Detection of hidden attacks to telecommunication company on the basis of the data mining technologies

V.N. Tarasov[a], M.E. Samarkin[a], E.M. Mezenceva[a]

[a] *Povolzhsky State University of Telecommunication and Informatics, 443010, 23 L. Tolstoy street, Samara, Russia*

Abstract

In this article we consider an approach of analysis the telecommunications company customers information using data mining technology. The purpose of this work is to develop a software solution for identifying customers who behave abnormally compared to ordinary customers. To solve the problem the method of cluster analysis is used, with k-means, method of support vector machines (SVM) and non-standard method of initializing k-means++ centroids. With the application of the developed programs, the analysis of the data obtained and the identification of customers who are suspected of hidden attacks on the company are identified. To classify the new data the classifier is developed, which is trained on large data sets and based on SVM.

Keywords: cluster analysis; k-means method; the method of initializing k-means++ centroids;

1. Introduction

The telecommunication companies (TC) collect a large amounts of information utilizing databases on a daily basis. It is necessary to extract valuable information from the data regarding to the company and its customers. For this purpose a DATA MINING technology is used - a multidisciplinary area, which is based on such sciences as applied statistics, classification and clustering, recognizing of images, artificial intelligence, database theory and others. It can be considered as the integration the theory of applied statistics, data analysis with data cleansing, training and visualization of results.

Regarding to TC information processing, the main problem is in the fact that it is often not known in advance how to group the data, how to find the connection between events, etc. To solve this problem partially using DATA MINING methods of classification and clustering. Thus it is possible to combine data for any selected criteria. For example, to find the relationship between the event and the the network load, or day of the week, etc.

2. Problem definition

As an example of the accumulation of large amounts of data, consider the telecommunication companies that every day gathering information about user traffic. The following problem is posed: using known methods of intellectual analysis of the entire amount of information obtained in one month of work, to reveal hidden attacks on the servers of the telecommunications company. Under hidden attacks we mean intentional or unintentional generation of the same type of traffic that does not carry any meaning, which loads the network equipment, thereby preventing the passage of normal traffic through network nodes.

The solution of the problem will include the following steps: 1) collection of traffic for a certain period during which "hidden attacks" were allegedly carried out; 2) using the methods of cluster analysis, it is necessary to identify anomalies (non-standard behavior of customers); 3) analyze the results obtained and select the signs by which such anomalies can be detected in the future; 4) label the data vector and train the machine for a further process of checking traffic in the future.

One of the telecommunication companies provided a large amount of data (135 GB in txt format) of traffic that customers generate on a daily basis. It contains, among others, the following information: Account ID, IP address, source, IP address, the recipient, the data packet size, source port number, destination port number, day and time of the event in unix format. A total of approximately 45 million lines of the following form (Figure 1.).

Please The choice of such attributes as: ID account, the amount of traffic per day in gigabytes, the day of the week, the ordinal number of the day, the type of the day (day off - 0, worker - 1), the number of requests per day is primarily justified by the task. The volume of traffic per day is an important indicator of activity for the client of the telecommunications company, the less it generates it, the less it will be necessary to pay the provider for the channel. In terms of traffic volume, you can see what the bandwidth of the acceptable channel the client uses, relative to what it purchased.

The day of the week, the ordinal number of the day and the type of day allow during the process of manual analysis to determine the "admissibility" of the behavior of a particular client, since the working day and day off are very different in their

activity of users. The number of requests per day, this is the indicator that will allow us to determine whether the number of requests made by the client is adequate relative to the time of the used channel.

timestamp	1456634876
account_id	0
source	192.168.101.3
destination	178.218.82.79
t_class	2210
packets	1
bytes	258
sport	53
dport	57577
date	Sun Feb 28 08:47:56 2016

Fig. 1. Example of a figure caption.

The classic problem of DATA MINING is when the most relevant and valuable information should be extracted from very large data set. It is not possible to process such amount manually, so the decision was to convert the data in the suitable format and apply the statistical methods for clustering and classification to the solution of the problem[2]. Moreover the standard statistical packages could not process such data volumes. The well-known classification method based on the correlation analysis also can not be applied here due to the fact that the correlation matrix obtained very high order. The k-means[1] algorithm was chosen by the authors as the simplest and well-described way of detecting emissions in the available raw data. Further the sequence of actions to obtain the practical results is provided.

