

Selecting Relevant Information for Medical Decision Support with Application in Cardiology

Jan Kalina^{1,2}, Libor Seidl^{1,3}, Karel Zvára^{1,3}, Hana Grünfeldová^{1,4}, Dalibor Slovák^{1,2,3}, Jana Zvárová^{1,2,3}

¹ European Center for Medical Informatics, Statistics and Epidemiology

² Institute of Computer Science of the Academy of Sciences of the Czech Republic

³ Charles University in Prague, First Faculty of Medicine

⁴ Municipal Hospital in Čáslav

Abstract

Objectives: The aim of our work was to implement a prototype of a decision support system which has the form of a web-based classification service. Because the data analysis component of decision support systems often happens to be unsuitable for high-dimensional data, special attention must be paid to the sophisticated selection of the most relevant variables before learning the classification rule.

Methods: We implemented a prototype of a diagnostic decision support system called SIR. The system has the ability to select the most relevant variables based on a set of high-dimensional measurements by means of a forward procedure optimizing a decision-making criterion. This allows to learn a reliable classification rule.

Results: The implemented prototype was tested on a sample of patients involved in a cardiology study. We used SIR to perform an information extraction from a cardiological clinical study containing both clinical and gene expression data. The classification performance was evaluated by means of a cross validation study.

Conclusions: The proposed classification system can be useful for clinicians in primary care to support their decision-making tasks with relevant information extracted from any available clinical study. It is especially suitable for analyzing high-dimensional data, e.g. gene expression measurements.

Keywords

Decision support system, web-service, information extraction, high dimension, gene expressions

Correspondence to:

Jan Kalina

EuroMISE Center, Institute of Computer Science AS CR
Address: Pod Vodárenskou věží 2, Prague 8, Czech Republic
E-mail: kalina@euromise.cz

EJBI 2013; 9(1):2–6

received: May 30, 2013

accepted: July 7, 2013

published: August 30, 2013

1 Introduction

Decision support systems (DSS) offer assistance with the decision-making processes in many areas. Their aim is to solve a variety of tasks and to analyze different information components. In medicine, a decision support system represents an inherent e-health technology tool for diagnostic and prognostic purposes capable to help during the therapy [1]. Recently proposed systems in some areas of medicine are required to extract information also from high-dimensional measurements in order to deduce conclusions for the diagnosis, therapy, or prognosis by comparing the risk corresponding to different alternatives. Some systems have been implemented in a web-based form [2, 3].

Decision support systems have established their place in current healthcare. Their potential for improving the quality of provided care and for generating economic benefits by reducing financial costs and saving human resources

have been described in literature [4]. They are known to bring the physician more comfort, a reduction of stress, a higher effectivity, and more time for the patient. The contribution of decision support systems to the patient safety has been summarized in [5]. Another aspect is a benefit for a less experienced physician in a complicated medical case. Also, the systems allow to exploit the level of knowledge reflecting the latest research developments in medicine.

The analytical component within a decision support often happens to be unsuitable for the analysis of high-dimensional data. Decision-making within a clinical decision support system is mostly based only on one of the standard classification procedures of multivariate statistics or machine learning that enables to construct objective classification rules for assigning individual observations to groups. Available classification methods include:

- Linear/quadratic discriminant analysis.
- Neural networks.
- Support vector machines.
- Classification trees.
- K-nearest neighbor.
- Cluster analysis.
- Knowledge-based rules (e.g. [6]).

However, these methods commonly suffer from a so-called curse of dimensionality if the number of variables (e.g. symptoms and signs) exceeds the number of patients [7].

In this work, we have proposed and implemented a prototype of a decision support system in the form of a web-based classification service for diagnostic decision support, which is particularly designed to address the needs for a reliable high-dimensional information extraction. Its organic part is a dimension reduction technique performed in the form of a variable selection. The statistical component of the system uses a variety of sophisticated classification rules, which are reliable also for the analysis of high-dimensional data sets.

We tested the prototype of the system on clinical data in a cardiology study, which includes a whole-genomic study of gene expression measurements. This paper presents the principles and advantages of the proposed system and summarizes results of the testing service on a cardiology study.

2 Methods

We propose a system called SIR (System for selecting relevant Information for decision support), which is an easy-to-use web-based generic service devoted to data collection and decision support with a sophisticated information extraction component. It is proposed for being used mainly for general practitioners in the primary care, but it is able to handle data from any area of medicine. The decision making of the SIR requires data from a (sufficiently large) clinical study in order to construct the optimal classification rule for the decision making problem.

Data collected within a clinical study represent the training database of the SIR. The SIR can import the whole data set from the clinical study automatically together with a data model. The system cleans the data e.g. by checking if the values of the imported quantitative variables do not exceed given bounds required by the data model.

