Kafkas Univ Vet Fak Derg
20 (3): 351-356, 2014
DOI: 10.9775/kvfd.2013.10154

Journal Home-Page: http://vetdergi.kafkas.edu.tr
Online Submission: http://vetdergikafkas.org

RESEARCH ARTICLE

Comparison of Classification Performance of Selected Algorithms Using Rural Development Investments Support Programme Data

Mehmet Ali ALAN [1] Cavit YEŞİLYURT [2] Saadettin AYDIN [3] Erol AYDIN [4]

[1] Cumhuriyet University, Faculty of Economics and Administrative Sciences, Department of Management Information Systems, TR-58140 Sivas - TURKEY
[2] Kafkas University, Faculty of Economics and Administrative Sciences, Department of Management, TR-36100 Kars - TURKEY
[3] Enstitute of Strategic Thinking, TR-06460 Ankara - TURKEY
[4] Kafkas University, Faculty of Veterinary Medicine, Department of Livestock Economics and Management, TR-36100 Kars - TURKEY

Makale Kodu (Article Code): KVFD-2013-10154

Summary

It is not always possible to solve a large size of data via traditional statistical techniques. In order to solve these kinds of data special tactics like data mining are needed. Data mining may meet these kinds of needs with both categorizing and piling tactic. In this study, we have used data mining by using Rural Development Investment Support Program (RDISP) data with various categorizing algorithms. The most prospering categorizing algorithm was tried to determine by using present data. At the end of analysis, it has been understood that MLP (multilayer perceptron), a nerve net model, is the best algorithm that makes the best categorizing.

Keywords: *Data mining, MLP, Nerve net model, RDISP, Rural development*

Kırsal Kalkınma Yatırımlarının Desteklenmesi Programı Verileri Kullanılarak Seçilen Algoritmalarının Sınıflandırma Performanslarının Karşılaştırılması

Özet

Kapsamlı verileri geleneksel istatistiksel teknikler yardımıyla değerlendirmek mümkün değildir. Bu tür kapsamlı verileri değerlendirmek için "Veri Madenciliği" gibi özel tekniklere ihtiyaç vardır. Veri madenciliği kapsamlı verileri hem kategorize ederek hem de kazık taktik kullanarak değerlendirmeyi kolaylaştırmaktadır. Bu çalışmada, Kırsal Kalkınma Yatırım Destekleme Programı (KKYDP) verilerinde çeşitli kategorize algoritmaları yardımıyla veri madenciliği tekniği kullanılmıştır. Çalışmada en uygun kategorize algoritma mevcut veriler kullanarak belirlenmeye çalışılmıştır. Sonuç olarak; analizlerde en iyi kategorizasyon yapan algoritma modelinin Çok Katmanlı Algılayıcı (ÇKA) yapay sinir ağ modeli olduğu belirlenmiştir.

Anahtar sözcükler: *ÇKA, Kırsal kalkınma, KKYD, Sinir ağ modeli, Veri madenciliği*

INTRODUCTION

Databases are rich with hidden information that can be used for intelligent decision making. Classification and prediction are two forms of data analysis that can be used to extract models describing important data classes or to predict future data trends. Such analysis can help provide us with a better understanding of the data at large [1].

The different disciplines on database's data make statistical, mathematical, machine learning and visual analyses with different purposes. One of those analyses techniques is data mining. There are a lot of algorithms in data mining.

Data mining is an interdisciplinary field, the confluence of a set of disciplines, including database systems, statistics, machine learning, visualization, and information science [1].

Data mining, the science and technology of exploring data in order to discover previously unknown patterns, is a part of the overall process of knowledge discovery in

İletişim (Correspondence)
+90 542 7254725
alan@cumhuriyet.edu.tr

databases (KDD). In today's computer-driven world, these databases contain massive quantities of information. The accessibility and abundance of this information makes data mining a matter of considerable importance and necessity [2].

Data mining is known as knowledge discovery process of analyzing data from different point of views and to work out into useful information which can be applied in various application, including advertisement, bioinformatics, database marketing, fraud detection, e-commerce, health care, security, web, financial forecasting etc [3].

