ORIGINAL RESEARCH

Comparison of three data mining algorithms for potential 4G customers prediction

Chun Gui*, Qiang Lin

College of mathematics and computer science, Northwest University for Nationalities, Lanzhou, China

Received: August 1, 2016 **Accepted:** September 26, 2016 **Online Published:** October 7, 2016
DOI: 10.5430/air.v6n1p52 **URL:** http://dx.doi.org/10.5430/air.v6n1p52

ABSTRACT

The size and number of telecom databases are growing quickly but most of the data has not been analyzed for revealing the hidden and valuable intellectual. Models developed from data mining techniques are useful for telecom to make right prediction. The dataset contains one million customers from a telecom company. We implement data mining techniques, *i.e.*, AdaboostM1 (ABM) algorithm, Naïve Bayes (NB) algorithm, Local Outlier Factor (LOF) algorithm to develop the predictive models. This paper studies the application of data mining techniques to develop 4G customer predictive models and compares three models on our dataset through precision, recall, and cumulative recall curve. The result is that precision of ABM, NB and LOF are 0.6016, 0.6735 and 0.3844. From the aspects of cumulative recall curve NB algorithm also is the best one.

Key Words: Data mining t algorithms, Predict 4G customers, Local outlier factor algorithm, AdaboostM1 algorithm, Naïve Bayes algorithm

1. INTRODUCTION

Many advanced data mining techniques are being used to reveal the hidden patterns and relationships with the rapidly increase of the telecom databases. The composite models come from these data mining techniques are useful for telecom industry to make right prediction or decision. For telecom industry customers play the most important role. Currently, telecom operators are facing increasing competition for 4G customers in China, therefore, prediction 4G customers is necessary and meaningful. This paper studies the application of data mining techniques to develop 4G customer predictive models and compares three models on our dataset through precision, recall, and cumulative recall curve.

At present, the number of articles on this subject is relatively small. We put forward the method of data mining applied to 4G customer forecast. Rahman *et al.* applied data mining techniques to the diagnosis of diabetes.[1] Ahmad LG *et al.* using data mining technology applied to prediction breast cancer recurrence.

In this paper 4G customer prediction problem was taken as the outlier detection problem. Those outliers may have particularly high APRU (Average Revenue Per User) value, high Flow or special advanced cell phone terminal. The goal is to help on verifying the outliers of these customers on the basis of the past experience and users' characteristics of the company. The output takes a ranking of the customers on the basis of the possibility of the potential 4G customers ranking. We required the models to get a score of outlier score for every observed value of the test set. The value of the score should between 1 and 0. The higher the score is, the higher it is in the ranking list, and the implication is that the observation is a potential 4G customer.

*Correspondence: Chun Gui; Email: guichun2103@163.com; Address: College of mathematics and computer science, Northwest University for Nationalities, Lanzhou 730000, China.

Our experiment is performed on one million customers based on the software R.[2–4] First, random forest was used to obtain the importance rank of the variables. Secondly, performance of three data mining algorithms was compared. The experiment result is that precision of AdaboostM1 (ABM) algorithm, Naïve Bayes (NB) algorithm and Local Outlier Factor (LOF) algorithm are 0.6016, 0.6735 and 0.3844. The results are achieved using hold out method for measuring the prediction precision of each model.

2. THE DATASET

This section introduces the dataset and data preprocessing. First, we present the data set we used. Secondly, we explore the distribution of the dataset, especially the usage and distribution of ARPU/Flow. Thirdly, data preprocessing will be introduced.

