Contemporary Engineering Sciences, Vol. 7, 2014, no. 11, 515 - 522
HIKARI Ltd, www.m-hikari.com
http://dx.doi.org/10.12988/ces.2014.4431

Data Mining Approach For Subscription-Fraud Detection in Telecommunication Sector

[1]P. Saravanan, [2]V. Subramaniyaswamy, [3]N. Sivaramakrishnan, [4]M. Arun Prakash and [5]T. Arunkumar

[1,2,3,4] School of Computing, SASTRA University, Thanjavur.
[5] School of CSE, VIT University, Vellore, India

Abstract

This paper implements a probability based method for fraud detection in telecommunication sector. We used Naïve-Bayesian classification to calculate the probability and an adapted version of KL-divergence to identify the fraudulent customers on the basis of subscription. Each user's data corresponds to one record in the database. Since, the data involves continuous numerical values, the Naïve-Bayesian classification for continuous values is used. This methodology overcomes the problem of existing system, which classifies the best customer as fraudulent customers, as it works on a threshold based method.

Keywords: Naive-Bayes classification, K-L divergence, Fraud detection, Tele-communication

1. Introduction

Fraud detection problems are found in many sectors such as Insurance, credit-cards, telecommunications, etc. Telecommunication sector is a wide sector with thousands of users. This sector broadly has two types of users – domestic and commercial. Connections are provided to the domestic users at an affordable rate while the commercial connections are given at a comparatively higher rate as the usage scale

is higher in the latter case. There are cases where the connections are bought under domestic category but used are on a commercial scale. This cause's substantial loss to the sector, as the connections, when bought under commercial category, will yield a greater income to the sector. There are many methods to identify such fraudulent customers. This paper implements a probability based method that classifies the customers as fraudsters and Non-fraudsters.

Naïve Bayesian Classification is a well-known method for classifying records based on, either, numerical or categorical. This method is used for calculating probability value for each record using the attributes that describe the record. The values of each record are numerical. Hence, it implements the formula for continuous values in Naïve-Bayesian method.

The accuracy of classification is achieved by an adapted version of KL-divergence, which is a well-known method for finding the divergence between two probabilistic quantities. This approach also reduces the requirement of huge number of records that is otherwise, needed for training the Naïve-Bayesian classifier.

The initial work on the problem of fraud detection has been explained in [3] and [4], while the problem of fraud detection in telecommunication sector is reviewed in [5] and [6].The authors of [1] present an adaptive and automatic design of user profiling methods for the fraud detection problem, using a series of data mining techniques. A significant contribution in the field of fraud detection in telecommunications belongs to Constantinos Hilas, who investigates the usefulness of applying different learning approaches to a problem of telecommunications fraud detection in [3], and constructs an expert system, which incorporates both the network administrator's expert knowledge and knowledge derived from the application of data mining techniques on real-world data in [4]. The recent study carried out in [2] aimed at identifying customers' subscription fraud by employing data mining techniques and adopting knowledge discovery process.

The rest of the paper focuses on the following topics. Section 2 explains the general methodology used in the existing system. Section 3 has described about the proposed system. Section 4 holds a sample dataset of ten records and the experimental results and Section 5 concludes the work.

2. Existing System

There are many systems that detect similar fraudulent problems in this sector using clustering, neural networks, decision tress and probability based methods. This paper

takes the base of probability based method for identifying fraudulent customers. The existing system is based on user profiling, using Latent Dirichlet Allocation (LDA) and KL-divergence, and threshold type classification. User profiling is done by implementing LDA probabilistic model for each user, thus, converting each user's data into a probability value, where each record in the database corresponds to one user.

KL-divergence is another method which is approximated to be used in this process. Firstly, KL-divergence is approximated between two LDAs. Here, one LDA corresponds to the user and the other corresponds to a reference model. The users' model is calculated in the first phase and the reference model has values that correspond to an average of the legal users' data. The resulting value is a measure of divergence between the reference model and the user model.

Similarly, KL-divergence is approximated for automatic computation of threshold using the steepest descends formula. This threshold value is, then, used for classification. If the above computed divergence value falls above the threshold, then the customer is classified as a fraud [1].

3. Proposed system

The proposed system solves the same problem as that of the existing system, but by using Naïve-Bayesian formula for continuous values, for calculating the probability and KL divergence. The disadvantage of the existing system is the means of classification. Further, the existing system uses only three attributes, namely, start-time, duration and destination. Although, the third attribute is considered as one of the prime factors for user profiling, it is unreliable. Figure 1 shows the architecture of fraud detection process.

This paper, also, follows a probability based method, ie., Naïve-Bayesian classification for probability calculation. The attributes that are taken for probability calculation are duration, number of calls made during day time, number of calls made during night time, number of calls in week days, number of calls in weekends and total number of calls.