According to the Gartner Group, "Data mining is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques [4].

Data mining is separated from the other statistical tactics in point of using the whole data. Instead of working with the small data that traditionally procured easier evaluation can be made and new independent data can be preferred [5].

Classification is a task that occurs very frequently in everyday life. Essentially it involves dividing up objects so that each is assigned to one of a number of mutually exhaustive and exclusive categories known as *classes*. The term "mutually exhaustive and exclusive"'simply means that each object must be assigned to precisely one class, i.e. never to more than one and never to no class at all.

Many practical decision-making tasks can be formulated as classification problems, i.e. assigning people or objects to one of a number of categories [6].

Classification is the operation of separating various entities into several classes. These classes can be defined by business rules, class boundaries, or some mathematical function. The classification operation may be based on a relationship between a known class assignment and characteristics of the entity to be classified. This type of classification is called supervised. If no known examples of a class are available, the classification is unsupervised. The most common unsupervised classification approach is clustering. The most common applications of clustering technology are in retail product affinity analysis (including market basket analysis) and fraud detection [7].

The concept of supervised classification in data mining is to learn a classification function or construct a classification model based on the known data, which is also called a classifier. This function or model maps the data in the data base to the target attribute, and can, therefore, be used to forecast the class of new data [8].

There are many data mining algorithms, such as association rules, clustering, decision trees, discriminant analysis, artificial neural networks, genetic algorithms, and so on. These algorithms are used to process data from various fields to retrieve information and discover knowledge that can drive an executive's decisions. Information is data associated with the past and the present. Knowledge provides a basis for the prediction of future trends based on original data and the necessary information extracted from the original data. Clearly, information and knowledge are communicated through data [9].

As a result of the data mining analysis, the algorithm of Multilayer Perceptron (MLP) has been the most successful classification algorithm. The algorithm of MLP is neural network model.

A neural network consists of a layered, feed-forward, completely connected network of artificial neurons, or nodes. Neural networks are used for classification or estimation [10].

Neural networks can be used for many purposes, notably descriptive and predictive data mining. They were originally developed in the field of machine learning to try to imitate the neurophysiology of the human brain through the combination of simple computational elements (neurons) in a highly interconnected system. They have become an important data mining method [11].

Representatives of machine learning methods are: artificial neural networks (ANN), self-organizing maps, Hopfield network, genetic algorithms, evolutionary algorithms, fuzzy systems, rough sets, rule-based systems, support vector machines, decision trees, Bayesian and probabilistic models [12].

Work on artificial neural networks (ANNs) has been motivated by the recognition that the human brain computes in an entirely different way from the conventional digital computer. It was a great challenge for many researchers in different disciplines to model the brain's computational processes. The brain is a highly complex, nonlinear, and parallel information-processing system. It has the capability to organize its components so as to perform certain computations with a higher quality and many times faster than the fastest computer in existence today. Examples of these processes are pattern recognition, perception, and motor control [13].

The backpropagation algorithm performs learning on a *multilayer feed-forward* neural network. It iteratively learns a set of weights for prediction of the class label of tuples. A multilayer feed-forward neural network consists of an *input layer*, one or more *hidden layers*, and an *output layer*. An example of a multilayer feed-forward network is shown by Han and Kamber [1]. Each layer is made up of units. The inputs to the network correspond to the attributes measured for each training tuple. The inputs are fed simultaneously into the units making up the input

layer. These inputs pass through the input layer and are then weighted and fed simultaneously to a second layer of "neuronlike" units, known as a hidden layer. The outputs of the hidden layer units can be input to another hidden layer, and so on. The number of hidden layers is arbitrary, although in practice, usually only one is used [1].

The weighted outputs of the last hidden layer are input to units making up the output layer, which emits the network's prediction for given tuples. The units in the input layer are called input units. The units in the hidden layers and output layer are sometimes referred to as neurodes, due to their symbolic biological basis, or as output units.

