Uncover the relations between the discretized continuous-valued features with multiple correspondence analysis in medical domain

Navasiolava NA, Tom IE, Krasko OV

United Institute of Informatics Problems of The National Academy of Sciences of Belarus, Minsk, Belarus

Abstract

Purpose: The aim of this study was to explore a way in which continuous-valued and categorical features can be treated simultaneously in multiple correspondence analysis for the solution of a particular medical problem.

Material and methods: The exploratory analysis of medical data from Belarussian Research Center for Pediatric Oncology and Hematology (BRCPOH) was performed using the developed combined method, consisted of prior continuous-valued data discretization with subsequent correspondence analysis of frequency tables.

Results: The developed method allows to perform supervised discretization of continues features and significantly decreases the number of its values, to treat continuous-valued and categorical features simultaneously in multiple correspondence analysis, to analyze the quality of discretization of continuous-valued features.

Conclusions: The proposed combined method of exploratory data analysis allows discovering interactions between the different features of any initial type.

Key words: exploratory data analysis, continuous-valued feature discretization, multiple correspondence analysis.

Introduction

The results of data analysis in medical domain frequently are required to be interpretable to the medical expert, therefore in the most cases it's very important to find the critical values of the features, that can explain the data analysis model output. It is more convenient for the medical expert to analyze the features, represented with a few values, categorical features. Different methods of the discretization of continuous-valued features are introduced in the literature [1,2] as a technique allowing to improve the accuracy and to simplify the structure of many data analysis models (decision tree, neural networks). In this paper the combined method of exploratory data analysis is proposed, consisted of prior continuous-valued data discretization with subsequent correspondence analysis of frequency tables [3]. This method not only allows to discover interesting interactions between the different features of any initial type, but also to analyze the quality of discretization of continuous-valued features.

Material and methods

The data of the children with acute lymphoblastic leukemia (ALL), collected in Belarussian Research Center for Pediatric Oncology and Hematology were analyzed with proposed combined method, consisted of prior continuous-valued data discretization with subsequent correspondence analysis of frequency tables. This method allows analyzing the interaction between the features of the different initial type.

The correspondence analysis of discretized medical data was performed using STATISTICA 6.0 software, StatSoft Inc.

Results

General

The main result of the study is the proposed combined method of exploratory data analysis, consisted of continuous-valued features discretization and the analysis of discretized and categorical features simultaneously with multiple correspondence analysis in order to reveal the interactions between them.

ADDRESS OF CORRESPONDENCE:
Navasiolava Natallia
Laboratory of Bioinformatics
United Institute of Informatics Problems
of The National Academy of Sciences of Belarus
6 Surganov Str., 220012, Minsk, Belarus
e-mail: novosel@newman.bas-net.by

Continuous-valued features discretization

Several methods of data discretization have been described in the literature as the preprocessing step of data mining process [1,2]. They can be divided into two classes:

- unsupervised methods, which deal with data without classification information available;
- supervised methods, which take the classification information into account.

In the proposed combined data analysis method we use the Chi^2 algorithm [4], extended with the prior feature ranking according to entropy measure.

In our interpretation continuous-valued features discretization process consists of two stages:

- ranking of initial features according to their information gain for classification;
- Chi^2 statistical method of discretization.

All the initial features of patients with ALL are ranked according to their information gain for classification, which is calculated on the basis of entropy value or expected information for classification:

$$I(U)=-\sum_{c=1}^{N_c}\frac{n_c}{n}\cdot\log_2\frac{n_c}{n}$$

where n_c – the number of objects (patients), belonging to class c, and n – the total object number in the data set U.

Information gain, contributed by feature X_i is following:

$$G_i=I(U)-E_i,$$

$$E_i=\sum_{k=1}^{N_i}\frac{n_{ik}}{n}\cdot I(U_{ik}),\ I(U_{ik})=-\sum_{c=1}^{N_c}\frac{n_{ikc}}{n_{ik}}\cdot\log_2\frac{n_{ikc}}{n_{ik}},$$

U_{ik} is the subset of U with feature X_i for all objects equals k, $I(U_{ik})$ is the expected information for subset U_{ik}, n_{ik} – number of objects from U_{ik}, n_{ikc} – number of objects from U_{ik} belonging to class c.

After that, the features in the ranking order are propagated to discretization process with Chi^2 method. The Chi^2 method is based on χ^2 statistics and consists of two phases. In the first phase each feature is sorted according to its values. Then the following is performed:

- calculating χ^2 value for every pair of adjacent intervals;
- merging the pair of adjacent intervals with the lowest χ^2 value;
- merging continues until all pairs of intervals have χ^2 values exceeding the parameter, determined by defined significance level;
- process is repeated with decreased significance level until an inconsistency rate exceeds the predefined δ level.

The χ^2 value is calculated according to the following formula:

$$\chi^2=-\sum_{i=1}^{2}\sum_{j=1}^{k}\frac{(A_{ij}-E_{ij})^2}{E_{ij}},$$

Table 1. **Results of discretization of features from the medical data base**

Feature name	No values before discretization	No values after discretization
WBC1	60	5
WBC8	44	5
AGE	11	1
BL_ABS_8	24	5
HEP_8	12	3
CD10	32	1
CD20	37	3
CD45_14	46	2

k – number of classes, A_{ij} – no cases in the ith interval, jth class, R_i – no cases in the ith interval, C_j – no cases in the jth class, N total no cases, E_{ij} – expected frequency of

$$A_{ij}=\frac{R_i\cdot C_j}{N}$$

Along with discretization the selection of features is accomplished by removing those, having only one discrete value. Both discretization and feature selection retain the discriminating power of processed data (*Tab. 1*).