2.1 The dataset

The data set comes from a telecom company and has been anonymized. It is about users' consumption characteristics in a given area net in six months. The data set has already been gone through some preprocessing at the company. The data set includes one million customers, each row of the data table contains information about every customer. Table 1 is about the customers' information consists of 12 attributes. As we can see (see Table 1), the dataset we use has the following columns:

Table 1. Attribute of the data set

NO.	Variable name	Definition
1	ID	A factor with a number of the customer
2	Flow	The flow used by the customer
3	ARPU	Average Revenue Per User
4	Type	The type of the customers' mobile phone (1-9)
5	Realname	Whether the real-name registration
6	Changephone	Whether the customer change his mobile phone in six months
7	Duration	How long the customer be in the net (1-6)
8	Guaranteefee	The guaranteed cost each month
9	Orderflow	Flow of order each month
10	Overflow	Beyond the flow of each month
11	Time	The total call time
12	Package	Package type used by the customer

2.2 Exploring the dataset

The data set consists of 12 attributes about one million customers' information. In this part we focus on the distribution of ARPU and Flow. Figure 1 shows the ARPU consumed by each customer. As you can see, the numbers are rather diverse across the customers. Figure 2 shows the flow used by each customer. Once again, we can observe strong variability and the widely range.

From the above two figures we can see that the data set has a lot of volatility. We can implement the experiment to find the customers who are the ones to produce less or more ARPU to the telecom company by outputs the top five maximum/minimum ARPU. It is significative to find the top one hundred customers on the list occupy nearly 40% (0.3995) of the ARPU of the telecom company, while the bottom ten thousand customers contribute less than 7% (0.065) of the income. This result shows that the data distribution and its imbalance, this can supply some insight into ultimate variation that needs to be implemented within the company.

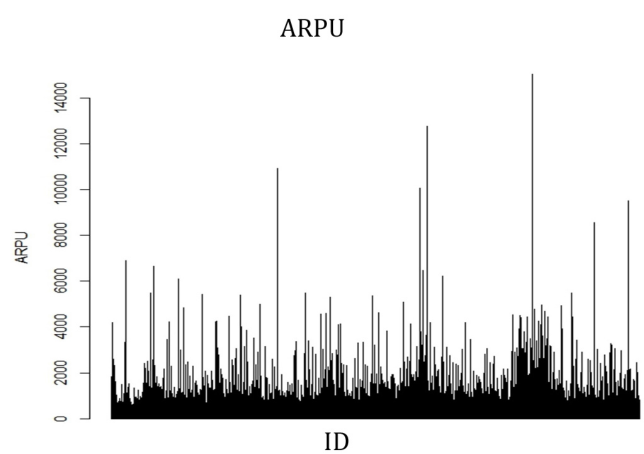

Figure 1. The ARPU consumed by each customer

2.3 Data preprocessing

In data preprocessing, handling of missing value is an important task, there are three common ways to handle missing value: 1) remove these few cases, 2) use some strategies to fill the missing value, 3) to process the missing values with the aid of tools.[5] In this paper about the missing values, we choose to remove these few cases. In our dataset only a few data has the missing values problem, and delete those a few data has no effect on the experimental results.

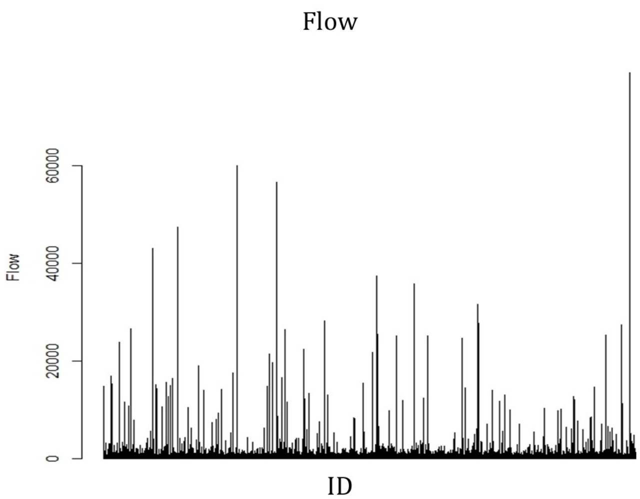

Figure 2. The Flow used by each customer

If a customer frequently changes the mobile phone, we consider him as a fanatic of fashionable digital products pursuer, who may take the lead in change and use 4G mobile phone, therefore ,we regard him as the potential 4G customer. On the opposite, we believe those people who do not change their handset frequently have little probability to become 4G customer and then delete them. In other words, customer with Changephone = 0 will be removed from the dataset. In addition, those users using unreal name are fraud users, they most likely to be with bad motivation. Only these customers who uses real name are eligible for 4G users, so we also delete those users whose Realname = 0.