Multilayer feed-forward networks are one of the most important and most popular classes of ANNs in real-world applications. Typically, the network consists of a set of inputs that constitute the input layer of the network, one or more hidden layers of computational nodes, and finally an output layer of computational nodes. The processing is in a forward direction on a layer-by-layer basis [13].

A multiplayer perceptron has three distinctive characteristics:

1. The model of each neuron in the network includes usually a nonlinear activation function, sigmoidal or hyperbolic.
2. The network contains one or more layers of hidden neurons that are not a part of the input or output of the network. These hidden nodes enable the network to learn complex and highly nonlinear tasks by extracting progressively more meaningful features from the input patterns.
3. The network exhibits a high degree of connectivity from one layer to the next one.

The multilayer perceptron is the most commonly used architecture for predictive data mining. It is a feed-forward network, with possibly several hidden layers, one input layer and one output layer, totally interconnected. It can be considered as a highly non-linear generalization of the linear regression model when the output variables are quantitative, or of the logistic regression model when the output variables are qualitative [11].

In this study, the algorithms named as MultiLayer Perceptron, Ridor, DTNB, ADTree, LADTree, SPegasos, SMO, Dagging, IBk, FT, LMT, LBR, Voted Perceptron, OneR, IB1, VFI, Decorate, Bayes Net, RBF Network, Naïve Bayes were used to select the algorithm has the best classification performances by benefiting the Rural Development Investments Support Programme (RDISP) data. In order to determine the best algorithm, Correctly Classified Instances, Kappa statistic, Mean absolute error, Root mean squared error, Relative absolute error, Root relative squared error, TP Rate, FP Rate, F-Measure, classification timing values of the algorithms has taken into consideration.

MATERIAL and METHODS

In this study the data from Sivas Provincial Directorate of Agriculture were used in the frame of Rural Development Investment Support Program. There are 14859 data belong to 1143 appeal. Data have been taken into Excel format. Then necessary regulations, transformations are made by using Excel macros and the file saved in the name of "kkydp.arff". Both individual and corporate applications of RDISP are kept as B/K (I/C). Lands owned by applicants have been identified as "up to 30 decare", "between 31-40 decare" and "more than 40 decare" and applicant's loan land has been identified similarly as "reach up 30 decare", "between 31-40 decare" and "more than 40 decare". Applicant's request type is set as "M1/M2" (M1: local, M2: Imported). For city or county "i" value and for village "k" value was assigned as application location. Utilization statement was assigned as "twice", "once" or "never" from RDISP. Results were determined to be "Positive" if the application has been accepted and "Negative" if it has not. It was assigned whether there is no value belong to a variable or it's undefined "?" value.

Variable definitions in pre-processing step of prepared dataset are given below: @relation kkydp, @attribute BT {B,K}, @attribute S30 {E,H}, @attribute S3040 {Y,N}, @attribute S40 {Y,N}, @attribute K30 {Y,N}, @attribute K3040 {Y,N}, @attribute K40 {Y,N}, @attribute MType {M1,M2}, @attribute location {i,K}, @attribute twice {Y,N}, @attribute once {Y,N}, @attribute notbenefited {Y,N}, @attribute Class {Positive, Negative}, @DATA B,N,N,Y,N,N,Y,M1,K,N,N,Y, Positive; B,N,Y,N,N,N,N,M1,K,N,N,Y, Negative.

RESULTS

In this study, WEKA (Waikato Environment for Knowledge Analysis) program's 3.6.9 version that was developed in Waikato University was used [14]. WEKA program is open source code software. This program supports a lot of categorizing, piling and coupling rules algorithm. Instead of text based arff., arff.gz, names, data, csv, c45, libsvm, dat, bsi, xrff, xrff.gz file types WEKA supports databases and URL addresses that include data.

An Intel i5 model and 1.7 Ghz CPU, 6 Gb RAM and 64 bit Win 8 operating systemic laptop was used during the application.

There are different but in equal number of values for every defined variables as it is shown in *Fig. 1*. In addition, every variable has been take by using two different values as yes or no {Y,N} and these values represented as two different colors.

