_{k'}$ feature vectors $\mathbf{x}_{j'}[n]$ with the highest values $\varphi_{j'}(\mathbf{w}_k^*[n])$ (21) from the data subset $C_k[n]$ allows to form the reduced data subset $C_{k'}[n]$:

$$C_{k'}[n] = C_k[n] / \sum_{j'} \mathbf{x}_{j'}[n] = \{\mathbf{x}_j[n]: j \in J_{k'}\} \quad (23)$$

where $J_{k'} (J_{k'} \subset \{1, \dots, m\})$ is a subset of the m_k indices j of the feature vectors $\mathbf{x}_j[n]$.

The number $m_{k'}$ of the neglected feature vectors $\mathbf{x}_{j'}[n]$ should be sufficient to achieve the below condition (11):

$$\min \Phi_k(\mathbf{w}[n]) = \Phi_{k'}(\mathbf{w}_k^*[n]) = \Phi_k^* = 0 \quad (24)$$

We can infer on the base of the *Theorem 1*, that the condition (24) results in a location of all vectors $\mathbf{x}_j[n]$ from the reduced set $C_{k'}[n]$ on the hyperplane $H(\mathbf{w}_k^*[n], 1)$ (12):

$$C_{k'}[n] = C_k[n] / \sum_{j'} \mathbf{x}_{j'}[n] = \{\mathbf{x}_j[n]: j \in J_{k'}\} \quad (25)$$

where $J_{k'} (J_{k'} \subset \{1, \dots, m\})$ is a subset of the m_k indices j of the feature vectors $\mathbf{x}_j[n_k]$.

$$(\forall \mathbf{x}_j[n] \in C_{k'}[n]) \quad (\mathbf{w}_k^*[n])^T \mathbf{x}_j[n] = 1 \quad (26)$$

The minimization (11) of the criterion functions $\Phi_{k'}(\mathbf{w}[n])$ (10) defined on elements $\mathbf{x}_j[n]$ of the subsets $C_{k'}[n]$ (23) allows to find the vector $\mathbf{w}_k^*[n] = [w_{k,1}^*, \dots, w_{k,n}^*]^T$ (11). The optimal vertex $\mathbf{w}_k^*[n]$ of the rank r_k (*Definition 1*) can be used to the identification of the r_k - dimensional feature subspace $F_{k'}[r_k] \subset F[n]$:

The k' -th feature subspace $F_{k'}[r_k]$ is composed of such r_k features x_i ($i \in I_{k'}$) which are linked to the such optimal weights $w_{k,i}^*$ (11) which are not equal to zero ($w_{k,i}^* \neq 0$). The reduced feature vectors $\mathbf{x}_j[r_k]$ ($\mathbf{x}_j[r_k] \in F_{k'}[r_k]$) are obtained from the feature vectors $\mathbf{x}_j[n]$ ($\mathbf{x}_j[n] \in F[n]$) (1) through neglecting of the $n - r_k$ components $x_{j,i}$ linked to the weights $w_{k,i}^*$ equal to zero ($w_{k,i}^* = 0$) similarly as in the *RLS* method of feature selection [4]:

$$(\forall i \in \{1, \dots, n\}) \quad (27)$$

$w_{k,i}^* = 0 \Rightarrow$ the component $x_{j,i}$ is reduced in all m feature vectors

$\mathbf{x}_j[n] = [x_{j,1}, \dots, x_{j,n}]^T$ from the data set $C[n]$ (1) and
the i -th feature x_i is reduced from the feature space $F[n]$

Definition 5: The set $C_{k'}[r_k] = \{\mathbf{x}_j[r_k]: j \in J_{k'}\}$ (25) of the m_k reduced feature vectors $\mathbf{x}_j[r_k]$ has the form of *collinear bicluster* of the rank r_k if and only if each of these vectors $\mathbf{x}_j[r_k]$ is located on the vertexical plane $P_{k'}(\mathbf{x}_{j(1)}[r_k], \dots, \mathbf{x}_{j(nk)}[r_k])$ (19), (20).

If the number m_k of the elements $\mathbf{x}_j[n_k]$ of the bicluster $C_k[n_k]$ is a sufficiently high, then these elements form the *collinear (flat)* pattern in the feature subspace $F_{k'}[r_k]$. Detection of flat patterns on the plane and in three-dimensional space has a rich tradition in computer vision. The Hough transformation is traditionally used for this purpose [6].

Acknowledgments: This work was supported by the project St/2016 grant from the Nałęcz Institute of Biocybernetics and Biomedical Engineering, Polish Academy of Sciences

References

1. Hand D., Smyth P. and Mannila H.: *Principles of data mining*, MIT Press, Cambridge (2001).

- Madeira S. C., Oliveira S. L.: Biclustering Algorithms for Biological Data Analysis: A Survey, *IEEE Transactions on Computational Biology and Bioinformatics* **1** (1): 24–45 (2004).
2. Bobrowski, L.: Discovering main vertexical planes in a multivariate data space by using *CPL* functions, *ICDM 2014*, Ed. Perner P., Springer Verlag, Berlin 2014
 3. Bobrowski, L.: Design of piecewise linear classifiers from formal neurons by some basis exchange technique, *Pattern Recognition*, 24(9), pp. 863-870 (1991).
 4. Bobrowski, L., Łukaszuk, T.: Relaxed Linear Separability (*RLS*) Approach to Feature (Gene) Subset Selection, pp. 103 – 118 in: *Selected Works in Bioinformatics*, Edited by: Xuhua Xia, *INTECH* (2011).
 5. Duda, O. R., Hart, P. E.: Use of the Hough Transformation to Detect Lines and Curves in Pictures, *Communications of Association for Computing Machinery*, 15(1):11-15, 1972.

A FRAMEWORK FOR COMBINING UNSUPERVISED AND SUPERVISED LEARNING PROCEDURES

P. Munro

University of Pittsburgh, Pittsburgh PA, USA

Abstract

A general learning rule, "BCM- δ ", is proposed that subsumes both unsupervised learning as a form of the BCM rule (Bienenstock, Cooper, Munro, 1982; Munro, 1984) and the delta rule (Rosenblatt, 1958; Rumelhart, Hinton, and Williams, 1986). The "BCM- δ " unit is composed of two subunits, T and L , each integrating distinct input streams across distinct sets of synapses. The two subunits follow a common Hebb-like learning procedure that reduces to an unsupervised rule for the T subunit and a supervised rule for the L subunit in which the T response is the training signal. This model suggests a neurally plausible mechanism for the shaping of concepts by labels.

Keywords: concept learning, connectionism, neural model

I. INTRODUCTION

Supervised Learning using the Delta Rule

Error driven synaptic learning rules are typically written in a "Hebb-like" form with a postsynaptic factor that has a positive (target) term and a negative (response) term. The *delta rule* (eg., Rosenblatt, 1960; Widrow and Hoff 1960; Rumelhart, Hinton, and Williams, 1986) has this property; see Eq. (1). Here, w_{ij} is the weight of the synapse connecting stimulus s_j to unit i , and the postsynaptic factor δ_i is expressed as the difference between the desired response d_i and the linear response r_i ; i.e. $r_i = \sum w_{ij}s_j$. The step size or learning rate is signified by η .

$$\Delta w_{ij} = \eta \delta_i s_j, \text{ where } \delta_i \equiv d_i - r_i \quad (1)$$

Unsupervised Learning using the BCM Rule

Bienenstock, Cooper, and Munro (1982) developed a synaptic modification rule to describe the development of ocular dominance and orientation selective cells in visual cortex. Like the delta rule, the BCM rule has a Hebb-like form. A modified version by Munro (1984) is given in Eq. (2) in terms of two bounded monotone increasing functions σ^+ and σ^- , where σ^+ increases more slowly than σ^- ; the precise condition is given in Munro (1984) as a theorem.

$$\begin{aligned} \Delta w_{ij} &= \eta_w \varphi_i s_j \\ \Delta q_i &= \eta_q (r_i^2 - q_i) \end{aligned} \quad (2)$$

where $\varphi_i(r, q) \equiv \sigma^+(r) - q\sigma^-(r)$

Oppositional Mechanisms

Both the unsupervised rule in Eq. (2) and the delta rule (1) are specific cases of a more general framework for synaptic learning; see Eq. (3), in which the postsynaptic factor is the difference between two terms.

$$\Delta w_{ij} = \eta (P_i - N_i) s_j \quad (3)$$

The function suggests that there are *separate oppositional associative mechanisms* for strengthening synapses and weakening synapses, resulting in LTP when $P_i > N_i$ and LTD otherwise. In the case of the delta rule, P_i is the training signal and N_i is the response, while in Equation 4, both terms are functions of the response r_i .

II. A UNIFIED FRAMEWORK

Consider a hypothetical neuron with multiple loci for accumulating PSPs from different parts of the dendritic complex. Here, two subunits labeled T and L are stimulated by different sets of stimuli \mathbf{s}^T and \mathbf{s}^L which are incident on the cell on separate sets of afferents, with corresponding synaptic efficacy vectors \mathbf{w}^T and \mathbf{w}^L .

The two subunits compute separated weighted sums r^T and r^L (Eq 4). The pyramidal cell type is a prime candidate for this kind of unit (Fig 1).

$$r_i^{(X)} = \sum_{j \in X} w_{ij}^{(X)} s_j^{(X)} \quad \text{where } X \in \{T, L\} \quad (4)$$

In this section, a *self-supervised* learning rule is presented. The T subunit synapses are trained according to the version of BCM in equation (2). The partial response from the T unit drives the P term for the L subunit learning procedure; thus, the L follows a form of the delta rule with r^T determining the training signal for L . The self-supervised procedure operates by selecting a specific region of the T stimulus space. Subsequently, the L subunit is trained to give a partial response that can be interpreted as predictive of the preferred T stimulus given the L stimulus.

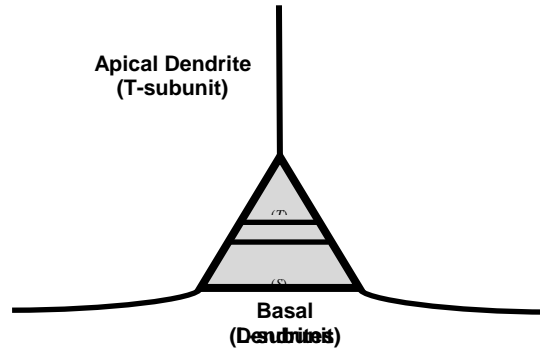


Fig 1: The pyramidal cell morphology as a model for the self-supervised framework. The two hypothetical partial responses could be integrated on mutually exclusive sets of afferents, such as those the apical dendrites and the basal dendrites.

$$\begin{aligned} \Delta w_{ij}^{(X)} &= \eta_w (\sigma(r_i^{(T)}, h_1) - q_i^{(T)} \sigma(r_i^{(X)}, h_2)) s_j^{(X)} \\ \Delta q_i^{(T)} &= \eta_q (r_i^{(T)} - q_i^{(T)}) \end{aligned} \quad (5)$$

Both sets of weights have the same P term driven by the T subunit and have N terms that are functions of their respective partial responses (Equation 5). This dynamical system suggests that a single synaptic modification based on oppositional mechanisms can subsume both unsupervised selectivity across a set of stimuli which can in turn drive a supervised learning procedure.

III. SIMULATIONS

Two scenarios are simulated: without T inputs and with T inputs. The L stimulus space is meant to simulate some primitive sensory space like visual space. The scenarios are therefore meant to compare concept learning with and without language.

Pattern Sets

The five patterns to the T subunit are nonorthogonal but linearly independent -- see Fig 2 (left). Input patterns to the L subunit are drawn from a set of 500 patterns in 10 clusters of 50 patterns. The first two principal components of the 500 patterns are shown in Fig 2 (right). The prototype vectors (larger circle) are randomly generated, with small random numbers added to each component.

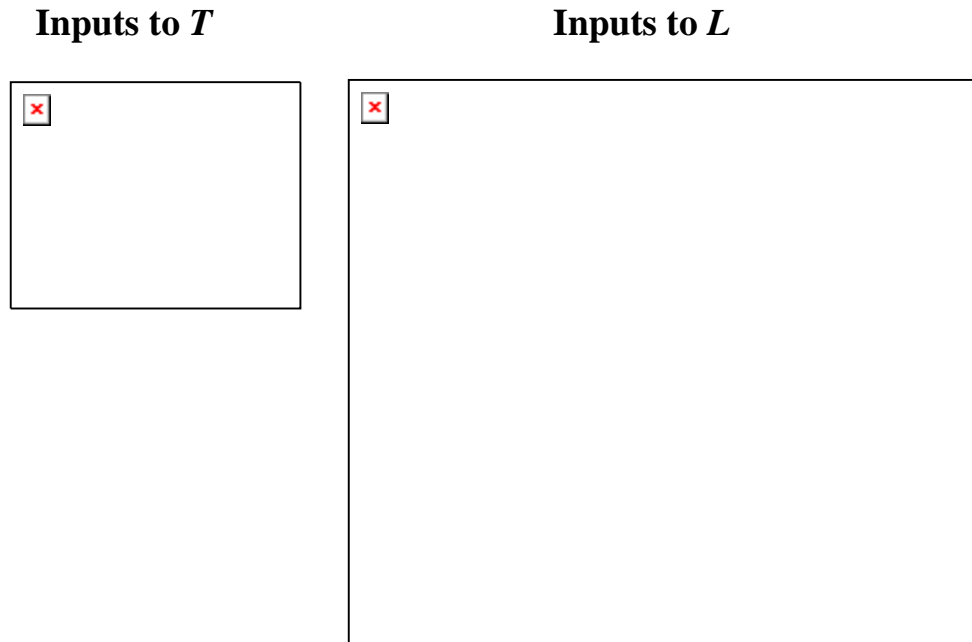


Fig 2: A PCA plot of the 10 clusters. The large spots are the prototype vectors. The color coding refers to the pairing with the five patterns from the T stimulus space for Experiment 2.

Experiment 1. L inputs alone.

In the first set of simulations the L subunit is trained without input to the T subunit. In every case, the subunit becomes responsive only to patterns from one of the 10 clusters. This demonstrates clustering by the unsupervised learning rule acting alone. Within cluster responses are tight, and responses to the selected cluster are well separated from the other clusters (Fig 3).

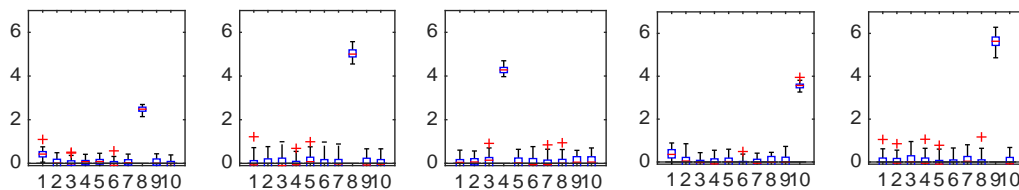


Fig 3: The box plot shows L subunit response profiles of 5 separately trained units to 20 patterns generated independently of the training set in each of 10 clusters *without* input to the T subunit.

Experiment 2. Joint input to both T and L

Here, inputs are presented to the T pattern set. The learning rule drives the T -subunit to choose one of the five patterns. Each pattern in this set is statistically correlated with two of the clusters from the L pattern set. These disjoint clusters are assigned independently to the T subunit patterns, and each "category" consists of two disjoint clusters with no similarity structure among the prototypes. The five pairs of clusters are color coded in Fig 2 (right). In Fig 4, the results of 5 simulations are displayed. The ordering in the Figure is pairwise to

highlight the responses of the simulated cell to the T pattern (green bar graph) and the associated clusters.

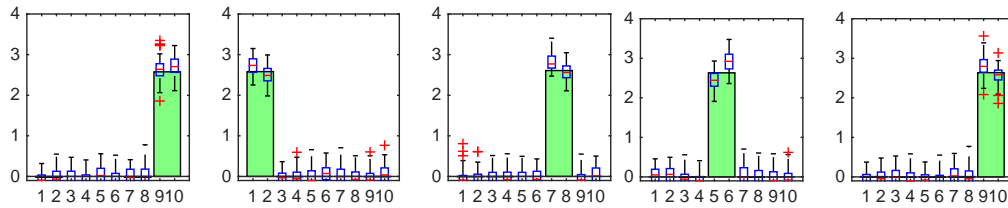


Fig 4: The box plots show L subunit response profiles of 5 separately trained units to 20 patterns generated independently of the training set in each of 10 clusters *with* input to the T subunit. Note that each cell becomes responsive to patterns from two of the clusters. The green bargraph shows the response to the 5 patterns from the T subunit. Note that one pattern of the 5 is chosen from the set, corresponding to the two associated clusters.

IV. DISCUSSION

This paper has demonstrated a system by which stimuli from one modality can shape the response properties of a unit to another modality using a framework that is biologically plausible and gives clues to the source of a teaching signal for supervised learning. This may lead to a neuron level explanation of the process by which language shapes concept formation.

Acknowledgements

The author would like to acknowledge insights of many former colleagues and mentors and the support of the University of Pittsburgh and the Center for the Neural Basis of Cognition.

