

Customer Segmentation of Bank based on Data Mining – Security Value based Heuristic Approach as a Replacement to K-means Segmentation

Shashidhar HV
Asst. Professor, School of Computing
SASTRA University
Tamil Nadu, India

Subramanian Varadarajan
Student, School of Computing
SASTRA University
Tamil Nadu, India

ABSTRACT

K-means segmentation algorithm can be applied to Customer Segmentation in Banks. If loan over-due amount of bank customers are normally distributed, then K-means can be used. In cases of significant outliers, K-means segmentation algorithm cannot be applied. In our proposed solution, bank loan customers are segmented based on security value and loan over-due amount. Proposed solution addresses segmentation issues on outliers and provides security value based heuristic approach as a replacement to K-means segmentation.

General Terms

Real-Time application of Data Mining

Keywords

Customer Segmentation, K-means outliers, Data Mining

1. INTRODUCTION

1.1 Banking Concepts

How does a Bank operate? Customers deposit their savings in banks for interests at 3%-9% and banks use this amount for giving loans at a higher rate of interest 8%-18%. Banks can run profitably only when customers pay on time (with interest monthly). Else banks will go in debt. So recovery must be high [1] [2] so that banks can thrive well.

Bank income can be categorized into interest income and non-interest income. Most income is via interest i.e. loans, deposits etc. Rest (~10%) of income can be grouped under non-interest products like rent on lockers, sale of gold coins etc. A bank's income is largely depends on loan recovery.

For securing a loan from a bank, a customer has to pledge / hypothecate his / her movable or immovable property as security [3]. After customer repays loan with interest, he/she gets back title of movable or immovable property.

Bank customers with a loan can be segmented based on loan over-due amount and security value. Customer segmentation parameters is from [4] [5] [6] and is modified according to this paper.

1. High risk customer
2. Medium risk customer and
3. Low risk customer

1.2 Data Mining

Data Mining is the process of analyzing large volumes of data and extracting important, useful information from them. Generally large volumes of data are stored in a database. Management use reporting developed from databases for Decision Making, improving performance by analyzing past successes / drawbacks and many more. It is also called as Knowledge Discovery in Database (KDD) [4].

Some of the activities involved in Data Mining is from [7] and are modified according to this paper.

1. *Segmentation*: In segmentation, data is analyzed and grouped. Grouping is done based on similarity (i.e.) each data in a group are similar to each other. E.g. loan customers groups are High, medium and low.
2. *Clustering*: There is no predefinition of groups. Data defines each cluster. Clustering is an unsupervised mining technology. Here each cluster contains similar data.
3. *Prediction*: In prediction, using past and present data, predicted a forecast of dependent variable is developed.
4. *Estimation*: Continuous valued outcome are dealt by estimation mining technology. Each and every record is scanned and data set is formed. Applying estimation to the data set to come up with outcome for continuous variable.
5. *Affinity Rules*: It deals with finding relationship among data. Using association rules, Model is developed which can identify the type of data associations. There is no guarantee that the same association rule will apply for future.

Data Mining follows two styles [8]

1. *Directed*: It is top-down approach. It comes under predictive modeling. In this approach, we do focus is not on how the model was designed, what the model is doing or how the process in the model is executed. Focus is on *most result* possible. Hence the model described as a "black box". Only the best prediction is needed.
2. *Undirected*: It follows bottom-up approach. Each data in this model speaks for itself. It finds pattern and

leaves the choice to the user, whether the resultant pattern is valid or not. In this approach, working of the model is taken care. To understand about the data, we need to know about the model. Hence the model is a semi-transparent box.

2. EXISTING SYSTEM

The existing system in bank loan customers' segmentation is based on loan over-due amount using K-means.

2.1 K-means Segmentation

Consider there are N- loan customers in bank, where N is very large. They need to be segmented into high, medium and low risk customers based on their loan over-due amount.

K-means segmentation is an iterative process. Initially, we must decide on number of customer segments. Let's consider a case where we segment customers into two categories.

First, we need to initialize mean values. Number of mean values needed is equal to the number of segments. Mean values $m_1, m_2 \dots m_n$ takes first K input values, where n, k is number of segments.