3. Initial processing of data

To determine the ratio of customer traffic and client activity data was converted to the following format: ID account, the volume of traffic per day in gigabytes, day of the week, the serial number of the day, type of day (day off - 0, working - 1), the number of requests per day. The 6 ordered attributes of each data vector in total. Thus, the feature space for the solution of the stated task includes the six characteristics listed above.

For such conversion the program is written in the C# programming language, which perfomed the following actions: read txt files consequently, dump daily traffic of each client and read the number of customer calls to the server. After all transformations the 60473 rows of information has been obtained, and, if we divide this number by the number of days in a month, we get the number of active subscribers of the company. Each line is information about the client per day. Below in Figure 2 shows an example of data for the five customers with the company's 6-ordered features.

```
0,30.0108,1,1,1,13883303
1,0.007,1,1,1,4573
10,0.0341,1,1,1,3333
29,0.7152,1,1,1,38790
66,0.0716,1 ,1,4737
```

Fig. 2. Data example after conversion.

4. Description of the data conversion program for clustering

The program algorithm is divided into several cyclical phases:

1. open text files with the information while all the files not passed,
2. read lines until the end of the file,
3. parse a string into components: ID account, the package size, etc,
4. summarize the traffic volume,
5. count the number of calls to the server;
6. convert data to CSV format and save the data to a file.

The scheme of the program algorithm is shown in the Figure 3. After the initial processing it is required to split the data into groups and analyze the result. At the same time it is supposed a priori that among company clients can be nonstandard records on two attributes (2nd and 6th). For this method clustering k- means was selected.

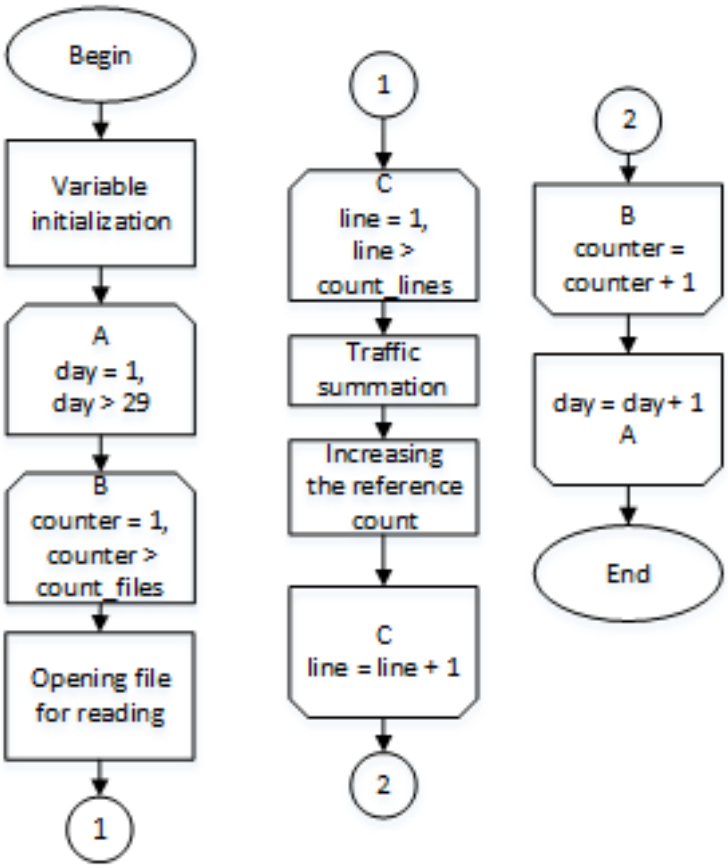

Fig. 3. Data example after conversion.

5. Clustering data

The accuracy of the vectors partition in clusters depends on the originally selected centroid. We've selected centroid initialization method called k-means++. It is an improved version of the clustering algorithm k-means. The essence of the improvements is to find a "good" initial values of the cluster centroid. The algorithm method includes 5 basic steps.