References

1. E. L. Bienenstock, L. N Cooper, and P. W. Munro (1982) Theory for the Development of Neuron Selectivity: Orientation Specificity and Binocular Interaction in Visual Cortex. *Journal of Neuroscience*. 2:32-48.
2. P. W. Munro (1984) A Model for Generalization and Specification by Single Neurons. *Biological Cybernetics*. 51:169-179.
3. Rosenblatt, F. (1958) The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological Review*, 65:386-408.
4. Rumelhart, D., Hinton, G., Williams, R. (1986) Learning internal representations by backpropagation. In: Rumelhart, D. E. and McClelland, J. L., editors, *Parallel Distributed Processing, Explorations in the Microstructure of Cognition*. Vol. 1 MIT Press, Cambridge MA.
5. Widrow, B. and Hoff, M. (1960) Adaptive switching circuits. Western Electronic Show and Convention, Institute of Radio Engineers, 4, 96-104.
6. Zipser, D. (1986) Feature discovery by competitive learning. In: Rumelhart, D. E. and McClelland, J. L., editors, *Parallel Distributed Processing, Explorations in the Microstructure of Cognition*. Vol. 1 MIT Press, Cambridge MA.

POSTER SESSION

COMBINING TWO CORRELATED VARIABLES INTO ONE FACTOR: AN APPLICATION TO OBESITY MEASUREMENTS

I. Chmiel¹, M. Górkiewicz¹

¹ *Faculty of Health Sciences, Jagiellonian University of Krakow (Poland)*

Abstract

Practical methodology for combining BMI and WHR into single variable with use of principal component analysis was proposed. Exemplary analyses were based on sample of N=92 children aged 14-15 years.

Keywords: principal components, Spearman, Kendall, overweight, BMI, WHR

I. INTRODUCTION

Development of strategies for counteraction against overweight, particularly among children, constitutes challenging problems in public health, [1, 2]. Besides, prospective study, [3], completed at N=2895 participants, supported hypotheses that waist to hip ratio (WHR) and waist to stature ratio (WSR) can be better risk factors for some illness than body mass index (BMI). Then, it is well-known that the BMI and WHR are correlated very significantly, [4]. For all these reasons, the current study was concerned on developing practical methodology for proper measuring obesity at children. In consequence it was proposed to combine BMI and WHR into a single variable with use of principal component analysis.

Principal component analysis is well-known technique for replacing original variables with smaller number derived variables, named principal components, [5]. For data-sets drawn from 2-dimensional normal distribution all computations are particularly uncomplicated, they can be made with any calculator or spreadsheet. In real-world circumstances some complications can arise with non-linear relationships and weighty departures from normality, [6]. However, with respect to departures from 2-dimensional normality, the difficulty can be pass over with use of rank correlation methods, [7], with proper caution, [8]. Tests of significance of correlation coefficients based on the ranks can be computed with formulas (1) and (2). Both statistics are approximately standard normal for large samples.

$$Z^2 = (N - 2)R^2 / (1 - R^2) \quad (1)$$

$$Z^2 = T^2(4.5N(N - 1)) / (2N + 5) \quad (2)$$

where: Z – standard normal variable; N – number of pairs of data; R , T – Spearman and Kendall coefficient of rank correlation, respectively, [7].

The next difficulties arise with psychological problems connected with obesity, [1, 9, 10]. Consequently, in current study focus was put also on self-esteem and allusions from environment.

II. METHODOLOGY

Participants of this study, N=92 children, included N = 35 boys and 57 girls were recruited at 2013 year in two schools in Krakow, Poland. The only inclusion criterion was

age, 14-15 years. All candidates invited to study didn't refuse. Body mass index (BMI) and waist-hip ratio (WHR) were measured by trained health professionals. Participants reported on their behavior and adherence to weight control using questionnaire, briefly presented in Table 1., [11].

item	variable	Likert scale
1	Gender	1=female; 2=male
3	Dwelling-place	1=town; 2=country
4	School	1=ordinary; 2=sporting
7	Self-esteem	1=No, I'm too underweight; 2=No, I'm somewhat underweight; 3=No, normal; 4=Yes, overweight; 5=Yes, too overweight
8	Contentment	1=Yes, I like my silhouette; 2=No, I don't
9	Anxiety	1=Yes, I'm anxious about my weight ; 2=No, I don't
10	Fitness	1=Yes, I practice exercises against obesity; 2=No, I don't
14	Diet	1=Yes, I practice diet against obesity; 2=No, I don't
12	Allusions	1=Yes, I hear allusions about my silhouette; 2=No, I don't
13	Self-efficacy	1=No, I'm submissive to persuasion ; 2=Yes, I don't
21	Knowledge	1=Yes, I'm convinced: obesity injures health; 2=No, I don't
33	BMEE	1=Yes, I very often have BMEE (Between Meal Eating Episodes); 2=Yes, rarely; 3=No, seldom; 4=No, very seldom
36	Fast-food	1=No, I don't eat fast-foods; 2=Yes, not rare than once a month; 3=Yes, 2-3 times per month; 4=Yes, once a week; 5=Yes, rarely than once a week; 6=Yes, everyday

Descriptive statistics of BMI and WHR except conventional N, mean, and standard deviation SD, included median and skewness, with aim to support normality of distributions. Rough values of BMI and WHR were transformed to distributions with mean=0 and SD=1 with use of formula (3). Linear regression BMI=f(WHR) models were calculated separately for boys and girls subgroups, for rough data and for transformed data.

$$x_N = (x - \text{mean}(x))/(\text{SD}(x)). \quad (3)$$

where: x, x_N - transformed variable before and after transformation.

Principal components, U and W, were estimated with well-known formulas (4) for samples drawn from 2-dimensional standard normal distributions, ($\text{mean}_1=\text{mean}_2=0$, $\text{SD}_1=\text{SD}_2=1$; R), where: mean_1 , mean_2 , SD_1 , SD_2 – parameters of normal distributions of BMI_N and WHR_N , respectively, and R – coefficient of Pearson's correlation between them.

$$U = \text{BMI}_N + \text{WHR}_N; V = \text{BMI}_N - \text{WHR}_N; \quad (4)$$

where: BMI_N , WHR_N – standardized BMI and WHR, respectively.

Three variables (Self-esteem, BMEE, and Fast-food) were considered as depended variables in linear regression models, with predictors: first principal factor U, Allusions, Knowledge, and Dwelling-place.

III. RESULTS

Linear regression between BMI and WHR was significant evidently. At boys subgroup statistics $F=19.4$ ($p=0,0001$) for rough data, and $F=20.1$ ($p=0,0001$) for standardized data. At girls subgroup statistics $F=104.7$ ($p<0,0001$) for rough data, and $F=7098.4$ ($p<0,0001$) for standardized data.

Table 2. presents descriptive statistics of BMI and WHR.

Table 2. Descriptive statistics of BMI and WHR					
BMI			WHR		
parameter	boys	girls	parameter	boys	girls
N	35	57	N	35	57
median	20,7	19,5	median	0,42	0,41
mean	20,9	20,1	mean	0,44	0,42
SD	2,96	2,98	SD	0,05	0,04
Skewness	1,28	1,13	Skewness	0,99	0,82

N – number of participants; SD – standard deviation; Skewness - Fisher-Pearson coefficient

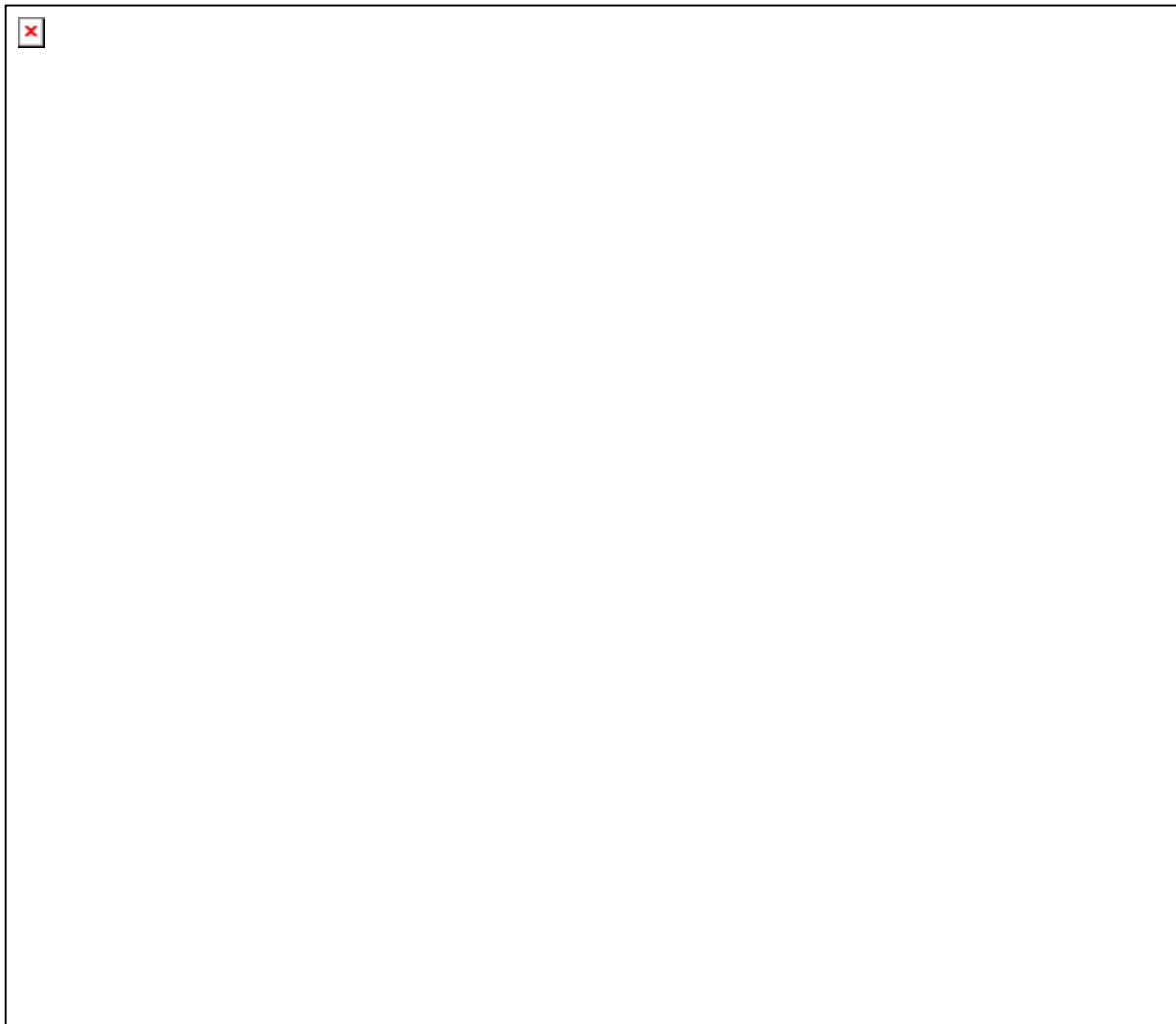


Fig. 1. Dependence $BMI=f(WHR)$ at boys and girls subgroups.

Table 3. presents distributions of categories of BMI in the study group, based on [12].

Table 3. Categories of BMI.					
BMI	underweight	normal	overweight	obese	Total
girls	7	39	8	3	57
boys	0	30	2	3	35
Total	7	69	10	6	92

Table 4. presents values of coefficients of determination, estimated with linear regression separately for boys and girls subgroups.

Table 4. Coefficients of determination for principal components.				
group	V(U)	V(W)	V(U)/2	V(W)/2
boys	1.61	0.39	80.5%	19.5%
girls	1.81	0.19	90.5%	9.5%

U, W – first and second principal components; V(U)/2, V(W)/2 – expressed in % from 2.

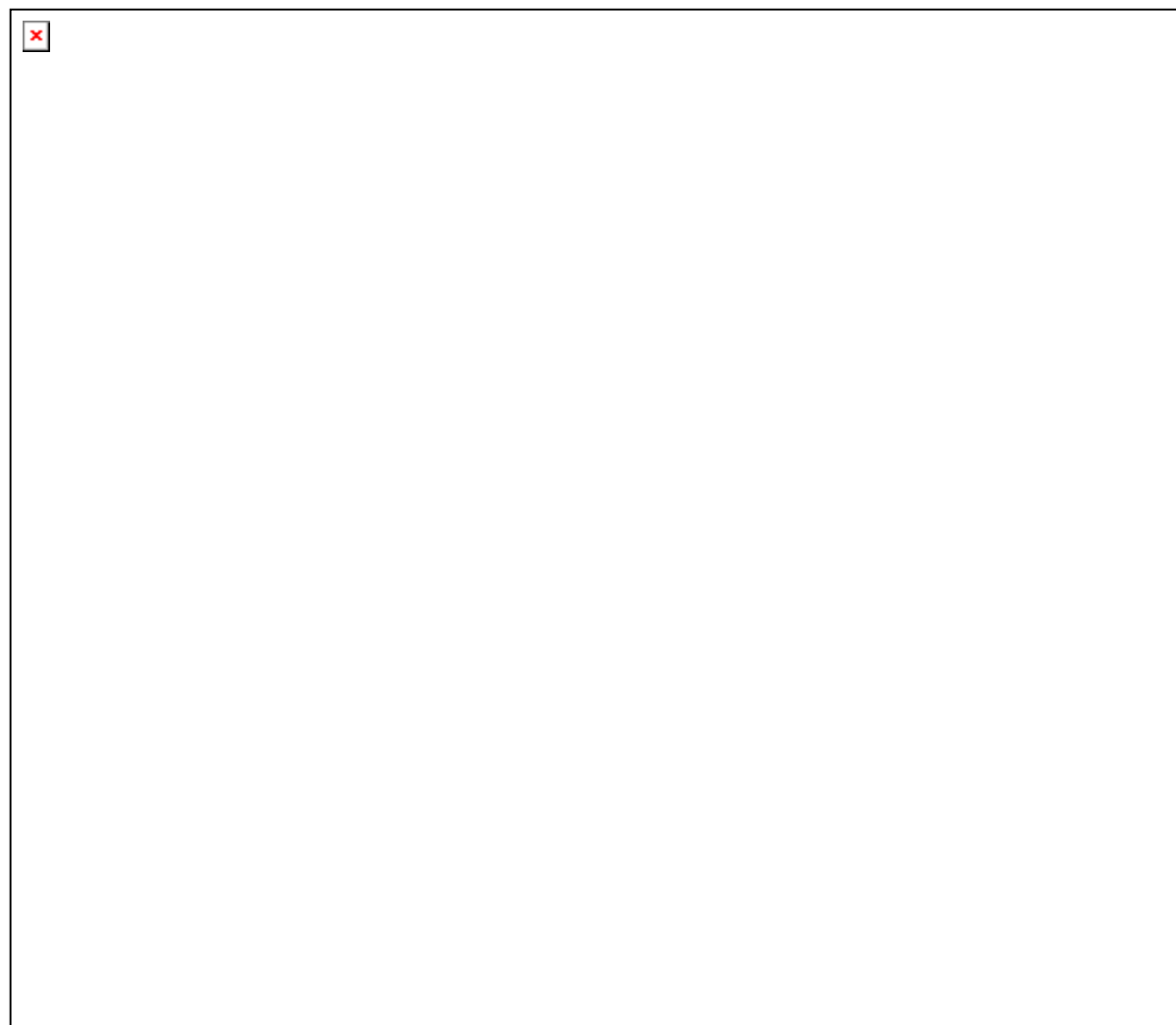


Fig. 2. Principal components at boys and girls subgroups.

Table 5. presents main results of current study: significance of hypotheses that Self-esteem, BMEE, and Fast-food depends on: principal factor U, Allusions, Knowledge, and Dwelling-place.

Table 5. Significance of influence of selected predictors on Self-esteem BMEE, Fast-food intake.

predictor	group	U	Allusions	Knowledge	Dwelling-place
Self-esteem	boys	p < 0,00001	-	-	-
	girls	p < 0,00001	p = 0,047	-	-
BMEE	boys	p = 0,003	p = 0,03	-	-
	girls	-	-	-	-
Fast-food	boys	p = 0,047	-	-	-
	girls	-	-	-	-

Self-esteem – approval of own weight; BMEE – frequency of Between Meal Eating Episodes; Fast-food - frequency of fast-food consumption; p – significance of linear regression coefficient.

IV. DISCUSSION

Use of BMI and WHR joined into a single principal component can explain 80.5% of variability at boys, and 90.5% at girls, instead of only 50% explained with a single variable.

It was stated that in study group, besides factual weight and silhouette, only allusions from some people can significantly influence health behaviour of children aged about 14-15 years.

V. CONCLUSION

Practical advantages of the proposed methodology were demonstrated on exemplary data on BMI and WHR at children aged about 14-15 years.

References

1. Moroshko I., Brennan L., O'Brien P.: Predictors of dropout in weight loss interventions: a systematic review of the literature. *Obes Rev.* 2011, 12, 912–934.
2. Daniels S.R.: The Consequences of Childhood Overweight and Obesity. *Childhood Obesity* 2006, 16, 47-67.
3. Ho S.Y., Lam T.H., Janus E.D.: Waist to stature ratio is more strongly associated with cardiovascular risk factors than other simple anthropometric indices. *Annals of epidemiology* 2003;13, 683–91.
4. Kuczmarski R.J., Flegal K.M.: Criteria for definition of overweight in transition: background and recommendations for the United States, *American Journal of Clinical Nutrition* 2000, 72, 1074-1081.
5. Jolliffe I.: *Principal Component Analysis*, New York, Willey, 2002.
6. Zhang Z., Zha H.: Principal Manifolds and Nonlinear Dimensionality Reduction via Tangent Space Alignment. *Journal of Shanghai University* 2004, 8, 406 -424.
7. Kendall M., Gibbons, J.D.: *Rank Correlation Methods*, London, Edward Arnold, 1990
8. Górkiewicz M., Gniadek A.: The practicality of any nonparametric statistical procedure should be confirmed thoroughly with regard to the data distribution under study. *Studies in Logic Grammar and Rhetoric* 2012, 29, 27-42.
9. Gromulska L., Piotrowicz M., Cianciara D.: Self-efficacy in health behavior models for health education. *Przegląd Epidemiologiczny* 2009, 63, 427 – 432.