Results that come from post prepared data set used in WEKA program is given in the chart. As it can be understood from the *Table 1*, Multilayer Perception algorithm is the

best algorithm that makes the best categorizing with 992 correctly classified instances. This algorithm's cappa statistic is 0.7321, True Positive rate is 0.868, and False Positive rate is 0.124 and F-measure is 0.869. This algorithm's categorizing time is 1.99 seconds. Ridor algorithm follows this algorithm with 973 true categorizing numbers.

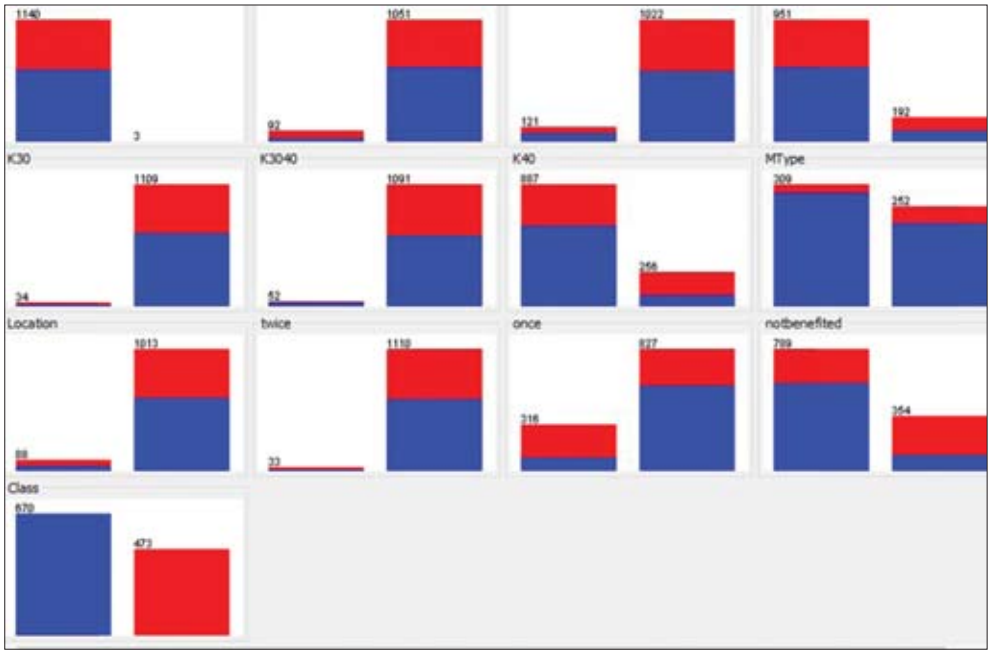

Fig 1. Variables Graphics
Şekil 1. Değişkenler grafiği

Table 1. *Performance ratings of selected algorithms in Decision Tree Analysis*
Tablo 1. *Karar Ağaçları Analizine ait bazı algoritmaların başarım dereceleri*