Initialization:

1. select the first random centroid (among all the points);
2. for each point find a square value of the distance to the nearest centroid (from selected ones) $(dx)^2$;
3. choose the next centroid from these points so that the probability of the point selection is proportional to the calculated square of the distance to it;
4. this can be done as follows. On the Step 2 with the calculation $(dx)^2$ we need to calculate the sum (dx^2). After accumulating the sum find the value Rand = random (0.0,1.0)*Sum. The Rnd generator randomly indicate the number in the interval [0; Sum), and we only need to determine which point it matches. To do this, start again to count the amount of $S(dx^2)$ as long as the amount does not exceed Rnd. Once this happens, the summation stops, and the current point can be taken as the centroid. When each of the following centroids is taken there is no need to ensure that it does not coincide with selected one, since the probability of re-selection of a point is equal to 0;
5. repeat steps 2 and 3 until all the centroids are found.
6. convert data to CSV format and save the data to a file.

Next, the basic k-means algorithm is performed. This is the most popular clustering method, which was invented back in the 1950s by mathematician Hugo Steinhaus and almost simultaneously by Stuart Lloyd.

The algorithm tends to minimize the total deviation of the points cluster from the centers of these clusters.

$$V = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - \mu_i)^2 \to \min,$$

where k - the number of clusters, Si - obtained clusters, i= 1,2, ...,k and μi – mass centers of vectors xj,, that belongs to Si.

6. The result analysis

For the cluster analysis a Python program was written utilizing the sklearn library, which contains the implementation of the majority of mathematical methods of intellectual Data analysis. The integrated diagram of algorithm of the program is shown in Fig. 4.

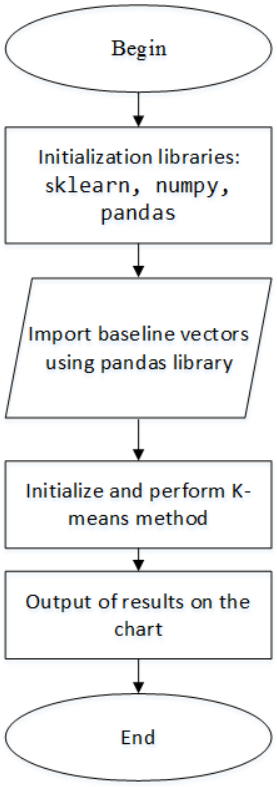

Fig. 4. The clustering program algorithm scheme.

Applying the combined method of k-means with the method of centroids initialization k-means++, and dividing the data into 2 clusters, we obtain the following results: 1st cluster includes 60427 vector; 2nd cluster - 46 vectors. Fig. 5 represents a graphical distribution of vectors (records) for clusters.