10. Michelini I., Falchi A.G., Muggia C., Grecchi I., Montagna E., De Silvestri A., Tinelli C.: Early dropout predictive factors in obesity treatment. *Nutr Res Pract.* 2014, 8, 94–102.
11. Chmiel I., Górkiewicz M., Zawada K.: [On need of joined considering the body mass index (BMI) and waist to hip ratio (WHR) in investigations on self-esteem and health behavior at people with different levels of obesity], in: Szaban D., Kurowska H., Wróbel R. (Eds.), *Stan zdrowia a procesy demograficzne.*, Zielona Góra (Poland), Urząd Marszałkowski Województwa Lubuskiego, 2015, 205-215.
12. Jarosz M.: [Rules of proper feeding children and adolescences, and recommendations on health behavior], Warszawa (Poland), Instytut Żywności i Żywienia, 2008.

ENSEMBLES OF VARIABLES SELECTION AND COMBINED CLASSIFIERS FOR MEDICAL DIFFERENTIATION ON THE BASIS OF GENES EXPRESSION DATA SET

M. Ćwiklińska-Jurkowska

Collegium Medicum, Bydgoszcz, Nicolaus Copernicus University, Toruń, Poland

Abstract

The usefulness of combining methods was examined on the example of microarray Colon data set, where expression levels of huge number of genes are reported. Discrimination problem into tumor and normal cases is examined. Cross-validation errors evaluated on half of whole data set, not used for selection of genes, were applied as measures of classifier. Frequent procedures of single selection of genes: Prediction Analysis of Microarrays (PAM) and Significance Analysis of Microarrays (SAM) were compared to different ensembles not including these selection results. Combining of genes selection methods was not essential in comparison to single PAM or SAM selection for any examined ensemble of classifiers.

On the other hand, combining the five classifiers: k nearest neighbours, SVM linear and SVM radial with parameter $c=1$, Shrunken Centroids Regularized Classifier and nearest mean classifier, significantly outperformed the resampling classifiers like bagging, double bagging or random forest. The previous step of combining ranking of variables was not essential for the performance for all examined ensembles of classifiers.

Keywords: combined methods, discriminant analysis, genes selection

I. INTRODUCTION

Because of the high number of investigated genes in one microarray, the pre-selection of features for inclusion into the classification rule is essential. Often, only a few tens of genes are really active; the remaining genes are not important for improvement of the discriminant procedure. In supervised classification, the variables with the biggest discriminant power are sought out.

For classification problems occurred for microarray data sets, typically bagging or boosting combined classifiers are applied. Ensembles of classifiers based on resampling, like bagging or boosting might improve stability. Those methods may be called the families of classifiers. Families are considered as homogenous ensembles, because all base classifiers that are merged in one decision, are of the same type, but are created on slightly different subsets (random subsets).

Ensemble of selection methods may also benefit in outcome ranking of the most discriminating genes. Thus, the usefulness of combining was examined for dimension reduction and also for building classifiers stage and for jointly both of them.

II. METHODOLOGY

In the *Colon* data set, containing expression levels of 2000 genes (variables) with 62 cases (patients), 40 tissues out of 62 are colon tumor tissues and 22 are normal. The whole *Colon* data set, was randomly divided into two subsets with similar number of patients G and D . Thus G , the subset of *Colon* data set has 2000 genes and 32 patients, where 20 are tumor cases. The data set was standardized by subtracting a gene average and dividing by a standard

deviation. For each discriminating problem, the first subset G, was applied for a selection of the most discriminating genes. The second independent set D of 30 patients included 20 tumor patients and it was used to assess generalization properties of selected subsets of genes. Subsequently, D set was divided into $k=10$ cross-validation subsamples and the 10 cross-validation error of discriminant functions was calculated. Cross-validation error, evaluated on D subset of whole data set, was the criterion for comparisons between correctness of selection procedures and classifiers.

Subsequent subsets of genes are increasing and include all genes selected in the previous set. Thus, after the selection, the variables are ranked according to the selection criterion. Selection criterion is connected with the discriminant power of variables set. Different reduction methods of dimension were applied. Single selection methods were: SAM (Significance Analysis of Microarrays) and PAM (Prediction Analysis of Microarrays). Ensemble of selection not including SAM and PAM were constructed and compared to single SAM and PAM. The sequence of genes indicates also the decreasing ranking. The genes are ranked from 1 to 100 and those sets are next applied for evaluation of examined classifiers errors.

To find relevant genes in another way, the variables selection of different nature was considered for merging into one ranking effect. The base rankings for combining are obtained according to Gini impurity measure, between to within groups ratio (BetweenWithinRatio), T and Wilcoxon test with Hochberg adjustment for multiplicity (denoted by Hochberg, HochbergWilc, respectively). Adjusted p-values for multiple testing procedures were also applied in permutation testes and permutation Welch T and Wilcoxon rank-sum tests are incorporated into the ensemble ranking. The permutation algorithm for the maxT and minP procedures is described in [1], and according to above information, the base selection methods are denoted as PermutAdjPmaxT, PermutAdjPminP, PermutAdjPmaxTWilcox, PermutAdjPminPWilcox. Such obtained merged rankings were investigated to compare with single SAM and PAM procedures.

Ensemble of variables selection procedure introduces the weight ranking of genes using base selection methods ranking in the way that higher joint ranking obtains a gene that occurs prior to others in most of combined base rankings. For sets, with increasing number of variables, the cross-validation classification errors are calculated, so classifier evaluation curves may be plotted for these values. Subsequent subsets of genes of increasing sizes from 1 to 100 are obtained for each classifier, so various discriminant methods are compared for ascending number of genes.

Because the number of considered variables can even reach 100, the classical discrimination fails for applied CV procedure constructed from D set. Thus, discriminant functions viable for high dimensionality were considered for merging. Various procedures have been discussed as alternatives to classical discriminant analysis. Some of them are: Shrunken Centroids Regularized Discrimination [2], k nearest neighbors discrimination (kNN), the uncorrelated linear discrimination and Support Vector Machines [3].

The Support Vector Machines (SVM) provides an optimally separating hyperplane in the sense that the margin between two groups is maximized. In the work, SVM [3] with linear kernel and also radial kernel was incorporated into classifiers ensemble and parameter $c=1$ was applied. Also k nearest neighbour (kNN) discriminant function was built into the ensemble. In kNN, the number of neighbours is optimized according to cross-validation error. Shrunken Centroids Regularized Discriminant Analysis [2] was also added to the ensemble. The last, fifth base classifier incorporated into the ensemble is a special case of linear diagonal (uncorrelated) discriminant function, assuming equal variances of genes, (i.e. nearest

mean classifier). Ensemble classification method is the classifier merging results of base classifiers, because those constituent methods are different in methodology, the majority vote was used for joint decision. The final joint classifier is heterogeneous, so is named in the work Heterogeneous Merge and denoted by HeterMerge2.

Researchers in classification tend to combine procedures, based on similar types or different base classifiers. Specifically, considerable attention has been paid lately to families of classifiers originating from two ideas: bootstrap aggregations and boosting. In the current work, the aggregation of classifiers constructed on data sets bootstrapping was applied. To this class belong typical bagging [4], modified double bagging with LDA and double bagging with SLDA [5]. Double bagging combine LDA (or singular LDA) and classification trees.

Also random subspace method use bootstrap aggregation with additional step of random selection of variables in each loop. Application of the decision trees gives the random forests procedure [6]. In the current work, random forests classifier is applied and accordingly, trees are also investigated as base classifiers in other bootstrap aggregation procedures like bagging, LDA double bagging and SLDA double bagging.

III. RESULTS

The misclassification rate assessment was completed for increasing number of variables: 2, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50, 55, 60, 70, and 100, because the usage of more than 100 genes occurred to be not constructive. For succeeding subsets of genes, ranked by the examined combined selection method, misclassification rates of different classification methods were assessed. The comparison of combined classifiers evaluation curves can be considered on the base of Fig. 1-4, where cross-validation technique of G set into 10 folds was used to assess generalization errors (Fig.1-4). Methods of applied combined selection are presented on Fig 2 and 4, where 10-CV errors of classification methods for succeeding subsets of genes, ranked by combined selection methods, are given.

Comparing Figures 1 and 2 with mean 10-CV errors with lines obtained by added and subtracted standard errors, we can observe significant outperforming of HeterMerge2 (solid line) over typical bagging and double bagging. It is especially distinct for about 60 genes. For random forest the difference is also observed, but the benefit coming from application of heterogeneously merged classifier is not so apparent as for other homogenous ensembles (Fig.1-2). The effect is hold for both single SAM selection (Fig. 1) and as well for combined ranking from Gini, PermutAdjPmaxT, PermutAdjPminP, PermutAdjPmaxTWilcox, and PermutAdjPminPWilcox (Fig. 2). Comparing Figs. 1 and 2 we can see very similar learning curves for single SAM selection and the selection ensemble.

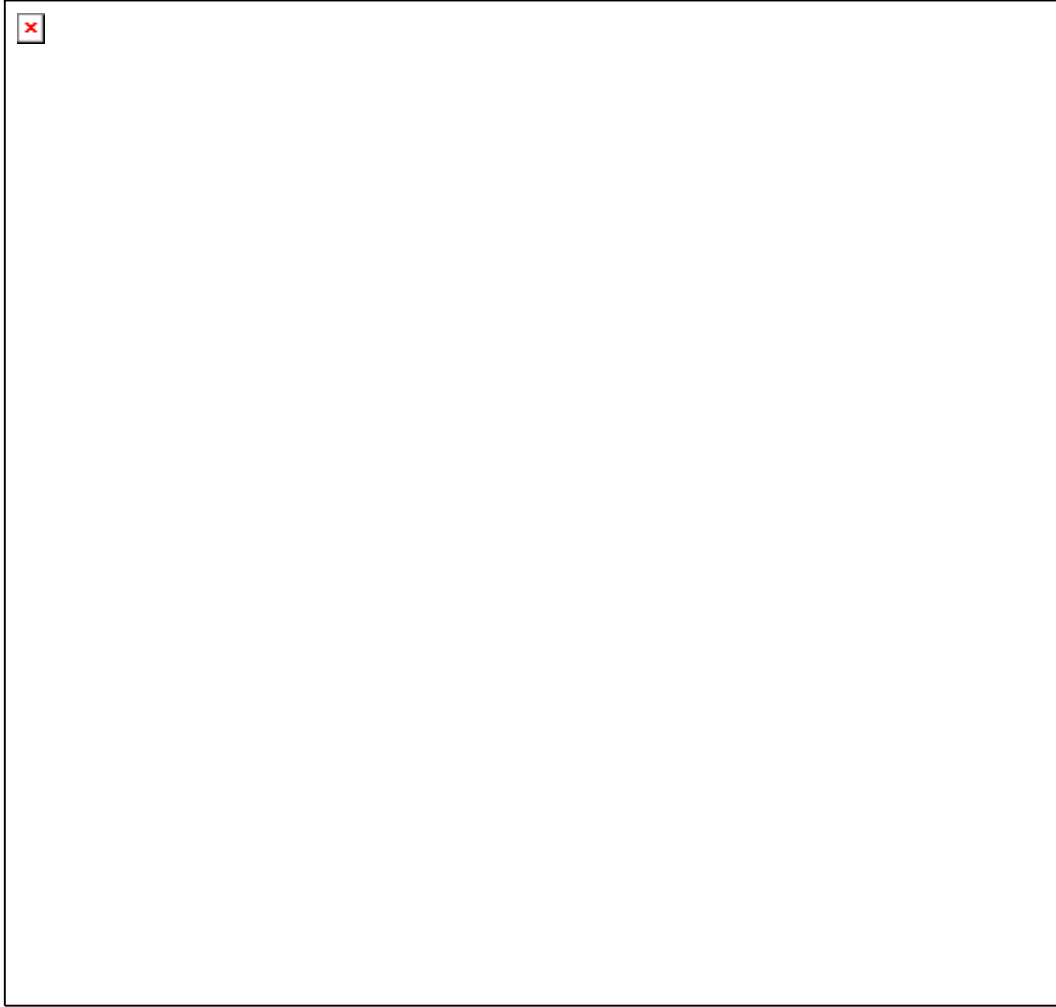


Fig.1. CV errors of homogenous and heterogeneous combined classifiers after SAM selection.



Fig.2. CV errors of homogenous and heterogeneous combined classifiers after selection obtained by combined ranking of five procedures: Gini, PermutAdjPmaxT, PermutAdjPminP, PermutAdjPmaxTWilcox, PermutAdjPminPWilcox.

Apparent difference can be observed between Heterogeneous Merge and homogenous classifiers based on bootstrap aggregating, for example bagging, LDA and SLDA bagging and random forest for more than 50 genes. For complement examination on the difference between Heterogeneous Merge and bagging, areas with standard errors were plotted (Fig. 3 for single SAM selection procedure) and Fig. 4 (for combined genes selection from Gini, PermutAdjPmaxT, PermutAdjPminP, PermutAdjPmaxTWilcox, PermutAdjPminPWilcox). The plots indicate significant difference between Heterogeneous Merge and bagging trees for 50-80 genes on both Figures 3 and 4. However, comparing Fig. 3 and 4, we can conclude that there is no difference between single PAM selection and combined selection procedure for types of all combined classifiers. Considered ensembles of selection procedures are not beneficial in comparison to popular SAM and PAM, but only part of results are displayed on classifier evaluation curves in the paper (Figs 1,3). Various ensemble techniques applied for selection of variables are not different according to evaluation by CV classification errors and additionally there is no essential difference between them and SAM or PAM selection.

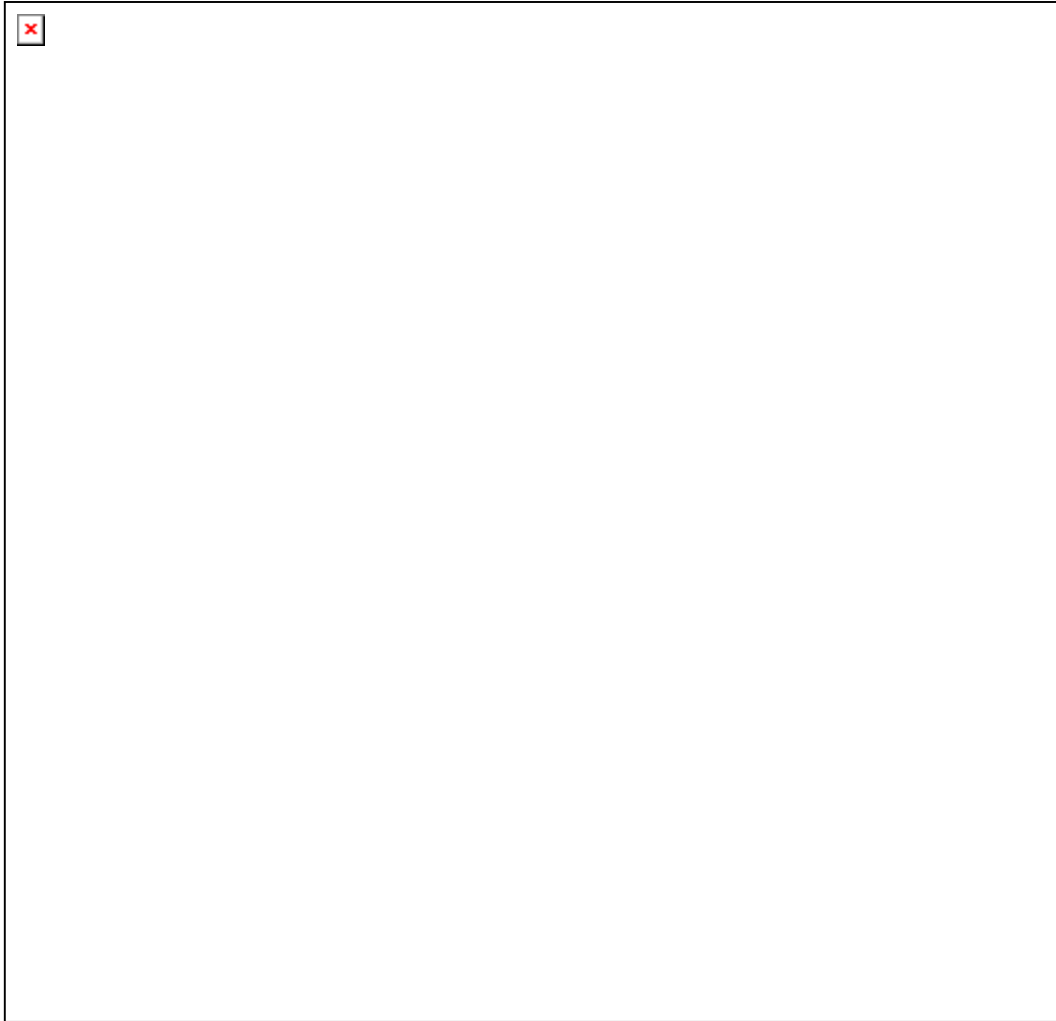


Fig. 3. Ten- fold CV classification errors with standard errors of classification methods: merged classifier (HeterMerge2 based on vote on k-nearest neighbour, regularized classifier, nearest mean classifier, linear SVM with $c=1$, radial SVM with $c=1$), and bagging tree (100 loops), for succeeding subsets of genes ranked by single PAM selection.

