Classification Of A Bank Data Set On Various Data Mining Platforms

Bir Banka Müşteri Verilerinin Farklı Veri Madenciliği Platformlarında Sınıflandırılması

Muhammet Sinan Başarslan
Computer Programming, School of Advanced Vocational Studies
Doğuş University
Istanbul, Turkey
mbasarslan@dogus.edu.tr

İrem Düzdar Argun
Industry Engineering Department, Engineering Faculty
Düzce University
Düzce, Turkey
iremduzdar@duzce.edu.tr

Abstract— **The process of extracting meaningful rules from big and complex data is called data mining. Data mining has an increasing popularity in every field today. Data units are established in customer-oriented industries such as marketing, finance and telecommunication to work on the customer churn and acquisition, in particular. Among the data mining methods, classification algorithms are used in studies conducted for customer acquisition to predict the potential customers of the company in question in the related industry. In this study, bank marketing data set in UCI Machine Learning Data Set was used by creating models with the same classification algorithms in different data mining programs. Accuracy, precision and f-measure criteria were used to test performances of the classification models. When creating the classification models, the test and training data sets were randomly divided by the holdout method to evaluate the performance of the data set. The data set was divided into training and test data sets with the 60-40%, 75-25% and 80-20% separation ratios. Data mining programs used for these processes are the R, Knime, RapidMiner and WEKA. And, classification algorithms commonly used in these platforms are the k-nearest neighbor (k-nn), Naive Bayes, and C4.5 decision tree.**

Keywords—data mining; banking; customer acquisition; data mining programs.

I. INTRODUCTION

Today, data mining is used in the solution of problems in many fields such as health, finance and education. Data mining studies are being carried out in the field of health for diagnosis of the disease, in customer-oriented industries such as telecommunication, insurance and banking to work on customer churn and customer acquisition. In this research, a forecasting study was carried out to see whether the campaign of a bank results in new customer acquisition. Another purpose of this study is to see the results of the same classification algorithms in different data mining programs. Obtained results were shown in tables in the results section.

There are many classification algorithms continuously developed for various applications in the literature on bank marketing data set. Bach et al. [2] have identified customers who responded positively to the campaigns by performing customer segmentation with a variety of methods such as artificial neural networks. Sumathi and Sivanandam have used data mining methods of financial institutions and other institutions to discover interrelationships between data [3]. Keramati et al. have used decision trees, artificial neural networks, k-nearest neighbors and support vector machines, among the machine learning algorithms, to predict existing customers who would prefer competing banks using a telecommunication company data located in Iran. They have identified the algorithm that gives the best result by comparing the algorithms used in the study [4].

II. OVERVIEW

This section addresses the data mining and classification algorithms and data mining programs used throughout the study.

A. Data Mining

Data mining is the process of extracting meaningful and structures information in the complex data sets. During this procedure, data mining methods such as classification, clustering and association rules are used. Data mining methods are used to analyze, categorize, summarize and determine the relationships using different dimensions of data [5]. These methods are divided into two groups as predictive or descriptive methods [7].

B. Classification Algorithms Used in the Study

In this study, bank marketing data set in UCI Machine Learning Data Set [1] was used. Models were created using classification algorithms on this data set. Classification algorithms used in the study are the k-nearest neighbor (k-nn), Naive Bayes (NB), and C4.5 decision tree. The classification algorithms used are addressed in this section.

978-1-5386-5135-3/18/$31.00 ©2018 IEEE

1) k-nearest neighbor algortihm (k-nn)

It is one of the most basic algorithms of sample-based learning algorithms. In this algorithm learning process is performed with the data in training set. The new samples are classified according to the similarity within the samples in the training set [8]. The k-nearest neighbor algorithm finds k samples that are closest to the unknown data by looking at the pattern space to find which class the unknown data belongs to. Distance is calculated by distance calculation methods such as Euclidean and Manhattan, and distance between neighbors is found. Unknown data are assigned to the class value that most closely resembles the nearest neighbors [9].

2) Naive Bayes algorithm

The Naive Bayes algorithm is named after the English mathematician Thomas Bayes. Bayesian algorithms are among the statistical classification techniques and are based on the statistical Bayesian theorem. Bayes classifier is a predictive model, easier to apply. Naive Bayes is a classification algorithm that shows the relationship between the independent variables and the target variable [10].

Let X = { $x_1, x_2, x_3, \ldots, x_n$ } is the sample set, and $C_1, C_2, C_3, \ldots, C_m$ is the class set.

$$P(X|C_i) = \frac{P(X|C_i)P(C_i)}{P(X)} \quad (1)$$

Probabilities are computed as seen in Eq. (1) for the sample to be classified. The data sample with the highest probability, calculated for each class, belongs to that class [11].