Correspondence analysis of frequency tables

Correspondence analysis is an exploratory technique for the analysis of frequency tables. It allows revealing the dependencies of medical features in order to fit the appropriate model to the problem in question. Another application of the correspondence analysis techniques is that it allows performing the equivalence of a multiple regression for categorical features, by adding supplementary data to the input frequency table. The display of the discretized continues-valued features along with categorical ones in reduced dimensional space provides an indication of the nature of the relationships between them.

The possibility to visualize positional relationship of different features, allows to select more significant ones for the task in question for further analysis with different methods, as for example neural networks or decision trees. The following combination of features was analyzed: three features with discretized values, viz Hepatomegaly, day 8 (HEP8, cm); Blasts, abs., day 8 (bl_abs8, 10^9/L); CD20 (%), CD45 (%) and one categorical feature Response on therapy on day 15 (RESP15).

The obtained coordinates of different features' categories in the selected two-dimensional space that preserves most of information about difference of frequency profiles of individual points are presented in *Tab. 2*.

The two-dimensional layout of different features' categories is presented on *Fig. 1*.

Category 0 of RESP15 corresponds to remission at the day 15, category 1 — absence of remission at the day 15. From the *Fig. 1* it can be concluded, that the relative row frequencies of categories of the feature RESP15 according to categories of other features are strongly different. There are relatively more patients with category 0 of RESP15 among patients with category 1 of CD45_14 and 2 of CD20, and with category 1 of RESP15

Figure 1. **Representation of features' categories in two-dimensional space**

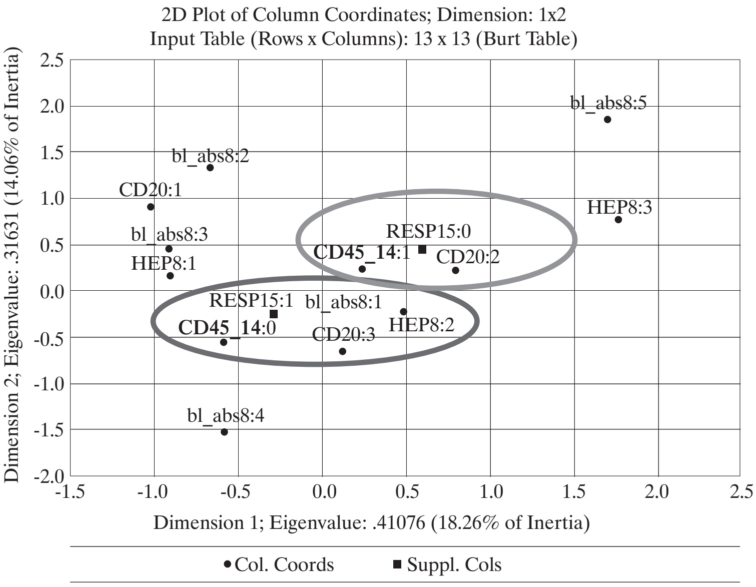

Table 2. **Coordinates and quality of representation of features' categories in two-dimensional space**

Feature	Row	Coordin1	Coordin2	Quality
HEP8:1	1	-0.88712	0.15889	0.584805
HEP8:2	2	0.48515	-0.23237	0.603148
HEP8:3	3	1.76494	0.75070	0.675892
bl_abs8:1	4	0.15641	-0.25785	0.661363
bl_abs8:2	5	-0.66448	1.32243	0.564276
bl_abs8:3	6	-0.90445	0.45886	0.566798
bl_abs8:4	7	-0.56815	-1.50782	0.730053
bl_abs8:5	8	1.70246	1.88259	0.883188
CD20:1	9	-1.01768	0.93388	0.696503
CD20:2	10	0.79206	0.20070	0.829504
CD20:3	11	0.14589	-0.63158	0.782826
CD45_14:0	12	-0.58271	-0.55839	0.666946
CD45_14:1	13	0.23882	0.22885	0.666946
RESP15:0		0.61981	0.47352	0.647287
RESP15:1		-0.29922	-0.22860	0.447287

among patients with category 0 of CD45_14 and 3 of CD20. Also the patients with category 1 of RESP15 strongly characterized with category 1 of bl_abs8. Therefore the interdependency of RESP15 and other more tightly related features can be further investigated with different statistical and data mining methods. Moreover the categories of the discretized feature that are close to each other in the two-dimensional space can be treated as one category without the loss of information, thus allowing to correct the results of discretization process.

Conclusions

The combined method of exploratory data analysis has been proposed, which allows solving the different tasks:

– to perform supervised discretization of continues features and significantly decrease the number of its values, keeping the discriminating power of a database;

– to treat continuous-valued and categorical features simultaneously in multiple correspondence analysis;

– to analyze the quality of discretization of continuous-valued features.

Acknowledgement

The financial support of European Community under the contract to The International Science and Technology Center (Project #B-522) is greatly acknowledged.

References

1. Dougherty J, Kohavi R, Sahami M. Supervised and unsupervised discretization of continuous features. [In:] Proc Machine Learning: Twelfth International conference, 1995, Morgan Kaufmann Publishers.

2. John G, Kohavi R, Pfleger K. Irrelevant features and the subset selection problem. [In:] Proc ML-94, 1994; 121-9.

3. Greenacre MJ, Hastie T. (1987) The geometric interpretation of correspondence analysis. Journal of the American Statistical Association, 82; 437-47.

4. Liu H, Setiono R. Feature selection via discretization of numeric attributes. IEEE Trans Knowledge and Data Eng, 1997; 9(4): 642-5.