Algorithms	Correctly Classified Instances	Kappa Statistic	Mean Absolute Error	Root Mean Squared Error	Relative Absolute Error %	Root Relative Squared Error %	TP Rate	FP Rate	F-Measure	Time (in seconds)
MultiLayer Perceptron	992	0.7321	0.1465	0.2767	30.1889	56.1843	0.868	0.124	0.869	1.99
Ridor	973	0.6972	0.1487	0.3857	30.6551	78.3030	0.851	0.145	0.852	0.06
DTNB	965	0.6737	0.2085	0.3161	42.9768	64.1741	0.844	0.181	0.843	0.47
ADTree	965	0.6796	0.2562	0.3243	52.7993	65.8508	0.844	0.163	0.844	0.06
LADTree	962	0.6584	0.2294	0.3200	47.2783	64.9635	0.842	0.212	0.835	0.13
SPegasos	927	0.6036	0.1890	0.4347	38.9500	88.2634	0.811	0.218	0.809	0.11
SMO	927	0.6036	0.1890	0.4347	38.9500	88.2634	0.811	0.218	0.809	0.17
Dagging	926	0.5991	0.2203	0.4106	45.4056	83.3628	0.810	0.225	0.807	0.14
IBk	924	0.5807	0.2188	0.3597	45.0871	73.0424	0.808	0.261	0.797	0.01
FT	923	0.5952	0.2644	0.3889	54.5024	78.9647	0.808	0.224	0.805	0.36
LMT	921	0.5923	0.2809	0.3758	57.8871	76.2947	0.806	0.224	0.803	1.78
LBR	916	0.5682	0.2457	0.3819	50.6500	77.5334	0.801	0.263	0.791	0.02
Voted Perceptron	914	0.5788	0.2003	0.4476	41.2942	90.8754	0.800	0.232	0.797	0.02
OneR	904	0.5831	0.2091	0.4573	43.0974	92.8438	0.791	0.188	0.792	0.01
IB1	897	0.5270	0.2152	0.4639	44.3597	94.1936	0.785	0.291	0.770	0.01
VFI	882	0.5119	0.4655	0.4675	95.9414	94.9301	0.772	0.278	0.765	0.02
Decorate	870	0.5007	0.4171	0.4361	85.9614	88.5406	0.761	0.267	0.759	1.38
Bayes Net	842	0.4456	0.3048	0.4151	62.8232	84.2842	0.737	0.300	0.733	0.05
RBF Network	840	0.4409	0.3252	0.4086	67.0371	82.9662	0.735	0.304	0.731	0.31
Naïve Bayes	806	0.3801	0.3323	0.4529	68.4844	91.9612	0.705	0.705	0.701	0.01

Although this algorithm's categorizing time is shorter than MultiLayer Perceptron the other values are worse. Other algorithms follow them.

DISCUSSION

Nowadays, the amount of stored data extremely increases. Data storage is performed by not only private sector, but also by public enterprises such as provincial directorate of agriculture. While private enterprises are achieved in increasing customer commitment to the enterprise and customer satisfaction by using these data, especially public enterprises could not use these data effectively time to time. These data could include some beneficial hide patterns for both public and private sector. One of the most important methods used in producing beneficial information from these data is data mining. Data mining is to produce beneficial and useful data from large scale of data.

In data mining; different methods such as; statistical methods, decision trees, genetic algorithm, fuzzy logic and artificial neural networks could be used. Despite traditional methods, in data mining, inferences could be deduced oriented to results by using the entire data. In this technique, not only numerical data but also alpha-numerical data is used in analyses. A data warehouse is formed by changing both numeric and alpha numerical data to required form. This study is performed by subjecting current data to required change and a data warehouse is prepared in text format by pruning.

Both numerical results and visual results are used in data mining, In this study, there is a graphic *(Fig. 1)* which shows the place of each variable in the entire data besides determination of the most successful classification algorithm. There are a lot of studies as the current study and most of these are from different data sets about the subject. One of these is Palaniappan et al.[15], categorized decision tree, Naïve Bayes method and artificial nerve nets by using Heart Disease Prediction System (HDPS)'s data and they stated that these results may help nursing and medicine students. In addition; Frank et al.[16] introduced how to use WEKA software in Bioinformatics, emphasizing that it supports important categorizing and regression techniques like decision trees, rule masses, bayes sorters, SVM (Support Vector Machines), logistic and linear regression, MLP (Multi-Layer Perceptron) and the closest neighbour. Another study is Ngai et al.[17] categorized Customer Relations Method (CRM) and data mining articles by scanning and probed using data mining in customer relations. Kirkos et al.[18], presented that fraudulent statements can be determined by using decision trees, Artificial Nerve Nets and Bayes Nets from data mining algorithms. DIMIC et al.[19] collected student data by using Moodle e-learning materials and made analyses with categorizing, piling and coupling rules technique by those data. Hsieh [20] made analyses with artificial nerve nets and coupling rules by using the data from bank database and presented the data mining's contribution on behavioral methods of a credit card customer in a bank. Hung et al.[21] tried to guess customer transfers between mobile companies through both artificial nerve nets and decision trees data mining, using data from a telecom company in Taiwan.