3) C4.5 Decision Tree Algortihm

C4.5 algorithm has been developed by Ross Quinlan. The Gain Ratio is used in the C4.5 decision tree. C4.5 algorithm can work with either categorical or numerical attributes. Decision Trees generated by C4.5 can be used for classification; therefore, C4.5 is generally called a statistical classifier [12].

C. Data Mining Programs

Numerous programs have been developed to implement data mining applications. Commercial programs such as SAS and open source programs such as RapidMiner (YALE), Waikato Environment for Knowledge Analysis (WEKA), R, Konstanz Information Miner (KNIME) can be given as examples of data mining programs developed [13]. In this section, the data mining programs used throughout the study are described briefly.

1) Knime

Konstanz Information Miner (Knime) is a data mining program developed by the Konstanz University data science team [14]. Knime can import data of various file extensions (such as .txt, .arff, .csv) [15].

2) RapidMiner (Yale)

It is a program developed by Ralf Klinkenberg, Ingo Mierswa and Simon Fischer in Artificial Intelligence Unit of Dortmund University of Technology. The Yale program has been developed at Yale University [16]. Yale has been reintroduced in 2007 with the RapidMiner name [17]. It works with 22 different file formats. It supports many databases such as Oracle, MS SQL Server, MySQL, IBM DB2 and text files [14]. It can run on MS Windows, Linux, and Mac OS X operating systems.

3) Weka

Waikato Environment for Knowledge Analysis (Weka) is an open source data mining program developed using Java in Waikato University under the GNU general public license [18]. It accesses the SQL database using Java Database Connectivity (JDBC) [19]. It includes all the data mining and machine learning algorithms. It works on the .arff (Attribute Relationship File Format) file format specially designed for WEKA.

4) R

It is a computer program developed for statistical calculation as well as being a programming language on its own. It has thousands of modules. With these packages, numerous operations can be performed such as data mining and data visualization. The R language, which has been developed by Ross Ihaka and Robert Gentleman in the New Zealand University of Auckland, is continually evolving due to the increased number of packages programmed in accordance with the needs [20]. It has been developed open source as an alternative to the S software [13].

III. Application

In this study, an application was carried out with classification algorithms of data mining methods in order to predict customer acquisition using the bank marketing data set in the UCI database.

A. Data Set

In this study, bank marketing data set in UCI Machine Learning Data Set [1] was used by establishing models with the same classification algorithms in different data mining programs. The bank marketing data set contains 17 attributes and 45211 customer records. Table 1 shows the data type and description of the attributes.

TABLE I. Bank Data Set

No	Attributes	Explanation of Attributes	Data Type
1	age	Customer's age	Numeric
2	job	Business status of the customers	Nominal
3	marital	Customer's marital status	Nominal
4	education	Customer's educational status	Nominal
5	default	Credit debt situation?	Nominal
6	balance	Average annual balance	Numeric
7	housing	Real estate debt situation?	Nominal
8	loan	Personal debt situation?	Nominal
9	contact	Type of communication	Nominal
10	day	Last interview day	Numeric
11	month	Last interview month	Numeric
12	duration	Last call duration	Numeric
13	campaign	Number of customers searching for the campaign during the campaign	Numeric
14	pdays	The number of times the customer has been called since the previous campaign	Numeric
15	Previous	How many times the customer is called before the campaign	Numeric
16	Poutcome	The end of the previous marketing campaign	Nominal
17	Customer	Is the customer a bank customer?	Nominal

B. *Model Performance Evaluation Criteria*

Evaluation of the model created by classification algorithms is carried out by various methods. One of these methods is the confusion matrix [21]. The actual values and the values predicted by the classification algorithm are shown in Table 2. Performance evaluation criteria of classification algorithms are shown in Table 2 below [21-23].

TABLE II. CONFUSION MATRIX

		Prediction	
		True	***False***
Actual	***True***	TT	TF
	False	FT	FF

Accuracy and error value of the model generated by the classification algorithms according to Table 2 are given by Eq. (2) and Eq. (3), respectively [22].

$$Accuracy = \frac{TT + \mathrm{FF}}{TT + TF + FT + FF} \tag{2}$$

$$Error\ = 1 - \mathrm{Accuracy} \tag{3}$$

Precision and sensitivity values of the model generated by the classification algorithms according to Table 2 are given by Eq. (4) and Eq. (5), respectively [22].

$$Precision\ = \frac{DD}{DD + YD} \tag{4}$$

$$Sensitivity\ = \frac{TT}{TT + TF} \tag{5}$$

Specificity and F-measure values of the model generated by the classification algorithms according to Table 2 are given by Eq. (6) and Eq. (7), respectively [22].

$$Specificity\ \ = \frac{YY}{YY + YD} \tag{6}$$

$$F - measure = \frac{2 \times Sensitivity \times Precision}{Sensitivity + recision} \tag{7}$$

Models are created using classification algorithms to make estimations on the data set. To see the performances of the classification models, the classification models are divided into training and test data. Various methods have been developed for this splitting process. Among these methods, the holdout method was used in this study. In the hold out separation, the test and the training data sets are divided once with a specific ratio. Figure 1 shows the flow of this method.