Algorithm:

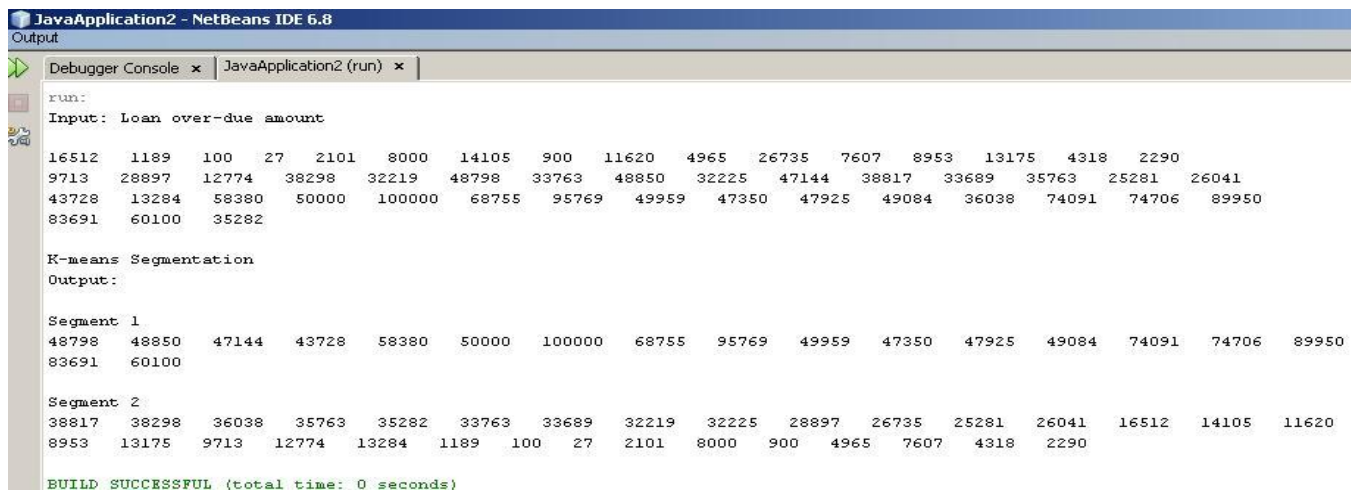
1. Assign first k loan over-due amount (here $k=2$) to mean values m_1, m_2 .
2. Assign each loan over-due amount to that segment which is having its nearest mean.
3. Calculate new mean for n segments (here $n=2$) until convergence criteria are met.

Convergence Criteria: Old mean and new mean of each segment are analyzed. If the two values for all segments are equal, then convergence criteria are met.

There are three ways of assigning initial mean values

1. First k loan over-due values
2. Randomly take k loan over-due values
3. Arbitrarily assigned

Algorithm is from [4] [5] [9] [10] and is changed according to this paper. In this paper, first way of assigning mean value is followed.



```
JavaApplication2 - NetBeans IDE 6.8
Output
Debugger Console x | JavaApplication2 (run) x |
run:
Input: Loan over-due amount
16512 1189 100 27 2101 8000 14105 900 11620 4965 26735 7607 8953 13175 4318 2290
9713 28897 12774 38298 32219 48798 33763 48850 32225 47144 38817 33689 35763 25281 26041
43728 13284 58380 50000 100000 68755 95769 49959 47350 47925 49084 36038 74091 74706 89950
83691 60100 35282

K-means Segmentation
Output:

Segment 1
48798 48850 47144 43728 58380 50000 100000 68755 95769 49959 47350 47925 49084 74091 74706 89950
83691 60100

Segment 2
38817 38298 36038 35763 35282 33763 33689 32219 32225 28897 26735 25281 26041 16512 14105 11620
8953 13175 9713 12774 13284 1189 100 27 2101 8000 900 4965 7607 4318 2290

BUILD SUCCESSFUL (total time: 0 seconds)
```

Figure 1: K-means Customer Segmentation (1)

First case: Let's assume that the loan over-due amount is normally distributed.

Figure 1 shows that loan customers are segmented adequately based on loan over-due amount.

In the above figure 1, customer's loan over-due amount varies normally from single digit to lakhs. Since the distribution is normal, loan over-due amount are segmented adequately.

Segment 1 contains customers' loan over-due amount ranging from 48thousand to lakhs and they belong to high risk customers group.

Segment 2 contains customers' loan over-due amount ranging from single digit to 48thousand and they belong to low risk customer group.