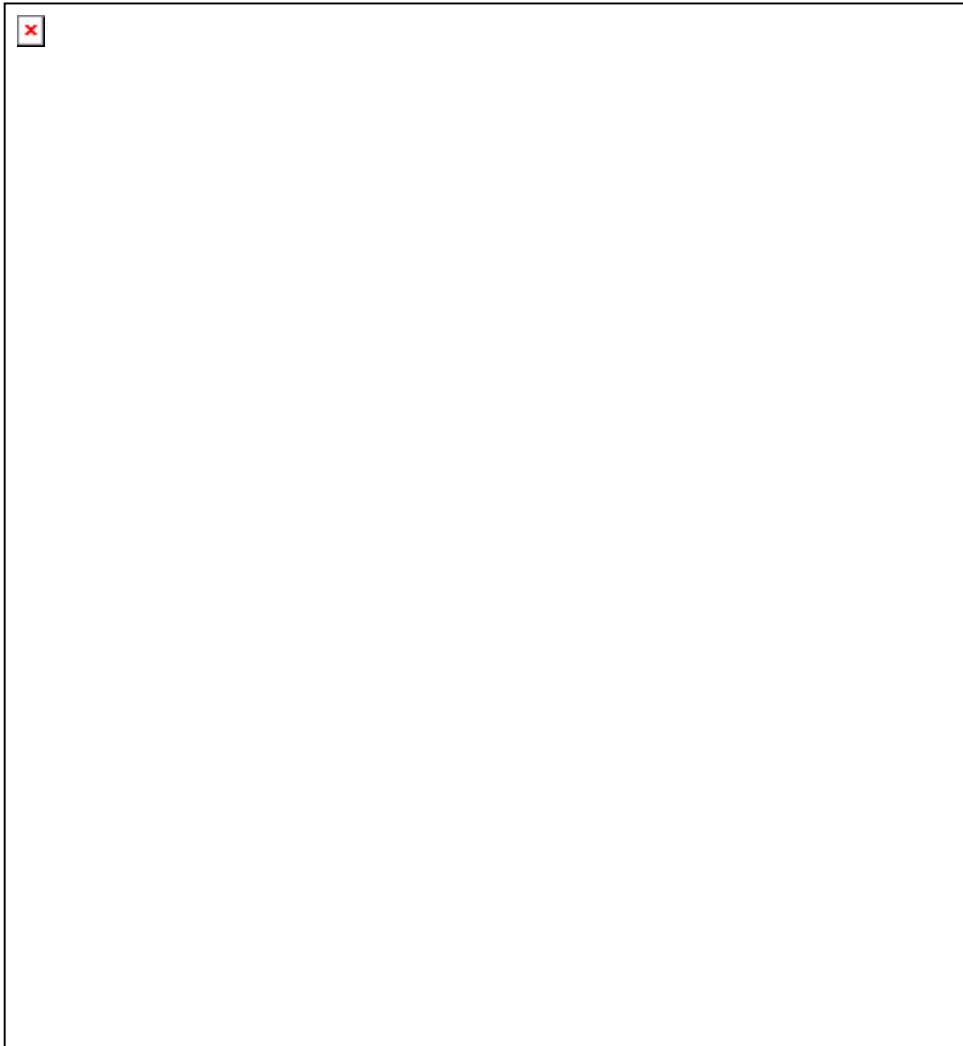


Fig. 4. Ten- fold CV classification errors with standard errors of classification methods: merged classifier (HeterMerge2 based on vote on k-nearest neighbour, regularized classifier, nearest mean classifier, linear SVM with $c=1$, radial SVM with $c=1$), and bagging tree (100 loops), for succeeding subsets of genes ranked by combined selection of eight procedures (BetweenWithinRatio, Gini, PermutAdjPmaxT, PermutAdjPminP, PermutAdjPmaxTWilcox, PermutAdjPminPWilcox, HochbergLS, and HochbergWilc).

IV. DISCUSSION

Different ensemble techniques examined for selection of variables are not essentially different according to evaluation by CV classification errors and also there is no important difference between them and SAM or PAM selection. In the set of joint selection methods, part of them comes from the same ideas- like permutation tests or Hochberg correction for multiplicity. The question arises if joining selections from wider set of diverse procedures might improve results for combined classifiers.

Heterogeneous Merge classifier obtained by majority voting on base decisions of five constitutive discriminant procedures, like k-nearest neighbour, regularized classifier, Euclidean classifier (nearest mean), linear SVM, radial SVM, occurred to outperform all four examined combined tree methods based on resampling of data set D. The base classifiers

joined into Heterogeneous Merge classifier come from several different ideas from wide pattern recognition methodology. Ensemble might use important and concurrent benefits of base methods.

Number of genes may be chosen as a trade-off between the size of genes set and the decrease of error. The optimal genes subsets are indicated by Heterogeneous classifier to be about 50-70 genes, where the classification error holds the stable level.

V. CONCLUSION

According to fold cross-validation errors, for microarray data set Prostate, heterogeneous merge of classifiers performs significantly better than homogenous ensembles like bagging, LDA double bagging and SLDA double bagging. Smaller 10-fold cross-validation errors were achieved for heterogeneous ensemble of regularized discriminant analysis, kNN, linear and radial SVM and nearest mean classifiers than for homogenous ensembles, independently on previous variables selection method. The difference between misclassification rates is significant for 50-80 genes. The advantage may come from different ideas of merged constituent classifiers, because all of them have different properties and benefits.

Hoverer, merging of ranking genes obtained from permutation procedures adjusted for multiplicity, Gini index, between-within groups diversity, parametric and nonparametric testing with Hochberg adjustment for multiplicity was not important for subsequent misclassification rates.

Thus, the usefulness of combining procedures over single procedures was beneficial for building classifiers stage but not for dimension reduction. The preliminary step of combining genes ranking was not essential for the performance for both heterogeneously and homogeneously combined classifiers.

References

1. Ge Y, Sandrine Dudoit S., Speed T.P. Resampling-based multiple testing for microarray data analysis. Jan. 2003. Technical Report 633.
2. Guo, Y. Hastie T., Tibshirani R. Regularized Discriminant Analysis and Its Application in Microarrays. *Biostatistics* 2005, 1, 1, pp . 1–18.
3. Cortes C. Vapnik V. Support-Vector Networks, *Machine Learning*, 20, 273-297, 1995.
4. Breiman L. Bagging predictions. *Machine Learning* 1996, 24 (2), 123-140.
5. Kropf Z. Hochdimensionale multivariate Verfahren in der medizinischen Statistik 2000, Shaker Verlag, Aachen.
6. Breiman, L. Random Forests. *Machine Learning* 2001, 45, 5–32.

ITEM RESPONSE THEORY METHODS CAN SUPPORT VALIDITY OF 4CornerSAT SCALE FOR CAREER SATISFACION OF PHYSICIANS

M. Górkiewicz¹, J.N Peña-Sánchez², I. Chmiel¹

¹Faculty of Health Sciences, Jagiellonian University of Krakow, Krakow, Poland;

²Dprt. of Community Health and Epidemiology, University of Saskatchewan, Saskatoon, Canada

Abstract

Evaluation of attitudes and individual satisfaction is one of the most important problem in real-world investigations. In this study the Master's partial credit model was successfully applied to each of four dimensions of the 4CornerSAT scale to measure career satisfaction of physicians.

Keywords: Rasch, partial credit model, career satisfaction

I. INTRODUCTION

In context of Item Response Theory (IRT) pseudo-Rash partial credit method (PCM) estimates a hidden linear ordering of a scale items along common axis of mean scores given by particular participants of a questionnaire survey to all scale items, [1].

PCM estimates thresholds between scale items with use of each pair of adjacent levels on applied Likert scale. Table 1 presents PCM estimates, obtained in studies [2] and [3].

Table 1: Ranks of seven chosen items of Physical Functioning scale from SF-36 questionnaire.

Accordingly to [3]			rank	Accordingly to [2]		
Thresh.:1/2	Thresh.:1/2&2/3	Thresh.:2/3		Thresh.:1/2	Thresh.:1/2&2/3	Thresh.:2/3
	PF09		1		PF09	
	PF05		2		PF08	
	PF08		3		PF05	
PF02		PF07	4	PF02		PF07
	PF04		5	PF07		PF02
PF07		PF02	6		PF04	
	PF01		7		PF01	

Thresh.:1/2 – threshold between score=1 and score=2; Thresh.:2/3 – threshold between score=2 and score=3; PF01: Vigorous activities; PF02: Moderate activities; PF04: Climbing several flights of stairs; PF05: Climbing one flight of stairs; PF07: Walking more than a mile; PF08: Walking several blocks; PF09: Walking one block.

SF-36 questionnaire includes ten items of physical functioning (PF) scale, scored with Likert scale from 1=(strong limitation), to 3=(no limitation), to assess how health limits physical functioning, [3]. In study [2] only seven PF items were considering.

It should be noted plainly that the both orderings showed at Table 1 violate the fundamental Rasch postulate: $range(PF02)=4$ was smaller than $range(PF07)$ for participants making a choice between score=1 and score=2, but for participants making a choice between score=2 and score=3, $range(PF07)=4$ was smaller than $range(PF02)$.

Despite this issue, any corrections of PF scale wasn't proposed, because authors in the field [2, 3] followed a pragmatic rule: the widely acknowledged effectiveness of PF scale prevails

over some little formal imperfection, However, from formal statistical perspective, in [3] it was proved that PF measured a unidimensional construct, and in [2] it was proved that the proposed orderings [2, 3], can both follow from the same hidden ordering.

The 4CornerSAT questionnaire to evaluate career satisfaction of physicians was originally created in English, [4], but later this was adapted to Polish and Spanish, [5, 6].

The 4CornerSAT had four scales, related to personal, professional, inherent and performance dimensions of career satisfaction. Each scale had four items, each scored on 6-point Likert scale: 1=very.dissatisfied; 2=satisfied; 3=somewhat.dissatisfied; 4=somewhat.satisfied; 5=satisfied; 6=very.satisfied.

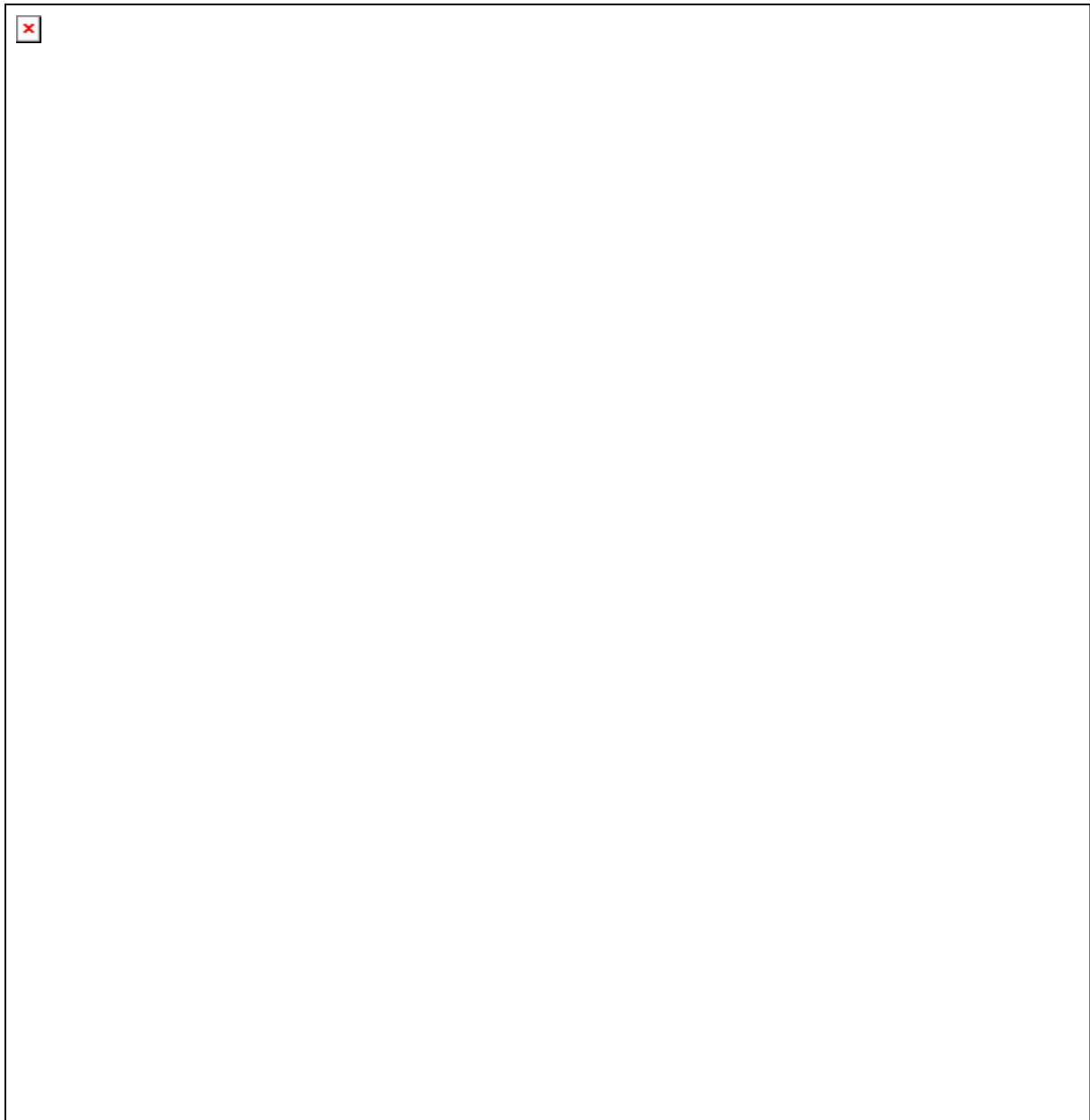


Fig. 1. Comparison of English and Polish versions of 4CornerSAT questionnaire, [5].

II. METHODOLOGY

Data were drawn from eligible physicians working in six hospitals of Andalusia, Spain, between 2009 and 2010. Participants were invited by e-mail to complete on-line questionnaire which included 4CornerSAT scale in Spanish, [7]. In relation to the sample size, N=121, the Likert scores 1 to 3 were merged into a single category “score123”.

The partial credit method was applied separately to each scale of the questionnaire. Intraclass Correlation and Analysis of Variance for difference among scale’s items, and for homogeneity of partial orderings were calculated using an on-line calculator: http://department.obg.cuhk.edu.hk/researchsupport/IntraClass_correlation.asp.

III. RESULTS

Of the N=299 eligible physicians, N=121 completed the questionnaire (40.7% response rate). The reliability of the questionnaire was supported with Cronbach’s alpha, $\alpha > 0,77$ for separate scales, and $\alpha = 0.92$ for the all four scales considered jointly. The further descriptive statistics and results of validation made with classical methods one can find in [6].

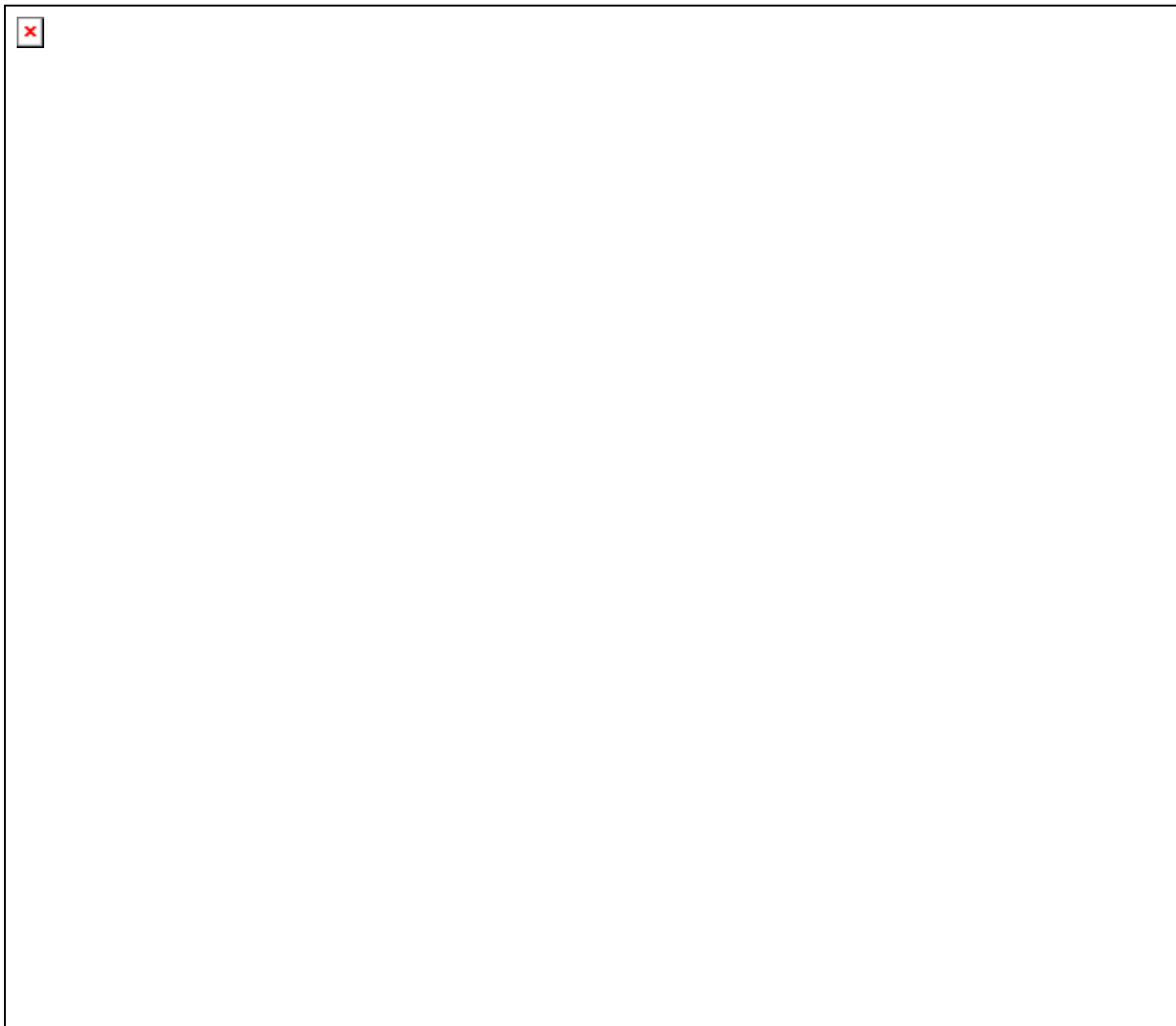


Fig.2. Factors associated with higher levels of satisfaction, (attained from source: [6]).