For these two attributes, Package and Type, we need to do is make them become numerical value, according to the cost of the packages or the release time of the type of the phone.

In our telecom dataset, ARPU is an important attribute for predicting 4G customers. If a user's ARPU is zero, considering from the perspective of telecom operators, the users do not bring any benefits for company. We think that he/she will never become a 4G customer. So we delete those customers whose ARPU = 0. In the same way, faster network speed and high quality call are two important factors to attract people using 4G. 4G customers could not have such features: usage in flow is zero or time on the phone call is zero. So we delete those uses whose Flow = 0 or Time = 0.

Generally, we make prediction is mainly to individual customer forecast, Flow or ARPU is very big generally considered the group customers, do not belong to the data of the experiment we need. therefore, we remove those very customer whose ARPU or Flow is greater than 1,000.

3. METHODS

In this section we explain the methods we used in our experiment, first, we used random forest to calculate the variable importance. According to the result we get 4G customers marked as ok and most unlikely 4G customers marked as bad (ok and bad are two categories of 4G customer). Secondly, we used one unsupervised algorithm and two supervised classification algorithms to acquire the ranking report. These three algorithms are LOF algorithm, NB algorithm and ABM algorithm.

3.1 Random forest

Random forest[6] is used to calculate the variable importance in telecom industry. Before using the random forest, we first use the expectation maximization cluster algorithm to cluster all the data, thus we acquire three clusters. Secondly, we use random forest to select the features in our dataset. We provide an original set of attributes and estimate the importance of all features by using a technique.

In order to use supervised classification approaches, we want to add an attribute called Tag, to identify whether the customer is the 4G customer. Tag has three values: ok, unkn, bad. So on the basis of how to label customer, we need a feature called SUM.

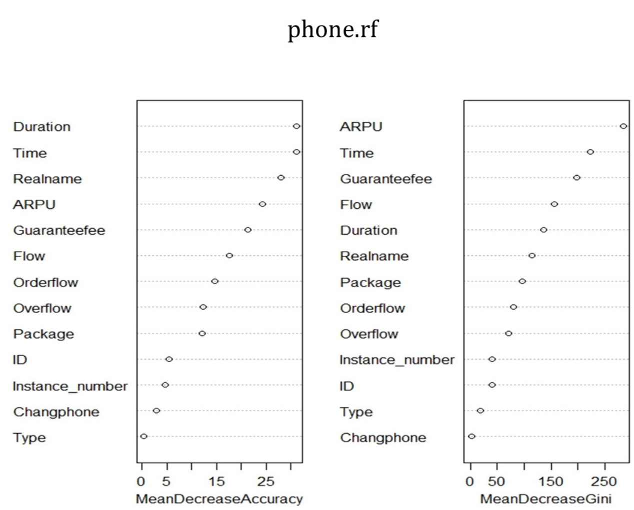

Figure 3. Importance of variables calculated by random forest

Figure 3 is the output about attributes' importance ranking using random forest. We use accuracy coefficient and Gini coefficient to rank variables. Table 2 is the importance scoring of each attribute, we calculate mean percentage of each attribute according to the accuracy coefficient and Gini coefficient.

In this paper according to the mean percentage of each attributes in Table 2, we get the following formula 1 to calculate SUM. After we acquire SUM, all customers are in descending order according to SUM values. We mark the Tag according to SUM, one part customers whose SUM greater than 80 are marked ok, we think that these people are 4G users; another part whose SUM less than 30 are marked as bad, we think that these people never become 4G users; the rest customers is marked as unkn.

