World Applied Sciences Journal 7 (Special Issue of Computer & IT): 216-223, 2009
ISSN 1818.4952

An Experimental Investigation of the Effect of Discrete Attributes on the Precision of classification Methods

Reza Entezari-Maleki, Seyyed Mehdi Iranmanesh and Behrouz Minaei-Bidgoli

Department of Computer Engineering, Iran University of Science and Technology (IUST), Tehran, Iran

Abstract: In this paper, the precision of logistic regression, naïve-Bayes and linear data classification methods, with regard to the Area Under Curve (AUC) metric have been compared. The effect of the parameters including size of the dataset, kind of the independent attributes, number of the discrete attributes and their values have been investigated. From the results, it can be concluded that in datasets consisting of both discrete and continuous attributes, the AUC of the three mentioned classifiers are the same. With increasing the number of the discrete attributes, the AUC of logistic regression is increased and the precision related to this classifier become more than the other two classifiers. Also considering the impact of the discrete attributes it can be seen that with increasing the number of values in discrete attributes the AUC related to the logistic regression classifier increases and linear classifier's AUC decreases, but the AUC of the naïve-Bayes classifier remains constant. Therefore, the results of this research can help data miners in selecting the most efficient classifier by considering the characteristics of the datasets.

Key words: Logistic regression . naïve bayes . linear classifier . Area under Curve (AUC)

INTRODUCTION

Data mining algorithms which carry out the assigning of objects into related classes are called classifiers. Classification algorithms include two main phases; in the first phase they try to find a model for the class attribute as a function of other variables of the datasets and in the second phase, they apply previously designed model on the new and unseen datasets for determining the related class of each record [1]. There are different methods for data classification such as decision trees (DT), rule based methods, Logistic Regression (LR), Naïve-Bayes (NB), Support Vector Machine (SVM), k-nearest neighbor (k-NN), Artificial Neural Networks (ANN), Linear Classifier (LC) and so forth [1-3]. The comparison of the classifiers and using the most predictive classifier is very important. Each of the classification methods shows different precision and accuracy based on the kind of dataset [4]. In addition, there are various evaluation metrics for comparing the classification methods that each of them could be useful depending on the kind of the problem.

Receiver operating characteristic (ROC) curve [5-10] is a usual criterion for identifying the prediction power of the different classification methods and the area under this curve is one of the important evaluation metrics that can be applied for selecting the best classification method [5-13]. Among the other criteria for comparing classification methods, G-means [14], RMSE [4, 15] and Accuracy [6, 16] can be mentioned.

In this paper, using a new method, three usual methods for data classification (logistic regression (LR), naïve-Bayes (NB) and linear classifier (LC)) have been compared based on the AUC criterion. These mentioned methods have been applied on the random generated datasets, which are independent from a special problem. This comparison is based on the effect of the numbers of existing discrete attributes in dataset and the numbers of their values.

The rest of this paper is organized as follows: In section II, previous works related to this area and the motivations of performing the new work have been presented. Section III provides an explanation about dataset generation and classification methods. Reporting the results of applying classification methods on the datasets and evaluating them are presented in section IV. Finally, section V concludes the paper.

Corresponding Author: Dr. Reza Entezari-Maleki, Department of Computer Engineering, Iran University of Science and Technology (IUST), Tehran, Iran

RELATED WORKS; BACKGROUND AND MOTIVATION

Lots of works related to comparison of classification methods are done. Each of these works compared variant classifiers with each other and regards to the test data and the evaluation criterion, reported the gained results.

Efficiency criterion RMSE has been used by Kim in [4] for comparing DT, ANN and LR. In [4], the effects of the kind of attributes and the size of the dataset have been investigated and the results have been reported. RMSE also has been used by Kumar in [15] for comparing ANN and regression. Regression and ANN have been applied on real and simulated data and the end results have been reported. These results show that if data has error and the real value of the attributes is not available, the statistical method of regression could be act better than the ANN method and its performance is much superior.

J. Huang *et al*. [6] have compared NB, DT and SVM using AUC criterion. In [6], by using the applying mentioned methods on real data; it is shown that the AUC criterion is better than accuracy for comparison of classification methods. Furthermore, it is shown that C4.5 implementation of DT has higher AUC compared to NB and SVM.