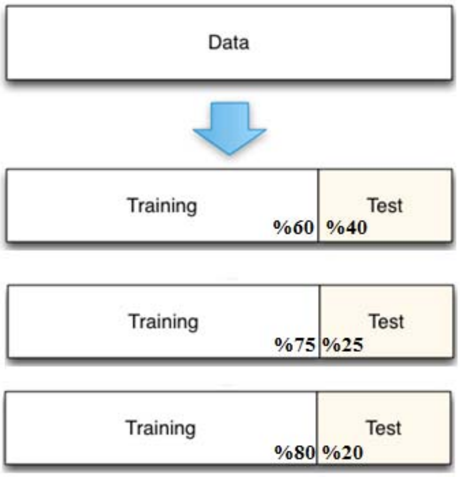

Fig. 1. Test and training set separation with the hold out method.

IV. RESULTS AND CONCLUSION

R, Knime, Weka and RapidMiner data mining programs were used on the bank marketing data set. Models were established with the k-nearest neighbor, C4.5 Decision Tree and Bayes classification algorithms commonly present in these programs. The performance of these models was evaluated with accuracy, precision and f-measure criteria. Within the scope of the study, the training and test sets were compared with 60%-40%, 75%-25%, 80%-20% and 90%-10% separations in each data mining program to test the performances of all the models applied. These splits are shown in Table 3, Table 4 and Table 5, respectively.

TABLE III. RESULTS OBTAINED IN R, KNIME, RAPIDMINER AND WEKA PROGRAMS WITH 60% HOLDOUT SEPARATION.

Criteria	**Accuracy**			**Precision**			**F-measure**		
Algorithms	***NB***	***k-nn***	***C4.5***	***NB***	***k-nn***	***C4.5***	***NB***	***k-nn***	***C4.5***
R	0.872	0.871	**0.904**	0.930	0.909	**0.933**	0.927	0.928	**0.946**
Knime	0.866	0.860	**0.902**	0.921	0.867	**0.938**	0.825	0.900	**0,935**
RapidMiner	0.861	0.846	**0.885**	0.916	0.857	**0.932**	0.880	0.890	**0,918**
Weka	0.881	0.864	**0.900**	0.936	0.913	**0.940**	0.933	0.924	**0.944**

In Table 3, Table 4 and Table 5, the same results were obtained for performance evaluation. In all three of the performance criteria, the best-performing was the C4.5 decision tree. In addition, the Weka program gave better result in the precision criterion, whereas the R program gave better results in the other three criteria.

TABLE IV. RESULTS OBTAINED IN R, KNIME, RAPIDMINER AND WEKA PROGRAMS WITH 75% HOLDOUT SEPARATION.

Criteria	**Accuracy**			**Precision**			**F-measure**		
Algorithms	***NB***	***k-nn***	***C4.5***	***NB***	***k-nn***	***C4.5***	***NB***	***k-nn***	***C4.5***
R	0.875	0.875	**0.906**	0.93	0.91	**0.935**	0.92	0.93	**0.948**
Knime	0.842	0.870	**0.905**	0.88	0.88	**0.933**	0.88	0.87	**0.90**
RapidMiner	0.869	0.842	**0.885**	0.88	0.88	**0.930**	0.86	0.86	**0.93**
Weka	0.882	0.865	**0.902**	0.93	0.91	**0.937**	0.93	0.92	**0.947**

Table 4 gives the accuracy, precision and F-measure performance criteria obtained with 75% training and 25% test data set in four different data mining programs (R, Knime, RapidMiner, Weka).

TABLE V. RESULTS OBTAINED IN R, KNIME, RAPIDMINER AND WEKA PROGRAMS WITH 80% HOLDOUT SEPARATION.

Criteria	**Accuracy**			**Precision**			**F-measure**		
Algorithms	***NB***	***k-nn***	***C4.5***	***NB***	***k-nn***	***C4.5***	***NB***	***k-nn***	***C4.5***
R	0.875	0.87	**0.906**	0.934	0.911	**0.935**	0.928	0.93	**0.947**
Knime	0.842	0.87	**0.905**	0.882	0.880	**0.933**	0.883	0.87	**0.90**
RapidMiner	0.869	0.84	**0.88**	0.880	0.882	**0.930**	0.865	0.86	**0.93**
Weka	0.882	0.86	**0.902**	0.931	0.913	**0.937**	0.933	0.92	**0.946**

Table 5 gives the accuracy, precision and F-measure performance criteria obtained with 80% training and 20% test data set in four different data mining programs (R, Knime, RapidMiner, Weka).