Table 2. The importance scoring of each attributes from Random Forest

Attributes	Accuracy coefficient	Gini coefficient	Mean percentage
Duration	31.152209	139.917827	0.155776338
Flow	14.046464	144.63575	0.102832053
Orderflow	12.758613	83.61098	0.074313588
Overflow	10.847556	73.345135	0.064084911
ARPU	23.63952	271.077703	0.184117645
Package	12.939335	100.574236	0.081672759
Type	3.112655	21.364165	0.018516011
Guarenteefee	20.003883	201.504234	0.144656943
Time	27.488828	214.967263	0.174029752

$$\begin{aligned} &SUM = 18\%ARPU + 17\%Time + 16\%Duration \\ &+ 14\%Guarenteefee + 10\%Flow + 8\%Package \\ &+ 7\%Orderflow + 6\%Overflow + 2\%Type \end{aligned} \tag{1}$$

3.2 Local outlier factor

There are some important applications of outlier detection, such as in the field of intrusion detection, fraud detection, and the robustness analysis in networks.[7] There are lots of definitions for outliers, the most common is proposed by Hawkins in 1980.[8] An outlier is a value which departs from other values so much to lead to doubts that it is produced by other mechanism. This is the most standard definition. In recent years, researchers proposed different outlier detection methods.[9–12] The traditional outlier methods are generally based on the calculation of distance. This paper uses LOF algorithm which is based on the density of local outlier detection algorithm. Because based on statistics or distance outlier detection algorithm based on a given set of data set global distribution. But the data are generally not homogeneous distributed. When analyzing the data whose density vary greatly, local outlier detection method based on the density shows satisfactory ability to be used to identify the local outliers. Breunig *et al.* proposed the LOF method which was usually considered as the most advanced outlier ranking method.[8] LOF method estimates the degree of the separate case according to its local neighborhoods to get an outlying score for each case. The method is on account of the local density of the observed values. Package DMwR includes the implementation of the LOF method. We could use this function directly. Then we use LOF algorithm to rank this data set. The goal of the function SoftMax() is to change the outlier factors into a range of 1 to 0 scale. The final step is to apply one hold-out procedure to get the approximate of our evaluation measures. Besides, this algorithm has a high requirement on the computing resources and may take a long time running.

3.3 Naïve Bayes

Considering the goal is to get a ranking of the reports, we need to limit the choice of models. We use the systems which can only produce probabilistic classifications. NB is a good probabilistic classifier which uses strong hypothesis about the independence between predictors. Nevertheless, NB model shows high precision and high efficiency, with minimum classification error rate, consuming less time. So NB is applied to many real life applications successfully. The Bayes formula is as following:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{2}$$

For a given test set, the probability of each class can be calculated by Naive Bayes classifier as the following equation:

$$P(c \mid X_1, \cdots, X_p) = \frac{P(c)P(X_1, \cdots, X_p \mid c)}{P(X_1, \cdots X_p)} \tag{3}$$

where c is one of the class, $X_1, \cdots, X_p$ are the observations of the predictors for the given test data.

$P(c)$ is the previous prospection of class c. $P(X_1, \cdots, X_p \mid c)$ is the probability of the test case given the class c. $P(X_1, \cdots X_p)$ is the possibility of the observed value. By using a few statistical methods on conditional probabilities and hypothesis conditional independence among predictors, the numerator of the fraction is reduced to:

$$P(c)P(X_1, \cdots, X_p \mid c) = P(c) \prod_{i=1}^{p} P(X_i \mid c) \quad (4)$$

There are some NB methods in R implementations, *e.g.* the function naiveBayes() from package e1071.

3.4 AdaBoostM1

AdaBoost is a learning method which belongs to the ensemble model class.[9] AdaBoost obtains the base models by using the adaptive boosting method. It is susceptible to outliers and noisy data, but is less sensitive to the overfitting problem.

