Proc. of the 1st International Conference on Electrical, Communication and Computer Engineering (ICECCE)
24-25 July 2019, Swat, Pakistan

Churn Prediction in Banking System using K-Means, LOF, and CBLOF

Irfan Ullah
Department of CS & IT Virtual University
Lahore, Pakistan
ms140400343@vu.edu.pk

Hameed Hussain
Department of CS University of Buner
Khyber Pakhtunkhwa, Pakistan
dr.hameed@ubuner.edu.pk

Iftikhar Ali
Department of CS & IT Virtual University
Lahore, Pakistan
iftikhar.ali@vu.edu.pk

Anum Liaquat
Department of CS & IT Virtual University
Lahore, Pakistan
anum.liaquat@vu.edu.pk

***Abstract*— Customer churn prediction helps in identifying those customers who are probable to stop a subscriptions, products or services, and is therefore very essential for any business. Churn predictions can be very valuable for customers retentions, as it helps in predicting customer that are at risks of sendoff. It is more challenging to put forth churn prediction in banking sector, as there are no contractual agreements between a customer and a bank regarding the duration of services. Loss of customers can be very costly as it is very expensive to obtain new customers in this age of competition.**

There are many churn prediction techniques however; K-Means, Local Outlier Factors (LOF) and Cluster-Based Local Outlier Factors (CBLOF) have not been used so far for this purpose. In this research, we apply these techniques for customer churn prediction. The results are evaluated and analyzed using Precision (Pr), Recall (Re) and F1-Measures to justify the efficiency and effectiveness of this research.

Keywords*— *Churn Prediction, Outlier Detection, Local Outlier Factor, Cluster Based Local Outlier Factor.

I. INTRODUCTION

In today busy world, the use of technology is essential in our daily life. One of the major examples is the banking system, which follows the technology deployment. The customers churn prediction has become one of the important problems in banking system all over the world. According to [5] "*survey on banking systems in 2012, 50% of customer, totally change their bank or were plans to switches to other banks. In USA and Canada, customer who changes their bank increase from 38% in 2011 to 45% in 2012*". The detection of churned customers, and gaining a more comprehensive understanding of their behavior is vital for both increasing your company's revenues and strengthening the relationship your brand has with your customers – two top priorities of any business (i.e. success and down). Furthermore, the given significance of customer as the most appreciated resources for banking system, customer's retentions looks to be vital; it is the very basic obligation of any company/organizations.

According to [6], Customers Relationship Management (CRM) is a commercial approach that's objective is safeguarding customer. Banking system that positively applies CRM to their business continuously improved their retention authorities that are the value of lost and gained customers. According to [3], in banking system, obtaining new customer can cost five times more than adequate and recalling present customer.

Researchers in [7], describes that recalling of old customers is more profitable for a bank than obtaining new customers. So, banks now a day requires to shifts their consideration from customers gaining to customers retaining, providing precise churns predictions model, and effectives churner preventions policies as additional customers retaining solutions to stop churners.

Authors of [8] states that the banking system can increase its income up to 85 % by refining the retaining rate up to 5 %. Customer retention in these days seems more significant than earlier.

The aim of the research work is to apply the existing data mining techniques; K-Means, LOF and CBLOF for churn prediction in the banking sector. A detailed comparison of the mentioned techniques (K-Means, LOF, and CBLOF) is carried-out to identify the most effective churn prediction technique in banking system.

The performance of outlier-detection / churn-prediction is generally evaluated by standard measures such as Precision (Pr) and Recall (Re) i.e. Precision (Pr) is the element of retrieve occurrences which are applicable, on the other hands Recall (Re) is the element of retrieved occurrences to all applicable occurrences [18].

The paper is organized as; related work is discussed in section two. Experimental results are depicted in section three. Section four is dedicated to evaluation metrics. Comparison and evaluation are done in section five. Finally, section six conclude the paper.