Since, all the attributes have continuous numerical values, the formula of Naïve-Bayesian classification used for continuous numeric value, is considered. With Naïve-Bayesian classification alone, the requirement of already classified records for training the classifier is huge. Hence, to improve accuracy and efficiency, the

idea of the KL divergence, by which it finds the divergence between two quantities, is used. This overcomes the requirement of huge training data. Further, the divergence value that is found with the Naïve-Bayesian probabilities shows a significant difference between a normal user and a suspected user.

3.1 Probability calculation

This phase involves probability calculation. This phase uses Naïve Bayesian classification as a method for calculating probability. The reason why Naïve-Bayesian classification alone is not used for entire classification is that it requires building a trainer that will work on already classified data. To train the classifier to achieve the desired level of accuracy, huge number of records is needed. Since the occurrence of a fraudulent customer in this sector is sparse compared to other sector, building a classifier that requires utmost training may be arduous. To overcome this problem, we combine Naïve-Bayesian classification with the idea of finding divergence. The general formula for calculating probability in Naïve-Bayesian classification is

$$p(x,Ci) = \prod_{k=1}^{n} p(X|Ci)$$

Here, x represents the tuple under consideration, Ci represents the class for which the probability is calculated and 'n' represents the number of attributes.
The frequent form of Naïve-Bayesian that is used is the one that deals with categorical data, where the probability calculation involves finding out the number of records of a particular value of a particular class divided by total number of records of that class. But, for continuous numeric values, the probability calculation goes by the following formula,

$$g(x,\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma(Ci)}} e^{\frac{(x-\mu(Ci))^2}{2\sigma(Ci)^2}}$$

$$p(x,Ci) = g(x,\mu(Ci),\sigma(Ci))$$

Where μ(Ci) represents the mean value corresponding to the class Ci. The mean value is calculated for each attribute with respect to each class; σ(Ci) represents the standard deviation corresponding to the class Ci. Similar to mean, the standard deviation is calculated for each attribute with respect to each class. The general formula for calculating standard deviation is

$$\sigma = \sqrt{(\frac{1}{N}\sum_{i=1}^{N}(Xi - \mu)^2)}$$

The above mentioned formulae are used in common for all records irrespective of the class label. Hence, the mean and standard deviation is calculated as a whole, for all attributes. The mean and standard deviation, calculated for a set of records is shown below in Fig.2 and Fig.3. Similarly, the over all probability is calculated only once with those common mean and variance values of the respective attributes. The above three intermediate steps appear in Fig.1, in order of the sequence number.

3.2 Calculating divergence

After calculating the probability of the user records, reference values are found for calculating reference probability based on which the divergence is calculated. The formula for finding divergence between two probability distributions using KL divergence is

$$\int p(x) * \log\left(\frac{p(x)}{q(x)}\right)$$

Where p(x) represents user probability and q(x) represents reference probability. This formula is directly applied where the user probability is substituted in the place of p(x) and the reference probability is substituted in the place of q(x). The divergence, thus, calculated yields the required difference to differentiate a suspected fraudulent customer from a normal customer. The divergence value for a sample data set with the difference in values for normal and suspected customers is shown in Fig.1.

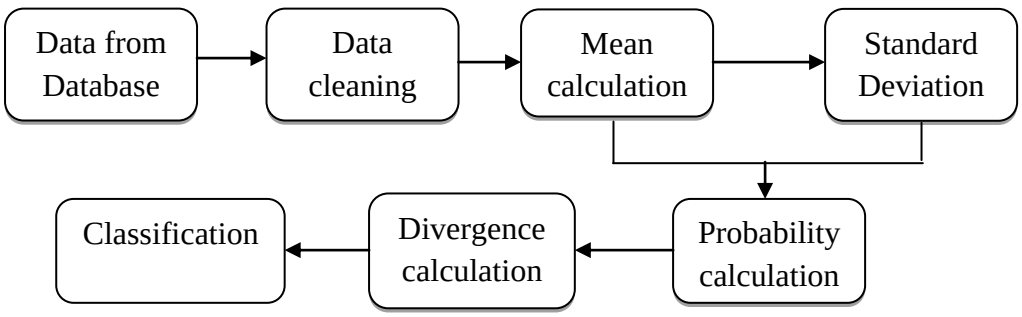

Figure 1 Flow diagram of the process steps

4 Experimental Results

In experimental results, the table holds ten records as a sample dataset which is taken from the original dataset taken for implementing the above mentioned method. It has eight fields in total out of which the last six fields are the data that is required for calculating the probability using Naïve-Bayesian formula. These records are then classified as fraudulent or non-fraudulent customers, after finding the divergence between the user probability and the reference probability. The category, 'total calls', is split up in two forms, with one combination being day time and night time and the other combination being weekdays and weekends. This kind of split up is taken because of the variation that is expected between a normal legal customer and a fraudulent customer. The fraudsters who use a domestic connection on a commercial scale may use more in day time and night times, in cases that involve night shifts and the difference in their usage during weekdays and weekends, compared to a non-fraudulent customer. These reasons justify the different kinds of number of calls taken for probability calculation. Table.1 shows a sample dataset with ten records.