Thus, when bank customers' loan over-due amount is normally distributed, K-means can be used to segment customers.

```

JavaApplication2 - NetBeans IDE 6.8
Output - JavaApplication2 (run) #2

run:
Input: Loan over-due amount

3617 7100 100 27 22472 8000 18195 900 22914 28532 18256 27611 23104 17609 28438 18099
23855 19297 14995 33709 37836 40973 40267 40418 35227 39162 43299 42022 33999 798360 670151
525347 805481 589047 50000 100000 60000000 521578 581581 739968 570862 926681 549633 860339 710386
943835 705691 523207

K-means Segmentation
Output: K-Means Defect

Segment 1
3617 100 27 900 7100 22472 8000 18195 22914 28532 18256 27611 23104 17609 28438 18099
23855 19297 14995 33709 37836 40973 40267 40418 35227 39162 43299 42022 33999 798360 670151
525347 805481 589047 50000 100000 521578 581581 739968 570862 549633 710386 705691 523207 929177 926681
860339 943835

Segment 2
60000000
BUILD SUCCESSFUL (total time: 0 seconds)

```

Figure 2: K-means Customer Segmentation (2)

2.2 Issues found in K-means Segmentation on Outliers

Second case: Let's assume that a few customers have a very large loan over-due amount when compared to other customers (Outliers).

Figure 2 shows that loan customers are not segmented properly. In the above figure 2, outliers exist i.e. customers' loan over-due amount varies from single digit to lakhs and very few in crores.

Segment 2 contains outlier i.e. 6 crores and Segment 1 contains rest of the input values, which means no optimal segmentation takes place

3. PROPOSED SYSTEM

In the proposed system, bank loan customers are segmented based on security value and loan over-due amount.

3.1 Customer Segmentation

Let's assume N- loan customers in bank, where N is very large. To segment customers into high risk, medium risk, and low risk customers, we propose - developing segmentation using security value and loan over-due amount.

Thus, when very few customers (Outliers) have very large loan over-due amounts compared to other customers, then optimal segments are not generated

This is a practical real-time issue which could arise when loan customers of bank are segmented using K-means algorithm. When over-due loan amount is not evenly distributed (Outliers), segments generated using K-means could be unreliable.

New proposed system addresses segmentation issues on outliers in the previous case and provides optimum segments

3.1.1 Loan Customer Database

Customer Database is from [9] and modified according to loan customers. Customer Database consists of following tables: customer profile, customer identifications, mode of operation, products, rate of interest, account opening, loan master, transaction, codes.

ID	cif_no	cust_name	loan_acc_no	product_name	security_value	due_loan_amt
1	1001	robert	hla001	homeloan	55000	50000
2	1002	brito	sal001	salaryloan	250000	1000000
3	1003	judith	vls001	vehicleloan	40000	50000
4	1004	jaleel	mtl001	medium term loan	350000	300000
5	1005	govind	mtl002	medium term loan	50000	70000
6	1006	rangan	hla002	homeloan	10000	26000
7	1007	suman	ret001	retail loan	50000	100000
8	1008	ajay	etl001	educational loan	10000000	1000000
9	1009	nirmal	sal002	salaryloan	100000	75000
10	1010	antony	vls002	vehicleloan	500000	70000

Figure 3: Customers Database

Loan master table is vital to customer segmentation. Loan master table contains following fields: Customer information file no., Customer name, Loan account no., Product name, Security value, Loan over-due amount.

3.1.2 Algorithm

Customers are segmented into high risk, low risk and medium risk. Instead of taking only the loan over-due amount, we can factor both security value and loan over-due amount into consideration for customer segmentation. This will make the customer segmentation more effective.

Analysis & Grouping:

Customer loan over-due amount, security amount are retrieved from the database.

If $security_value > (2 * due_loan_amt)$, then customer is low risk.

Customer Database is given above (Figure 3).

Figure 3 shows all the fields and first 10 loan customer records in loan master table.

Else if $security_value < due_loan_amt$, then customer is high risk.

Else if $security_value \leq (2 * due_loan_amt)$ and $security_value \geq due_loan_amt$, then customer is medium risk.

Repeat the above analysis for all loan customers and group them. Segment 1: High Risk, Segment 2: Low Risk, Segment 3: Medium Risk

Update the Database with creation of high risk, low risk, and medium risk tables.