Table 2. Presents the distribution of mean scores of the participants.

Table 2: Distribution of the mean scores from all participants (N=121)

Y/4	1-1.5	1.75	2	2.3	2.5	2.8	3	3.3	3.5	3.8	4	4.3	4.5	4.8	5	5.3	5.5	5.75	6
N	0	2	2	0	1	1	0	3	3	7	11	11	18	30	20	6	5	1	0

Y – sum of scores from four scales of 4CornerSAT questionnaire; Y/4 – mean score; N – number of participants

Partial medians of mean scores were shown in Tables 3, 4, 5, and 6.

Table 3: Partial medians of the mean scores from the personal satisfaction scale.

Item	score123	score4	score5	score6
11	2,88	4,00	4,75	5,25
12	3,25	4,25	4,5	5,25
16	3,25	4,5	4,75	4,75
15	3,25	4,25	4,75	5,38

Item – number of item in the questionnaire;
score - level of applied Likert scale.

Table 4: Partial medians of the mean scores from the professional satisfaction scale.

Item	score123	score4	score5	score6
10	1,75	4,00	4,75	5,25
8	3,5	3,75	4,75	5,25
9	3,25	4,5	5	5,5
13	3,75	4,75	5	5,75

Item – number of item in the questionnaire;
score - level of applied Likert scale.

Table 5: Partial medians of the mean scores from the performance satisfaction scale.

Item	score123	score4	score5	score6
4	3,25	4,25	4,75	5,00
6	3,25	4,25	4,75	5,00
5	3,75	4,25	4,75	5,25
7	3,5	4,5	4,75	5,13

Item – number of item in the questionnaire;
score - level of applied Likert scale.

Table 6: Partial medians of the mean scores from the inherent satisfaction scale.

Item	score123	score4	score5	score6
1	1,88	3,75	4,50	5,00
2	3,25	4,25	4,75	5,13
3	3,75	4,25	4,75	4,88
14	4,25	4,625	5	5,00

Item – number of item in the questionnaire;
score - level of applied Likert scale.

Partial items ranks from the personal, professional, performance, and inherent satisfaction, were shown in Tables 7, 8, 9, and 10.

Table 7: Partial items ranks from items at personal satisfaction scale.

thresholds	item11	item12	item16	item15
score123 score4	1	2	4	3
score4 score5	1	2	3,5	3,5
score5 score6	3	1	2	4

score - level of applied Likert scale; Intraclass Correlation ICC=0,67: high;

Analysis of Variance:

homogeneity of partial orderings $p(F < 0,0001) < 0,0001$;

difference among scale's items $p(F = 5,25) = 0,69$.

Table 8: Partial items ranks from items at professional satisfaction scale.

thresholds	item10	item8	item9	item13
score123 score4	1	3	2	4
score4 score5	2	1	3	4
score5 score6	1,5	1,5	3,5	3,5

score - level of applied Likert scale; Intraclass Correlation ICC=0,89: high;

Analysis of Variance:

homogeneity of partial orderings $p(F < 0,0001) < 0,0001$;

difference among scale's items $p(F = 19,00) = 0,83$.

Table 9: Partial items ranks from items at performance satisfaction scale.

thresholds	item4	item6	item5	item7
score123 score4	1,5	1,5	4	3
score4 score5	2	2	2	4
score5 score6	2,5	2,5	2,5	2,5

score - level of applied Likert scale; Intraclass Correlation ICC=0,68: high;

Analysis of Variance:

homogeneity of partial orderings $p(F < 0,0001) < 0,0001$;

difference among scale's items $p(F = 5,50) = 0,70$.

Table 10: Partial items ranks from items at inherent satisfaction scale.

thresholds	item1	item3	item2	item14
score123 score4	1	3	2	4
score4 score5	1	2,5	2,5	4
score5 score6	1	2,5	2,5	4

score - level of applied Likert scale; Intraclass Correlation ICC=0,73: high;

Analysis of Variance:

homogeneity of partial orderings $p(F < 0,0001) < 0,0001$;

difference among scale's items $p(F = 6,70) = 0,73$.

IV. DISCUSSION

Homogeneity of partial orderings, shown in Tables 7, 8, 9, and 10, was supported for each scale separately, with high Intraclass Correlation $ICC > 0,67$ and with Analyses of variance: $p(F < 0,0001) < 0,0001$. Substantial difference among scale's items was also supported by Analyses of variance: $p(F > 5,25) > 0,69$.

V. CONCLUSION

It was proved that in the context of the Item Response Theory (IRT), the Master's partial credit method can support validity of 4CornerSAT questionnaire to evaluate career satisfaction of physicians.

References

1. Scheiblechner H.: Rasch and pseudo-Rasch models: suitability for practical test applications. *Psychology Science Quarterly* 2009, 51, 2009, 181 – 194.
2. Górkiewicz M., Chmiel I.: Zastosowanie podejścia Rasch do porównawczej analizy wyników pomiaru jakości życia za pomocą polskiej wersji kwestionariusza SF-36. In: Sienkiewicz Z, Fidecki W, Wójcik G [Eds.], *Wybrane Determinanty Pielęgniarstwa; Część II*. Warszawa (Poland), Warszawski Uniwersytet Medyczny, 2010, 128–136.
3. Martin M., Kosinski M., Bjorner J.B., Ware J.E., MacLean R., Li T.: Item response theory methods can improve the measurement of physical function by combining the modified health assessment questionnaire and the SF-36 physical function scale. *Quality of Life Research* 2007, 16, 647–660.
4. Lepnurm R, Danielson D, Dobson R, Keegan D. Cornerstones of career satisfaction in medicine. *Can J Psychiatry* 2006, 51, 512-22.
5. Peña-Sánchez J. N., Domagala A., Górkiewicz M., Targowska M., Oleszczyk M. Adapting a tool in Poland for the measurement of the physicians' career satisfaction, *Problemy Medycyny Rodzinnej* 2011, 12, 58–65.
6. Peña-Sánchez JN, Delgado A, Lucena-Muñoz JJ, Morales-Asencio JM. [Adapting and Validating in Spanish the 4CornerSAT Questionnaire to Measure Career Satisfaction of Specialized care Physicians. Andalusia, Spain]. *Rev Esp Salud Pública* 2013, 87, 181-189.
7. Peña-Sánchez J.N, Lepnurm R., Morales-Asencio J.M., Delgado A., Domagala A., Górkiewicz M. Factors identified with higher levels of career satisfaction of physicians in Andalusia, Spain. *Health Psychology Research* 2014, 2, 58-62.

SOFTWARE FRAMEWORK FOR VALIDATION OF SEGMENTATION RESULTS (VOS)

A. Korzyńska¹, L. Roszkowiak¹, J. Zak¹, D. Pijanowska¹

¹*IBBE PAS, Warsaw, Poland*

Abstract

The papers introduced software called VoS for validation results of image segmentation using reference image. This software is for objective and quantitative assessment of the performance of segmentation methods and their comparisons. It is general framework for revealing the performance of any segmentation algorithm despite the examples of its usefulness presented in paper concerns segmentation of images of the immunohistochemically stained tissue section.

Keywords: empirical goodness measure, image processing, image segmentation, quantitative assessment, unsupervised evaluation, validation framework

I. INTRODUCTION

The image segmentation is an important process of image analysis with a great influence on the results [1]. Determined by the aim of analysis it divides image into areas that correspond to real-world objects of interest [2]. There are thousands of proposed image segmentation algorithms [3 – 5] but still there is no general performance measure of segmentation that would allow comparison of different methods or different parameterizations of single method. Generally, all proposed methods are evaluated by their authors using their own image database and/or using reference images prepared by specialists under authors guidelines. It is called supervised or relative evaluation method. There is a need of a method to compare methods/parameterizations in an application-independent way [6], so-called unsupervised also known as stand-alone or empirical goodness methods. This type of comparison quantifies how well it matches a broad set of characteristics or patterns prepared for testing according to the researchers needs. There are special databases of simple artificial or natural patterns with various features. Each pattern composition should be accompanied by the corresponding information about its location in the image. Such database is proposed by Zhang Y. J. [7] and by Brodatz album [8, 9].

The proposed software called Validation of Segmentation (VoS) can be used to compare images after segmentation with reference image or information about patterns. The software automates the procedure of comparison between reference image or patterns location and result image, then it calculates the statistics that describe performance of segmentation and creates visualization of the comparison.

II. METHODOLOGY

For binary (black and white) images, the VoS calculates the number of pixels in four standard categories: true positive, true negative, false positive and false negative. The pixels of segmented objects that are also present in the reference image, are called true positives (TP). False positives (FP) are the pixels belonging to objects in the segmented image but to the background in the reference image. On the other hand pixels classified by segmentation as background that belong to objects in reference image are true negatives (TN). False negatives

(FN) are pixels classified as background in the segmented image and in the reference image. Based on the number of pixels in those four categories it is possible to calculate more complex evaluation. The VoS software implements the statistical measurements of performance such as sensitivity, specificity and accuracy as well as coefficients developed by Dice, Jaccard, Sokal and Sneath, and Rogers and Tanimoto. All these features are calculated according to definitions in [10].

Additionally to calculating parameters, this software can produce visual comparison of the binary images. Objects from the result of segmentation and reference image are presented as object boundaries imposed in various colors on white background, as it is presented in fig. 2 in [11].

Moreover, the VoS software calculates the measurements of image similarity and quality such as: mean square error (MSE), root mean square error (RMSE), normalized root mean square error (RMSE_{norm}), peak signal-to-noise ratio (PSNR) and normalized peak signal-to-noise ratio (PSNR_{norm}). Additionally, measurement matched to perceived visual quality is calculated. It is a perception-based model of structural similarity index (SSIM) defined by Wang and coworkers [12] that compares separately three components: luminance, contrast and structure. These parameters can be calculated for binary as well as full color (RGB) images.

III. RESULTS

The software is developed in MATLAB R2015b, implemented on Intel(R) Core(TM) i7-4710MQ@2.50GHz CPU, 16.0GB RAM.

The VoS software has been used by authors on many occasions with great results. It produces reliable results of comparison and accumulates different features into one dataset. For example, all the coefficients are presented in tables for each single image in test database as it is presented in tables 4-6 in [10] or averaged version for all compared methods as it is presented in tables II and III in [11].

IV. DISCUSSION AND CONCLUSION

The proposed framework for validation of image segmentation results is efficient and can be used to compare methods/parameterizations in both application dependent and independent ways. It is a general framework for evaluation of the performance of any segmentation algorithm despite the examples of its usefulness presented in this paper concerning images of immunohistochemically stained tissue sections. It creates an alternative to use of general image manipulation software, for example Photoshop or ImagePro Premier supported by Excel, where this type of operations are possible but requires plenty of human direct data manipulation.

Acknowledgements

This study was supported by statutory funds of Nałęcz Institute of Biocybernetics and Biomedical Engineering Polish Academy of Sciences.

References

1. Gonzalez R. C., Woods R. E.: Digital image processing 2008, Upper Saddle River, NJ, USA: Pearson Education, Inc.
2. Zhou S. K.: Medical image recognition, segmentation and parsing; Machine learning and multiple objects approaches 2016, Amsterdam, Elsevier.

3. Korzyska A., Strojny W., Hoppe A., Wertheim D., Hoser P.: Segmentation of Microscopic Images of Living Cells. *Pattern Analysis and Application*, 2007, 10: 301-319.
4. Neuman U., Korzyska A., Lopez C., Lejeune M.: Segmentation of stained lymphoma tissue section images. In Pietka E., Kawa J. [EDS.], *Information Technology in Biomedicine 2, Advances in Intelligent and Soft Computing 2012*, 101-113.
5. Korzyska A., Iwanowski M.: Multistage morphological segmentation of bright-field and fluorescent microscopy images. *Opt-Electronics Review* 2012 20(2), 87-99.
6. Zhang H., Fritts E. J., Goldman S. A.: Image Segmentation Evaluation: A Survey of Unsupervised Methods. *Computer Vision and Image Understanding* 2008, 110(2), 260–280.
7. Zhang Y. J.: Evaluation and comparison of the different segmentation algorithms. *Pattern Recognition Letters* 1997, 18 963-974.
8. P. Brodatz, *Textures, a photographic album for artistes and designers*. Dover, New York, 1966.
9. Base of 300 synthetic images, <http://www.ensibourges.fr/LVR/SIV/interpretation/evaluation/>.
10. Korzyska A., Roszkowiak L., Lopez C., Bosch R, Witkowski L., Lejeune M.: Validation of various adaptive threshold methods of segmentation applied to follicular lymphoma digital images stained with 3,3'-Diaminobenzidine&Haematoxylin. *Diagnostic Pathology* 2013, **8**:48 1-21.
11. Roszkowiak L., Korzyska A., Pijanowska D.G.: Short survey: adaptive threshold methods used to segment immunonegative cells from simulated images of follicular lymphoma stained with 3,3'-Diaminobenzidine&Haematoxylin. In: Ganzha, M., Maciaszek L., Paprzycki M. [EDS.], *Proceedings of the 2015 Federated Conference on Computer Science and Information Systems 2015. Annals of Computer Science and Information Systems, AVSIS*; 5, 291-295.
12. Wang Z., Bovik, A.C., Sheikh, H. R., Simoncelli, E.P.: Image Quality Assessment: From Error Visibility to Structural Similarity. *2004 IEEE Transactions on Image Processing*, 13(4), 600–612.

OBLIQUE SURVIVAL TREES AND COMPETING RISKS

M. Krętowska

Faculty of Computer Science, Bialystok University of Technology, Poland

Abstract

Tree-based predictors are recently quite common approaches to analysis of survival data. In the paper an oblique survival tree is proposed for prediction of CIF function for an event of interest in the presence of competing risks. Induction of the tree is based on minimization of dipolar criterion function; pruning phase is conducted according to split-complexity method proposed by LeBlanc and Crowley. The predictive ability of the received tool is measured by c-index with its extension for competing risks.

Keywords: competing risks, dipolar criterion, survival tree

I. INTRODUCTION

Tree-based predictors are recently quite common approaches to analysis of survival data. They may be considered as alternative tools to statistical methods, that usually require many assumptions to meet. Survival trees dedicated to competing risks are narrowed to models with single variable tested in each internal node. The test usually takes a form of simple inequality: $v < a$ and $v \geq a$, where v is a variable and a is an real value. In case of oblique trees each internal node has a form of a hyperplane $H(w, \vartheta): \{v: w_1v_1 + w_2v_2 + \dots + w_Nv_N - \vartheta = 0\}$, where v_1, v_2, \dots, v_N are variables and $w_1, w_2, \dots, w_N, \vartheta$ are coefficients. The test divides the data into two subsets: the feature vectors which are placed on positive ($w_1v_1 + w_2v_2 + \dots + w_Nv_N - \vartheta \geq 0$) or on negative ($w_1v_1 + w_2v_2 + \dots + w_Nv_N - \vartheta < 0$) side of the hyperplane.

In the paper the algorithm of induction of survival trees dedicated to competing risks is presented.

III. METHODOLOGY

Let us assume, that we have a learning sample L with M observations described by (x_i, δ_i, t_i) , where x_i is a vector of N variables $x_{i1}, x_{i2}, \dots, x_{iN}$ describing i th subject, $\delta_i \in \{0, 1, \dots, K\}$ is a failure indicator representing the type of event occurred, t_i is a survival time. For censored subjects the failure indicator is equal to 0.

