Available online at www.sciencedirect.com

SciVerse ScienceDirect

Procedia Economics and Finance 3 (2012) 420 – 425

Procedia Economics and Finance

www.elsevier.com/locate/procedia

Emerging Markets Queries in Finance and Business

A Comparison of Data Classification Methods

Raluca-Mariana Ștefan[a,]*

[a]*Academy of Economic Studies, Bucharest, Romania*

Abstract

Categorizing data in order to use them at their highest level of effectiveness and efficiency is called data classification and it is used to perform complex and varied actions in many different fields including the financial field. Accurate classifications can lead to accurate predictions so, the applied classification method is very important. This paper describes and compares the performances of some data classification methods applied for a real dataset. A list of pros and cons is made for each of the used method. The obtained results show that an important role for the method's level of accuracy is played by the choice of features for the considered data to classify.

Keywords: data classification; supervised classification; linear discriminant analysis (LDA); neural networks (NN); support vector machines (SVM)

1. Introduction

The Maastricht Treaty imposes to every country, component or emergent in the European Union, to apply a certain set of measures that are needed in economic field in order to achieve the nominal and real convergence imposed by a specific level of economic development and prosperity due to globalization process. Lately, the necessity to study extremely large data populations that have heterogeneous features has appeared far more

* Corresponding author. Tel.: +0-40744487477;
E-mail address: rstefan2012@yahoo.com

doi:10.1016/S2212-5671(12)00174-8

often in economic field. So, data analysis and classification data analysis are two complex processes that must be performed in order to be obtained relevant information and knowledge.

Very large and increasingly complex volumes of electronic economic data makes it impossible for the researchers to create an absolute method of classification that can be used regardless the data to extract important information needed to elaborate economic predictions and decisions.

Data analysis represents a set of techniques based on mathematical modeling of data source Cocianu, State, 2008 applied with the help of a searching, validation and classification procedures set. Supervised classification is such a technique of data analysis.

Determining the supervised classifier supposes that a set of labeled samples exists and the classes they belong to. The supervisor observes a training set of samples and based on them it can make predictions on the categorization of some other new presented samples.

Linear discriminant analysis, neural networks and support vector machines are classification methods that were randomly selected to be applied on a real economic data set. Each of these methods was applied in the economic field especially for marketing and banking and they have proved efficient. It is important that these methods are tested on another economic datasets hoping some good results can be found.

By testing these classification methods in order to classify countries based on a set of values regarding a set of economic indicators we can choose the best.

2. Data classification methods

Generally speaking, linear discriminant analysis solves the following types of problems:

- Determining the set of optimal features of objects that allows the best discrimination between two or more object types;
- Utilizing the variables from the optimal set of features to deduct some criteria or rules based on which a separation of the studied population can be done by group or class;
- Using the set of features having the highest discriminating power and the identified separation criteria in order to classify a set of objects, whose belonging is not known, in one of the classes of the observed population Ruxanda, 2009.

A method of determining a classifier or an estimator is the based on model approach Heijden et al., 2004. The set of samples is called training set or learning data. The selection of these samples must be made randomly.

Linear discriminant analysis uses a linear discriminant function that depends on a set of parameters that can be determined based on the training set during the application of the algorithm.

Linear discriminant analysis procedure provides a series of results that contains statistics regarding the studied population of objects. If a parametric method is used then the discriminant function is found in the data set in order to classify future observations.

The sign of the discriminant function determines the label class prediction. The objective is minimizing the error resulted in case of new presented samples to the classifier by adjusting the function set of parameters. There are many methods derived from the classical linear discriminant analysis.

All of these improved methods are used to overcome two of the greatest difficulties of the linear discriminant analysis, dimensionality reduction and small sample size. Among these we mention only a few: Linearly Optimized Discriminant Analysis – LODA Zhang, Chow, 2012, 1D-LDA and 2D-LDA that are algorithms based on vectors Zheng, Lai, Li, 2008, Least Squares – Incremental Linear Discriminant Analysis - LS – ILDA is proposed as an approach of Linear Discriminant Analysis Liu et al., 2009, Least Squares – Online Linear Discriminant Analysis - LS – OLDA Wang, Zhang, 2012 updates exactly the solution to linear discriminant analysis of least squares in case of inserting or deleting an instance, a Graph-based Fisher Analysis algorithm was elaborated Cui, Fan, 2012 and Null Linear Discriminant Analysis – NLDA.

