STATISTICAL ANALYSIS AND A CONCEPTUAL CLUSTERING METHOD TO RANK CLIENTS OF A FINANCIAL BANK

RESUMEN
En este artículo, se presenta una propuesta de clasificación para la categorización de clientes del banco Tatra Bank (Eslovaquia). En esta propuesta se realiza un análisis estadístico para extraer información útil acerca de los clientes del banco, seguido por una estrategia de clasificación basada en el algoritmo LAMDA (Learning Algorithm for Multivariable Data Analysis). Mediante esta propuesta, los datos de los clientes del banco son pretratados para conformar un archivo de entrenamiento, el cual es usado como entrada en el proceso de aprendizaje. La clasificación de los clientes se desarrolla durante el proceso de reconocimiento. La estrategia planteada tiene dos posibilidades, auto aprendizaje y aprendizaje supervisado.

PALABRAS CLAVES: Análisis estadístico, clasificación, caracterización de datos, algoritmos de aprendizaje.

JUAN JOSE MORA
Ingeniero Electricista, Ph.D. (c)
Profesor
Universidad Tecnológica de Pereira
jjmora@utp.edu.co

JOAN COLOMER LLINÁS
Físico, Ph.D.
Profesor Titular
Universitat de Girona
colomer@eia.udg.es

ABSTRACT
In this paper, a classification approach to establish a ranking of the clients of the Tatra Bank (Slovakia), is given. The paper presents the statistical analysis of data in order to select useful information about the clients, followed by a classification strategy based on LAMDA (Learning Algorithm for Multivariable Data Analysis). Because of this approach, the client data provided by Tatra Bank is analyzed to obtain a training file. It is used as input file in learning process. The client classification is performed by the recognition process. This approach has two possibilities: self and supervised learning.

KEYWORDS: *Statistical Analysis, classification, data characterization, learning algorithms.*

1. INTRODUCTION

Guided by the goals of the European Network on Intelligent Technologies for Smart Adaptive Systems – EUNITE, several projects related to artificial intelligence have been conducted for educational and research centers. Particularly in problems of data classification, several approaches using techniques as: neural networks, fuzzy logic, expert system and genetic algorithm, as well as signal processing techniques like Fourier transform, Markov models and wavelet analysis have been used.

In special case, this research is related to Tatra Bank of Slovakia. This bank wants to have a software tool in order to perform a selection or classification of its clients in two defined classes (0 and 1), which means "good" or "not good" clients. In this paper, the statistical analysis of data performed in order to select useful information about the clients, followed by a classification strategy based on LAMDA (Learning Algorithm for Multivariable Data Analysis) is presented. The data of clients provided by Tatra Bank are analysed to obtain a training file used as input file in learning process. The client classification is doing during the recognition process. It is an excellent opportunity to show how the approach presented in this paper performs a good classification using real data.

As content in the second part of this paper, the proposed approach is presented. As third part, the performance of the classification tool is given. Following and as fourth part, the approach of adaptive learning is described. In the fifth part, some conclusions are presented, and finally as last part, the bibliography used is given.

2. THE PROPOSED APPROACH

The approach proposed here is composed by two clear defined processes. The first one is related to data preprocessing and it is based on a statistical approach. The second one is related to the classification tool used to assign a class (0 or 1) to each client of the bank.

2.1. Statistical Approach

In order to select the more significant variables to be used for classification purpose, a statistical analysis of the client data is performed. This analysis is based on different techniques as: duplicate data filtering, outliers filtering, box plot analysis, independency test, factorial analysis, logistics regression and finally, fitting test using the Kolmogorov-Smirnov method.

Fecha de recepción: 31 Mayo de 2004
Fecha de aceptación: 23 Julio de 2004

Duplicate data filtering. Considering that several clients described in the file "client_train_2001.txt" have equal characteristics described by its 36 dimensional data but different output class, a duplicate data filtering was performed. All clients with the characteristics described above were removed form the original training file.

