Third International Conference
Modelling and Development of Intelligent Systems
October 10 - 12, 2013
"Lucian Blaga" University
Sibiu - Romania

Computational intelligence in medical data sets

Ionela MANIU, George MANIU, Daniel HUNYADI

Abstract

In recent years, collection of data, regardless of the field, has become a normal phenomenon. In the activity and evolution of an organization is imperative to take into account the data collected in order to achieve decision process. As the volume and complexity of data are in constant growth is necessary to use intelligent methods and fundamental tools for storing, processing, filtering and obtaining information from these data.

1. Introduction

Extraction of information/learning from data/knowledge discovery from data is the primary goal of intelligent computational methods [4]. Learning from data can be done supervised or unsupervised. The objective of supervised learning is to predict the amount of output data based on the input data, and in unsupervised learning the objective is to describe associations and characteristics/structures of the input data.

2 Strategies to achieve the knowledge discovery

2.1 Descriptive and exploratory phase

First step in knowledge discovery consists in data exploration [1][5]. This first phase is descriptive and exploratory and analyse elements such as distribution, identification of atypical values, data transformations required by the distribution form or data standardization, means, cluster variance, correlation, classification, etc. This approach has as result an achieving of data descriptions that establish relationships among variables providing a first general idea of the data.

In this phase can be considered two main objectives. The first objective is to explore one-dimensional and multidimensional or reduce the data dimension and the methods and the instruments used are: factor analysis, principal component analysis, analysis of simple correspondences. These methods consist in analysing a weighted point cloud into a space with a special metric, the cloud forms characterizing the nature and intensity of the relationships between variables and revealing information contained in data structures. The second objective is the classification or segmentation and

it can be achieved by the following methods: hierarchical ascending classification (cluster progressive elements), the k-means (iterative aggregation elements around mobile canters) or mixed methods[3]. In this case we want the division and distribution into classes or categories by optimizing a criterion, each class having property that is as homogeneous in its entirety and report more distinctive compared to other classes.

The k-means clustering algorithm [6] is a straightforward and effective algorithm for finding clusters in data. The algorithm proceeds as follows.

- Step 1: Ask the user how many clusters k the data set should be partitioned into.
- Step 2: Randomly assign *k* records to be the initial cluster center locations.
- Step 3: For each record, find the nearest cluster center. Thus, in a sense, each cluster center "owns" a subset of the records, thereby representing a partition of the data set. We therefore have *k* clusters, C_1, C_2, . . . , C_k.
- Step 4: For each of the *k* clusters, find the cluster centroid, and update the location of each cluster center to the new value of the centroid.
- Step 5: Repeat steps 3 to 5 until convergence or termination.

The "nearest" criterion in step 3 is centroid distance (distance between the centroids of each cluster), although other criteria may be applied as well.

Technically speaking, the algorithm steps are:

- Assume the existence of N vectors $x^l = (x_1, x_2, .., x_n)$;
- Identify a representative set of *k* vectors c_j, where $j = 1, 2, ..., k$;
- Partition data in *k* disjoint subsets *Sj* containing N_j points, so to minimize the clustering function given by:

$$J = \sum_{j=1}^{k} \sum_{l \in S_j} \| x^l - c_j \|^2 \qquad (1)$$

where c_j is the average centroid data from the set *Sj*, given by:

$$c_j = \frac{\sum_{l \in S_j} x^l}{N_j} \qquad (2)$$

One attractive classification method involves the construction of a decision tree, a collection of decision nodes, connected by branches, extending downward from the root node until terminating in leaf nodes. Beginning at the root node, which by convention is placed at the top of the decision tree diagram, attributes are tested at the decision nodes, with each possible outcome resulting in a branch. Each branch then leads either to another decision node or to a terminating leaf node.