In Table 3, Table 4 and Table 5, the same results were obtained for performance evaluation. In all three of the performance criteria, the best-performing was the C4.5 decision tree. In addition, the Weka program gave better result in the precision criterion, and the R program gave better results in the other two criteria.

V. DICCUSSION

In this study, the performances of different data mining programs were examined by establishing models with classification algorithms. Different results were obtained in the four programs used. However, the algorithm that gives the best result in all programs was the decision tree algorithm. This result suggests that decision tree method gives better performance regardless of the program used. Further studies are also needed to support this result by working with data other than the bank data set. This is a subject of another research to be further investigated in the future.

REFERENCES

[1] S. Moro, P. Cortez, and P. Rita, “A data-driven approach to predict the success of bank telemarketing,” *Decision Support Systems*, vol. 62, s. 22-31, 2014

[2] M.P. Bach, S. Juković, K. Dumiči, and N. Šarlija, “Business client segmentation in banking using self-organizing maps,” *South East European Journal of Economics and Business*, vol. 8, no. 2, s. 32-41, 2013.

[3] S. Sumathi, S. Sivanandam, Introduction to Data Mining and Its Applications, *Springer Science & Business Media,* 2006.

[4] A. Keramati, R. Jafari-Marandi, M. Aliannejadi, I. Ahmadian, M. Mozaffari, and U. Abbasi, “Improved churn prediction in telecommunication industry using data techniques,” *Applied Soft Computing,* vol. 24, s. 994-1012, 2014.

[5] M.S. Başarslan, F. Kayaalp, “Customer churn analysis with classification algorithms in telecommunication sector. ICAT’17, Istanbul, Turkey, 2017

[6] J. Han, M. Kanber, Data Mining: Concepts and Techniques, *Morgan Kaufmann*, 2006

[7] S. Akyokuş, “Veri Madenciligi Yöntemlerine Genel Bakış,” TBD Veri Madenciliği Günü sunumu, Doğuş Üniversitesi, 2006

[8] T. Mitchell, Machine Learning, *McGraw Hill*, New York, 1997.

[9] P. Harrington, Machine Learning In Action, *Manning*, New York 2012.

[10] H. Arslan, “Sakarya üniversitesi web sitesi erişim kayıtlarının web madenciliği ile analizi,” Yüksek lisans tezi, Elektronik-Bilgisayar Eğitimi, Sakarya Üniversitesi, Sakarya, Türkiye, 2008.

[11] J. Han, M. Kamber, Data Mining: Concepts and Techniques, Morgan Kaufmann Publisher, San Francisco, USA, 2000.

[12] I.H. Witten, E. Frank, Mark Data Mining, Morgan Kaufmann, Elsevier, San Francisco, 2005.

[13] M.S. Başarslan, F. Kayaalp, “Telekomünikasyon Sektöründe Müşteri Kayıp Analizi,” Bilgisayar Mühendisliği Ana Bilim Dalı, Düzce Üniversitesi, 2017.

[14] KNIME, Bilgisayar Programı, Konstanz Üniversitesi, Zürih Teknopark, 2004.

[15] T.T. Bilgin, “Veri akışı diyagramları tabanlı veri madenciliği araçları ve yazılım geliştirme ortamları,” Akademik Bilişim’09 Konferansı, Şanlıurfa, Türkiye, 2009.

[16] YALE, Bilgisayar Programı, Yale Üniversitesi, 2001.

[17] RAPIDMINER, Bilgisayar Programı, Dormunt Teknoloji Üniversitesi Yapay Zeka Birimi, 2006.

[18] M. Dener, M. dörterler, ve A. Orman, “Açık kaynak kodlu veri madenciliği programları: Weka’da örnek uygulama,” Akademik Bilişim’09 Konferansı, Şanlıurfa, Türkiye, 2009.

[19] K. Dahiya, S. Bhatia, “Customer churn analysis in telecom industry,” Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions), 2015 4th International Conference on IEEE, Noida, India, s. 1-6, 2015.

[20] A. Demirci. (2015, 28 Eylül). Data Driven Kavramı [Online]. Erişim: http://devveri.com/kategori/haberler.

[21] N. Japkowicz, “Performance evaluation for learning algorithms,” International Conference on Machine Learning, Edinburg, Scotland, 2012.

[22] M. Clark, “An Introduction to machine learning with Applications in R,” Lecture Notes, University of Notre Dame, 2015.

[23] P. Flach, “The many faces of ROC analysis in machine learning,” Lecture Notes, University of Bristol, 2004.