Outliers filtering. Due the normalization performed by the classification tool, data outliers were removed. The outliers are defined as values not-included in the statistic distribution of data. It was done using the box plots, and considering outliers at these values out of the interval defined as [q3+1.5RI, q1-1.5RI]. q1 and q2 are the first and the third quartiles respectively, and RI is the inter-quartil range (RI=q3-q1) [1]. The figure 1 shows the plot of v33 vs v37 before the outliners were removed, while in the figure 2 shows the same plot after the outliers were removed.

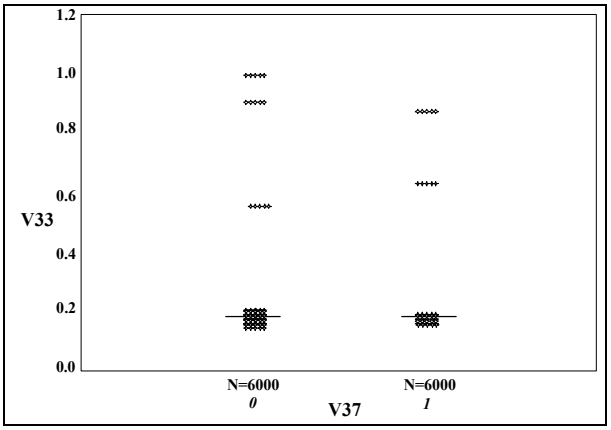

Fig.1. Box plot before applying the outlier removing process

Box plot analysis. According to this analysis and as presented in the figure 2, the distribution of group 0 is biased to topside, while the group 1 has a symmetric distribution. According to this analysis, the variables with significant difference are: v12, v13, v19, v25, v31 and v33.

Independency test. The purpose of this test is to determine if the clients of the bank classified as class 1 are independent of the clients classified as class 0. The specific technique is different depending of the qualitative or quantitative nature of data [2].

In case of qualitative data, an independency χ^2 test is performed. The hypotheses are: H0 in case of these variables vi and v37 were independents (i = 1, 2,…6) and H1 in case of these variables were not-independents. According to this study, the variables not-independents are v1, v3, v4, and v6.

In case of quantitative data, variance equality Lavene test is performed at first time, following by an equalities average T-student test as presented in table 1.

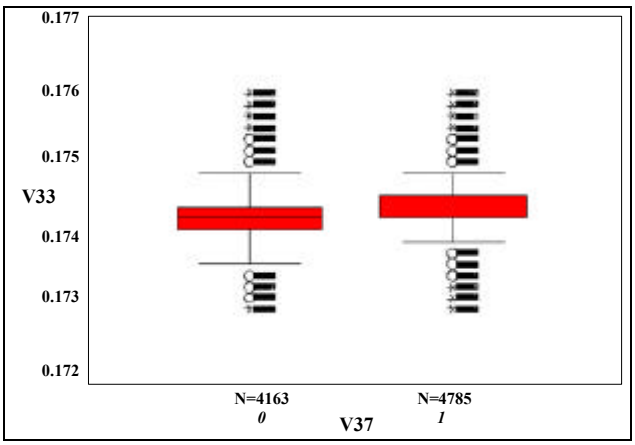

Fig.2. Box plot after applying the outlier removing process

Lavene Test	T-Student test
$\mathbf{H_o}: s_i^2 = s_{37}^2$	$\mathbf{H_o}: m_i = m_{37}$ $\mathbf{H_o}: m_i \neq m_{37}$ (Considering equal variances)
$\mathbf{H_1}: s_i^2 \neq s_{37}^2$	$\mathbf{H_o}: m_i = m_{37}$ $\mathbf{H_o}: m_i \neq m_{37}$ (Considering different variances)

Table1. Test performed to quantitative data of "client_train_2001" file

Where s_i^2 and m_i are the variance and average value of variable *i*, respectively.

According to this test, the variables with a significant difference in its average values are all variables except v7, v8, v9, v11 and v13.

Factorial analysis. This test is performed in order to find hidden variables denoted as factors. These factors explain the correlation configuration in a set of variables observed. The analysis described helps to reduce the amount of data allowing to identify a small number of factors used to explain the variance observed in a great number of original variables [3].

The analysis of principal components is implicit in the factorial analysis. Principal components are characterized by: a) Lineal combination of original variables, b) itself interrelated and c) not directly observable.