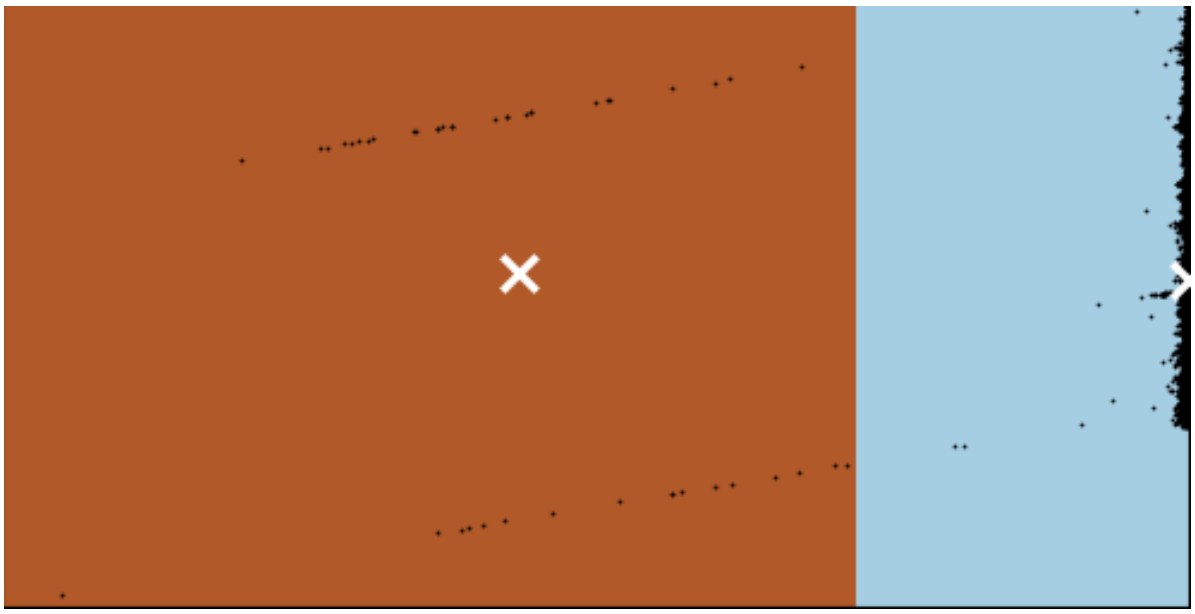

Fig. 5. Clusters obtained by method k-means.

Here on the X axis the number of responses from the server is indicated, and on the Y axis – account id. White crosses represent centroid clusters. How the findings can be interpreted? The first cluster (right) contains information about standard records that do not stand out from the bulk. This is the default behavior of customers during the whole month. The second

cluster (left) contains information about the separated from the bulk of non-standard entries that contain hidden attacks on the company. This is abnormal behavior for clients, most likely on the side of customers with such features can detect viral activity, or even some "bad" software, causing a very large number of calls to the TC server.

7. Building and training the classifier

To make the software in the future determine the status of a record to already isolated clusters obtained above automatically, it is necessary to apply classification method. For this we consider a support vector machine[3].

The support vector method initially refers to binary classifiers. It belongs to the family of linear classifiers. It is easier to explain the idea of the method by a simple example: points are given on the plane and they must be divided into two classes. Then, as the separating line between the classes, we choose that distance from which to each class is maximal. In the general case, instead of the dividing line, there will be a hyperplane with the maximum distance to each class. In general case there will be a hyperplane with the maximum distance to each class instead of the dividing line. Since a hyperplane can be specified in the form $\langle \mathbf{w}, \mathbf{x} \rangle + b$ for some **w** and *b*, it is necessary to define such **w** and *b* that maximize the distance to each class.

The support vector method constructs the classifying function F in the form

$$F(x) = \text{sign}(\langle \mathbf{w}, \mathbf{x} \rangle + b),$$

where is the scalar product of the vectors, w is the normal vector to the separating hyperplane, and b is the intermediate parameter.

To apply the classification method to the data, each record has been manually labeled to a particular class. We will consider these data as training sample, since the data are marked in accordance with manual classification, and not using machine learning algorithms. Marking of vectors was made in accordance with the subjective view of the authors on "hidden attacks". The data is visualized as follows (Fig. 6). The new field ‘target’ was added, which stores an information about record belonging to a particular class.

account_id	0
traffic	30.0108
count	13883303
dayOfWeek	1
n_day	1
type_day	1
month	2
target	1

Fig. 6. Type of data.

Here the target field is classifier: 1 - refers the data to the first cluster. 0 – to the second.

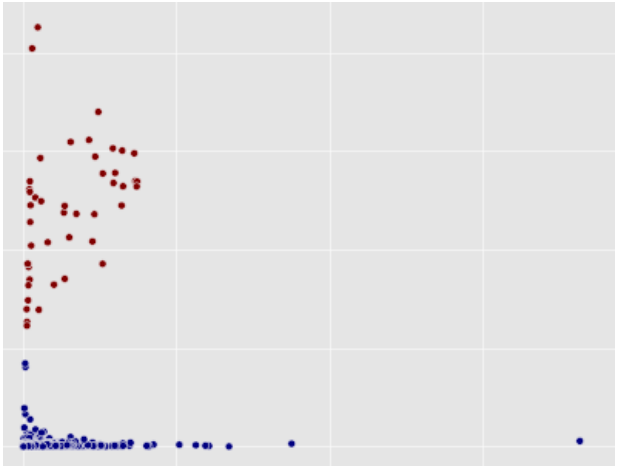

Fig. 7. Graphical representation of classified data.