Classification and decision trees are used to forecast membership of objects / instances in different categories, based on their measures in relation to one or more predictor variables. Classification tree analysis is a major data mining techniques. The flexibility of this technique makes it especially attractive, particularly because the benefit of present and suggestive views (tree which summarizes the classification obtained).

Conceptually, the construction algorithm and decision tree classification is as follows:

- Let D_t training set which is at node *t*;
- If D_t is the empty set, then *t* is a leaf labeled default C_Φ;
- If D_t contains instances belonging to the same class C_t, where *t* is a leaf labeled C_t;
- If D_t contains several instances belonging to one class, then use an attribute node test to divide D_t in smaller subsets. The procedure is applied recursively for each node.

The strategy underlying the optimal partitioning of a node type is a greedy method, a recursive construction "top down" *divide et impera* type.

In principle, the methodology for classification and decision tree induction consists of two phases:

- Construction of the original tree, using the available training set until each leaf is "pure" or almost "pure".
- "Forming" tree as "increased" to improve the accuracy obtained by the test set.

Briefly, the algorithm behind the building and decision tree classification is as follows:

```
Build tree (training data T)
{
        Partition (T)
}
Partition (S data)
{
        if (all points of S are in the same class) then
                returns
        for each attribute A do
                evaluates the split on attribute A;
        using the best split found for partitioning S in S1 and S2
        Partition(S1)
        Partition(S2)
}
```

2.2 Inferential and confirmatory phase

The first phase is preceded by a second inferential step. This step uses the results obtained in the first stage as assumptions in statistical tests or probabilistic models that explain that to predict a certain variable with one or more explanatory variables.

The main objective at this phase is modelling and deduction of a predictive model. To achieve this you can use methods such as linear regression, ANOVA, ANCOVA, neural networks, classification and regression trees, SVM, models based on statistical observations series: spectral analysis, seasonality analysis.

3 Case study

In the database used in this paper are analysed indicators such as blood sugar, cholesterol, systolic and diastolic blood pressure, indicators that we see that are closely correlated with body mass index (IMC) based indicator which are established cases of overweight and obesity. The correlation between IMC and age was also studied [2]. A first set of data from the database is represented by the values of these indicators observed in the case of a factory employee (N=433), the second set are the values seen from a hospital employee (N=300) and a third for employees in the administrative field (N=70).

During the first phase of knowledge discovery, an descriptive and exploratory phase, the objective is to analyze the shape of distribution indicators above, identify outliers or data entry errors, missing values identification, variables transformation.

Cases studied distribution shapes and clinical characteristics are represented in figure 1 and table 1. As can been seen from the graph and histogram respectively after applying the test “Kolmogorov-Smirnov”, in the case of 6 variables analyzed distribution differs significantly from a normal distribution ($p < 0.05$). In this situation, if you want to compare values of variables is indicated using operations transformation of values or using nonparametric tests. In the analysis were compared and values of these variables on lots has been rechecked their distribution form for each lot. After the checks were encountered situations of normality and abnormality of the distribution.