A special instantiation of the AdaBoost method is ABM.[13] It is used as base learner classification trees with a fewer number of vertices. For ABM algorithm, this is assigned with a weight for every training samples, the size of the weight represents the probability of a sample was chosen to be the next weak classifier's training sample set. If a sample can be accurately classified by the current weak classifier, when constructing next weak classifier's training sample set, the probability of the sample will be chosen is very low; On the contrary, if a sample could not be the current classifier classified correctly, its weigh increased accordingly.

ABM method is carried out in function AdaboostM1() of the extra package adabag.[14] However, the predict method cannot return class probabilities which is a fearful limit to application. Fortunately, packages of RWeka provide the function AdaBoostM1() of this algorithm. The predict method can output a probabilistic classification, therefore, in this paper we used the packages RWeka.

4. Experiment and analysis

In this section, first we talk about how to evaluate the models. Secondly, we described the experimental methods which will be used to get the dependable estimates of the selected evaluation measures. In this paper we chose the holdout methodology. A detailed introduction will be introduced about this methodology in the following sections. At last we compared the experimental results from different aspects.

4.1 Evaluation criteria

In this paper we used precision and recall, PRcure and cumulative recall chart to compare three models. A good model should get a ranking which contains as many as possible known oks at the top positions of the ranking in the application. The ok values are a minority in the experimental data set. As we all known, when the goal is to predict a series sets of scarce events (in our case is oks), recall and precision are appropriate for the evaluation of the indicators. The recall and precision the k top positions of the ranking can be calculated by the given inspection effort limit k. The proportion of the k top values can be obtained by the value of precision, which are labeled as oks. The value of recall is a good measure to obtain the scale of oks cases in the test set which are contained in the k top ones.[15]

Precision is the value of correctly classified ok customers in all customers which are classified as selective by a model, and can be defined as:

$$Precision = \frac{true\ positive}{true\ positive + false\ positive} \quad (5)$$

Recall is the fraction of correctly classified ok customers over all ok customers in the testing data by a model, and it can be defined as:

$$Recall = \frac{true\ positive}{true\ positive + false\ negative} \quad (6)$$

PR (means Precision/recall) curve is a visual representation of the property for a model according to the precision and recall values. By appropriate insertion of the values of the statistics at diverse working points we can obtain the PR curves. In the paper this working points correspond to different effort limits which are used to the ranking given by the algorithms. The different values of the precision and recall can be obtained by iterating over different limits. In this paper we chose to inspect 10% reports. Each model iterated three times using Hold Out experiment Methodology.

Cumulative recall chart is a more interesting graph. It can be carried out by using the ROCR package.[16] For cumulative recall charts, the model is the better when the curve more near to the left top corner on the graph. The function of CRchart() is also obtained by the DMwR package.

4.2 Hold out methodology

As mentioned before, our data set is imbalance among the distributions of customers. A stratified sampling method is recommended to adopt to solve the imbalanced class distributions problem. The effect of the Hold out method is similar to the Monte Carlo and cross-validation experiments. But the reason why we choose holdout method is that we can be specified using stratified sampling strategy by setting the parameters in the holdout() function. There is a function called holdout() in the DMwR package.[13] This function is used to run hold out experiment.

Hold Out method can randomly split the dataset into two partitions (about 30%-70% proportions). One of the partitions is used to obtain the models, while the other one is used to test. The process can be repeated several times to make certain the reliability. In our experimental we choose repeat three time.

4.3 Experimental results

In this paper we compare three models to predict 4G customers used a telecom dataset. Table 3 shows the average precision and recall for three different data mining models. Figure 4 compares the three data mining models in the form of a histogram.