According to this test v7, v11, v13, v25, v27, v31, and v33 were extracted as useful variables for the classification purpose.

Logistical regression. This technique is useful to predict the value of a dichotomist variable Y (0 or 1), dependent of m explicative variables (Xj), using the following probability model [2].

$$P\{Y=1\}=\frac{1}{1+\exp[-(b_0+b_1X_1+....b_mX_m)]} \quad (1)$$

This iterative method finishes when the error is sufficient small. In this particular case the following regression is obtained:

$$P\{Y=1\}=\frac{1}{1+\exp[-(Z_{estimated})]} \quad (2)$$

Where

$$Z_{estimated}=(8.911v7+14.84v8+762.288v13+66.637v36-62.122) \quad (3)$$

Because of this test, the variables selected for classification purpose are v7, v8, v13 and v36.

Fitting test. This test is performed if the analyzed data follow a specific distribution (normal, uniform, or exponential), in order to select the best alternative for classification purpose. Because of this test, the data are susceptible to follow a normal distribution, due to χ^2 statistic has the smallest value. This test is performed using the Kolmogorov-Smirnov method [1].

2.2. Learning Algorithm for Multivariable Data Analysis - LAMDA

Basic definition. Classification methods based on hybrid connectives combine both the pure numeric and the pure symbolic classification algorithms, taking profit from the generalizing power of fuzzy logic and the interpolation capability of hybrid connectives [4] [5]. A classification technique called LAMDA (Learning Algorithm for Multivariate Data Analysis) is based on implementation of these possibilities as a fuzzy method of conceptual clustering. It uses data mining techniques under the supervision of an expert to obtain a process or system model [6]. Classification is performed using LAMDA under a supervised learning approach whereas the unsupervised one allows clustering.

General functionality. LAMDA is a classification tool (decision algorithm to assign an object to a class). It has a learning mechanism (based on inductive reasoning using examples) to conform the system model. In addition, it has a recognition mechanism (mechanism to find the best class to be assigned to the object under classification) used to recognize one specific object as a member of one particular class. In this specific case, the data of each client are called "objects"

According to figure 3, LAMDA has two fundamental steps: Learning and Recognition.

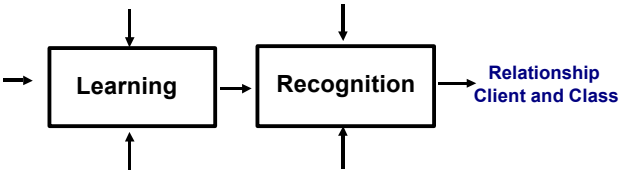

Fig.3. General functionality of LAMDA classification tool

Learning. According to figure 4, at the first stage of learning step (Learning-1) the parameters used to construct the knowledge are updated using training data and its pre-assigned classes. These classes are result of an expert definition knowing as supervised-learning; or a definite number of allowable classes due to a self-learning process. The classes and updated learning parameters are the output of this initial learning stage.

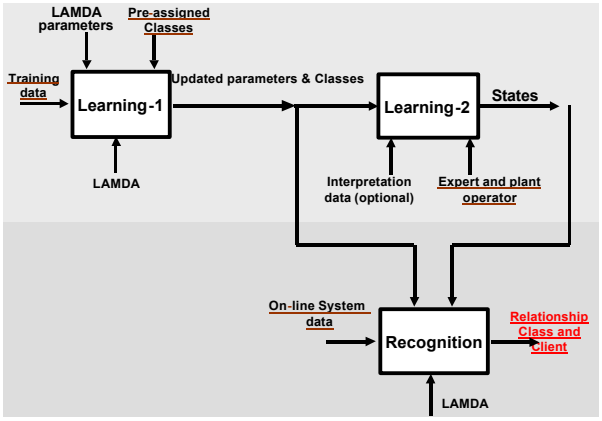

Fig. 4. Detailed functionality of LAMDA classification tool

On the second stage (Learning-2), the output classes are analyzed by an expert. This analysis is performed in order to establish a relationship between states and classes. States are defined as special characteristics that have meaning for the result of the classification. This is an optional stage and is accomplished specially if a self-learning process is performed. As an example, if a self-learning is performed, would appear more than two classes [0,1], and a relationship between the resulting classes and the allowable classes must be established[1].