J.H. Song *et al*. [7] have compared LR and ANN for breast cancer detection by using the experimental medical data. In [7], it has been shown that LR and ANN almost have the same precision, but in this situation and sensitivity of detection, using ANN compared to LR is prior. In [9], S.M. Rudolfer *et al*. have compared LR and DT and have reported that the precision of the LR and DT methods are the same. Consequently, they have presented a synthesis method which has higher order of precision compared to other previous methods. W.J. Long *et al*. [12] have compared LR and DT in medical application considering AUC criterion. The comparison has been done in this work show that; two mentioned methods have almost the same precision, but in the tasted data in this article, LR partly has more precision compared to DT.

Amor *et al*. [13] have compared DT and NB in intrusion detection systems. This comparison has been done on KDD'99 and the obtained results express that the estimated predictions with NB are better than DT's predictions. Also the same comparison has been done between SVM and ANN classifiers, by W.H. Chen *et al*. in [11]. The reported results show that in considered case, SVM acts better than ANN.

Le Xu *et al*. [14] have compared LR and ANN for finding the source of the error in power distribution by using the G-mean criterion. According to this article, ANN has better results compared to LR and therefore; using neural networks has been proposed. Amendolia *et al*. [16] have compared k-NN, SVM and ANN for talasemi detection by using accuracy criterion. This test has been done for real data and the results obtained from the test show that ANN acts better than the other two methods. B. Karacali *et al*. [17] have compared SVM and KNN methods by using error rate and finally by combining these two methods and by using the power of SVM and simplicity of k-NN have expanded a synthesis classifier which has the advantages of two methods. M. O'Farrell *et al*. [18] have compared k-NN and ANN in classification of spectral data. The results gained from testing show that if values of data have deviation from real values, using ANN is better, otherwise using the simple k-NN classifier is more advised.

All of the mentioned researches have compared different classifiers with each other. The problem which the most of these works engage is that experiment has been done on a special dataset. Since the special datasets are related to the specific problem, the results obtained from experiment are haywire and decision making based on them is not true. Therefore, it makes different observations about the priority of one method to the others. As an example, the reported results in [6, 9, 11-13, 16] are not matched with each other, because the datasets tested in these researches are associated to a special problem.

Furthermore, the most of works which have been done in this field have ignored parameters like; size of the datasets, kind of the attributes and the number of discrete and continuous attributes which affect on the precision of classification methods.

DATA ANALYSIS

In this section, the approach of datasets generation is expressed and then applying of classifiers on the generated datasets is explained.

Random dataset generation: Linear data creation model [4] has been used for generating datasets. Class label, as a linear function of set of the discrete and continuous attributes has been supposed. Class label is calculated from equation (1) for each record i which has n continuous attributes with symbol x and m discrete attributes with symbol c.

Table 1: Properties of datasets having 3 variables

ID	Number of continuous variables	Number of discrete variables
DS_1	3	0
DS_2	2	1
DS_3	1	2
DS_4	0	3

Table 2: Properties of datasets having 5 variables

ID	Number of continuous variables	Number of discrete variables
DS_5	5	0
DS_6	4	1
DS_7	3	2
DS_8	2	3
DS_9	1	4
DS_{10}	0	5

Table 3: Properties of datasets having 7 variables

ID	Number of continuous variables	Number of discrete variables
DS_{11}	7	0
DS_{12}	6	1
DS_{13}	5	2
DS_{14}	4	3
DS_{15}	3	4
DS_{16}	2	5
DS_{17}	1	6
DS_{18}	0	7

$$Y_i = 1 + 3 \times \sum_{j=1}^{n} x_j + 2 \times \sum_{j=1}^{m} c_j \quad (1)$$

which x is a continuous variable and has monotonic distribution in interval [0,1]. Variables c and Y had been continuous and then by using equation (2) they have been categorized and changed to the discrete variables.

$$Y_{Discrete} = Y_{Continuous} \bmod M \quad (2)$$

With regard to above explanation, datasets with different size could be made. These datasets in addition to independency of special problem have capability of variation in discrete and continuous variable numbers and in the number of discrete values. Properties of the datasets which have been generated are in Table 1-3.