A cumulative incidence function (CIF) is used to describe the failure time distribution in case of competing risks. The CIF for the i th type of event is defined as:

$$F_i(t) = \int_0^t f_i(t) dt$$

and may be interpreted as the probability that an event of type i occurs before or at time t . The estimate of CIF is given as:

$$\hat{F}_i(t) = \sum_{j: t_j \leq t} \frac{dN_{ij}(t_j)}{n_j}$$

Our purpose is to build a binary, oblique survival tree which would be able to predict CIF for event of interest for a new subject.

We build a tree starting from the root node. On the base on the learning data, the optimization algorithm searches for a test which divides the feature vectors into two subsets. In case of oblique trees, a test has a form of a hyperplane. For each subset a new tree node is generated and the procedure is repeated. If certain conditions are fulfilled, the node becomes a terminal one. Terminal nodes are final nodes and do not contain any tests, they comprise only sets of subjects, that have reached the node. The subjects belonging to a terminal node constitute the output of the tree; in our approach the output is defined as a CIF function.

The optimization procedure which searches for a best test in each internal nodes is focused on dividing the feature space into areas, which would include the patients with similar survival times. The test hyperplane in a given internal node is obtained by minimizing a dipolar criterion function being a sum over some specified criterion functions connected with dipoles - pairs of different feature vectors ($\mathbf{x}_i, \mathbf{x}_j$) from the learning set [1]. Mixed dipoles are created between two subjects, that should be divided, while pure ones – between two subjects that should remain undivided. Taking into account censored cases and assumptions made by Fine and Gray [2], who treated the subjects failed for other causes as at risk at any time, I propose the following rules of dipole construction:

- a pair of feature vectors $\{\mathbf{x}_i, \mathbf{x}_j\}$ forms the pure dipole, if
 - $\delta_i = \delta_j = 1 \wedge |t_i - t_j| < \eta, z=1,2,\dots, p$
- a pair of feature vectors $\{\mathbf{x}_i, \mathbf{x}_j\}$ forms the mixed dipole, if
 - $\delta_i = \delta_j = 1 \wedge |t_i - t_j| > \zeta, z=1,2,\dots,p$
 - $(\delta_i = 0, \delta_j = 1 \wedge t_i - t_j > \zeta)$ or $(\delta_i = 1, \delta_j = 0 \wedge t_j - t_i > \zeta)$
 - $(\delta_i = z, \delta_j = 1)$ or $(\delta_i = 1, \delta_j = z), z=2,\dots, K$

Parameters η and ζ are equal to quartiles of absolute values of differences between survival times for all evaluable pairs of subjects. Based on earlier experiments, the parameter η is fixed as 0.3 quantile and ζ - 0.6.

The final tree is received by application of pruning algorithm, which reduces a part of tree branches and causes better prognostic ability. I use the split-complexity pruning [3], with Grey's splitting statistics and 10-fold cross-validation procedure.

The predictive ability of the received tool is measured by the c-Index with its extension for competing risks proposed by Wolbers at al. [4].

Acknowledgements

The present study was supported by a grant S/WI/2/2013 from Bialystok University of Technology and founded from the resources for research by Ministry of Science and Higher Education.

References

1. Bobrowski L., Kretowska M., Kretowski M. (1997) Design of neural classifying networks by using dipolar criteria, Proceedings of the Third Conference on Neural Networks and Their Applications, Czestochowa, pp. 689-694.
2. Fine J. P., Gray R.J., A proportional hazards model for the subdistribution of competing risk. Journal of American Statistical Association 1999, 94, 496-509.

3. LeBlanc M., Crowley J., Survival trees by goodness of split. *Journal of American Statistical Association* 88 (422), 457-467.
4. Wolbers M., Koller M.T., Witteman J. C. M., Steyerberg E.W., Prognostic models with competing risks. *Methods and application to coronary risk prediction. Epidemiology* 2009, 20(4), 555-561.

APPLICATION OF THE RELAXED LINEAR SEPARABILITY METHOD FOR THE ANALYSIS OF GENOMIC DATA

T. Łukaszuk¹, A. Gyenesei^{2,4}, L. Bobrowski^{1,3}

¹*Faculty of Computer Science, Bialystok University of Technology, Bialystok, Poland*

²*Bioinformatics & Scientific Computing, Vienna Biocenter Core Facilities (VBCF GmbH), Vienna, Austria*

³*Institute of Biocybernetics and Biomedical Engineering, PAS, Warsaw, Poland*

⁴*Center for Bioinformatics and Data Analysis, Medical University of Bialystok, Bialystok, Poland*

Abstract

Feature selection techniques are used for reduction as many as possible number of irrelevant or redundant features in a given classification or regression problem. Feature selection problem is particularly important and challenging in the case when the number of patients is low in comparison to the number of features used to characterise these patients. Such situation appears typically in exploration of genomic data sets. The Relaxed Linear Separability (RLS) is the method of feature subset selection developed by us currently. New results of an application of the RLS method to important genomic data set is described in the paper.

I. INTRODUCTION

Data mining tools are aimed at extraction of useful patterns in learning data sets [7,9]. Particular data mining tools which are based on minimization of the convex and piecewise linear (CPL) criterion functions are developed and applied by us for the solution of various pattern recognition problems [5,6]. The perceptron criterion function belongs to the considered family of the CPL functions. This criterion function originated from the concept of linear separability of multivariate data sets.

Feature selection techniques are expected to allow finding such feature subsets, which are beneficial for the solution of practically important problems of classification or prognosis. The Relaxed Linear Separability (RLS) method of feature subset selection is based on the minimization of the perceptron criterion function and used for evaluating the degree of linear separability of learning data sets in various feature subspace [3,4]. Using the RLS for feature selection from large, multidimensional data sets can be based on computational techniques, which are called the basis exchange algorithms [1]. Basis exchange algorithms are similar to the Simplex algorithm from linear programming. They are applied by us in efficient minimization of the CPL criterion functions.

The feature selection problem is particularly important and challenging in the cases of genomic data sets, when the number of cases is low in comparison to the number of features (genes) used to characterise these patients. The presented paper contains new results of application of the RLS method to selection of important feature from genomic Type 1 Diabetes data set (T1D) obtained from a genetic screening study for T1D susceptibility.

II. THE RELAXED LINEAR SEPARABILITY (RLS) FEATURE SELECTION METHOD

The Relaxed Linear Separability (RLS) is the method of feature selection. This approach to the feature selection problem refers to the concept of the linear separability of the learning sets C^+ and C^- , containing m ($m = m^+ + m^-$) feature vectors \mathbf{x}_j [2].

$$\mathbf{x}_j = [x_{j1}, \dots, x_{jn}]^T \quad (j = 1, \dots, m)$$

The sets C^+ and C^- are linearly separable if and only if they can be fully separated by some hyperplane $H(\mathbf{w}, \theta)$:

$$(\exists \mathbf{w}, \theta) \quad (\forall \mathbf{x}_j \in C^+) \quad \mathbf{w}^T \mathbf{x}_j > \theta \quad \text{and} \quad (\forall \mathbf{x}_j \in C^-) \quad \mathbf{w}^T \mathbf{x}_j < \theta$$

The term ‘‘relaxation’’ means the deterioration of the linear separability due to the gradual neglecting of selected features. The considered approach to feature selection is based on repetitive minimization of the CPL criterion functions $\Phi_\lambda(\mathbf{w}, \theta)$.

$$\Phi_\lambda(\mathbf{w}, \theta) = \sum_{\mathbf{x}_j \in C^+} \varphi_j^+(\mathbf{w}, \theta) + \sum_{\mathbf{x}_j \in C^-} \varphi_j^-(\mathbf{w}, \theta) + \lambda \sum_{i=1, \dots, n} |w_i|$$

$$\varphi_j^+(\mathbf{w}, \theta) = \begin{cases} 1 + \theta - \mathbf{w}^T \mathbf{x}_j & \text{if } \mathbf{w}^T \mathbf{x}_j < 1 + \theta \\ 0 & \text{if } \mathbf{w}^T \mathbf{x}_j \geq 1 + \theta \end{cases}$$

$$\varphi_j^-(\mathbf{w}, \theta) = \begin{cases} 1 - \theta + \mathbf{w}^T \mathbf{x}_j & \text{if } \mathbf{w}^T \mathbf{x}_j > -1 + \theta \\ 0 & \text{if } \mathbf{w}^T \mathbf{x}_j \leq -1 + \theta \end{cases}$$

The RLS feature selection method consists of three stages [3,4]. The first stage is to determine an optimal hyperplane $H(\mathbf{w}^*, \theta^*)$ separating objects \mathbf{x}_j from the learning sets C^+ and C^- . This stage results in the optimal hyperplane $H(\mathbf{w}^*, \theta^*)$ and an initial feature set F_k composed of k features.

In the second stage, the value of the cost level λ in the criterion function $\Phi_\lambda(\mathbf{w}, \theta)$ is successively increased. This causes the reduction of some features x_i as a result of the zeroing of corresponding weights w_i . As a result of the second stage, we obtain the descended sequence of feature subsets F_k with decreased dimensionality.

The last step is the calculation of the classifier accuracy in each reduced dataset corresponding to the subsets of features in sequence F_k, F_{k-1}, \dots, F_1 . As a result of the third stage and the whole RLS method, we obtain the feature set F^* . This is the feature set characterized by the greatest accuracy of classifier.

III. COMPUTATIONAL EXPERIMENT

1. Data set

The T1D data set contains information about 3463 patients diagnosed with type 1 diabetes. The data are well balanced, because 1705 (49.23%) objects concern disease person and 1758 (50.77%) objects representing healthy persons. Each object is described by 154 features, among which we can distinguish Gender, Age at Diagnosis, HLA genotype, HLA subgroup, HLA riskgroups and 128 single nucleotide polymorphisms (SNPs, genetic information).

2. The course of the experiment

To be able to use the RLS method for data set, a data set should be represented in the form of a matrix of numerical values. The matrix should not contain missing values. Since the T1D

data set did not meet the above conditions, it had to be subjected to a preliminary preparation. Preprocessing included missing data imputation and coding SNP features.

The raw T1D data set was missing a total of about 25% values (see Figure 1). To obtain the data set without missing values, we performed two preprocessing steps. At the beginning we removed the 1637 objects that have more than 35% of missing values and 4 features that have more than 35% of missing values. The remaining about 4% missing values we filled out using 1NN imputation [8].



Fig. 1. Fractions of objects containing missing values, e.g. value $\sim 52\%$ for maximum percent of missing values equals 35% means that there are 52% objects containing at most 35% missing values each.

Features representing SNPs are single-nucleotide polymorphisms; variation in a single nucleotide that occurs at a specific position in the genome. Since the RLS method requires that the features have only numeric values, the features representing SNPs have to be converted into numerical values. In order not to favor any of the original feature value representing the SNP, we have replaced each of these features by a few new binary features. One primary feature was replaced by n new features, where n equals the number of different values which the primary feature adopts. The values of the new features are 1 or 0: 1 where the value of the original feature was equal to the value represented by the new feature, and 0 otherwise. The example of described transformation is shown in Figure 2.

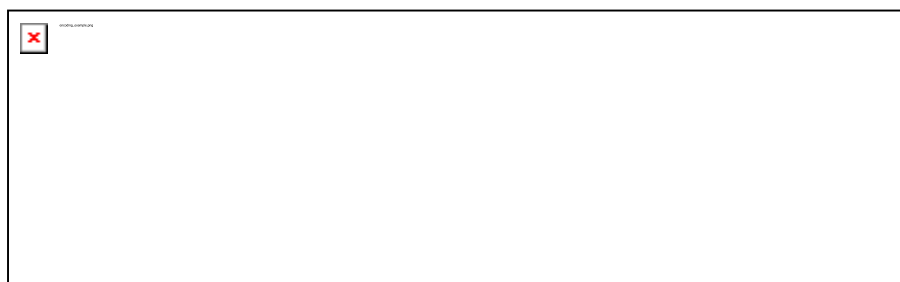


Fig. 2. The example of coding genetic feature. $rs479955_chr10_NA$ is an original feature, the $rs479955_chr10_NA$ (CT), $rs479955_chr10_NA$ (CC) and $rs479955_chr10_NA$ (TT) are new coded features.

After the preprocessing procedures, we received data set contains 1826 objects, each object described by 396 features. On the prepared data set we used the RLS method of features selection.

3. Results

Figure 3 illustrates both the apparent error (AE) and the cross validation error (CVE) values in feature subspaces tested by the RLS procedure. A significant change in the trend of errors is visible in the feature subspace of size 6. In connection with this, the feature subspace of size 6 has been selected as the result of our feature selection procedure.

6 selected features:

- HLA riskgroup
- rs689 chr11 INS (AA)
- rs2476601 chr1 PTPN22 (GG)
- rs1409338 chr10 PLXDC2 (GG)
- rs3087243 chr2 CTLA4 (GG)
- rs1701704 chr12 NA (AA)

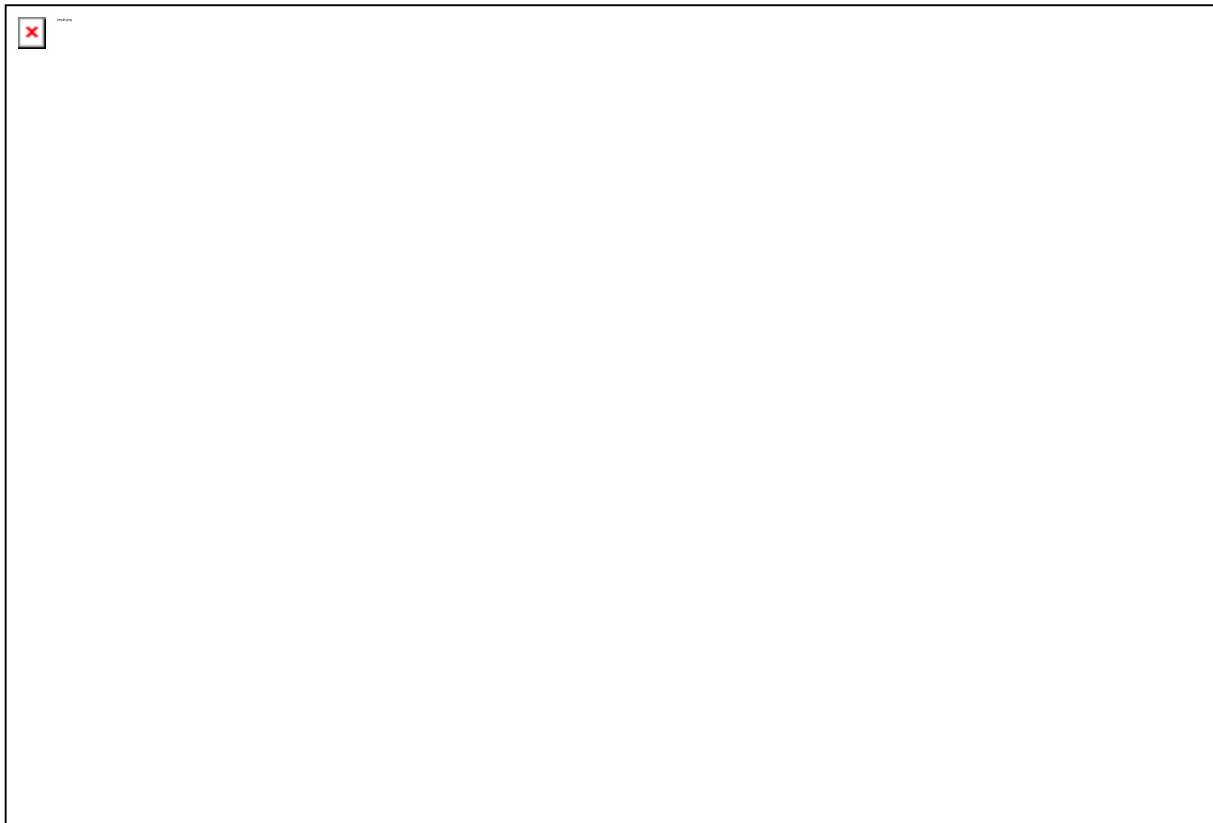


Fig. 3. The apparent error rate (AE) and the cross-validation error (CVE) in different feature subspaces F_k .

IV. CONCLUDING REMARKS

The above-described research is a preliminary attempt to analyze the T1D data set. The results must be thoroughly evaluated by experts in the field of medicine and genetics to found practical application in the future. Currently, a cursory evaluation of the set of selected features indicates that these are the features (genes) which have a significant association with diabetes. This relationship was also confirmed in other independent studies.