Two learning step (Learning-1 and Learning-2) are accomplished off-line, but the option of learn online is also available in this methodology as explained in the fourth part of this document.

[1] Defined in this case as two *States*: 0 and 1

Recognition. It is performed using updated learning parameters, classes and states generated in previous stages. In this stage, as presented in the figure 5, one object (Xi) has a number of characteristics called "descriptors". These descriptors are used to describe the object. Every object is assigned to a "class" in the classification process. Class (Ci) is defined as the universe of descriptors, which characterize one set of objects.

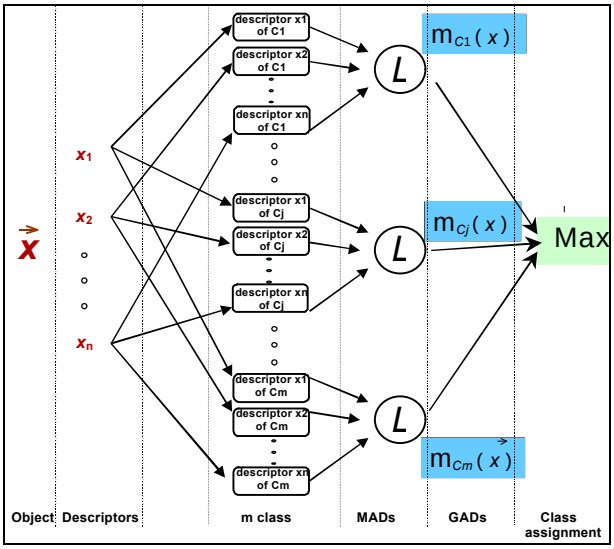

Fig. 5. Basic LAMDA recognition methodology

The MAD (Marginal Adequacy Degree) concept is a term related to how similar is one object descriptor to the same descriptor of a given class, and GAD (Global Adequacy Degree) is defined as a membership degree of one object to a given class [7] [8].

Classification, in LAMDA, is performed according to a similarity criteria computed in two stages. First MAD to each existing class is computed for each object descriptor. Second, these partial results will be aggregated in order to get a GAD of an individual to a class.

The former implementation of LAMDA included only a possibility function to estimate the numeric descriptors distribution [5]. It was a "fuzzification" of the binomial probability function computed as:

$$MAD_{c,d} = r_{c,d}^{X_{i,d}} \ (1 - r_{c,d})^{(1-X_{i,d})} \qquad (4)$$

Where

$r_{c,d}$ = Learning parameter (Ro) for class ***c*** and descriptor ***d***

$X_{i,d}$ = Descriptor ***d*** of individual ***i***

Other implementation, frequently used when the volume of the observed data is important, it is very likely to follow a Gaussian or semi-Gaussian distribution [5]. The marginal adequacy will be computed as:

$$MAD_{c,d} = e^{-\frac{1}{2}\left(\frac{X_{i,d} - r_{c,d}}{s^2}\right)^2} \qquad (5)$$

Where s = Standard deviation

GAD computation is performed as an interpolation between t-norm and t-conorm by means of the β parameter. β = 1 represents the intersection and β = 0 means the union.

$$b\,T(MAD) + (1-b)\,S(MAD) \qquad (6)$$

In table 2, some connectors used for GAD computation are presented

Name	**T-Norm**
Min-max	$T = \min\{x_1, x_2, \ldots x_n\}$
Product	$T = \prod_{i=1}^{n} x_i$
Frank	$T = \log_s\left(1 + \frac{\prod_{i=1}^{n}(s^{x_i} - 1)}{(s-1)^{n-1}}\right)$
Name	**S- Conorm**
Min-max	$S = \max\{x_1, x_2, \ldots x_n\}$
Product	$S = 1 - \prod_{i=1}^{n}(1 - x_i)$
Frank	$S = 1 - \log_s\left(1 + \frac{\prod_{i=1}^{n}(s^{x_i} - 1)}{(s-1)^{n-1}}\right)$

Table 2. T-norms and S-conorms

3. RESULTS

The obtained results correspond to LAMDA supervised-learning process, using preprocessed data as explained in section 2.1. After of this preprocessing (Duplicate data and outliner filtering) were selected 8887 clients data, in order to train the classification tool.