II. RELATED WORK

The customer churn-prediction/outlier-detection in bank sector tries to indicate the switching of customer from one bank to another. Outliers in customer form are known as churners. There is a little difference between outlier/churner and noise. The noise is unwanted data while outliers/churners are the data of interest. In banking sector, the churner/outlier customers are the customer whose finishes all her/his account and stopover doing business with a particular bank. There are numerous motives for customers to close their accounts e.g. when somebody makes an account for a precise purpose and close it directly

978-1-7281-3825-1/19/$31.00 ©2019 IEEE

after the purposes is achieved, alternatively if someone is moved/transferred to another place and hence closes his/her account. This leaves banks in a situation where they essentially think, which kind of churned customers are probable to distinguish.

Authors of [1] in their research clustered dataset and apply existing data mining techniques for outliers' detection. They also proposed a novel algorithm called Relative Outlier Cluster Factors (ROCF) without top-n parameter, which can automatically be figuring outs the outliers' rates of a datasets via creating the decisions graphs.

According to [2], clusters and outliers are significant data analysis task. They proposed the K-Means with outlier removals (KMOR) algorithm by spreading the K-Mean algorithms to deliver clusters data and outliers detections concurrently. In the KMOR algorithms, three parameters k , n_0 and Υ is use to controls the amounts of outlier, where k is the preferred amount of cluster, while n_0 is the maximums amount of outlier, and Υ parameter are used to categorize normal point and outliers. In generals, whenever the values of Υ increase, the number of outliers will be decrease.

Researcher in [4] in their research, addressed the problems of customer's churn detection in micro-blogs. They intended to apply user made fillings to predict churner customers. In their work, they only focus on the language modelling and assessment features of customer's churner in micro-blogs for tweets and churner pointer demonstration. Furthermore, they exposed that the task is basically different from sentimentality investigation and as such new technique and method needs to be established for targeted-dependents churner detection in micro-blogs. In addition, they presented three types of customer pointers: demographic-pointers, contents-pointers and contexts-churner pointers. The demographic-pointers features are removed from user's profile, whereas contents-pointer and contexts-pointers are removed from the contents of micro-posts and conversation thread correspondingly.

There is a plethora of knowledge about outlier-detection/churn-prediction. According to [9], the outlier detection is used in numerous applications areas such as, industrial damaged detection, image process, loan applications process, weather prediction, marketing and customer segmentation. A-lots of works has been done in the fields of statistics (i.e. Statistical-based algorithms) on the detection of anomaly/outliers. There are many outlier detection / churn-prediction techniques in data mining namely; classification-based, nearest neighbor-based, naïve-Bayes based and decisions tree-based etc. However, as proposed, in this research work, K-Means, LOF and CBLOF will be compared from churn prediction perspective. Therefore, authors would prefer to give concise description of the techniques from the plethora of knowledge.

The K-Mean distance-based method is the popularly use method for outlier's detection. According to, [10] the K-Means distance-based are calculating the distances between each data items and cluster centers in each iteration could be calculated using linear data structure list. Finally, in the re-calculating cluster center phase, to modify the center vector updating procedure of the basic K-Means that reduces the formation of empty clusters.

According to [11], K-Mean is one-of simple un-supervised machine learnings algorithm, which explain the distinguished clusters problems very efficiently. In addition, K-Means clustering does not need calculation of all-possible pair wise distances of circumstances and only needs loops step of computing centroids of new cluster and reallocating cases to neighboring cluster. Therefore, it is straightforwardly appropriate to very huge datasets and is broadly used in data mining (DM). In short, in choosing initial k-centroids phase the early clusters center have obtain by the use of divide and conquers method see, [11]. Here, similarity of the sample to the cluster center, K-Mean has unsuccessful when cluster are of different sizing, density, non-globules shape this entire problem are well solve through Local Outlier Factors (LOF).

Zeng you He et al. [13] proposed LOF. The LOF is also as an un-supervised algorithm that detects the outlier's trough local-density deviations of a given data-points with respects to his neighbors. It reflects as outlier's sample that have a significantly low-density than their neighbors i.e. see Fig 1.