References

1. Bobrowski L. Design of Piecewise Linear Classifiers from Formal Neurons by Some Basis Exchange Technique. *Pattern Recognition*, 24(9):863–870, 1991.
2. Bobrowski L. Feature subsets selection based on linear separability, Lecture notes of the VII-th ICB seminar: statistics and clinical practice, 2008.
3. Bobrowski L, Łukaszuk T. Feature selection based on relaxed linear separability. *Biocybern Biomed Eng* 2009; 29(2):43–59.
4. Bobrowski L, Łukaszuk T. Relaxed linear separability (RLS) approach to feature (gene) subset selection. In: Xia X, editor. *Selected works in bioinformatics*. Rijeka, Croatia: InTech; 2011. URL: <http://www.intechopen.com/articles/show/title/relaxed-linear-separability-rls-approach-to-feature-gene-subset-selection>.
5. Bobrowski L., Łukaszuk T., Lindholm B., Stenvinkel P., Heimburger O., Axelsson J., Bárány P., Carrero J.J., Qureshi A.R., Luttrupp K., Debowska M., Nordfors L., Schalling M., Waniewski J. Selection of genetic and phenotypic features associated with inflammatory status of patients on dialysis using relaxed linear separability method. *PloS one*. 2014.
6. Bobrowski L., Łukaszuk T. Prognostic modeling with high dimensional and censored data. In *Industrial Conference on Data Mining 2012 Jul 13* (pp. 178-193). Springer Berlin Heidelberg.
7. Han J., Pei J., Kamber M. *Data mining: concepts and techniques*. Elsevier; 2011.
8. Nguyen D.V., Wang N., Carroll R.J. Evaluation of missing value estimation for microarray data. *Journal of Data Science*. 2004; 2(4):347-70.
9. Witten I.H., Frank E. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

COMPARISON OF SUPPORT VECTOR REGRESSION AND CLASSICAL TRANSFORMED LINEAR MODEL IN CONSTRUCTED WETLAND TREATMENT CONTEXT

P. Malinowski¹; W. Dąbrowski²; B. Karolinczak²

¹*Medical University of Białystok, Białystok, Poland*

²*Białystok Technical University, Białystok, Poland*

Abstract

Due to inherent complex structure of underlying process, it may be difficult to build a successful model of process, using traditional statistical inference. Search through non-linear combination of variables usually suffers from curse of dimensionality, and requires great experience and data insight. On the other hand, data mining methods, like SVM may offer increased performance and flexibility. This work presents a comparison of classical linear regression, and SVR in study of organic nitrogen removal from waste water.

Keywords: data mining, regression, SVM

I. INTRODUCTION

Even when feature number in dataset is small with compare to observations, simple reasoning using traditional statistical methods may be difficult. For regression methods, a number of assumptions must be met to make a valid inference from obtained model. Those assumptions are checked after the model is created. Standard linear regression model is not sufficient to many kind of data, therefore multiple extensions were added to it. Data may be transformed according to a link function, or extended to fractional power, variance structure may be extended, and so on. When appropriate regression model is found, however, it usually tells something about the nature of the studied process. Some processes are inherently complicated, therefore searching for good model take some time. Data mining regression models, like support vector regression, or deep networks usually have less specific assumptions, and can be way more flexible. This flexibility often comes with a price of excessive tuning to observed data, reducing generality of obtained solution. Overfitting can be controlled, but this requires additional computing power. If depth of learning process is sufficient, it may be impossible to infer relations between original variables and the result. Such trained deep algorithm has semantics similar to a black box. In some cases one can, however, reconstruct some meta-features and their relation to a solution. When price of flexibility is less significant, than potential gains, choice of data mining regression may be preferred.

II. THE DATASET

Reject water is typical byproduct in every biological waste water treatment plants (WWTP). It is usually returned to the main sewage line causing problems with stable and effective work of treatment plants. Constructed wetland method is well known for treatment of different kind of sewage. The biggest advantage of this method is simplicity, low operating and constructing costs and high efficiency, especially in nitrogen removal. Full scale application of constructed wetland system would result in a significant decrease of pollutants load in rejected water.

The dataset was collected in research installation located in dairy WWTP. Three kinds of vertical flow constructed wetlands (VFCW) were checked to prove their efficiency for treatment of sewage (reject water) with high ammonia concentration [2]. Hydraulic load was the same for each VFCW kind (0.1 m^3 per 1 m^2 of bed surface per day). VFCW were filled with gravel and sand and planted with reeds (*Phragmites australis*). Fig. 1 presents installation and cross section of VFCW. Tab. 1 presents also thickness of the VFCW layers.



Fig. 1. Installation 3 and VFCW cross section scheme [2]

Tab. 1. Details of VFCW vertical cross section [2]

Layer	Material	Thickness in VFCW 1	Thickness in VFCW 2	Thickness in VFCW 3
A	sand (0-2 mm)	0.15 m	0.30 m	0.15 m
B	gravel (2-8 mm)	0.15 m	0.25 m	0.35 m
C	gravel (8-20 mm)	0.20 m	0.30 m	0.15 m
D	gravel (20-80 mm)	0.15 m	0.15 m	D layer not present

During the study course (years 2009 and 2010) such parameters as: BOD_5 ; COD; nitrogen, including organic and inorganic; phosphorus and suspended solids were measured. Obtained data was used to calculate pollutants loads and efficiency of their removal, including organic nitrogen. Additional environmental parameters as temperature of air and reject water were measured on daily basis, to evaluate their influence on the process.

Analyzed dataset consists of 224 observations, each described by 7 features:

- growing season ("1" for april - october period, "0" otherwise)
- temperature of air and reject water
- organic nitrogen intake load
- indicators of beds A and B ("1" for data obtained from given bed)
- organic nitrogen removal efficiency (dependent variable), defined as in [3]

Increase in prediction power is especially desired for selected dependent variable because of low predictive power of original regression model ($R^2 = 0.4$) [3], partially due to lack of dependency on intake load, and complex nature of nitrogen removal process.

III. METHODOLOGY

Organic nitrogen removal efficiency was estimated using the ν support vector regression algorithm (ν -SVR). ν -SVR [4] belongs to a family of linear classifiers and regression algorithms [1]. In case of regression, the algorithm tries to build a hyperplane as close to the observed data, as possible. This results in convex optimization problem, with can be effectively solved. Obviously, non linear patterns cannot be represented by a linear-one. ν -SVR operates on transformed feature space by utilizing kernel trick, where such nonlinear

pattern may become linear. Dimensionality of this transformed space can vastly outnumber original one. In this work, a Gaussian kernel is used, due to its key properties in original feature space [5]. Three free parameters were tuned:

- a control parameter ν
- a cost parameter C
- a kernel flatness parameter γ

To address issue of overfitting, cross validation procedure was performed on top of ν -SVR. Additionally, 30% of the data, selected at random, was retained to from training part of cross validation loop to form validation set. Best model was obtained by minimizing mean squared error (MSE) between observed and estimated value of dependent feature.

IV. RESULTS

Efficiency of organic nitrogen removal was modeled using ν -SVR, with 10-fold cross validation. R language implementation of ν -SVR in package e1071 [6] was used. Following parameters range was checked:

Tab. 2. Tuned parameters range

Parameter	Range
ν	0.30-0.95 by 0.05, linear scale
C	$2^{-20} - 2^{+20}$, base 2 logarithmic scale, multiplied by 4
γ	

Fig. 2 presents result of training for optimal value of $\nu = 0.3$, along with optimal values of other parameters and mean squared error (MSE). Due to large variation in MSE, value of this parameter is presented here in the logarithmic scale (natural based).

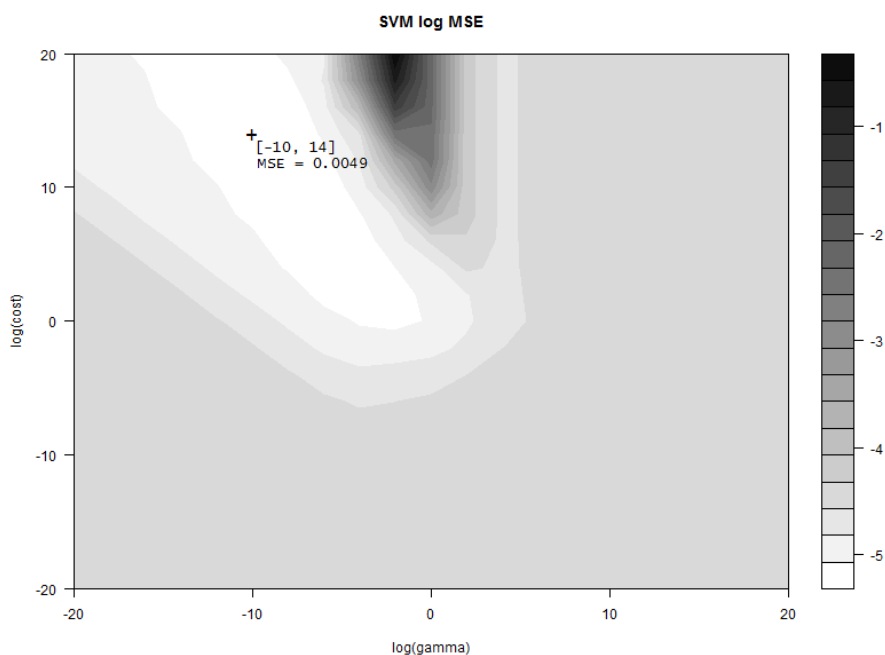


Fig. 2. Logarithm of MSE for the optimal ν value

Tab. 3 presents comparison of MSE obtained from linear model with transformation [3] and that obtained from validation part of data on trained ν -SVR.

Tab. 3. Results comparison

	Linear model with transformation	ν -SVR
MSE	0.0062	0.0051

V. DISCUSSION

MSE obtained from validation part of dataset was almost the same as reported by training procedure. To fully compare results of ν -SVR and original model one must take into consideration, that latter was trained without cross-validation. While it is good for statistical inference, model overfitting was not determined. Even with such advantage ν -SVR model has better accuracy by about 10% (on scale of dependent variable). Selected model has rather low bias and sharp details, due to combination of high cost, and low gamma parameters value. ν is the lower bound for proportion of support vectors w.r.t. full dataset. It also controls the upper bound on observations count with non-acceptable deviance from calculated pattern. Small value of ν is result of good accordance of data to the selected model.

VI. CONCLUSION

This work presents application of ν -SVR algorithm to analysis of organic nitrogen removal in VFCW, and compares it with earlier results. Together with additional information in for of original pollutant load, trained algorithm outperformed linear transformed model, increasing accuracy of prediction by 10% (on scale of dependent variable). Further research are needed to obtain even better results. Better conformance can lead to better predictability of work of treatment plant, connected to VFCW, even further decreasing its operating costs.

References

1. Boser B. E., Guyon I. M., Vapnik V. N. A training algorithm for optimal margin classifiers. In: Haussler D. (Ed.), Proceedings of the Annual Conference on Computational Learning Theory. Pittsburgh, PA, ACM Press, 1992, 144–152.
2. Dąbrowski W. Oczyszczanie odcieków z oczyszczalni mleczarskich w systemach hydrofitowych. Białystok, Polska, Oficyna Wydawnicza Politechniki Białostockiej, 2004, 97-100.
3. Dąbrowski W. Oczyszczanie odcieków z oczyszczalni mleczarskich w systemach hydrofitowych. Białystok, Polska, Oficyna Wydawnicza Politechniki Białostockiej, 2004, 162-179.
4. Scholkopf, B., Smola A. J., Williamson R. Shrinking the tube: A new support vector regression algorithm. In Kearns M. S., Solla S. A., and Cohn D. A. (Eds.), Advances in Neural Information Processing Systems. Cambridge, MA. MIT Press, 1999 (11).
5. Smola A. J., Scholkopf B.: A tutorial on support vector regression. Statistics and Computing , 2004, vol. 14, 199-222.
6. Meyer D., Dimitriadou E., Hornik K., Weingessel A., Leisch F. e1071: Misc Functions of the Department of Statistics (e1071), TU Wien. R package version 1.6-4. 2014. Retrieved from <http://CRAN.R-project.org/package=e1071>.

ASSESSING AGREEMENT BETWEEN TWO MEASUREMENT METHODS OF CORE BODY TEMPERATURE

K. Sałapa¹, T. Darocha^{2,3,4}, J. Majkowski³, T. Sanak^{3,5,6}, P. Podsiadło^{4,7}, S. Kosiński^{2,3,8},
R. Drwiła^{2,3}

¹*Department of Bioinformatics and Telemedicine, Jagiellonian University Medical College, Cracow, Poland*

²*Severe Accidental Hypothermia Center, Department of Anesthesiology and Intensive Care, the John Paul II Hospital, Jagiellonian University Medical College, Cracow, Poland*

³*Foundation Heat for Life, Cracow, Poland*

⁴*Polish Medical Air Rescue, Warsaw, Poland*

⁵*Department of Disaster Medicine and Emergency Care, Medical College of Jagiellonian University, Krakow, Poland*

⁶*Department of Combat Medicine, Military Institute in Warsaw*

⁷*Polish Society for Mountain Medicine and Rescue, Szczyrk, Poland*

⁸*Department of Anesthesiology and Intensive Care, Pulmonary Hospital, Zakopane, Poland*

Abstract

The aim of the study is assessment of accuracy of new mobile system to estimate core body temperature in comparison with the reference thermometer. Concordance correlation coefficient (CCC) and limits of agreement analyses were applied to verify the agreement of those two measurement methods. CCC was estimated based on the variance components of a linear mixed model, the variance components and fixed effects were estimated using restricted maximum likelihood (REML). Furthermore, the linear regression was performed to describe the reference method given the various temperatures measured by the new one and to calculate 95% prediction intervals. The latter analysis was applied because the assumptions of the limits of agreement analysis were not valid. The results showed that the new system performed well in vitro, provided a good correlation and a clinically acceptable agreement, in comparison with the reference thermometer.

Keywords: agreement assessment, concordance correlation coefficient, limits of agreement, linear regression

I. INTRODUCTION

The accidental hypothermia is a life-threatening situation and patient's core body temperature is crucial information. Esophageal temperature provides key data for deciding whether or not to continue or to withhold resuscitative efforts in asystolic patients. The aim of the study is assessment of accuracy of new mobile system to core body temperature monitoring in comparison with the reference thermometer.

Assessing the agreement between two measurement methods is a common statistical goal. Among the methods to verify agreement when data is continuous, the concordance correlation coefficient (CCC), introduced by Lin [1], has risen as one of the most used approaches. It measures the variation of their linear relationship from the 45° line through the origin (concordance line, perfect agreement) and has components of precision (how far each observation deviates from the line fit to the data) and accuracy (how far this line deviates from the 45° line through the origin) [2]. Pearson correlation coefficient can be very misleading in such situations. Any departures from the concordance line would produce CCC below 1 even if Pearson correlation coefficient is equal to 1. This is because the latter

coefficient measures a linear relationship between two continuous data but fails to detect any departures from the equality line [1]. Lin expressed CCC as a function of the means, variances and covariances of the bivariate distribution of two methods [1]. Carrasco and Jover [3] showed that it may be expressed in terms of the variance components of a linear mixed model, the variance components and fixed effects are estimated using restricted maximum likelihood (REML). One important feature of the REML is that it gives unbiased and asymptotically normally distributed estimates of the variance components [3, 4]. This approach allows the CCC to be defined for more than two methods, for repeated measures and can be adjusted by adding the covariates to the model [4].

Furthermore, the analysis of limit of agreement was applied as a graphical technique to analyse the assessment of repeatability of two methods, proposed by Bland and Altman [5, 6]. The idea of this analysis is to plot the individual difference between measurements taken from two different methods against their individual mean. It allows to investigate the size of discrepancies between two measurements and to judge whether they are clinically acceptable or not. Clinically acceptable limits of agreement should be defined a priori. It is assumed that there is no relation between the individual differences and the means and that the differences are normally distributed [6]. In case of lack independence between differences and means a logarithmic transformation to the raw data was proposed to reduce the significant relationship [5, 6]. If it fails the alternative analysis is proposed by Bland and Altman [5, 7], i.e. least squares regression to predict the measurement obtained by the old methods from the measurement obtained by the new one, and calculate 95% prediction interval for the old methods depending on the various temperatures measured by the new one. This approach gives something similar to the limits of agreement [7]. This is a calibration approach and does not directly answer the question of comparability.

II. METHODOLOGY

The new system allowing to measure the core body temperature were tested with a reference thermometer. The thermometers were tested simultaneously in a the same water bath at different temperatures between 10°C and 42°C. The temperature increments were .5°C for each measurement point (N=65).

Agreement between two continuous data measured from two different measurement methods was evaluated by means of Concordance Correlation Coefficient (CCC) [1]. It was estimated by variance components to reduced bias of the moment-methods estimator proposed by Lin [4]. Furthermore, the analysis of limit of agreement was applied [5-7]. Clinically acceptable limits of agreement were a priori defined to be $\pm 1^\circ\text{C}$. The assessment of relationship between the individual differences and means was based on Spearman rank correlation coefficient because both features were not normally distributed. The Shapiro-Wilk test was applied to check normality. In case of lack independence between differences and means a logarithmic transformation was applied to the raw data to reduce this significant relationship [5-7]. Unfortunately, this approach failed and the alternative analysis proposed by Bland and Altman, i.e. least squares regression was applied. The measurements obtained by the reference thermometers were predicted on the measurement obtained by the new system, then 95% prediction interval for the reference methods were calculated depending on the various temperatures measured by the new one system [7]. Bonferroni correction was applied to calculate prediction limits for a few new observations.

The statistical analysis were performed by means of R software (packages: cccrm, BlandAltmanLeh and lmtest).

III. RESULTS

Firstly, agreement between the new system and the reference thermometer was evaluated. Measurements of the first one are plotted against the measurements of second one in *Fig. 4* (the left panel). It can be seen, within a tolerable error, that the measurements fall on the 45° line through the origin. This indicate that the reference thermometer is highly reproducible by the new system. The concordance coefficient with 95% confidence interval is equal to .9999 (.9998, .9999). The source of slight disagreement is the random error (variance estimate of .0082) rather than the systematic differences between thermometers (variance estimate of .005).