As is shown in Table 1-3, datasets DS_1 to DS_{15} have different number of continuous and discrete variables. Also for investigating the effect of the size of dataset on the precision of classifiers, the samples with size of 200, 500, 1000, 3000 and 5000 records have been made from datasets. These numbers have been selected to simulate the datasets with small, medium and large sizes. In each of datasets DS_1 to DS_{18}, the numbers of discrete variable values have been considered to be equal to 2, 5 and 10. These numbers have been selected for investigating the effect of the number of the existing values in discrete attributes on precision of the classifiers. As the following and according to the gained results in the current work, increasing the number of the values in discrete attributes affects the precision of classification methods.

Data classification methods: Linear regression is used to model continuous-valued functions. It is widely used, owing largely to its simplicity. Generalized linear models represent the theoretical foundation on which linear regression can be applied to the modeling of categorical response variables. Common types of generalized linear models include logistic regression and Poisson regression. Logistic regression (LR) models the probability of some event occurring as a linear function of a set of predictor variables. Count data frequently exhibit a Poisson distribution and are commonly modeled using Poisson regression [19]. In this paper, LR has been used as a one of common classification methods for comparing.

Bayesian classifiers are statistical classifiers. They can predict class membership probabilities [19]. Naïve Bayes (NB) probabilistic classifiers are commonly studied in machine learning. The basic idea in NB approaches is to use the joint probabilities of words and categories to estimate the probabilities of categories given a document. The naïve part of NB methods is the assumption of word independence, i.e. the conditional probability of a word given a category is assumed to be independent from the conditional probabilities of other words given that category. This assumption makes the computation of the NB classifiers far more efficient than the exponential complexity of non-naïve Bayes approaches because it does not use word combinations as predictors [20].

Generalized linear models are currently the most frequently applied statistical techniques. They are used to describe the relationship between the trend of one variable and the values taken by several other variables. The relationship that fits a set of data is characterized

by a prediction model called a regression equation. The most widely used form of the regression model is the general linear model formally written as equation (3).

$$Y = \alpha + \beta_1.X_1 + \beta_2.X_2 + \beta_3.X_3 + ... + \beta_n.X_n \quad (3)$$

Applying equation (3) to each of the given samples we obtain a new set of equations i.e. equation (4):

$$y_j = \alpha + \beta_1.x_{1j} + \beta_2.x_{2j} + \beta_3.x_{3j} + ... + \beta_n.x_{nj} + \varepsilon_j \quad \text{and} \quad j = 1,...,m \quad (4)$$

where e_j's are errors of regression for each of m given samples. The linear model is called linear because the expected value of y_j is a linear function: the weighted sum of input values [2].

For using classifiers, two datasets should be used. First dataset for training (training set) and the second one for testing (test set). In this article, Cross Validation method with fold value equal to 10 has been used for training and testing phases. It causes each of the Learners to be trained by 10 stages with 90% of data and to be tested by 10 stages with 10% of data. Consequently, all of data will affect the training and testing of classifiers. For implementing classification methods on dataset, Orange tools and its programming language (python) [21] have been used. Also AUC criterion has been used for comparing the accuracy of the classifiers.

EXPERIMENTAL RESULTS

Classifiers LR, NB and LC have been implemented on datasets DS_1 to DS_{18} and with discrete attribute values equal to 2, 5 and 10. Because of reducing the amount of tables and plots in this paper, only the gained results of datasets with five variables have been reported. Table 4 shows the results of applying the classifiers on datasets DS_5 to DS_{10} when each of the discrete attributes could have 0, 1 values.

Table 5 and 6 show the results of applying the classifiers on the same datasets with discrete attribute values 0, 1, …, 4 and 0, 1,…, 9, respectively.

Because data has been generated randomly, for being sure about the truthfulness of the results, some samples have been generated from each of the datasets. After applying the classifiers on datasets, their averages have been calculated from the gained results. With regard to the Table 4, it could be concluded that the existing fluctuation is high in AUC value for datasets with small sizes. With increasing the size of datasets, or with increasing the number of records in datasets, more stable results can be achieved and the intensity of the fluctuations has been reduced.

