A COMPARATIVE STUDY OF CONSUMER CLAIMS DATA SET BY CLUSTERING ALGORITHMS

Dhyan Chandra Yadav
E-mail: dc9532105114@gmail.com.
S.N.P.G. College, Narahi, Balia (U.P.)
Department of Computer Application

ABSTRACT:

This paper discusses the development of an application for both consumer and companies or banks. The product, pricing and policy differ from country to country. In case of purchasing each company or bank has its rules and regulations by the help of clustering algorithms we easily analyzed dispute YES or NO from consumer claim dataset. So we use three major clustering algorithms: K-Means, Hierarchical clustering and Density based clustering algorithm and compare the performance of these three major clustering algorithms on the aspect of correctly class wise cluster building ability of algorithm. Performance of these techniques are presented and compared using a clustering tool WEKA.

Keywords: Clusterer: K-means algorithms, Hierarchical clustering, Density based clustering algorithms, Weka.

[1] INTRODUCTION

When your credit card due date doesn’t align with your payday you can find yourself constantly rearranging bills to make your payment on time. If this is your problem call your credit card issuer and ask it to change your due date. Keep in mind that your due dates will fall on the same day every month. Credit reporting companies must investigate your dispute, forward all documents to the furnisher, and report the results back to you unless they determine your claim is frivolous. If the consumer reporting company or furnisher determines that your dispute is frivolous, it can choose not to investigate the dispute so long as it sends
you a notice within five days saying that it has made such a determination. If the furnisher corrects your information after your dispute, it must notify all of the credit reporting companies it sent the inaccurate information to, so they can update their reports with the correct information [1].

J. Han and M. Kamber introduced about Clustering. It can be considered the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A loose definition of clustering could be “the process of organizing objects into groups whose members are similar in some way”. A cluster is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters [2].

J Han and M Kamber introduced about K-means. It is a widely used partitional clustering method in the industries. The K-means algorithm is the most commonly used partitional clustering algorithm because it can be easily implemented and is the most efficient one in terms of the execution time [3].

Fig.1. Representation of Instances by K-Means clusterer.

M Verma, M Srivastava, N Chack, A K Diswar and N Gupta introduced about Hierarchical Clustering. Hierarchical Clustering builds a cluster hierarchy or, in other words, a tree of clusters, also known as a dendrogram [4].

Agglomerative (bottom up)
1. Start with 1 point (singleton).
2. Recursively add two or more appropriate clusters.
3. Stop when k number of clusters is achieved.

Divisive (top down)
1. Start with a big cluster.
2. Recursively divides into smaller clusters.
3. Stop when k number of clusters is achieved.
Dhyan Chandra Yadav

Fig. 2. Visualized of Instances by Hierarchical Clusterer.

T. C. Havens introduced about Density-based clustering algorithms. Density-based clustering algorithms try to find clusters based on density of data points in a region. The key idea of density-based clustering is that for each instance of a cluster the neighborhood of a given radius (Eps) has to contain at least a minimum number of instances (MinPts). One of the most well-known density-based clustering algorithms [5].

Fig. 2. Visualized of Instances by Make Density Based Clusterer.

[2] RELATED WORKS:

H, K analyzed that Credit cards fraudulence arises at very high level scale so we cannot easily detect and predict the related attributes but by the help of data mining classifier tool to prevent the activity of fraudsters in the misuse of credit cards uses the algorithms of neural
networks. This system predicts the probability of fraud on an account by comparing the current transactions and the previous activities of each holder [6].

D C Yadav and S Pal discussed that classifier algorithms provide very accurate result in software error detection by J48, ID3 and Naïve Bayes data mining algorithms correctly classified instances will be partition in to numeric and percentage value, kappa statics, mean absolute error and root mean square error will be at numeric value only ID3 and J48 time taken to build model: 0.2 seconds and test mode :10 fold cross validation. Here Weka compare all required parameters on given instances with the classifiers respective accuracy and prediction rate based on highest accuracy of J48 is 100% without error also Naïve Bayes 100% correctly classified but with some error and ID3 95% correctly classified, so it is clear that J48 is the best in three respective algorithms so it is more accurate [7].

D C Yadav and R Kumar discussed that association algorithms provide very accurate result in the frequent and relationship between data object and find the percentage of confidence, support, of data object by the help of apriori, predictive apriori and filtered associate algorithms. Therefore these algorithms can be used in other domains to bring out interestingness among data present in the origin [8].