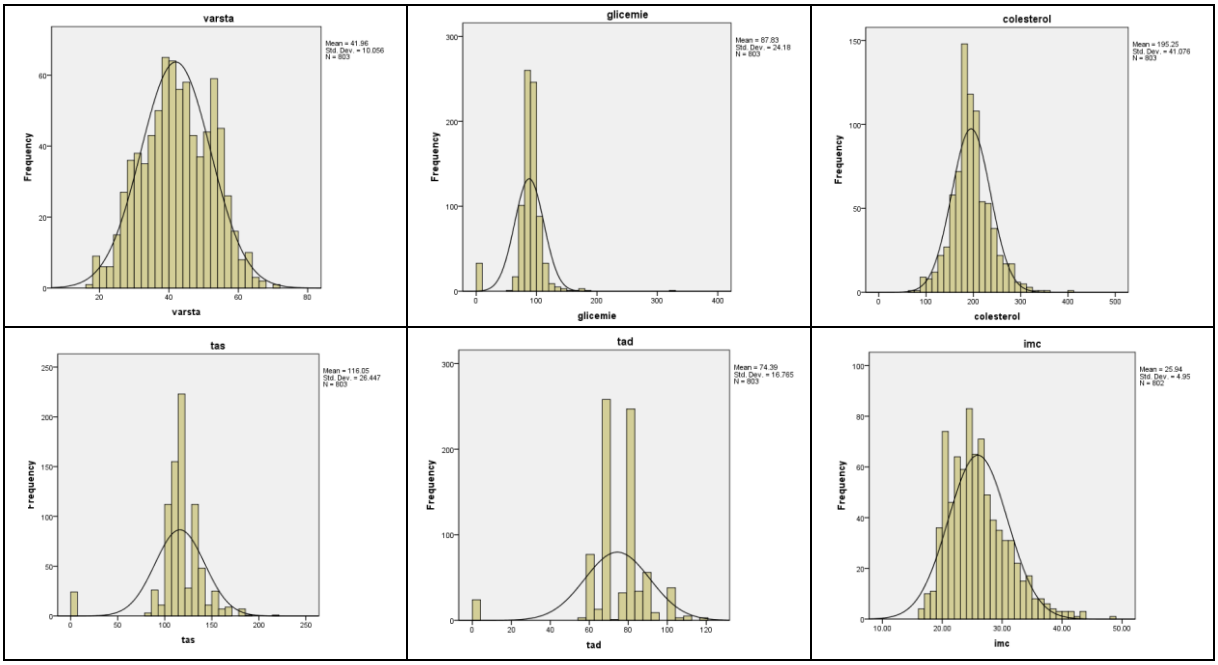

Figure 1 - Cases studied distribution shapes

Characteristics	Mean ±SD	Range	Skewness / Kurtosis
Age (varsta)	41.96 ± 10.05	17 -70	-0.04 / -0.58
Glucose (glicemie)	87.83 ± 16.25	58 - 321	4.82 / 5.25
Cholesterol (colesterol)	195.25 ± 41.07	68 - 409	0.49 / 1.58
TAS	119.62 ± 17.11	80 - 220	1.07 / 2.69
TAD	76.68 ± 10.66	55 - 120	0.75 / 1.01
IMC	25.93 ± 4.95	16.32 – 48.27	0.87 / 0.94

Table 1. Cases studied clinical characteristics

Obesity groups are defined by: normal IMC<25, overweight (suprapondere) 25<=IMC<30, obese (obezitate) IMC >=30. Data are expressed as Mean ± SD. Comparation of normal vs. overweight or normal vs. obese is done by either the parametric "One way ANOVA" with "Post Hoc Tests (Bonferroni)" or the nonparametric Kruskal-Wallis test. The significance is indicated by (♦) p<0.05, (♦♦) p<0.01, (♦♦♦) p<0.001 for normal vs. overweight and by (*) p<0.05, (**) p<0.01, (***) p<0.001 for normal vs. obese.

Clinical characteristics of the different adiposity groups defined by IMC are represented in the table 2.

Characteristics	Normal N=387	Overweight N=259	Obese N=156
Age (varsta)	37.96 ± 9.42	44.85 ± 9.06 ♦♦♦	47.10 ± 9.20 ***
Glucose (glicemie)	88.55 ± 11.78	93.06 ± 19.95 ♦♦	96.80 ± 11.66 ***
Cholesterol (colesterol)	186.20 ± 35.84	200.99 ± 41.74 ♦♦♦	208.27 ± 46.89 ***
TAS	113.11 ± 13.47	123.75 ± 17.09 ♦♦♦	129.19 ± 18.77 ***
TAD	72.21 ± 8.31	79.09 ± 10.02 ♦♦♦	83.96 ± 11.73 ***

Table 2. Clinical characteristics of the different adiposity groups

As can be seen from the table above, glucose, cholesterol, SBP, DBP were significantly higher in overweight or obese persons compared to persons with normal BMI. This trend was the same in case

of group of people who work in the factory and the hospital, while in case of the people from administrative group is found glucose values, cholesterol, SBP, DBP slightly higher in overweight to obese.