Fig. 7 shows the classified data on the target field: red marks are the vectors belonging to the cluster 1, blue - to the cluster 2. The program in Python language is written, utilizing the sklearn library, which classifies the new data using the SVM

method[4], and displays the information which data belongs to one or another class. The aggregative diagram of the program algorithm is shown in Fig. 8.

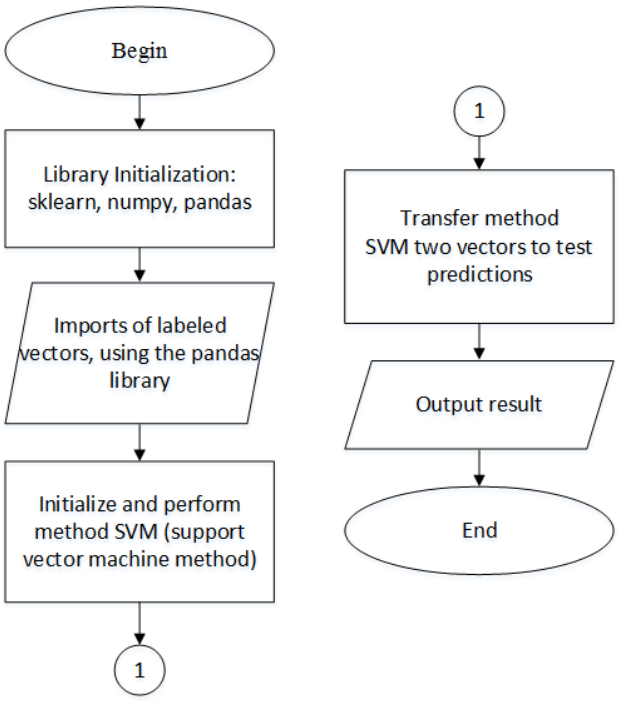

Fig. 8. The scheme of data classification program algorithm.

Based on the authors' experience of finding hidden attacks in the manual mode, records are taken from the history that correspond to the "hidden attack" and normal behavior. Accept these data for the control sample. Only two entries are taken, so that the reader can clearly see the values of the attributes that are transferred to the machine, rather than the dry data on how many records were obtained belonging to one or another class.

To prove the approach the following records (traffic size in gigabytes, the number of responses from the server, day of week, month, type of day, month) are examined for belonging to the cluster:

1. 30.0108,13883303,1,1,2,1;
2. 0.007,4573,1,1,2,1,0.

The answer provided by the program is: 1st record belongs to the first cluster, 2nd – to the second cluster, it confirms the classifier correctness. On the properties of the first record, it is clear that a lot of requests to the server have been performed, so let's attribute this to a "hidden attack". The second record is equivalent to the normal behavior of the client.

8. Conclusion

The full range of analysis of the company’s data is performed. The company provided the raw data for the February, containing the service information of IP-streams. For comprehensive analysis the program written in C# programming language, which aggregate data by day. Thereby the tests were discovered, which be the basis for the further classification according to the known methods. The cluster analysis was conducted utilizing the author’s program written in Python programming language. The two clusters are identified: the regular customers and a "special", which are the threat to the telecommunications company. The clustering information is displayed in diagrams[5].

Using the initial data the the learning procedure was completed on the large data set (60427 vectors) with the ‘target’ field added. The support vector machine method was applied to classify the new data received in real time. The trained classifier can be further used for the analysis of new TC data.

References

[1] Mandel, I. Klasternyi analis [Cluster analysis] / I. Mandel - M.: Finansy i Statistika, 1988. – 176 p.
[2] Barsegyan, A. Analis dannykh i protsessov [Analysis of data and processes] / Barsedyan, A. i dr.: uch. posobie 3-e isd., SPb.: BKHB-Peterburg, 2009. — 512 p.
[3] Nello, C.. An Introduction to Support Vector Machines and Other Kernel-based Learning Methods / Nello C. , Taylor J. — Cambridge University Press, 2000.
[4] Richert W., Building Machine Learning Systems with Python / Richert W., Coelho L. - Packt Publishing, 290 p.
[5] McKinney W., Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython / McKinney W. - O'Reilly Media, 400 p.