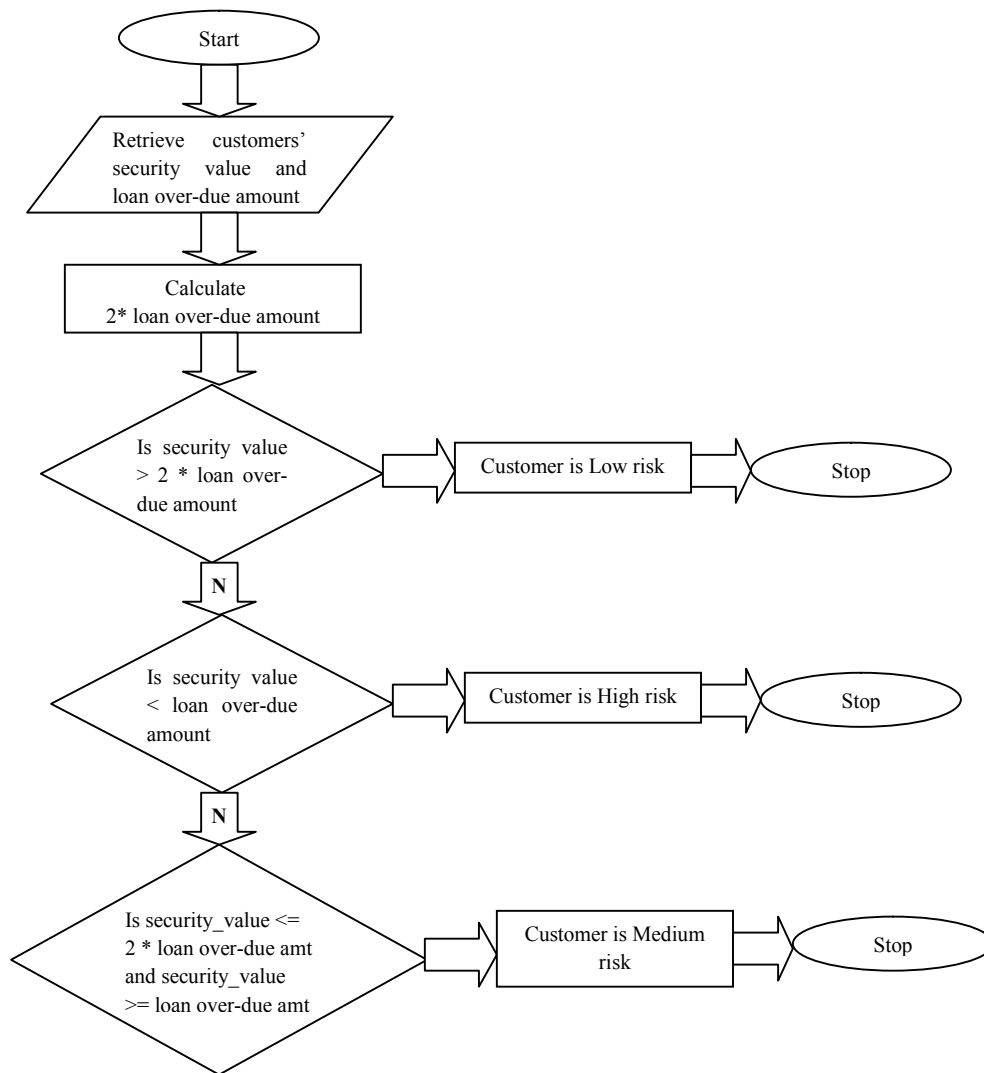


Figure 4: Flow Chart – Analysis and Grouping for single customer

Figure 4 describes pictorially/graphically, analysis and grouping process for single customer in proposed segmentation algorithm.

Tables
accounts_master
codes
customer_identification
customer_profile
high_risk
low_risk
medium_risk
mode_of_operation
products
rate_of_interest
transaction

Figure 5: Customer Database tables (after Segmentation)

Figure 5 shows newly created 3 tables: high risk, low risk, medium risk and all the other tables in customer Database. Vital tables are highlighted.

cif_no	cust_name	loan_acc_no
1002	brito	sal001
1003	judith	vls001
1005	govind	mtl002
1006	rangan	hla002
1007	suman	ret002
1011	sangili	amt001
1019	p a construction	mtl003
1020	xyz enterprise	ssi001
1023	agarwal	hla007
1027	sky fabricators	ssi004
1032	sundar	hla009
1033	sunil	amt005

Figure 6: High Risk customers (Group 1)

cif_no	cust_name	loan_acc_no
1008	ajay	etl001
1010	antony	vls002
1012	muniyandi	amt002
1014	shanmugam	etl002
1016	srinivasan	hla003
1018	jacob	hla005
1021	mannar co	ssi003
1022	roy gupta	hla006
1024	rasi and co	ssi004
1026	manikandan	hla008
1028	kulkarni	amt004
1031	bhel	lsi004

Figure 7: Low Risk customers (Group 2)

cif_no	cust_name	loan_acc_no
1001	robert	hla001
1004	jaleel	mtl001
1009	nirmal	sal002
1013	abc industry	ssi001
1015	ekambaram	etl003
1017	ravi	hla004
1025	durga society	amt003
1029	ss matric school	mtl004
1030	zion industry	ssi005
1034	rainbow industry	lsi002
1035	raja	hla010
1036	saradas textile	ssi005

Figure 8: Medium Risk customers (Group 3)

Customer Database is created with significant outliers i.e. customers' loan over-due amount varies from single digit to lakhs and few in crores.

