ISSN (Online) 2393-8021
ISSN (Print) 2394-1588

International Advanced Research Journal in Science, Engineering and Technology
Vol. 2, Issue 7, July 2015

A Comprehensive usage of Enhanced K-Medoid Clustering Algorithm in Banking Sector

B. Kalaiselvi

Assistant Professor, Department of Information Technology, NGM College

Abstract: This paper is used to cluster the various components of a bank customer details and segregate potential customers eligible for loan. The application of this technique helps the banker to scale the potentiality of their customers and take necessary steps to decide for loan approval. The classic difficulty of recognizing the customer's potentiality is solved using this thesis and helps them to judge the good will of their customers.Identifying the eligible customers for loan in this modern society is complex. The identified customer must be able to repay his loan in the proper installments throughout the tenure. The Banker can identify the loyalty of the customer with the help of this thesis. After completing the collection of data, it is clustered according to the monthly salary, movable and immovable assets. Since K-Medoid algorithm is an unsupervised algorithm, we specify the number of clusters.

Keywords: K-Medoid, Clusters, Unsupervised Algorithm, Potential Customers, Loan Installment, Goodwill.

1. INTRODUCTION

Data mining refers to extracting or mining knowledge from large amounts of data. Data mining functionalities are used to specify the kinds of patterns to be found in data mining task. In general, data mining tasks can be classified into two categories, Descriptive and Predicative. Descriptive mining task characterize the general properties of the data in the database. Predictive mining task perform inference on the current data in order to make prediction.

In some cases, users may have no idea regarding what kinds of patterns in their data may be interesting, and hence may like to search for several different kinds of patterns in parallel. Thus it is important to have a data mining system that can mine multiple kinds of patterns to accommodate different user expectation or applications. Further more, data mining systems should be able to discover patterns at various granularity (i.e. different levels of abstraction). Data mining systems should also allow users to specify hints to guide or focus the search for interesting patterns. Because some patterns may not hold for all of the data in the database, a measure of certainty or "trustworthiness" is usually associated with each discovered pattern.

2. RELATED WORK

In the paper, Dileep B. Desai and Dr. R. V. Kulkarni[1], gave a view on data mining algorithm is used for customer segmentation, to predict bank profitability, predict payment from customers, marketing, detecting fraud transactions. K-means clustering is used to segregate the customers according to their information. In the paper, Aneta Hryckiewicz, and Lukasz Kozlowski [2], describes that they are able to identify true banking strategies consisting of different combinations of bank asset and funding sources and assess their impact on the mortgage crisis. They then estimate how distinct strategies have affected bank profitability and risk before the crisis, and what impact they have put on the mortgage crisis. Their results prove that the asset structure of banks was responsible for the systemic risk before the mortgage crisis, whereas the liability structure was responsible for the crisis itself. Finally, they show that countries with banks that rely on investment activities experienced a greater but more short-lived drop in GDP compared to countries that have a predominantly traditional banking sector. Goran Radonic [2007] [3] gave a clear view about business intelligence approaches to key business factors in banking sector. The important goals that need to be accomplished in order to achieve data consistency are (1) Timeliness (2) Accuracy (3) Acceptance. This paper user OLAP tools to present the data in multi-dimensional format. The data must be accessible for analysis and knowledge extraction. Finally the paper presented a review of typical Business Intelligence techniques and their applications in the banking industry. Waminee Niyagas, Anongnst srivihok and sukumal kitisin [2006] [1] described that study of e-banking is scantly due to the limitation of data confidentiality. This study analyses the historical data of e-banking usages from a commercial bank in Thailand. It uses apriori algorithm to detect the relationships within the features of e-banking services. Finally, the experimental result is concluded with the eight different clusters and by using the apriori algorithm it predicts the usage of e-banking services.

3. METHODOLOGY

The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering. A cluster is a collection of data objects that are similar to one another within in the same cluster and are dissimilar to the objects in other clusters.

Partitional Clustering

The partition clustering techniques partition the database into a predefined number of clusters. They attempt to determine k partitions that optimize a certain criterion function. The partition clustering algorithms are of two types: k-means algorithms and K-Medoid algorithms. Partitioning algorithms construct partitions of a database on N objects into a set of k clusters. The construction involves

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Vol. 2, Issue 7, July 2015

determining the optimal partition with respect to an objective function. There is approximately kN/k! Ways of partitioning a set of N data points into k subsets.

K-Medoid Algorithm

The basic strategy of K-Medoids clustering algorithms is to find k clusters in objects by first arbitrarily finding a representative object for each cluster. Each remaining object is clustered with the Medoid to which it is the most similar. K-Medoid method uses representative objects as reference points instead of taking the mean value of the objects in each cluster. The algorithm takes the input parameter k, the number if clusters to be partitioned among a set of n objects. A typical K-Medoids algorithm for partitioning based on Medoid or central objects is as follows:

A Typical K-Medoid Algorithm

- Use real object to represent the cluster
- Select k representative objects arbitrarily repeat
- ➢ Assign each remaining object to the cluster of the nearest Medoid
- ➢ Randomly select a non Medoid object
- ➢ Compute the total cost, S, of swapping o_j with o_{random}
- ➢ If S<0 then swap o_j with o random
- ➢ Until there is no change

Build phase

1. Choose k entities to become the Medoids, or in case these entities were provided use them as the Medoids;
2. Calculate the dissimilarity matrix if it was not informed;
3. Assign every entity to its closest Medoid;

Swap phase

4. For each cluster search if any of the entities of the cluster lower the average dissimilarity coefficient, if it does select the entity that lower the most this coefficient as the Medoid for this cluster;
5. If at least the Medoid from one cluster has changed go to (3), else end the algorithm.