Table 3. Average precision and recall of three data mining models

	Local outlier factor	Naïve Bayes	AdaboostM1
Precision	0.3844	0.6735	0.6016
Recall	0.1921	0.3366	0.3007

According to Table 3 and Figure 4, we can see that the highest precision is 67.35% belongs to NB algorithm and lowest precision is 38.4% belongs to LOF algorithm. Besides, the highest recall is 33.66% belongs to NB algorithm. Our results show that supervised NB algorithm outperforms both LOF algorithm and ABM algorithm in the parameters of precision and recall. NB is the best predictor of 4G customers.

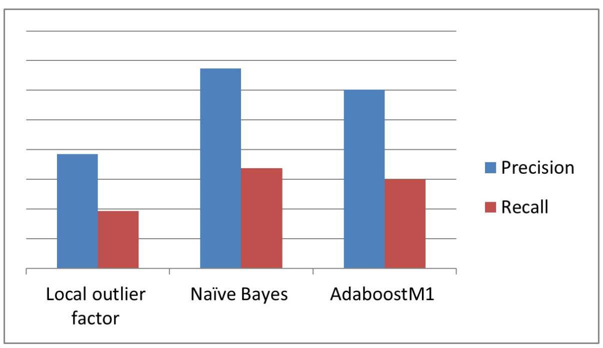

Figure 4. Comparison of three data mining models

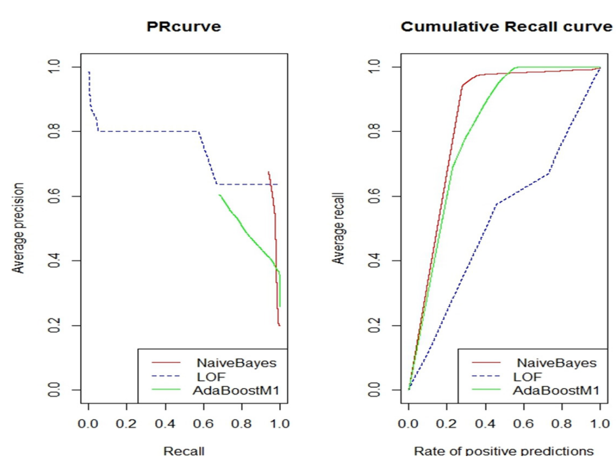

Figure 5. The PR and cumulative recall curves for the three models

By the PR and cumulative recall curves can obtain a more global perspective. In order to get a better comparison among the three methods, the curves of those methods are plotted as Figure 5.

The left of Figure 5 of PR curves shows that for smaller recall values, the LOF generally reaches a very lower precision. For values of recall above 90%, NB becomes much better. According to recall reached by inspection effort (right of Figure 5). Obviously, LOF algorithm is also the worst method. But we can say that generally the NB method dominates the Adaboost for inspection efforts below 50%.

Random forest was used to obtain the importance rank of the variables firstly, and then the performance of three data mining algorithms was compared in order to predict the potential customers. The experiment result shows that precision of ABM algorithm, NB algorithm and LOF algorithm are 0.6016, 0.6735 and 0.3844. NB algorithm is much better. NB model needs few parameters and less sensitive to missing data, and it is relatively simple. So it is more suitable for the analysis of telecom datasets. As the benefits of the companies are clearly on fewer efforts to reduce costs, we say that the NB algorithm is much better. In fact, with the effort about 20% to 50%, one can retain around 75%-98% of the 4G customers.

4.4 Experimental analysis

In this paper, we put forward applying data mining technology to solve customer forecast problem for the first time. According to some indicators, the final output result is a potential 4G customer ranking report. The experiment combined the ID, Flow, ARPU, Type, Realname, Changephone, Duration, Guaranteefee, Orderfolw, Time, Package (The specific meanings are in Table 1) seven attributes for 4G customers to do the churn rate prediction, the guidance for telecommunication industry development was provided. On the other hand, the obtained results allow us to conclude that is needed to analysis the topic in more depth about the following challenges:

(1) It is needed to test these algorithms in other data set of different domains, such as medical diagnostics, fault detection, property refinance prediction and intrusion detection and so on.
(2) It is necessary to analysis big data situation, such as flow compute, MapReduce framework. When faced with the flow imbalanced data set, how should we deal with through combined models.
(3) It is obvious that the datasets of telecom customers are usually imbalanced. The traditional data mining approaches are not able to cope with the new requirements imposed by imbalanced data set. It is necessary to make the prediction with combination models to resolve the inherent defects of imbalanced data set.