Table 4: AUC values for dataset DS_5 to DS_{10} with 2 distinct values for discrete attributes

ID	Sample size	AUC		
		LR	NB	LC
DS_5	200	0.4945	0.5046	0.4875
	500	0.4734	0.4908	0.4698
	1000	0.5353	0.5003	0.5206
	3000	0.5158	0.4905	0.5023
DS_6	200	0.6591	0.6791	0.7071
	500	0.6894	0.6852	0.6825
	1000	0.7035	0.7085	0.7090
	3000	0.6794	0.6886	0.6853
DS_7	200	0.5527	0.5557	0.5434
	500	0.5516	0.5263	0.5932
	1000	0.6876	0.6920	0.6894
	3000	0.5715	0.5784	0.5875
DS_8	200	0.6628	0.6236	0.6661
	500	0.7245	0.7124	0.7265
	1000	0.7408	0.6787	0.7153
	3000	0.7288	0.6794	0.7046
DS_9	200	0.8047	0.6585	0.7266
	500	0.8278	0.7208	0.7403
	1000	0.8970	0.7123	0.7451
	3000	0.9121	0.7411	0.7474
DS_{10}	200	0.9362	0.8316	0.8247
	500	0.9232	0.8068	0.7718
	1000	0.9301	0.7637	0.7892
	3000	0.9530	0.7878	0.7881

Figure 1 and 2 show the reported results in table (4) which has been represented for graphical comparison of the three classifiers. Here, diagrams of the datasets with 1000 and 3000 data have been depicted.

As it can be realized from the table (4) and diagrams (1) and (2), when the ratio of continuous attributes to discrete attributes is high, the AUC of the three methods LR, NB and LC are equal. Gradually, with increasing discrete attributes, the AUC of the two methods NB and LC remains equal, but the AUC of LR increases. It can be concluded that when the ratio of

Table 5: AUC values for dataset DS_5 to DS_{10} with 5 distinct values for discrete attributes

ID	Sample size	AUC		
		LR	NB	LC
DS_5	200	0.4945	0.5046	0.4875
	500	0.4734	0.4908	0.4698
	1000	0.5353	0.5003	0.5206
	3000	0.5158	0.4905	0.5023
DS_6	200	0.7664	0.7185	0.5889
	500	0.7440	0.6914	0.4592
	1000	0.6907	0.6853	0.5168
	3000	0.6980	0.6873	0.5306
DS_7	200	0.7046	0.5834	0.4397
	500	0.6962	0.6792	0.5179
	1000	0.7226	0.6981	0.5379
	3000	0.6936	0.6923	0.5282
DS_8	200	0.7484	0.6325	0.4825
	500	0.7387	0.7024	0.5363
	1000	0.7303	0.6894	0.5345
	3000	0.7190	0.6722	0.5335
DS_9	200	0.8839	0.6450	0.4342
	500	0.9385	0.7234	0.5000
	1000	0.8591	0.6951	0.5202
	3000	0.8673	0.7188	0.5324
DS_{10}	200	0.9789	0.8188	0.5331
	500	0.9740	0.7792	0.5413
	1000	0.9103	0.7341	0.5347
	3000	0.9184	0.7572	0.5480

Table 6: AUC values for dataset DS_5 to DS_{10} with 10 distinct values for discrete attributes

ID	Sample size	AUC		
		LR	NB	LC
DS_5	200	0.4945	0.5046	0.4875
	500	0.4734	0.4908	0.4698
	1000	0.5353	0.5003	0.5206
	3000	0.5158	0.4905	0.5023
DS_6	200	0.7317	0.7087	0.6200
	500	0.7326	0.6844	0.5864
	1000	0.7017	0.6965	0.5585
	3000	0.6821	0.6882	0.5119
DS_7	200	0.6408	0.5582	0.5063
	500	0.5766	0.5378	0.5377
	1000	0.5529	0.5588	0.5323
	3000	0.5509	0.5428	0.5423
DS_8	200	0.7038	0.6326	0.5374
	500	0.7350	0.7096	0.5800
	1000	0.7332	0.6875	0.5828
	3000	0.7334	0.6670	0.5690
DS_9	200	0.8512	0.6486	0.6078
	500	0.6327	0.5830	0.5599
	1000	0.5555	0.5392	0.5535
	3000	0.5245	0.5102	0.5165
DS_{10}	200	0.9683	0.8399	0.5918
	500	0.9544	0.7932	0.6049
	1000	0.9666	0.7829	0.6245
	3000	0.9865	0.8067	0.6094