The Loop Hole in the Classical K-Medoid Algorithm

The original K-Medoid algorithm stops working when the previously calculated cost is lesser than the currently calculated cost. This algorithm is not checking whether the non-selected Non-Medoids may also be the best Medoid and can hold the data point with minimum cost. Because it stops the iteration, when it find the minimum cost from the currently calculated costs.

Enhanced K-Medoid Algorithm

This new enhanced K-Medoid algorithm assumes all the data points as Medoids and calculates the costs for individual points. After calculating the total cost of all the data points it specifies the number of clusters in which the original data to be grouped. Since K-Medoid algorithm is an unsupervised algorithm, we specify the number of clusters. The Medoids are selected from the data points in which that data point scored the least minimum cost. For example we need ten clusters, the first 10 least minimum cost points are selected as Medoids. This algorithm overcomes the problem of possibility to check all the data points as Medoids. Manhattan distance metric is used to calculate the distance between the cluster points.

Attributes in Banking Domain

The following are the attributes used in the algorithm related to banking domain.
Serial Number, Name, Sex, Account Number, Type Of Account, Date Of Birth, Age, Type Of Deposits, Cheque Book Availability, Deposit Name, Father Name, Permanent Address, Official Address, Qualification, Profession, Designation, Monthly Income, Email Id, Pan Number, Movable Assets, Immovable Assets, Mobile Number, Availed Loans.

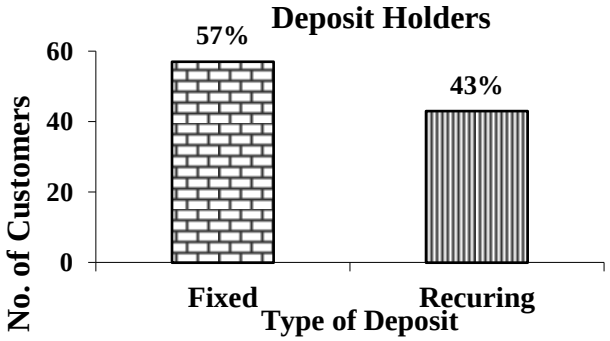
4. RESULTS AND DISCUSSIONS

Analysis of Clustering In Banking Domain

The banking domain has the large amount of data to be presented as useful information. But there are no systematic techniques to produce the information. So, we apply the data mining techniques to mine the data related to customers in order to provide the new mined information to the banker. Most of the banks have the customer related information as a whole. Here clustering one of the data mining techniques can be used to group the customer related information into different clusters. Each cluster represents the similar group of customers. For example in our application, the customers are clustered according to their profession, i.e. the professions of the customers are categorized into three groups. That is, Employee's of Government organizations, Employee's of Non-Government Organizations and Individual's of Non-Resident of India are clustered, the value of monthly salary is greater than or equal to fifty thousand, monthly salary less than or equal to fifty thousand, the value of movable assets is greater than five lakhs, the value of immovable assets is greater than ten lakhs etc. can be clustered separately to know about the economic status of the customers. There are many attributes that can be used as parameters to cluster the customer related data. Some of the sample attributes are mentioned above. In this implementation the data is clustered by using record number as a parameter with the help of the enhanced K-Medoid clustering algorithm.

Graphical View of Customer's Data

In banking system, there are two types of deposits, i.e. fixed and recurring deposits. The following graph represents the number of customers who have fixed and recurring deposits in percentage.



Graph 4.1 Percentage of fixed and recurring deposits

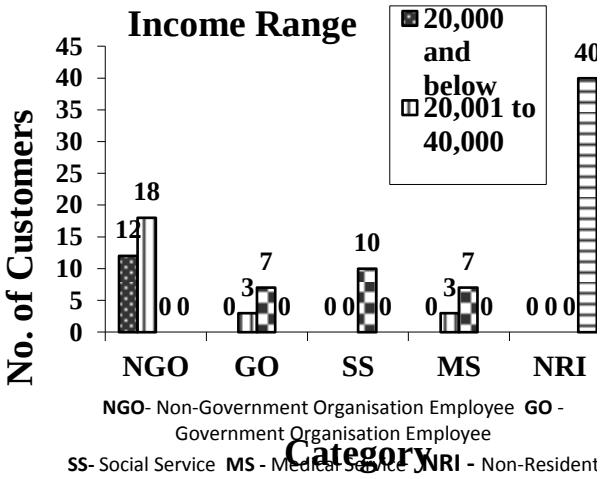
ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Vol. 2, Issue 7, July 2015