The next step in the analysis of the data from the clinical study is dimension reduction. Statistics distinguishes between variable selection and feature extraction [8], where the latter searches for a smaller set of linear

combinations of all variables. Here, we perform the variable selection, which reduces the set of all measured symptoms or laboratory measurements to a smaller set of relevant symptoms. This step, which is necessary especially for high-dimensional data obtained in genetic studies [7], is performed by a forward procedure optimizing a decision-making criterion.

Categorized data are considered and the contribution of a given variable (say X) to explaining the uncertainty in the response Y (i.e. in the separation among the groups) is quantified by means of the conditional Shannon information, which is denoted as $d(Y|X)$. The first variable (say X_1) fulfils

$$d(Y|X_1) = \max d(Y|X) \quad (1)$$

over all variables X . Thus, X_1 is the most relevant variable for explaining the classification. Further on, the method successively selects the most relevant variables with the maximal value of the statistical dependence. In other words, other variables are iteratively added to the set with the best improvement of the conditional relevance. If variables X_1, \dots, X_s have been selected as the most relevant, the next variable (say X_{s+1}) is selected as the variable fulfilling the requirement

$$d(Y|X_1, \dots, X_s, X_{s+1}) = \max d(Y|X_1, \dots, X_s, X), \quad (2)$$

where all variables X not present in the set $\{X_1, \dots, X_s\}$ are considered. Finally, we consider only such variables for the consequent analysis which contribute to explaining more than 90 % of the inter-class variability. The system allows quantifying the influence of an additional examination (variable) on the diagnostic decision. Additionally, the dimension reduction procedure may take into account the cost of obtaining each clinical or laboratory measurement by using the information theoretical approach of [9].

The process of learning of the classification rule within the SIR has the ability to decide automatically for one of several different methods. A criterion of optimality is adaptively chosen to minimize the risk of a wrong classification result due to special properties of the data and the sample sizes. The implemented methods include the linear discriminant analysis (LDA), which is a multivariate statistical method for separating groups by means of a linear function [10]. The same covariance structure is assumed in each group. Another approach implemented in the SIR is the empirical Bayes inference mechanism, which minimizes the aposterior Bayes risk across all groups of samples. Let us say that the task is the classification to K groups. It uses a discretization of data and we denote the levels of a discrete variable X by X_1, \dots, X_r . The method assumes a conditional independence of the levels of X for each group $k = 1, \dots, K$.

The construction of the classification rule in the system SIR may additionally allow a combination of data with medical knowledge. To be specific, a clinician may interfere manually the system in order to incorporate ad-

Universal Entry Form for Decision Support System.
Created with support of project #1M06014 of Ministry of Education, Youth and Sports in Czech Republic

CVA vs AMI

Date: December 29, 2011
Published by: Libor Seidl
Description: Prediction of AMI vs CVA based on 160 patients study. Dimension reduction has been from 24 to 13 dimensions with 0.91103 degree of reliability.
Technical Details: model=CBI-AMIsCVA, data=160 patients

[\[Back to list of models \]](#)

Admin. Gender:

Education:

Smoking:

PCI in anamnesis:

Diabetes Mellitus in anamnesis:

HN in anamnesis:

Beta blocker (chronic medication):

Statins (chronic medication):

ASA (chronic medication):

Aggrenox (chronic medication):

Warfarin (chronic medication):

Tidopidin (chronic medication):

Other chronic medication:

Your preliminary decision:

Advice provided by the Service:

Your final decision:

Reasons for Final Decision:

Description of Reasons:

Why have you changed your decision:

Reasons:




Figure 1: Illustration of the prototype of the system SIR.

ditional expert knowledge based on education, experience, or intuition and can e.g. eliminate a certain diagnosis for a specific combination of symptoms and signs, if their joint occurrence is known to have a zero probability.

From the implementation point of view, the prototype can be understood as four subsystems: an Administration System, a DSS SOAP Frontend, a DSS Web Frontend, and a DSS Backend. The Administration System is devoted to a model creation, data gathering and manipulation, dimension reduction, and DSS publication. The SOAP Frontend provides the classification of a patient on a request by an end-user system (including on-fly generation of WSDL) by working with an internal XML description of the decision support system, which is generated during the process of publication in the Administration System. The DSS Web Frontend provides a HTML Form based user-friendly interface to the SOAP Frontend. All these three subsystems are programmed using PHP5. The main two tasks of the DSS Backend are to learn a new dataset during the process of DSS publication and to classify a new patient during a SOAP Frontend request. The computation in the DSS Backend consists of several R scripts, which are called on demand from PHP5.