Table.1 Sample Dataset

S.no	Name	Duration	Day time	Night time	Week days	Week ends	Total calls
1.	David	172	58	4	34	28	62
2.	Joseph	567	87	22	67	42	109
3.	Ram	1800	300	0	146	154	300
4.	Lalitha	76	23	12	15	20	35
5.	Gayathri	932	193	172	245	120	365
6.	Suvashini	117	44	16	37	23	60
7.	Varshini	2100	57	24	52	29	81
8.	Kayalvizhi	473	50	17	35	32	67
9.	Sathish	1971	114	59	119	54	173
10	Sriram	396	44	9	26	27	53

Fig.2 shows the mean and the standard deviation of the six attributes, used in probability calculation. Fig.3 illustrates the divergence values, apparently indicating the difference between the normal using customers and fraudulent customers. Fig.3

shows the final results with the fraudulent customer details and their call details to assist the telecom officials check the customers' activities.

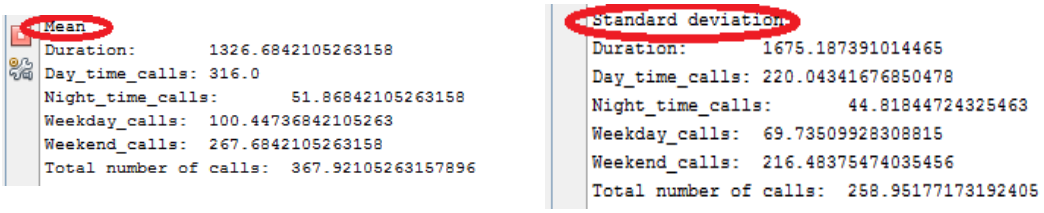

Fig.2 Mean values and standard deviation of the attributes

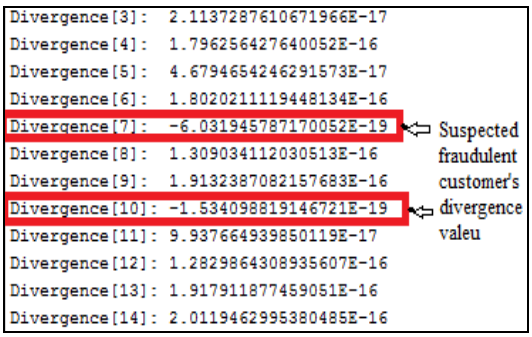

Fig.3 Final outcome with fraudulent customer details

5. Conclusion

Thus, this paper overcomes the problem of identifying fraudulent customers in telecommunication sector by classifying the true fraudulent customers alone. The level accuracy and with optimal number of records that is available, the classification is carried out. Thus, this reduces the loss, the sector faces by identifying fraudulent customers and not classifying the best customers as fraudulent customers, thus, eliminating the risk of losing such customers. Further work in this domain can implement other probability-distribution based algorithms for a greater accuracy on results.

References

1. D.Olszewski, A probabilistic approach to fraud detection in telecommunications, Knowledge Based Systems, Volume 26, pp. 246-258, 2012.

2. H.Farvaresh, M.M.Sepehri. "A data mining framework for detecting subscription fraud in telecommunication", Engineering applications of artificial intelligence, Volume 24, No. 1 pp. 182-194, 2011.

3. R.J. Bolton, David, Statistical fraud detection: A review, Statistical Science, Volume 17, pp. 235–255, 2002.

4. Y.Kou, C.-T.Lu, S.Sinvongwattana, Y.-P.Huang, Survey of fraud detection techniques, Proceedings of the IEEE International Conference on Networking, Sensing & Control, Volume 2, pp. 749-754, 2004.

5. P.Burge, J.Shawe-Taylor, C.Cooke, Y.Moreau, B.Preneel, C.Stoermann, Fraud detection and management in mobile telecommunications networks, Proceedings of the European Conference on Security and Detection ECOS97, pp.91–96, 1997.

6. T.Fawcett, F.Provost, Adaptive fraud detection, Data Mining and Knowledge Discovery, Volume 1, No. 3, pp. 291–316, 1997.

Received: April 7, 2014