The learning process for a neural network means adjusting the weights by using three types of samples: training, validation and testing samples. A special principle is that the training samples are used to support neural network in its process of adjusting weights, by using the error resulted after a iteration has been made. Neural networks are able to approximate any continuous function and that is why they can provide nonlinear models for time series allowing some efficient predictions Vesely, 2011.

A neural network has layers: input layer, hidden layer and output layer; every layer contains a number of neurons that are interconnected. Given a vector of inputs, the neuron activity implies a function, named activation function that is applied to a linear combination of input values to who a constant value, bias or threshold, is added. The most used activation functions are the step function, sigmoid function or logistic function Vesely, 2011.

The algorithms used for the learning process of a neural network are based on a method of adjustment the weights. One of these methods is based on gradient method that implies the local minimum of a function. Among the most used algorithms we mention Levenberg-Marquart, used to estimate the optimal size of the weights. The weights help the network to accelerate the learning process due to the fact that information are in the weights, not in the neurons.

A study was made regarding the used methods to classify and predict economic data Guresen et al., 2011. The conclusion was that using neural networks for economic data classification and prediction their performances were proved to be very good or excellent being able to compete with diverse econometric models.

The neural networks were used to develop a hybrid model of neural networks and linear regression models in order to have better classification and prediction results Mehdi et al., 2012. This hybrid was applied to a set of banking data. The performances of a neural network were compared to the performances of an econometric model method applied for a set of data regarding the inflation rate Moshiri and Cameron, 2000.

Learning machines that are building decision functions are named support vector machines. These support vector machines have the advantage of using the two-class classification to make a n-class classifier by applying n rules of classification for a set of objects.

Support vector machines represent the state-of-the-art for machine learning and they are linear classifiers for multidimensional spaces. The complexity of building a support vector machine is due to the number of support vectors not to the features space dimension Vapnik, 1998.

Training a support vector machine means solving a quadratic programming problem. Some decomposition methods are used in order to solve this problem: Sequential Minimal Optimization – SMO and SVM^{light} methods of decomposition. Another class of algorithms created to solve the quadratic problem was proposed by Osuna, 1997.

Sequential Minimal Optimization algorithm Platt, 1999 allows solving the quadratic problem easier, decomposing it in many smaller sub problems of the same type, based on Osuna theorem.

A gradient-based learning algorithm was presented and its performances were described considering applying it on a dataset Cocianu et al., 2011.

3. Classifying countries based on economic indicators data

Linear discriminant analysis, neural networks and two algorithms based on Support Vector Machine were applied to a set of numeric observations regarding the registered values for an economic indicators group that supplies the level of economic prosperity and development for a country. The data were taken from the European Commission site based on an analysis made by Golden Mind & Spirit company. Each of the 26 chosen countries for classification is first characterized by 6 numeric features representing the values for the following indicators: exchange rate, inflation rate, long-term interest rate, budget deficit/surplus, public debt and general domestic product.

The second application is also considering 26 countries characterized by 6 numeric features: energy consumption, employment rate, resource productivity, unemployment rate by age, unemployment rate by education level and unemployment rate by gender. Two classes were considered for classification: developed country and emergent country.

Table 1. National economic prosperity indicators group and economic development indicators group

Indicators for the first application	Indicators for the second application
Exchange rate	Energy consumption
Inflation rate	Employment rate
Long term interest rate	Resource productivity
Budget deficit/surplus	Unemployment rate by age
Public debt	Unemployment rate by education level
Gross domestic product	Unemployment rate by gender

Source: www.buzznews.ro

After a Matlab code was applied to classify the countries using linear discriminant analysis method to the first group of data, the percentage error equal to a rate of incorrect classifications that was obtained is 28.53%, so the correct classification rate was 71.47%. A confusion matrix shows, for each of the two considered classes, the number of correct respectively, incorrect classified countries.

For the second case we obtained a correct classification rate equals to 52.56%, this result being the weakest of all classification methods.

We have been surprised to have the very same result when we have applied Support Vector Machine technique using Sequential Minimal Optimization.

Confusion matrix elements showed for the linear discriminant analysis a big result difference between the countries that were classified based on the national prosperity indicators group and the countries classified based on the values of economic development indicators.

Confusion matrix resulted for the neural network technique displays three types of data: training data, validation data and testing data.