Healthy estimation are very important in the studies. One of the most widely used techniques in data mining is the classification. Estimation techniques based on machine learning have been proved to be more successful than the traditional estimation techniques in parallel to developments in information technologies.

As a result of the study, MultiLayer Perceptron algorithm is the best classification algorithm which is an artificial neural networks model. In this study it has been introduced, artificial neural network classification performance in data mining algorithms is higher than that of the other algorithms.

REFERENCES

1. Han J, Kamber M: Data Mining: Concepts and Techniques. Second ed., Morgan Kaufmann Publications, San Francisco, 2006.

2. Rokach L, Maimon O: Data Mining With Decision Trees. World Scientific, New Jersey, 2008.

3. Jain YK, Yadav VK, Panday GS: An efficient association rule hiding algorithm for privacy preserving data mining. *IJERT*, 3 (7): 2792-2798, 2011.

4. Larose DT: Discovering Knowledge in Data. Wiley Publication, New Jersey, 2005.

5. Weiss SM, Zhang T: Performance Analysis and Evaluation. The Handbook of Data Mining. Arizona State University Press, Lawrence Erlbaum Associates, Mahwah, 425-440, 2003.

6. Bramer M: Principles of Data Mining. Springer, London, 2007.

7. Nisbet R, Elder J, Miner G: Handbook of Statistical Analysis and Data Mining Applications. Elsevier Inc, Burlington, 2009.

8. Dong-Peng Y, Li JL, Lun R, Chao Z: Applications of Data Mining Methods in the Evaluation of Client Credibility. **In.** Soares C, Peng Y,Meng J, Washio T, Zhou ZH (Ed): Applications of Data Mining in E-Business and Finance IOS Press, Amsterdam, 2008.

9. Wu T, Li X: Data- Storage and Management. **In,** Nong YE (Ed): The Handbook of Data Mining. New Jersey: Lawrence Erlbaum Associates, Inc., 393-407, 2003.

10. Larose DT: Data Mining Methods and Models, A John Wiley & Sons, Inc., Publication, New Jersey, 2006.

11. Giudici P, Figini S: Applied Data Mining For Business and Industry, Second ed., Wiley Publicition, West Sussex, 90-91, 2009.

12. Lappas G: Machine Learning and Web Mining: Methods and Applications in Societal Benefit Areas. **In,** Rahman H (Ed): Data Mining Applications for Empowering Knowledge Societies, 76-95, 2009.

13. Kantardzic M: Data Mining: Concepts, Models, Methods and Algorithms, John Wiley & Sons J. B. Speed Scientific School, University of Louisville IEEE Computer Society, 2003.

14. WEKA: Waikato Environment for Knowledge Analysis. http://www.cs.waikato.ac.nz/ml/weka/downloading.html, *Accessed:* 06.09.2013.

15. Palaniappan S, Awang R: Intelligent heart disease prediction system using data mining techniques. *IJCSNS*, 8 (8): 343-350, 2008.

16. Frank E, Hall M, Trigg L, Holmes G, Witten H: Data mining in bioinformatics using WEKA. *Bioinformatics*, 20 (15): 2479-2481, 2004.

17. Ngai EWT, Xiu L, Chau DCK: Application of data mining techniques in customer relationship management: A literature review and classification. *Expert Systems with Applications*, 36, 2592-2602, 2009.

18. Kirkos E, Spathis C, Manolopoulos Y: Data Mining techniques for the detection of fraudulent financial statements. *Expert Systems with Applications,* 32, 995-1003, 2007.

19. Dimić G, Kuk K, Ahorjanski M: Mining Student's data for analyze electronic learning materials available on the moodle course. *Metalurgia International*, 16 (12): 78-82, 2011.

20. Hsieh NC: An integrated data mining and behavioral scoring model for analyzing bank customers, *Expert Systems with Applications,* 27, 623–633, 2004.

21. Hung SY, Yen DC, Wang HY: Applying data mining to telecom churn management. *Expert Systems with Applications,* 31, 515-524, 2006.