The left panel of *Fig. 4* presents comparison of measurements by means of the limits of agreement plot. The mean difference is .102 (95%CI: .069, .133), hence, the new system overestimate the measurement of temperature, by between .069 and .133. The new system gave higher results in 38 measurement points (58.46%), exactly the same in 25 cases (38.46%) and lower results in 2 cases (3.08%) than the reference one. The limits of agreement lie in the range of -.149 (95%CI: -.204, -.095) and .353 (95%CI: .298, .408). It should be noticed that the assumptions of this analysis are not met. The differences are not normally distributed ($p < .001$) and the differences are not independent on the mean, i.e. the individual differences decrease as the averages of measurements increase ($R = -.64$, $p < .001$). The logarithmic transformation was applied to the data, but unfortunately it did not improve the results.

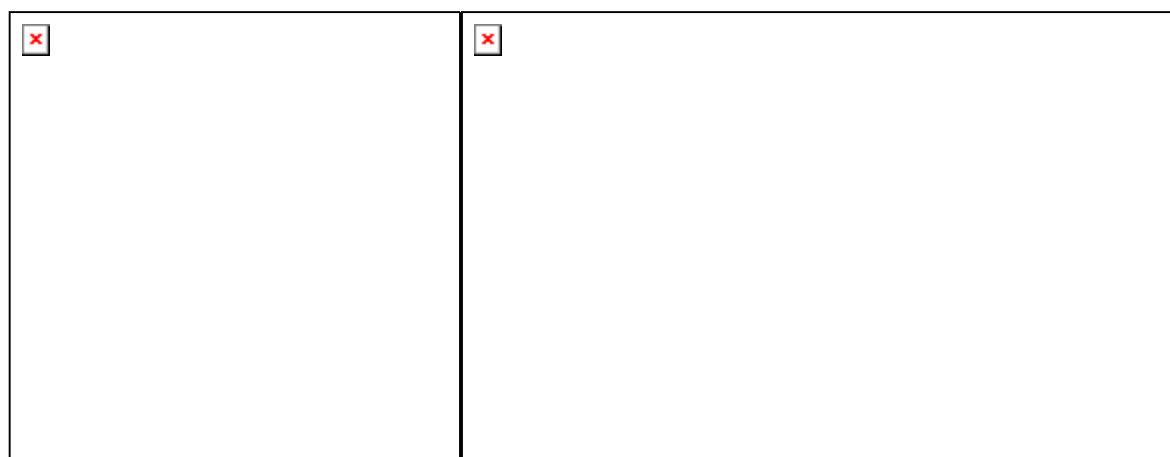


Fig. 4. New system measurements versus spacelab measurements in przelyk (left panel). Dashed lines represent the 45° line through the origin. Bland-Altman plot for przelyk (right panel).

The regression analysis was applied to describe the negative trend and to calculate 95% prediction intervals for the reference method given the various temperatures measured by the new methods. It was found that the measurement of the reference thermometer lie between 9.50°C and 10.04°C with probability of .95 when the new method reading is equal to 10°C, between 19.58°C and 20.12°C when the new method measurement is equal to 20°C, between 29.66°C and 30.20°C when the new method reading is equal to 30°C and between 39.74°C and 40.28°C when the new method reading is equal to 40°C. Thus, the presented prediction intervals for the old method were within the limit of $\pm .54^\circ\text{C}$.

IV. DISCUSSION & CONCLUSION

The early hypothermia recognition allows to perform suitable management and to reduce mortality among trauma patients. The treatment of a patient in accidental hypothermia should

be modified, depending on patient's core body temperature. The study showed that the new system proposed by the authors to estimate core body temperatures performed well in vitro, provided a good correlation and a clinically acceptable agreement, in comparison with the reference thermometer. The main limitation of this study is that the new system was evaluated with only one reference thermometer.

References

1. Lin L. I.-K., A concordance correlation coefficient to evaluate reproducibility, *Biometrics* 1989, 45, 255–268.
2. King T.S., Chinchilli V. M., Carrasco J. L.: A repeated measures concordance correlation coefficient. *Statistics in medicine* 2007, 26, 3095–3113.
3. Carrasco J. L., Jover L., Estimating the Generalized Concordance Correlation Coefficient through Variance Components, *Biometrics* 2003, 59, 849–858.
4. Carrasco J. L., Phillips B. R., Puig-Martinez J., King T. S., Chinchilli V. M., Estimation of the concordance correlation coefficient for repeated measures using SAS and R, *Computer methods and programs in biomedicine* 2013, 109, 293–304.
5. Altman D. G., Bland J. M., Measurement in Medicine: the Analysis of Method Comparison Studies, *Statistician* 1983, 32, 307–317.
6. Bland J. M., Altman D. G., Statistical methods for assessing agreement between two methods of clinical measurement, *Lancet* 1986, 1, 307–310.
7. Bland J. M., Altman D. G., Applying the right statistics: Analyses of measurement studies, *Ultrasound in Obstetrics & Gynecology* 2003, 22, 85–93.

The special session addressed to medical doctors

**STATISTICAL STANDARDS
IN MEDICAL RESEARCH PROJECTS
AND PUBLICATIONS**

*The session organized by the Polish National Group
of the International Society for Clinical Biostatistics (ISCB)*

SCALE OF CAREER SATISFACTION IN MEDICINE

J. N. Peña-Sánchez¹, M. Górkiewicz²

¹*Department of Community Health and Epidemiology, University of Saskatchewan,
Saskatoon, Canada*

²*Faculty of Health Sciences, Jagiellonian University of Krakow, Krakow, Poland*

Evaluation of attitudes and individual satisfaction is one of the most important problem in real-world investigations. In this study the comparison of English and Polish versions of 4CornerSAT questionnaire to measure career satisfaction of physicians is presented. The 4CornerSAT questionnaire originally was created in English, but later this was adapted to Polish and Spanish. The 4CornerSAT had four scales, related to personal, professional, inherent and performance dimensions of career satisfaction. Each scale had four items, each scored on 6-point Likert scale: 1=very.dissatisfied; 2=satisfied; 3=somewhat.dissatisfied; 4=somewhat.satisfied; 5=satisfied; 6=very.satisfied. The practical usefulness and validity of 4CornerSAT questionnaire were supported in several studies in the matter.

META-ANALYSIS IN BIOMEDICAL RESEARCH WYKORZYSTANIE METAANALIZY W BADANIACH BIOMEDYCZNYCH

M. Polak

*Institute of Public Health, Faculty of Health Sciences, Jagiellonian University Medical
College, Krakow, Poland*

Meta-analysis is a statistical technique for combining the findings from independent studies. Outcomes from a meta-analysis may include a more precise estimate of the effect of treatment or risk factor for disease, than any individual study contributing to the pooled analysis. As with any statistical procedure, meta-analysis has its strengths and limitations, but is now one of the standard tools for providing transparent, objective, and replicable summaries of research findings in the social sciences, medicine and other fields. The objectives of the presentation are to provide an introduction to meta-analysis and to discuss the rationale for this type of research and other general considerations using a medical data. We will then delve into fixed-, and random/mixed-effects models for combining the observed outcomes and for examining whether the outcomes depend on one or more moderator variables.

THE SELECTION OF A REPRESENTATIVE SAMPLE IN MEDICAL RESEARCH

U. Cwalina¹, D. Jankowska¹, A. J. Milewska¹, D. Citko¹, R. Milewski¹

¹*Department of Statistics and Medical Informatics, Medical University of Bialystok, Poland*

Abstract

The data in medical research should be carefully selected. Properly prepared study requires representative sample, what should be taken into account on the stage of planning the study. There are many aspects that should be considered before beginning of the data collection process. Insufficient attention at an early stage can lead to very serious consequences. In an extreme case database may become useless in the study. However, sometimes research must be carried out despite the fact that sample is unrepresentative. In such a situation, conclusions have to be fairly drawn and carefully formulated. It should also be noted that each research requires an individual approach. Copying scheme of procedures from other studies may lead to draw wrong conclusions.

I. INTRODUCTION

The term “representative sample” is far from clear. In medicine it depends on the type of the research. For example, representative sample in cohort study means something else than in clinical trial. The definition of this term depends on the context in which it is used. In 1979 Altman and Mosteller carried out an analysis of ways of defining “representative sampling” [1], [2], [3]. They studied the use of this term in scientific and non-scientific literature. Despite the passage of years meaning of “representative sampling” does not change.

There may be enumerated following meanings of the term “representative sample”:

- Sample selected at random;
- Absence of confounding factors;
- Miniature of the population;
- Coverage of the population;
- Allows for good estimation of population parameters;
- Good enough for particular purpose.

However, regardless of what the definition researcher accepts, checking if sample meets this definition would be very difficult.

II. STAGES OF SAMPLING AND ITS REPRESENTATIVENESS

A lot of factors affect on the representativeness of the sample. Mistakes made at each stage of sampling may cause that the proposed analysis will be impossible to carry out. Hypotheses that have been formulated can remain unresolved. Population should be defined in a precise, accurate and clear way. It is important to prepare current and complete sampling frame. The next step is to choose the sample selection process. In the case of random sampling method (such as simple random sampling, systematic sampling, stratified random sampling, and cluster random sampling) each unit has the same chance to be chosen to the sample. Samples selected in this manner are more representative and less biased compared to a non-random samples. It is also important that the sample should have adequate number of elements. The

minimum sample size should be determined based on the fixed level of significance and power of the test. It is well known that the higher the number of units the better the quality of the study. Note, however, that “Samples which are too small can prove nothing; samples which are too large can prove anything” [4].

III. SELECTION BIAS

The sample which is not representative can be called biased sample. There are a number of factors that may cause the sample biased [4]. Several examples of biased samples are presented below. Furthermore it is shown as the bias may affect that the results become misinterpreted.

In medicine there are often compared two or more interventions. Patients should be assigned to groups at random. Unfortunately, sometimes this decision depends on investigator, what may lead to occurrence of the selection error.

An example of such situation would be a Simpson paradox [5]. The authors compared the efficacy of different treatments of renal calculi. Table 1 shows the number of successes obtained during the open surgery and percutaneous nephrolithotomy. Success was defined as no stones at three months or stone reduced to particles lower than 2 mm in size.

Table 1. A comparison of the number of successes between open surgery and percutaneous nephrolithotomy.

	Open surgery		Percutaneous nephrolithotomy	
	n	%	n	%
Success	273	78%	289	83%
Failure	77	22%	61	17%
Total	350	100%	350	100%

It can be concluded that percutaneous nephrolithotomy has greater efficacy (83%) compared to open surgery (78%). In the analyzed example confounding variable is the size of the stones. It turns out that taking into account this variable, the conclusion is opposite. Open surgery was characterized by greater efficiency in the case of small (93% vs. 87%) and large (73% vs. 69%) stones (Table 2 and 3).

Table 2. A comparison of the percentage of successes between open surgery and percutaneous nephrolithotomy for stones of a mean diameter of less than 2 cm.

	Open surgery		Percutaneous nephrolithotomy	
	n	%	n	%
Success	81	93%	234	87%
Failure	6	7%	36	13%
Total	87	100%	270	100%

Table 3. A comparison of the percentage of successes between open surgery and percutaneous nephrolithotomy for stones of a mean diameter of 2 cm or more or with multiple stones.

	Open surgery		Percutaneous nephrolithotomy	
	n	%	n	%
Success	192	73%	55	69%
Failure	71	27%	25	31%
Total	263	100%	80	100%

Patients were not assigned for treatment at random. The method of treatment was dependent on the medical condition of patients. In the cases of poorer health more frequently surgical procedures were performed. Patient who were treated with percutaneous nephrolithotomy were characterized by better general health and smaller stones. Comparing these two methods of treatment without taking into account the size of the stone is a serious mistake.

Another example of a non-random assignment of patients to the groups is described below. The study concerns two methods of dental treatment. Assume that the researchers want to compare two methods of dental filling: A and B. Then they have decided to use a material A always on the right side of the mouth, while the material B always on the left side. One year after intervention it turns out that material B is stronger. One would conclude that the material A is poorer quality. However, it should be taken into account that majority of the respondents are right-handed. It can be find that not the difference between the materials would affect the results. The difference might have been caused by the fact that right-handed people better care of hygiene on the left side of the mouth.

Medical data are often collected in a particular health center. In this case the results should be generalized very carefully. For example, hospital patients are not representative for the whole population from which they come. They were not randomly selected to the sample. Joseph Berkson in 1946 wrote about the limitations that occur analyzing hospital data [6]. The following example illustrates the mechanism of Berkson paradox. It can be considered the population of 2500 people and two diseases: A and B. Assume that during the relevant period 250 of them went to the hospital. Tables 4 and 5 present the incidence of these diseases in the sample and in the population, respectively.

Table 4. Incidence of diseases A and B among hospital patients.

	Disease A YES	Disease A NO	Total	% people with disease A
Disease B YES	10	20	30	10/30=33%
Disease B NO	30	190	210	30/210=14%
Total	40	220	250	40/250=16%

Table 5. Incidence of diseases A and B among whole population.

	Disease A YES	Disease A NO	Total	% people with disease A
Disease B YES	30	200	230	30/230=13%
Disease B NO	250	2020	2270	250/2270=11%
Total	280	2220	2500	280/2500=11%

Analysis of data collected in a hospital suggests the existence of a statistically significant association between the occurrence of disease A and B ($\chi^2=7.62$, $p=0.0058$). However, it turns out that in a population this relationship is not found ($\chi^2=0.87$, $p=0.3522$). This discrepancy comes from the fact that patients with more diseases more often go to the hospital (Table 6).

Table 6. The percentage of people admitted to the hospital with regard to the occurrence of diseases A and B.

	Disease A YES	Disease A NO
Disease B YES	10/30=33%	20/200=10%
Disease B NO	30/250=12%	190/2020=9%

Researchers who analyze hospital data should be aware that this kind of material may not be representative for the whole population. Unfortunately, without examine the different sample (randomly selected from the population) it is not possible to assess whether the conclusions observed in the hospital are true for the population.

IV. CONCLUSIONS

Selecting a representative sample from the study population remains a main challenge in medical research. Essentially, sampling has to be juxtaposed with a study context. The samples do not need to be representative in all the aspects. They need to match key characteristics important for an investigated problem. Some of research studies rely on unrepresentative sampling which cannot be indicative and require a special attention when interpreting results.

References

1. Kruskal, W., & Mosteller, F. (1979). Representative sampling, I: non-scientific literature. *International Statistical Review*, 47, 13–24.
2. Kruskal, W., & Mosteller, F. (1979). Representative sampling, II: scientific literature, excluding statistics. *International Statistical Review*, 47, 111–127.
3. Kruskal, W., & Mosteller, F. (1979). Representative sampling, III: the current scientific literature. *International Statistical Review*, 47, 245–265.
4. Sacket, D. L. (1979). Bias in analytic research. *Journal of Chronic Disease*, 32(1-2), 51–63
5. Charig, C. R., Webb, D. R., Payne, S. R., & Wickham, J. E. A. (1986). Comparison of treatment of renal calculi by open surgery, percutaneous nephrolithomy, and extracorporeal shockwave lithotripsy. *British Medical Journal*, 292, 879–882.
6. Berkson, J. (1946). Limitations of the application of fourfold table analysis to hospital data. *Biometrics Bulletin*, 2, 47–53.

MULTIVARIABLE ANALYSIS. THE ESSENTIALS

K. Szafraniec

Institute of Public Health, Faculty of Health Sciences, Jagiellonian University Medical College, Krakow, Poland

Multivariable analysis is used frequently in studies of clinical outcomes. The most important methods include linear regression for continuous outcomes, logistic regression for binary outcomes, Cox regression for time-to-event data, and Poisson regression for frequencies and rates. These statistical models can admit a mixture of categorical and continuous variables that are collected to determine factors affecting an outcome of interest or to investigate relationships among variables.

Variable selection and model performance assessment are the crucial steps in the process of model building. Unfortunately, nowadays a user-friendly statistical software doesn't require mathematical background to perform analysis; therefore, careless application of modelling procedures may result in models that poorly fit within the data or inaccurately predict studied outcome.

The aim of this presentation is to provide practical advice on how to set up and interpret multivariable models. We will discuss step by step the process of performing multivariable analysis: 1/ choosing the correct model, 2/ selecting set of independent variables, 3/ setting up the model (interaction, missing data, convergence), 4/ interpreting model's parameters, 5/ checking underlying assumptions, and 6/ validating the model.

STATISTICAL INFERENCE IN MULTIVARIATE ANALYSIS OF MEDICAL DATA

A. Wolińska-Welcz

Maria Curie-Skłodowska University, Lublin, Poland

In the lecture we will return to the diagram of a typical statistical inference for multivariate data analysis in medical research. We will discuss the subsequent stages of statistical inference - starting from the determination of the aims of the research, description of initial data for analysis, through the formalization of assumptions about the model and data, selection and implementation of appropriate procedures of statistical inference, and ending with the verification of how the results are related to the medical problem, and therefore whether they can give us adequate answers to the posed questions.

Medical data is often difficult to analyze and during the process of the statistical inference it is important to find a common ground for clinicians and statisticians. Moreover, it appears crucial to develop the awareness of how the assumptions about the selected model and about the data may affect the choice of the inference procedures, the subject of interest and the final results. Assumptions underlying many statistical methods are usually not fulfilled in clinical practice.

A careful search for truth, memory of the methodological remarks of our masters, honesty, integrity and professionalism at every stage of statistical analysis, openness to other methods, departure from an automatic use of statistical packages, not yielding under the pressure of data manipulation - all of this will certainly lead to valuable results of research that will serve our health.