We note the confusion matrix cm_{ij}, where i and j belong to {1, 2, 3} so we can make the next notes for the considered cases:

- cm_{11} and cm_{22} shows the number and corresponding percentage of correct classified objects;
- cm_{12} and cm_{21} shows the number and corresponding percentage of misclassified objects;
- cm_{33} displays overall number and percentage obtained for correct classified objects and misclassified objects.

For the first application neural network technique we obtained another interesting result because all of the countries were correct classified, having 100%.

For the second application a percentage of 94.20% correct classified objects, was obtained. In order to make a good classification for a set of countries it is very important the phase of choosing the features that characterize them.

Percentage of correct classification resulted after we applied Support Vector Machine with the default algorithm (QP) to the first group of indicators is 68,59%, that is lower than the one with the sequential minimal optimization. That indicates the importance given by the choice of the algorithm.

The same method was applied for the second group of economic indicators and 53.21% of the countries were correct classified.

Our opinion is that the differences that were obtained are due to the highest level of complexity, heterogeneity and instability that economic indicators data have. In order to make a right decision based on classification techniques we have to consider using more than one method.

4. Conclusions

Table 2. Comparison of the results obtained for economic data classification

Classification technique	Correct classifications rate (%) Economic prosperity data	Correct classifications rate (%) Economic development data
LDA	71.47	52.56
NN	100.00	94.20
SVM (QP)	68.59	53.21
SVM (SMO)	73.08	52.56

Source: Matlab results of author's applications

We can note the fact that the best correct classification rate of the countries was found when neural network method was applied, according to economic indicators values used for both cases. We cannot help wonder if this is due to *over fitting*, that is when the neural network learns by heart the samples and the classes they belong to. So, we need to expand our research to larger samples applications.

We have chosen to classify a certain category of economic data that characterize the level of economic prosperity and development because we wanted to expand the researches made until now. Macroeconomic indicators values can be influenced by their components, other economic indicators. It is necessary to use these methods in order to provide essential and strong information to build a foundation on which an economic decision can be made.

So, is very important for us to test the influence of the chosen features by applying data classification methods on a larger data set. Further future research regarding the use of macroeconomic variables for data classification can be useful to overcome economic difficulties based on the prognosis results.

There are other techniques that can be used for classification and prediction, but they have the disadvantage of not being robust, offering more than one optimal solution. This disadvantage is overcome more or less by the used techniques and that is why we recommend the use of more than one method for the macroeconomic domain, so that the best technique can be chosen to be a strong support for classification and prediction problems.

Acknowledgements

Many sincere thanks to my supervisor, Prof. dr. Cătălina-Lucia Cocianu, who accepted me as PhD. Sudent and who guided me and offered me a great support in my research.

References

Agathon, D.M.; Sava, C., 2012, A Golden Mind & Spirit analysis – The figures that tear us apart. How far are we from the developed countries from E.U. Available at www.buzznews.ro.

Alpaydin E., Introduction to Machine Learning, MIT Press, 2004.

Chen P.-H., Fan R.-E., Lin C.-J., 2006, A study on SMO-type decomposition methods for support vector machines, IEEE Trans. Neural Networks, 17:893-908.

Ching, W.-K., Chu, D., Liao, L.-Z., Wang, X., 2012, Regularized orthogonal linear discriminant analysis, Pattern Recognition, Elsevier, Article in press.

Cocianu C. L., State L., Uscatu C. R., Ştefan R.-M., 2011, Learning From Data – A SVM Approach, Economy Informatics, vol. 11, no. 1.

Ganatr, A.; Kosta, Y.P. (2010) Spiking Back Propagation Multilayer Neural Network Design for Predicting Unpredictable Stock Market Prices with Time Series Analysis, International Journal of Computer Theory and Engineering, 2 (6), p. 963 – 971.

Guresen, E., Kayakutlu, G., Daim, T.U., 2011, Using artificial neural network models in stock market index prediction, Expert Systems with Applications 38, Elsevier, pag. 10389-10397.

Guyon, I.., Stork, D., 2000, Linear Discriminant and Support Vector Classifier: Advances in Large Margin Classifiers, Cambridge, MA, MIT Press. (Schőlkopf, B., Burges, C.J., Smola, A.J. eds.).