A clinician as the user of the decision support service is not required to understand the background of the methods. His/her aim is to determine the diagnosis of a new patient (not included in the clinical study), who can be examined on a distant place. All variables selected by the variable selection procedure are required to enter the decision support system, which can be performed through the automatically generated interface from an electronic health record (EHR) or health information system (HIS), although a manual input of data is also possible, as illustrated in Figure 1. The clinician must specify the prior diagnosis before entering the data to the SIR, because he/she is the only one to carry the legal responsibility for the clinical decision. Now the SIR can be used through the web service to obtain a diagnosis support. Then, the clinician is asked to manually select his/her final decision and only if it is not in accordance with the SIR, the clinician writes a short text justifying the decision.

3 Results

We implemented the prototype of the system SIR and evaluated its performance on a real clinical study of cardiovascular diseases, which incorporated the measurement

Table 1: Sets of personal and clinical variables in the cardiological clinical study.

Set A	Sex, height, weight, education, smoking, diabetes, systolic blood pressure, cholesterol.
Set B	Height, weight, education, systolic blood pressure, cholesterol.

of gene expressions across the whole genome. The study was performed in the years 2006-2011. The aim of the study was to identify a small set of genes and clinical variables associated with excess genetic risk for the incidence of a cardiovascular disease. Clinical and gene expression measurements are measured on a set of 59 patients having an acute myocardial infarction (AMI), 45 patients having a cerebrovascular stroke (CVS), and 77 control persons (CP) chosen as individuals without a manifested cardiovascular disease with the same risk factors as the patients. These 181 individuals serve as a training database for constructing an efficient classification rule for assigning a new individual to one of the groups (AMI, CVS, CP).

A set of 4 personal and 4 clinical variables recorded for each patient is shown in Table 1. The gene expressions of all genes (>39 000 gene transcripts) are measured for each patient using Illumina BeadChip microarrays. We will describe the training of the SIR to classify the samples to one of three groups (AMI, CVS, CP). A routine statistical analysis of a subset of these data was performed in [11]. There, gene expressions AMI patients are compared to those of CPs and values of sensitivity and specificity are presented.

We used the dimension reduction method to select a set of 10 most relevant genes from the high-dimensional set of measurements. We categorized each continuous variable into 4 categories (if possible) and assume an equal importance of each of the variable. The set A was reduced to 5 most relevant variables (set B) shown in Table 1. Set B contains significant instruments of the life style of a particular patient and explains 97.9 % of the intra-class variability of the set A.

Further, the SIR used the linear discriminant analysis on the original data (without categorization) to learn

a classification rule into one of three groups (AMI, CVS, CP). Table 2 presents results of an independent validation study performed by leave-one-out cross-validation using various sets of measurements. Thus, the set of all genes has the ability to determine the diagnosis correctly for 85 % of patients.

A reduced set of variables can retain a relatively high classification performance, which is a consequence of redundancy of the remaining variables or their multicollinearity (cf. [12]). Moreover, we have verified the results also with other statistical validation criteria, e.g. leave-10-out cross-validation or bootstrap.

4 Discussion

We implemented an easy-to-use system called SIR (System for selecting relevant Information for decision support), which has the ability to select the variables relevant for a reliable information extraction from high-dimensional measurements. The system allows a diagnostic decision support by means of a web technology and can be characterized as a practical tool for evidence-based medicine [13]. We believe that a reliable decision support system should be always equipped with a statistical component allowing to extract information from very complex measurements. Without the help of such specialized tool, a clinician would never be able to extract the information from such high-dimensional measurements e.g. in the molecular genetic context.

The SIR simulates a decision making process as performed by a clinician. The system can be used as a purely assistive technology to the clinician, who carries the responsibility for the diagnosis decision making in combina-

Table 2: Evaluation of the system SIR in the task of a diagnostic decision support based on the data from the cardiology clinical study. Percentage of correctly classified samples to one of three groups (AMI, CVS, controls) in the leave-one-out cross-validation procedure using the linear discriminant analysis.

Variables used in the classification rule	Classification performance
Set A (8 personal and clinical variables)	0.56
Set B (5 personal and clinical variables)	0.56
All genes	0.85
All genes + set A	0.85
All genes + set B	0.85
10 genes	0.65
10 genes + set A	0.72
10 genes + set B	0.72

tion of a scientific and empirical knowledge to infer the interpretation in all steps of healthcare provision. The clinician determines a prior diagnosis and has the possibility to decide for a different aposterior diagnosis based on the recommendation of the system. In such case, however, the SIR collects a feedback from the clinician.