According to different alternatives proposed by the statistical analysis, several test were performed. The file composed by descriptors proposed in factorial analysis was selected as a training file, in order to perform a supervised learning process [2]. In addition, due to the

[2] File composed by qualitative variables: v1, v3, v4, and v6; and quantitative variables: v7, v11, v13, v25, v27, v31, and v33.

values adopted for the variables of the training file, all of these were consider as qualitative nature. The training file and learning strategy were selected because of the good global performance of the classification tool in the recognition stage.

Table 3. Results of class recognition of client_train file

MAD	GAD	β	Error (%)
$r_{c,d}^{X_{i,d}} \ (1-r_{c,d})^{(1-X_{i,d})}$	Min_max	0	10.75
$r_{c,d}^{1-\lvert X_{i,d}-c\rvert} \ (1-r_{c,d})^{\lvert X_{i,d}-c\rvert}$	Product	0.4	10.05
$e^{-\frac{1}{2}\left(\frac{X_{i,d}-r_{c,d}}{s^2}\right)^2}$	Frank (-5.0)	0.2	8.25

In table 3, the error total quantification in client class recognition process is presented. The client data used for recognition were data of client_train file. The error computation is performing taking in to account the number of clients wrong classified, divided by the total amount of clients under recognition.

Several tests were realized, using different approaches, especially in MAD computation using some variations of its characteristic function presented in equations 4 and 5. In addition, several tests were performed using variations in GAD calculation, as presented in table 2. Finally, some tests were performed using different β values (Equation 6).

4. ADAPTIVE STRATEGY

As explained in section 2.2, LAMDA has the option to learn online as an adaptive-learning system. This option will be consider only if the user is completely sure if the new bank client is well classified.

In order to introduce new data, this data must have the same structure as the original train file structure (as explained in the third section of this document). All learning parameters will be updated performing a learning stage with a new training file. This new training file includes both, the original training file and the new client data file.

The possibility above explained is implemented considering each new client data as an object to be classified or to be added to the LAMDA knowledge base.

5. CONCLUSIONS

In most of the cases, a previous filtering and pretreated process of data were necessary. In the particular case here presented, the outliers and duplicate data of the clients of the bank were removed in order to have a better train file.

LAMDA tool, has several possibilities to perform the best classification, even if the objects under analysis are located in a n-dimensional space were there are not linearity separable. It is possible due to class-state matching strategy.

Finally, using this strategy the bank has established its own client ranking useful to determine a value of "trust", which is very important in business.

6. REFERENCES

[1]. ALVAREZ, M: "Análisis estadístico con spss". University of Deusto. 2001.

[2]. BAQUERIZO, R: " Estadística multivariable" http://es.geocities.com/r_vaquerizo/. 2000.

[3]. MILTON, J.S : "Estadística para biología y ciencias de la salud" Interamericana-McGraw-Hill. 1999

[4]. MOORE, K. et al: "Using neural nets to analyse qualitative data". A Marketing Research, vol. 7, nº. 1, p.. 35-39, Winter 1995.

[5]. AGUILAR-MARTÍN and R. López de Mántaras: "The process of classification and learning the meaning of linguistic descriptors of concepts". Approximate Reasoning in Decision Analysis. p. 165-175. North Holland, 1982.

[6]. PIERA, N. : Connectius de lògiques no estandard com a operadors d'agregació en classificació multivariable i reconeixement de formes. N. Piera. Doctorate dissertation in the Universitat Politècnica de Catalunya,1987.

[7]. AGUADO, J.C.: "A Mixed Qualitative-Quantitative Self-Learning Classification Technique Applied to Situation Assessment in Industrial Process Control". Ph. D. Thesis Universitat Politècnica de Catalunya, 1998.

[8]. AGUADO, J.C. et al: "Comparison of structure and capabilities between a non-standard classification technique and the radial basis function neural networks". Proceedings of the 13th European Simulation Multiconference (ICQFN 99), Vol. II, pp. 442-448, Warsaw, Poland, 1999.