Hastie, T., Tibshirani, R., Discriminant Analysis by Gaussian Mixtures, J.R. Statist. Soc 1, pag. 155-176, 1996.
Haykin, S., Neural networks and learning machines, Prentice Hall, Third edition, 2009.
Heijden, van der F., Duin, R.P.W., de Ridder, D., Tax, D.M.J., 2004, Classification, Parameter Estimation and State Estimation: An Engineering Approach Using MATLAB, Wiley.
http://epp.eurostat.ec.europa.eu/.
Joachims, T., 2001, A Statistical Learning Model of Text Classification with Support Vector Machines. Proceedings of the Conference on Research and Development in Information Retrieval (SIGIR), ACM.
Kecman, V., Huang T.-M., Vogt M., Iterative Single Data Algorithm for Training Kernel machines from Huge Data Sets: Theory and Performance, In L. Wang, editor, Support Vector Machines: Theory and Applications, pag. 255-274, Springer Verlag, 2005.
Mehdi, K., Hamadani, A.Z., Bijari, M., 2012, A novel hybrid classification model of artificial neural networks and multiple linear regression models, Expert Systems with Applications 39, p. 2606-2620.
Moshiri S.; Cameron N. (2000) Neural Network versus Econometric Models in Forecasting Inflation, Journal of Forecasting, 19, p. 201 – 217.
Osuna, E., Freund, R., Girosi, F., 1997, An improved training algorithm for support vector machines. In Neural Networks for Signal Processing VII – Proceedings of the 1997 IEEE Workshop, New York.
Pao, H.-T., 2008, A comparison of neural network and multiple regression analysis in modelling capital structure. Available at www.sciencedirect.com.
Platt, J., 1999, Fast Training of Support Vector Machines Using Sequential Minimal Optimization. In: Advances in Kernel Methods – Support Vector Learning, Cambridge, MA, MIT Press., Schőlkopf, B., Burges, C.J., Smola, A.J. eds.
Rafiei, F.M., Manzari, S.M., Bostanian, S., 2011, Financial health prediction models using artificial neural networks, genetic algorithm and multivariate discriminant analysis: Iranian evidence, Expert Systems with Applications 38, Elsevier, pag. 10210-10217.
Ruxanda, G., 2009, Analiza multidimensională a datelor, curs Şcoala Doctorală, Academia de Studii Economice, Bucureşti.
Schőlkopf, B., Burges, C.J., Smola, A.J., 1999, Advances in Kernel Methods – Support Vector Learning, Cambridge, MA, MIT Press.
Smola, A.J., Schőlkopf, B., Müller, K.-R., 1998, General Cost Functions for Support Vector Regression. In Proc. of the Ninth Australian Conf. in Neural Networks, Brisbane, Australia.
Ştefan, R.-M., 2012, An Overview of Frequently Used Algorithms to Build Clusters, Journal of International Scientific Publications: Materials, Methods & Technologies, Volume 6, Sunny Beach, Bulgaria.
Ştefan, R.-M., Şerban, M., 2012, Neural Network Principles to Classify Economic Data, 19th International Economic Conference – IECS June 15, 2012, The Persistence of the Global Economic Crisis: Causes, Implications, Solutions, Lucian Blaga University of Sibiu.
Ştefan, R.-M., Şerban, M., 2012, Linear Discriminant Analysis for Data Classification, Arad Academic Days 22nd Edition, Universitatea de Vest Vasile Goldiş.
Ştefan, R.-M., Şerban, M., 2012, Tehnici de instruire SVM pentru rezolvarea problemelor de clasificare a datelor, Conferinţa Economică Naţională "Echilibre şi dezechilibre ale pieţei româneşti în perioada actuală", 8 mai 2012, Universitatea Spiru Haret.
Vapnik, V., 1998, Statistical Learning Theory, John Wiley, N.Y.
Vesely, A., 2011, Economic classification and regression problems and neural networks, Agricultural Economics, 3, p. 150 – 157.
Wang, Q., Zhang, L., 2012, Least Squares Online Linear Discriminant Analysis, Expert Systems with Applications 39 (2012), pag. 1510-1517, Elsevier.
Youn, E., Koenig, L., Jeong, M.K., Baek, S.H., 2010, Support vector-based feature selection using Fisher's linear discriminant and Support Vector Machine, Expert Systems with Applications 37, p. 6148-6156.
Zhang, Z., Chow, T.W.S., 2012, Robust Linearly Optimized Discriminant Analysis, Neurocomputing, Vol. 79, pag. 140-157, Elsevier.