The prototype version of our system has not been released for a public usage on the internet yet. We plan an intensive validation stage allowing exposing the system repeatedly to real situations starting with formulating requirements, implementation of modifications, multi-level testing under artificial conditions, and testing. Only this will allow tuning all parameters of the system, which must be maintained, supervised and monitored for a long-term period in cooperation with clinicians before introducing a fully public version to real applications following necessary rules of data safety. This will also require a secure access in concordance to current legislation (public-key infra-structure, service versioning, etc.).

In general, the system can be used to analyze different data sets in various areas of medicine. The knowledge from recent medical research can reach clinicians quickly by means of the system, which can assist them as a supporting tool within the decision making process. At the same time, the system SIR is designed to be convenient for a data collection e.g. within a hospital, while the classification rule can be learned continuously during its operation.

So far, we have applied the system SIR to real cardiology data. The system determined a set of 10 crucial genes among more than 39 000 gene transcripts. The selected genes are believed to be associated with the risk of a manifestation of AMI or CVS for a particular patient in the population in the Czech Republic. The paired design of the study allowed eliminating the influence of known risk factors (e.g. systolic blood pressure) on the discrimination. Thus, we revealed the added value of including the gene expression data to the study. A clinician with access to the web classification service may obtain a prediction of the risk of a more severe prognosis or a relapse for new patients. The clinician has the information about the classification reliability of the system. We are preparing other studies for validating the ability of the SIR to select the relevant information from high-dimensional measurements for a reliable decision support.

Acknowledgements

The research was supported by the project 1M06014 MŠMT ČR. We are thankful to Martin Horáček for the help with the implementation of the classification analysis. The system SIR was first presented at the EFMI Special Topic Conference 17-19 April 2013, Prague, Czech Repub-

lic and a short paper was published in the Proceedings of the conference [14].

References

- [1] D.J. Power, *Decision support systems: Concepts and resources for managers*, Quorum Books, Westport, 2002.
- [2] M.J. Romano, R.S. Stafford, *Electronic health records and clinical decision support systems: Impact on national ambulatory care quality*, *Archives of Internal Medicine* 171 (2011), 897-903.
- [3] F. Sicurello, M. Gündel, A. Donzelli, *Data analysis web service using statistical packages*. *International Journal of Advanced Statistics and ITC for Economics and Life Sciences* 1 (2009), 3-7.
- [4] K. Kawamoto, C.A. Houlihan, E.A. Balas, D.F. Lobach, *Improving clinical practice using clinical decision support systems: A systematic review of trials to identify features critical to success*, *BMJ* 330 (2005), 330:765.
- [5] J. Kalina, J. Zvárová, *Decision support systems in the process of improving patient safety*. In A. Mourtoglou, A. Kastania (Eds.): *E-Health Technologies and Improving Patient Safety: Exploring Organizational Factors*. IGI Global, Hershey, Pennsylvania, 2013, 71-83.
- [6] P. Berka, J. Rauch, M. Tomečková, *Data mining in the atherosclerosis risk factor data*. In Berka P., Rauch J., Zighed D.A. (Eds.): *Data Mining and Medical Knowledge Management: Cases and Applications*, IGI Global, Hershey, 2009.
- [7] J. Kalina, *Classification analysis methods for high-dimensional genetic data*. *Biocybernetics and Biomedical Engineering* (2013). Accepted.
- [8] W.L. Martinez, A.R. Martinez, J.L. Solka, *Exploratory data analysis with MATLAB*. 2nd edn. Chapman Hall/CRC, London, 2011.
- [9] J. Zvárová, M. Studený, *Information theoretical approach to constitution and reduction of medical data*, *International Journal of Medical Informatics* 45 (1997), 65-74.
- [10] A.C. Rencher, *Multivariate statistical inference and applications*, Wiley, New York, 1998.
- [11] Z. Valenta, I. Mazura, M. Kolář, H. Grünfeldová, P. Feglarová, J. Peleška, M. Tomečková, J. Kalina, D. Slovák, J. Zvárová, *Determinants of excess genetic risk of acute myocardial infarction-a matched case-control study*, *European Journal for Biomedical Informatics* 8 (2012), 34-43.
- [12] C. Ding, H. Peng, *Minimum redundancy feature selection from microarray gene expression data*. *Journal of Bioinformatics and Computational Biology* 3 (2005), 523-528.
- [13] H. Chen, S.S. Fuller, C. Friedman, W. Hersh, *Medical informatics, Knowledge management and data mining in biomedicine*, Springer, New York, 2005.
- [14] J. Kalina, L. Seidl, K. Zvára, H. Grünfeldová, D. Slovák, J. Zvárová: *System for selecting relevant information for decision support*. In: B. Blobel, A. Hasman, J. Zvárová (Eds.): *Data and Knowledge for Medical Decision Support, Studies in Health Technology and Informatics* 186, IOS Press, Amsterdam, 2013, 83-87.