In future study, it is necessary to study in span and depth the described challenges so that the companies can spend minimum cost to get maximum income. Furthermore, it is also advisable to improve the current approaches considering the big sample size problem so that the telecommunications to meet the challenges of the era of big data.

ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China (No. 61562075), the Fundamental Research Funds for the Central Universities 2014 (No. 31920140089), the Natural Science Foundation of Gansu Province (No. 1506RJZA269), the Fundamental Research Funds for the Gansu Universities (No. 2015B-02) and the Fundamental Research Funds for the Central Universities (No. 31920140058).

REFERENCES

[1] Rahman RM, Afroz F. Comparison of Various Classification Techniques Using different Data Mining Tools for Diabetes Diagnosis. Journal of Software Engineering and Application. 2013; 06: 85-97. http://dx.doi.org/10.4236/jsea.2013.63013

[2] Kabacoff R. R in action. Manning Publications. 2015; 6: 1-608.

[3] James G, Witten D, Hastie T, et al. An introduction to statistical learning: with applications in R. Springer-Verlag New York. 2014; 19(4): 419-20.

[4] Bell J. Machine learning with R. John Wiley & Sons, Inc. 2015: 315-48. PMid:25303537. http://dx.doi.org/10.1002/9781119183464.ch12

[5] Torgo L. Data Mining with R: Learning with Case Studies. Chapman & Hall/CRC. 2010: 189-92. http://dx.doi.org/10.1201/b10328

[6] Breiman L. Random forests. Machine Learning 2001; 45: 5-32. http://dx.doi.org/10.1023/A:1010933404324

[7] Aggarwal CC, Yu PS. Outlier Detection for High Dimensional Data. Acm Sigmod Record. 2002; 30(5): 37-46.

[8] Hawkins DM. Identification of Outliers. Springer. Chapman & Hall. 1980: 1-188. PMid:10297823. http://dx.doi.org/10.1007/978-94-015-3994-4_1

[9] Breuning MM, Kriegel HP, Ng RT, et al. LOF: Identifying Density-Based Local Outliers. ACM SIGMOD Conference Proceedings. 2000; 29(2): 93-104. http://dx.doi.org/10.1145/335191.335388

[10] Aruing A, Agrawal R, Raghavan P. A Linear Method for Deviation Detection in Large Databases. KDD Conference Proceedings. 1999: 1-6.

[11] Knorr EM, Ng RT. Algorithms for Mining Distance-Based Outliers in Large Datasets. Proceedings of the 24rd International Conference on Very Large Data Bases. Morgan Kaufmann Publishers Inc. 1998: 392-403.

[12] Ramaswamy S, Rastogi R, Shim K. Efficient Algorithms for Mining Outliners from Large Data Sets. ACM SIGMOD Conference Proceedings; 2000.

[13] Freund Y. Experiments with a New Boosting Algorithm. Thirteenth International Conference on Machine Learning. 1996; 13: 148-56.

[14] Alfaro E, Gámez M, Garc'Ia N. Adabag an r package for classification with boosting and bagging. Journal of Statistical Software. 2014; 54(2): 1-35.

[15] Torgo L. Data Mining with R: Learning with Case Studies. Chapman & Hall/CRC. 2010: 188-9. http://dx.doi.org/10.1201/b10328

[16] Sing T, Sander O, Beerenwinkel N, et al. Visualizing the performance of scoring classifiers. 2015.