Correlation between IMC and other clinical characteristics is then analysed and the results are represented by scatterplots. Data are expressed as "Pearson correlation coefficient" and the significance is the same as upstairs. Results are presented in table 3 and figure 2.

Pearson correlation	Normal N=387	Overweight N=259	Obese N=156
Age (varsta) vs. Glucose (glicemie)	0.169 **	0.178**	0.152
Age (varsta) vs. Cholesterol (colesterol)	0.124 *	0.090	0.069
Age (varsta) vs. TAS	0.332 **	0.235**	0.233**
Age (varsta) vs. TAD	0.266 **	0.216**	0.155
Age (varsta) vs. IMC	0.204 **	-0.035	-0.062
Glucose (glicemie) vs. Cholesterol (colesterol)	0.041	0.130*	0.128
Glucose (glicemie) vs. TAS	0.090	0.243**	0.190*
Glucose (glicemie) vs. TAD	0.094	0.160*	0.160
Glucose (glicemie) vs. IMC	0.070	0.023	-0.081
Cholesterol (colesterol)vs. TAS	0.077	0.137*	0.247**
Cholesterol (colesterol)vs. TAD	0.136 **	0.188**	0.166**
Cholesterol (colesterol)vs. IMC	0.095	0.061	-0.036
TAS vs. TAD	0.731 **	0.753**	0.796**
TAS vs. IMC	0.187**	0.230**	0.155*
TAD vs. IMC	0.192**	0.185**	0.247**

Table 3. Correlation coefficient between IMC and other clinical characteristics

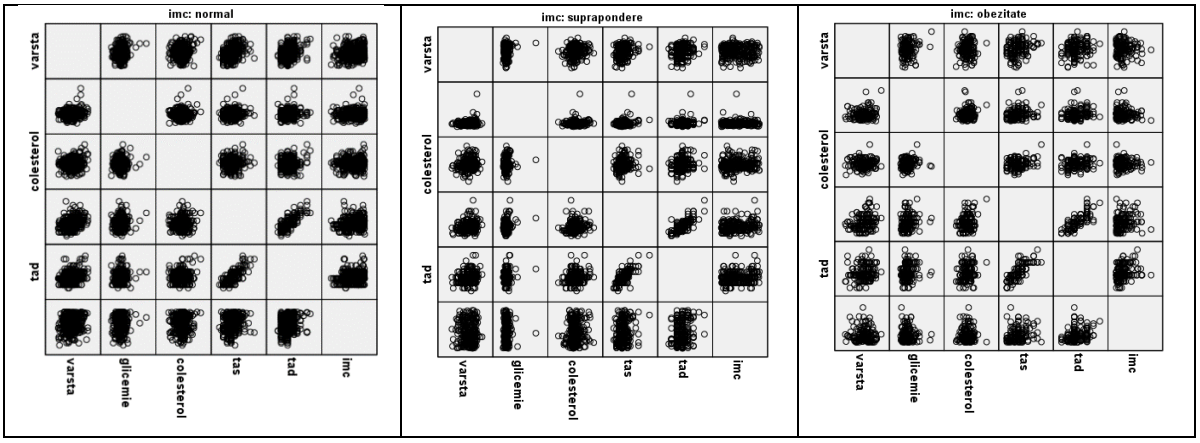

Figure 2. Correlation scatterplot

If normal BMI, it is significantly positively associated with age while if overweight and obesity have a negative association, but insignificant.

There was no significant correlation between BMI and blood sugar, cholesterol, except that in both cases the association was positive at people with normal BMI and overweight and the obese identified a negative association. A positive association was observed between BMI and blood.