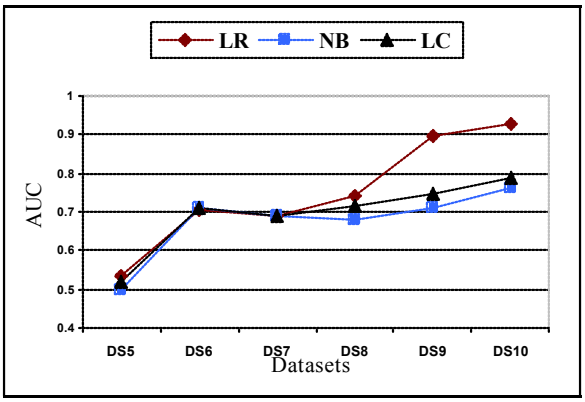

Fig. 1: Classification AUC for datasets with 1000 records shown in Table 4

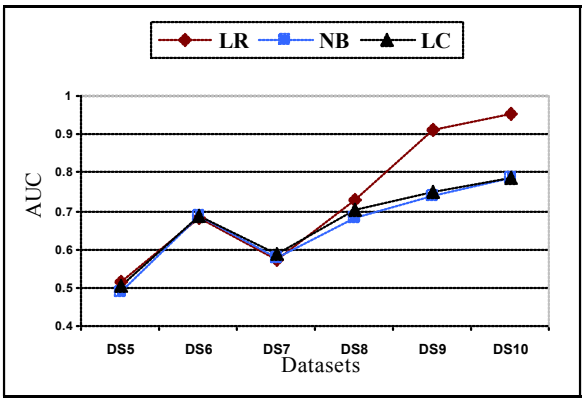

Fig. 2: Classification AUC for datasets with 3000 records shown in Table 4

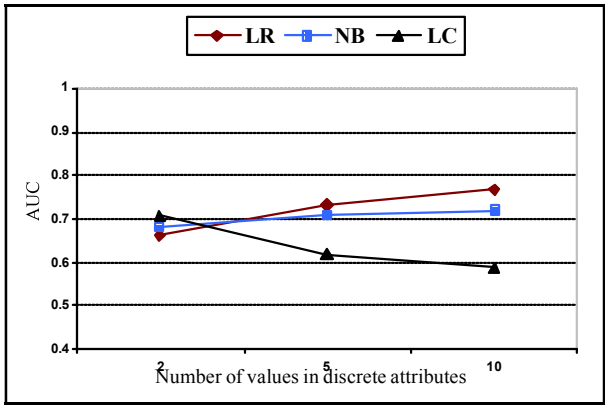

Fig. 3: Classification AUC for datasets with 200 records and having 4 continuous attributes and 1 discrete attribute

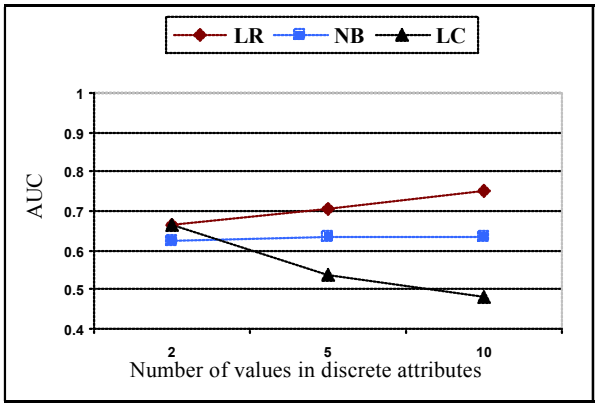

Fig. 5: Classification AUC for datasets with 200 records and having 2 continuous attributes and 3 discrete attributes

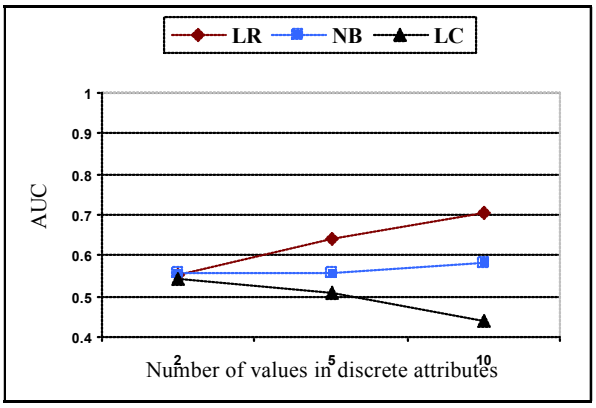

Fig. 4: Classification AUC for datasets with 200 records and having 3 continuous attributes and 2 discrete attributes