D C Yadav and R Kumar discussed that three major clustering algorithms: K-Means, Hierarchical clustering and Density based clustering algorithm and compare the performance of these three major clustering algorithms. Author compared using a clustering tool and find result: K-Means algorithm is better than Hierarchical Clustering and Make density based algorithm because all the algorithms have some ambiguity in some (noisy) data when clustered [9].

R Sukanya and K Prabha discussed that back propagation Neural Network, Support Vector Machine is used for rainfall prediction. ANN improves the efficiency of Rainfall prediction by analyzing the historical and current facts to make accurate predictions about future [10].

D C Yadav analyzed that in statistical analysis of binary classification, the F1 score is a measure of a test's accuracy. It considers both the precision and the recall of the test to compute the score. In this analysis author computed the best score for F1 by the help of data mining classifier algorithms and choose the ID3 Tree is the best data mining classifier algorithms to be applied over selected datasets. Because ID3 Tree has highest F1 score and take less time to build a node [12].

D C Yadav analyzed that the Matthews correlation coefficient is used in machine learning as a measure of the quality of binary (two-class) classifiers. It takes into account true and false positives and negatives and is generally regarded as a balanced measure which can be used even if the classes are of very different sizes. Author computed the MCC is in essence a correlation coefficient between the observed and predicted binary classifications by
the help of data mining classifier algorithms and ID3 Tree is the best data mining classifier algorithms to be applied over selected datasets. Because ID3 Tree has highest MCC value and minimum number of time in second 0.00 to build a model [13].

D C Yadav analyzed that the informedness of a prediction method as captured by a contingency matrix is defined as the probability that the prediction method will make a correct decision as opposed to guessing and is calculated using the bookmaker algorithm. Their correlation is the generated by LAD Tree, ID3, and J48 data mining algorithms and find ID3 is the best data mining classifier algorithms to be applied over selected datasets [14].

D C Yadav analyzed that the FDR-controlling procedures provide less stringent control of Type I errors compared to class wise errors. In this analysis we choose the ID3 Tree is the best data mining classifier algorithms to be applied over selected datasets. Because ID3 Tree has minimum time to build a model [15].

T, R and Liu discussed that a framework was presented on the base of security systems and Case based reasoning for fraud detection. First, a set of normal and fraud cases are made from labeled data. Then, the primary detectors are made with random or genetic algorithms. Then, negative selection and clonal selection operations are applied on primary detectors in order to obtain a set of detectors with different algorithms that can detect a variety of frauds [17].

M, S, B and Saira discussed that many fraud detection systems that have been presented so far, have used data mining and neural network approaches. While no fraud detection system with the combination of anomaly detection, misuse detection and decision making system have been used so far for fraud detection in credit cards. Then, a system was proposed that used Hidden Markov Model to detect the fraudulent transactions [18].

A John analyzed that hybrid feature selection and anomaly detection algorithm in order to detect fraud in credit cards. The authors have noted that fraud detection on the internet must be done online and immediately. Since the use of credit card by card holders follows a fixed pattern, this fixed pattern can be extracted from a usual legal activity of card holders in 1 or 2 years. thus, this pattern is compared to the use of process of card holder and in case of non-similarity in the pattern, the activity is considered illegal. It should be noted that the neural networks were used to teach the patterns detection in the model in this study [19].

A P, M K and A N discuss that data mining as one of the most efficient tools of data analysis has attracted the attention of many people. The use of different techniques and algorithms of this tool in various fields like customer relationship management, fraud management and detection, medical sciences, sport and etc. Due to the large number of data in banks, data mining has had lots of functions in financial and monetary affairs so far. Credit
risk management, fraud detection, money laundering, customer relationship management and banking services quality management are some examples of data mining function in banks [20].

In this paper we use K-Means, Hierarchical, Make Density Based algorithms by which to describe data and to explain the relationship between one dependent binary variable and one or more nominal, ordinal, interval or ratio-level independent variables also we analyzed comparison between them.

[3] METHODOLOGY:

Our research approach is to use K-Means, Hierarchical, Make Density Based algorithms on consumer claim data set. The research methodology is divided into 5 steps to achieve the desired results:

Step 1: In this step, prepare the data and specify the source of data.

Step 2: In this step select the specific data and transform it into different format by weka.

Step 3: In this step, implement data mining algorithms and checking of all the relevant dispute is perform.

Step 4: The decision is taken on the presence of dispute in source code. If dispute is present then proceed further, otherwise it will stop. We classify the relevant dispute using K-Means, Hierarchical, Make Density Based algorithms.

Step 5: At the end, the results are display and evaluated.

Now we have study about consumer complaints of bank different type’s loans and relate dispute.