Dendrogram obtained by cluster analysis, using the *Between-groups linkage* and the metrics *Squared Euclidian distance* shows that the variables are clustered in two groups, in fist group age (varsta) is clustered with IMC, and in the second one glucose (glicemie) is clustered with TAS, TAD, and cholesterol is not a part of any of the clusters (figure 3).

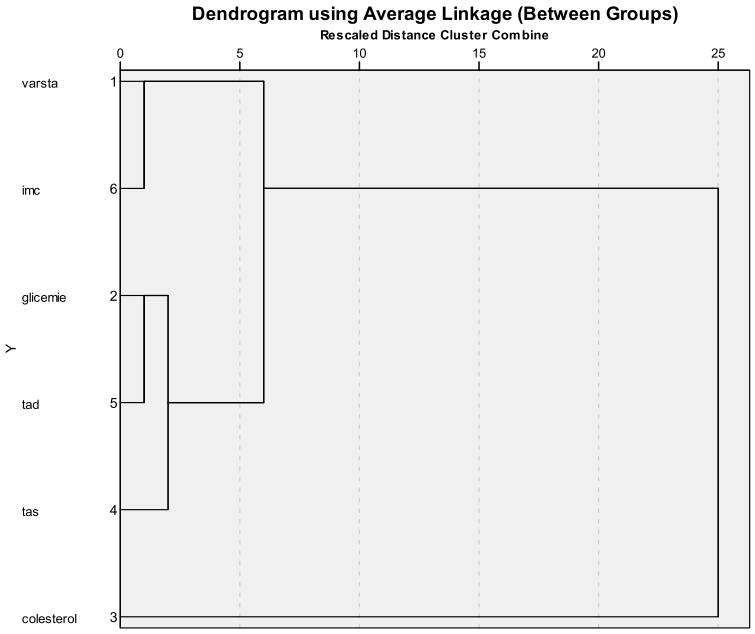

Figure 3. Dendrogram obtained by cluster analysis

4 Conclusions

Exploratory and explanatory methods are basic tools for data exploration. There is no best method, but experience in the choice of the method has an important role that is adapted to the types of database variables and having very good experience clarified the objectives of the study.

5 References

[1] Vapnik, V., The nature of statistical learning theory, Springer-Verlag, 1995
[2] Bhargava, S.K., Sachdev, H.S., Fall, C., Osmond, C., Lakshmy, R., Barker, D.J.P., Biswas, S.K.D., Ramji, S., Prabhakaran, D. & Reddy, K.S., Relation of serial changes in childhood body-mass index to impaired glucose tolerance in young adulthood. New England Journal of Medicine, 2004
[3] Han, J. & Kamber, M., Data Mining: Concepts and Techniques (2nd ed.). Ed. Morgan Kaufmann, 2006
[4] Baccini, A., Besse, P., Data mining / Exploration Statistique. Toulouse INSA, 2010
[5] Lepadatu, C., Explorarea datelor si descoperirea cunostiintelor – probleme, obiective si strategii, RRIA, vol. 4, nr. 4, 2012
[6] Daniel T. Larose (2005), Discovering Knowledge in data. An Introduction to Data Mining, John Wiley & Sons, Inc

Ionela MANIU
"Lucian Blaga" University of Sibiu
Faculty of Sciences
Sibiu, Dr. Ioan Rațiu St. No. 5 - 7
ROMÂNIA
E-mail: ionela.maniu@yahoo.ro

George MANIU
"Lucian Blaga" University of Sibiu
Faculty of Sciences
Sibiu, Dr. Ioan Rațiu St. No. 5 - 7
ROMÂNIA
E-mail: costelmaniu@yahoo.com

Daniel HUNYADI
"Lucian Blaga" University of Sibiu
Faculty of Sciences
Sibiu, Dr. Ioan Rațiu St. No. 5 - 7
ROMÂNIA
E-mail: daniel.hunyadi@ulbsibiu.ro