The following graph represents the salary of the customers who are working in Teaching NGO (Non-Government Organizations), Teaching GO (Government Organizations), Social Service, Medical

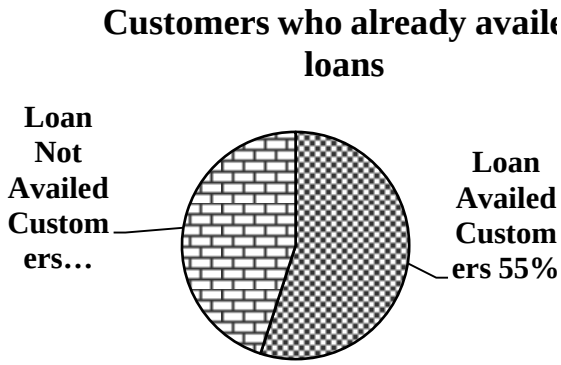
The existing customers of the bank may have availed loans. The following chart represents the number of customers who have already availed loans in percentage.

Findings Based On Clusters in Banking Sector

This algorithm will cluster any kind of numeric data. Since, the data can be clustered according to all attributes. The clustering algorithm clusters the bank data according to the customer record number.



Graph 4.2 No. of customers who have the salary in ranges



Graph 4.3 Percentage of customers who already availed loans

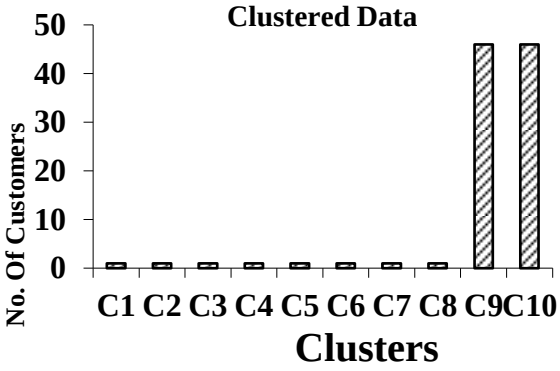
The hundred record numbers are given as input and partitioned into ten clusters. The result is shown below. In that **C** represents the Clusters.

Objectives Obtained By Applying K-Medoid

When K-Medoid algorithm is applied for banking sector, the banker will definitely benefited because the banker will not have this kinds of clusters to know about the characteristics of his customers. The clusters will be helpful to know about the taste and preference of the particular group of customers. This result will be helpful when the banker need to mine the customer related data at the time of starting the new schemes. i.e. it will provide the required information to the banker, so that the banker will know that what are the areas he should concentrate more when his product should be reachable to all the customers. When the customer is approaching the banker for loan, the banker can easily determine the goodwill of the customer, whether the customer will definitely repay the amount in given durations or not. When the new customer is entered into the bank, he will fall to any one of the cluster according to the details he furnish at the time of account creation. When continuing the other transactions, the bank will come to know that this customer falls in which cluster.

Clus. No.	M	DATA POINTS							
C1	50	75							
C2	51	25							
C3	49	26							
C4	52	76							
C5	48	74							
C6	53	24							
C7	47	27							
C8	54	77							
C9	46	50	51	49	52	48	53	47	54
		46	55	45	56	44	57	43	58
		42	59	41	60	40	61	39	62
		38	63	37	64	36	65	35	66
		67	34	33	68	32	69	31	70
		30	71	29	72	28	73		
C10	55	23	78	22	79	21	80	20	81
		19	82	18	83	17	84	16	85
		15	86	14	87	13	88	12	89
		11	90	10	91	9	92	8	93
		7	94	6	95	5	96	4	97
		3	98	2	99	1	100		

According to the cluster the customer is assigned the bank will decide that the policies to be followed with the customers.



Graph 4.4 Clusters obtained based on the application of K-Medoid algorithm

5. CONCLUSION

The limitations with the existing K-Medoid clustering algorithm is found through the literature review and enhanced through this thesis. It is applied for banking sector and found the results are best suite. This algorithm clusters the bank customer details with one parameter. In future, all

ISSN (Online) 2393-8021
ISSN (Print) 2394-1588
International Advanced Research Journal in Science, Engineering and Technology
Vol. 2, Issue 7, July 2015

the attributes concerned for sanctioning the loan will be considered as a parameter for this algorithm and it will be implemented in the banking sector.

REFERENCES

1. Dileep B. Desai and Dr. R. V. Kulkarni, "A Review: Application of Data Mining Tools in CRM for Selected Banks, (IJCSIT) International Journal of Computer Science and Information Technologies, Vol. 4 (2) , 2013, 199 – 201, ISSN: 0975-9646
2. Aneta Hryckiewicz, and Lukasz Kozlowski, "Banking business models and the nature of financial crisis", (SSRN) Social Science Research Network, SSRN-id2601325.
3. Goran Radonic, "A review of business Intelligence approaches to key business factors in banking" published in journal of knowledge management practice, Volume 8, SI 1 May 2007.
4. Waminee Niyagas, Anongnart Srivihod and Sukumal Kittisin, "Clustering e-banking customers using data mining and marketing segmentation", published by department of computer science, faculty of science, kesetsart university, Bangkok 10900, Thailand.