1. Data Preparation:

When your credit card due date does not align with your payday you can find yourself constantly rearranging bills to make your payment on time. If this is your problem call your credit card issuer and ask it to change your due date. Keep in mind that your due dates will fall on the same day every month. Credit reporting companies must investigate your dispute, forward all documents to the furnisher, and report the results back to you unless they determine your claim is frivolous. If the consumer reporting company or furnisher determines that your dispute is frivolous, it can choose not to investigate the dispute so long as it sends you a notice within five days saying that it has made such a determination. If the furnisher corrects your information after your dispute, it must notify all of the credit reporting companies it sent the inaccurate information to, so they can update their reports with the correct information.
Table 1. Representation of Computational Variables of Consumer Claims

<table>
<thead>
<tr>
<th>Property</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Complaint Type</td>
<td>Consumer Complaint Database Attributes for Financial Problem (Bank, Lender &amp; Company etc.)</td>
</tr>
<tr>
<td>Sample Size</td>
<td>1000 Total: 197 Consumer dispute and 803 non dispute</td>
</tr>
</tbody>
</table>

2. Data Selection and transform:
The database generally updates daily, and contains certain information for each complaint, including the source of the complaint, the date of submission, and the company the complaint was sent to for response. The database also includes information about the actions taken by the company in response to the complaint, such as, whether the company’s response was timely and how the company responded. If the consumer opts to share it and after we take steps to remove personal information, we publish the consumer’s description of what happened. Companies also have the option to select a public response. Company level information should be considered in context of company size. Data from those complaints...
helps us understand the financial marketplace and protect consumers. If you identify an error on your credit report, you should start by disputing that information with the credit reporting company. You should explain in writing what you think is wrong, why, and include copies of documents that support your dispute. You can also use our instructions and template letter as a guide.

3. Data mining Implementation-
Weka is the data mining tools. It is the simplest tool for classify the data various types. It is the first model for provide the graphical user interface of the user. For perform the clustering we used the promise data repository. It is provide the past project data for analysis. With the help of figures we are showing the working of various algorithms used in weka. weka is more suitable tool for data mining applications. This paper shows only the clustering operations in the weka, we will try to make a complete reference paper of weka. Clustering is a main task of explorative data mining, and a common technique for statical data analysis used in many fields ,including machine learning. I am using Weka data mining tools for this purpose. It provides a batter interface to the user than compare the other data mining tools.

[4] RESULT AND DISCUSSION-
Above section involves the study of each of the three techniques introduced previously using Weka Clustering Tool on a set of e-commerce data consists of 8 attributes and 199 instances. Clustering of the data set is done with each of the clustering algorithm using Weka tool and the results are:

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Prior Probability</th>
<th>NO. of Iterations</th>
<th>Sun of Square Error</th>
<th>Time</th>
<th>Clustered Instances</th>
<th>Log Likelihood</th>
<th>Classified Instances</th>
<th>Log Likelihood</th>
<th>Classified Instances</th>
</tr>
</thead>
<tbody>
<tr>
<td>K-Means</td>
<td>-</td>
<td>3</td>
<td>691</td>
<td>0.00</td>
<td>Zero 172(86 %) One 27(14%)</td>
<td>-</td>
<td>YES</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Hierarchical</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>0.05</td>
<td>Zero 198(99 %) One 1(1%)</td>
<td>-</td>
<td>YES</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Make Density Based</td>
<td>0.139</td>
<td>3</td>
<td>691</td>
<td>0.11</td>
<td>Zero 173(87 %) One 26(13%)</td>
<td>-15.07</td>
<td>YES</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Dhyan Chandra Yadav
Performance of K-Means for dispute acceptance consumer claim data is 14%.
Performance of Hierarchical algorithms for dispute acceptance consumer claim data is 1%.
Performance of Density based algorithms for dispute acceptance consumer claim data is 13%.
Performance of K-Means for rejection dispute in consumer claim data is 86%.
Performance of Hierarchical algorithms for dispute rejection consumer claim data is 99%.
Performance of Density based algorithms for dispute rejection consumer claim data is 87%.

From the above table we discussed the result of consumer claim dispute and find the performance of these the algorithms.

[5] CONCLUSION

After analyzing the results of testing the algorithms we can obtain the following conclusions:
- The performance of Hierarchical algorithms is better than K-Means and Density based algorithms in Clustering.
- All the algorithms have some ambiguity in some (noisy) data when clustered.
- Density based clustering algorithm is not suitable for data with high variance in density.

REFERENCES

A COMPARATIVE STUDY OF CONSUMER CLAIMS DATA SET BY CLUSTERING ALGORITHMS