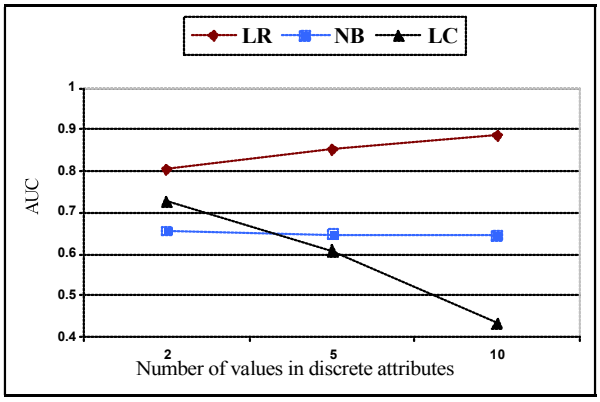

Fig. 6: Classification AUC for datasets with 200 records and having 1 continuous attribute and 4 discrete attributes

discrete attributes rather than continuous attributes is high; among these three methods, LR is the best and its AUC is the largest one. Moreover, the precision of the three methods increase, when the ratio of the discrete attributes increases rather than the continuous attribute.

With drawing diagrams related to Table 4-6 together, the effect of the increment in the values of discrete attributes can be observed. Here, the diagrams of datasets with 200 records have been investigated. Figure 3-7 have compared the three methods LR, NB and LC considering the AUC and the number of values in discrete attributes.

As it has been shown in Fig. 3-7, with increasing the number of the values in discrete attributes from 2 to 5 and 10, the AUC of LR increases and the AUC of the LC decreases, but the AUC of NB method remains constant. These variations occur in datasets which the numbers of continuous attributes are more than the discrete attributes. Also these variations occur for LR with more slopes and for LC with fewer slopes. Gradually with decreasing the number of continuous attributes, the vice versa state occurs and the decrease slope of LC becomes more than the increase slope of LR.

CONCLUSION AND FUTURE WORKS

In this paper, the precision of LR, NB and LC have been investigated and the efficiency of these three methods is compared to each other. In this comparison, the size of datasets, attribute types, the number of discrete and continuous attributes and the number of values in discrete attributes have been considered. The AUC criterion has been calculated for all classifiers. The

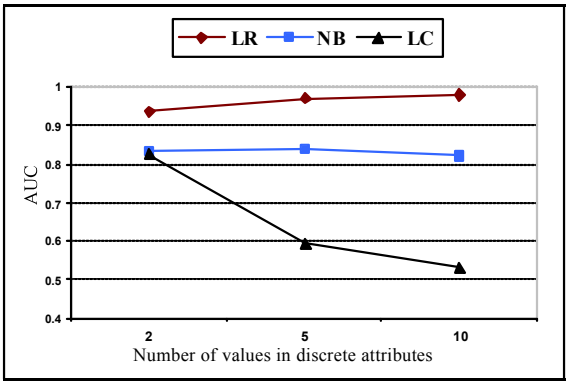

Fig. 7: Classification AUC for datasets with 200 records and having 0 continuous attribute and 5 discrete attributes

above analysis show that if the number of continuous attributes is more than the number of discrete attributes, three methods LR, NB and LC have the same AUC. Increasing the number of discrete attributes increases the AUC of LR more notably compared to other mentioned methods. Also it has been observed that; by increasing the number of discrete attribute values, AUC of LR method increases while the AUC of the LC method decreases and the AUC of NB method remains constant.

Lastly, it should be mentioned that the current research is based on simulation data and the generated datasets which are not dependent to a special problem. Consequently, the above results can be extended to a wide range of problems and these datasets are suitable for comparing the mentioned methods. Following the proposed procedure we can choose the best classifier according to data type and continuous or discrete attributes.

In future studies, it is possible to compare the efficiency of other classifiers by using the current method. Furthermore, using other evaluation criteria and applying new classifiers on datasets with more variables and datasets which contain missing values, could be as open problems in this field.