Figure 6 shows the content in high risk customer table - Customer segmented group 1

Customer Database shows that loan over-due amount is widely distributed i.e. Outlier is taken into consideration.

Figure 7 shows the content in low risk customer table - Customer segmented group 2

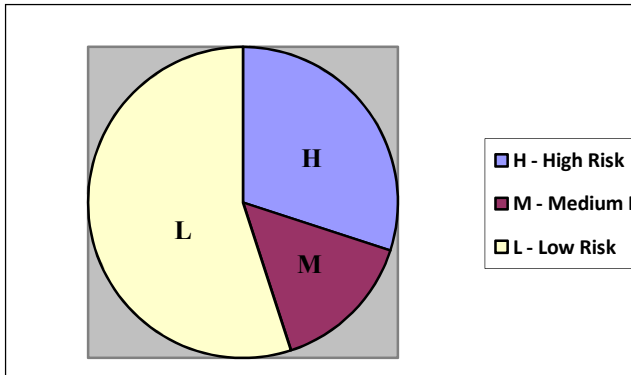


Figure 9: Pie Chart: Segmented Customers

Figure 8 shows the content in medium risk customer table - Customer segmented group 3

Figure 9 describes segmented customers - pictorially/graphically Security based heuristic approach of customer segmentation, will result 3 tables as shown above. High risk, low risk and medium risk tables shows that loan over-due amount are segmented adequately. Each data in a particular group i.e. all data in high risk/ low risk/ medium risk group are similar to each other i.e. optimum segments are generated.

Thus, when very few customers (Outliers) have very large loan over-due amounts compared to other customers, security based heuristic can be used to segment customers'.

In this algorithm, security value and loan over-due amount are used to segment customers effectively. This system answers the issue found in k-means segmentation on outliers. Proposed system will thus segment customers even when outliers exists i.e. when loan over-due amount are unevenly distributed. This system will give optimum perfection in customer segmentation.

4. CONCLUSION

The proposed system segments customers effectively based on security value and loan over-due amount. The security value based heuristic addresses the outlier issue with K-means segmentation. Another approach to overcome the issue on outliers is a) To separate the outliers from master table and generate K-means segmentation on remaining data b) Our proposed security value based heuristic can be applied on the outlier population, thus segmenting the overall population in two steps. Future work may concentrate on customer segmentation

based on non-performing asset and credit risk factors to better segment customers.

5. REFERENCES

- [1] S. Peter and Rose, "Commercial Bank Management," 4th, Liu Z.Y. Translator, China Machine Press, Beijing, China, 2001.
- [2] Xiaohua Hu, "A Data Mining Approach for Retailing Bank Customer Attrition Analysis," Kluwer, Applied Intelligence 22, pp47-60, 2005
- [3] Lu Haiyan, Fu YingLiang and Xing Cuifang, "Data Warehouse in the banking customer relationship management application," Journal of Dalian Maritime University vol.33, Jun. 2007, pp.181-186.
- [4] Shuxia Ren, Qiming Sun and Yuguang Shi, "Customer Segmentation of Bank Based on Data Warehouse and Data Mining," IEEE, 2010.
- [5] Wang Yinghui, "Improvement Research of Customer Segmentation in Knowledge Intensive Business Services" IEEE, 2010.
- [6] Ganglong Duan, Zhiwen Huang and Jianren Wang, "Extreme Learning Machine for Bank Clients Classification," IEEE, 2009.
- [7] Zhang Rong, "An application of Data warehouse and data mining in customer relationship management of bank," Computer and Information Technology, Feb. 2006, pp.79-81.
- [8] Zhang Guozheng, "Customer segmentation in customer relationship management based on data mining," Commercial Research, vol.345, 2006(13), pp153-155.
- [9] Wei Li, Xuemei Wu, Yayun Sun and Quanju Zhang, "Credit Card Customer Segmentation and Target Marketing Based on Data Mining," IEEE, 2010.
- [10] Wu, J., and Lin, Z., "Research on customer segmentation model by clustering," ACM International Conference Proceeding Series, 2005.