REFERENCES

1. Tan, P.-N., M. Steinbach and V. Kumar, 2006. Introduction to Data Mining. Addison-Wesley Publishing.
2. Kantardzic, M., 2003. Data Mining: Concepts, Models, Methods and Algorithms. John Wiley and Sons Publishing.
3. Witten, I.H. and E. Frank, 2005. Data Mining: Practical Machine Learning Tools and Techniques. Morgan Kaufmann Publishing, Second Edition.
4. Kim, Y.S., 2008. Comparision of the decision tree, artificial neural network and linear regression methods based on the number and types of independent variables and sample size. Journal of Expert Systems with Application, Elsevier, pp: 1227-1234.
5. Fadlalla, A., 2005. An experimental investigation of the impact of aggregation on the performance of data mining with logistic regression. Journal of Information and Management, Elsevier, pp: 695-707.
6. Huang, J., J. Lu and C.X. Ling, 2003. Comparing Naïve Bayes, Decision Trees and SVM with AUC and Accuracy. Proceedings of the Third IEEE International Conference on Data Mining.
7. Song, J.H., S.S. Venkatesh, E.A. Conant, P.H. Arger and C.M. Sehgal, 2005. Comparative Analysis of Logistic Regression and Artificial Neural Network for Computer-Aided Diagnosis of Breast Messes. Journal of Academic Radiology, 12 (4): 2005.
8. Bradley, A.P., 1997. The Use of the Area Under the ROC Curve in the Evaluation of Machine Learning Algorithms. Journal of Pattern Recognition, 30 (7): 1145-1159.
9. Rudolfer, S.M., 1999. A Comparison of Logistic Regression to Decision Tree induction in the Diagnosis of Carpal Tunnel Syndrome. Journal of Computers and Biomedical Research, 32: 391-414.
10. Hajian-Tilaki, K.O. and J.A. Hanley, 2002. Comparison of Three Methods for Estimating the Standard Error of the Area under the Curve in ROC Analysis of Quantitative Data. Journal of Academic Radiology, Vol: 9 (11).
11. Chen, W.H., S.H. Hsu and H.P. Shen, 2005. Application of SVM and ANN for intrusion detection. Journal of computers & operations research, Elsevier, 32: 2617-2634.
12. Long, W.J., J.L. Griffith, H.P. Selker and R.B. D'Agostino, 1993. A Comparison of logistic regression to decision-tree induction in a medical domain. Journal of Computers and Biomedical Research, 26: 74-97.
13. Amor, N.B., S. Benferhat and Z. Elouedi, 2004. Naïve Bayes vs Decision Trees in Intrusion Detection Systems. ACM Symposium on Applied Computing, Cyprus.

14. Xu, L., M.-Y. Chow and X.Z. Gao, 2005. Comparisons of Logistic Regression and Artificial Neural Network on Power Distribution Systems Fault Cause Identification. IEEE Mid-Summer Workshop on Soft Computing in Industrial Applications, Finland.
15. Kumar, U.A., 2005. Comparison of neural networks and regression analysis: A new insight. Journal of Expert Systems with Applications, 29: 424-430.
16. Amendolia, S.R., G. Cossu, M.L. Ganadu, B. Golosio, G.L. Masala and G.M. Mura, 2003. A comparative study of K-Nearest Neighbour, Support Vector Machine and Multi-Layer Perceptron for Thalassemia screening. Journal of Chemometrics and Intelligent Laboratory Systems, 69: 13-20.
17. Karacali, B., R. Ramanath and W.E. Snyder, 2004. A comparative analysis of structural risk minimization by support vecto machines and nearest neighbor rule. Journal of Pattern Recognition Letters, 25: 63-71.
18. O'Farrell, M., E. Lewis, C. Flanagan, W. Lyons and N. Jackman, 2005. Comparison of k-NN and neural network methods in the classification of spectral data from an optical fiber-based sensors system used for quality control in the food industry. Journal of Sensors and Actuators B, pp: 354-362.
19. Han, J. and M. Kamber, 2006. Data Mining: Concepts and Techniques. Elsevier, Second Edition.
20. Yang, Y. and X. Liu, 1999. A re-examination of text categorization methods. Annual ACM Conference on Research and Development in Information Retrieval, USA, pp: 42-49.
21. http://www.ailab.si/orange/