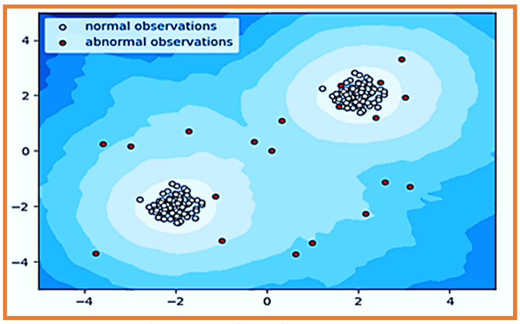

Fig 1: Local Outliers Factors

Furthermore, Zeng you He et al. [13] provided some strategies for selecting the limits of the neighborhood sizes ranges. Furthermore, the LOF algorithm can be broken down into four parts: i.e. (a). K-Distance and p-Neighbors (b). Reachability-Distance (c). Local Reachability Density and (d). Local Outlier Factor calculation. Major drawback of LOF is that, sometimes it detects normal objects as outliers, and vice versa.

Cluster-Based Local Outlier Factors (CBLOF) was proposed by Deng, S et al. [12]. In their work, they used clusters in orders to control dense regions in a given dataset and executes density-estimate for apiece clusters. In theories every clusters-based algorithm, can be use to clusters the dataset in a 1st step. After clustering a dataset, the CBLOF is uses to categorize the resultant clustering into-large and small cluster. Lastly, anomaly/outliers scores are calculated by the distances of each object to its cluster centers multiply by the object belongs to its clusters. For small cluster the distances to the closest' large clusters are used. The main thoughts in the CBLOF detection technique suggested by [13] can be sensible to describe the outlier from the points of views of clustering and classify those object that don't lies in any-larger cluster as outlier.

According to [13] Fig 2 illustrate the concept of CBLOF. The points P lies in the small's cluster C2 and therefore the score would be equals to the distances to C1, which is the near larger clusters multiplied by five which is the size of C2.

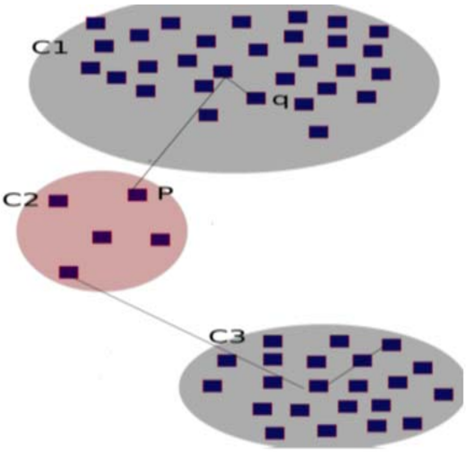

Fig 2: Cluster Based Local Outlier Factor (CBLOF)

III. Churn Prediction Using K-Means, LOF and CBLOF

In the first phase of the experiential analysis, authors apply the banking system rule of dormancy status of customers in Pakistan, As the detection of churner/outliers (e.g. dormant) customer duration is maximum 06 month or 180 days. Furthermore, in this research we detect those customers who have not use their account in last e.g. (no-transactions) in past 06 months and called them churner/outliers. also, we check their duration since last usage of their account and after that we may chose it normal/active and churner/outlier customers. Furthermore, the customer who has closed/change their bank or shutdown account in a particular bank authors called them is outlier customer or churner.

A. K-Means Algorithm:

For outlier's detection we applied the K-Means clustering technique on our dataset through distance-based. this technique shows their prediction mechanisms through distance of each customer to thrashed value of distance = 180.

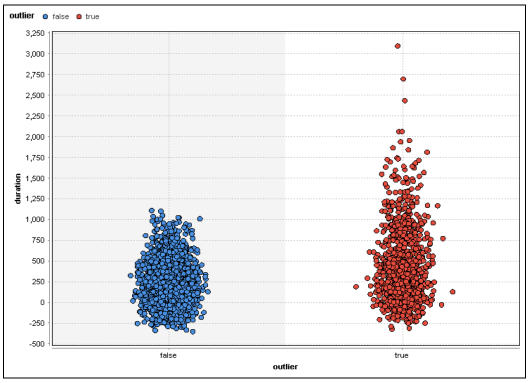

Fig 3: K-Means implies into two clusters for outliers

In Fig: 3 some values are shown in negative due to jitter. The jitter job is very obliging, particularly for datasets which not only contains number but also nominal values.

Also, the number of true outliers/churners is 1160 and normal customer (false outlier) is 3361 see Fig: 4.

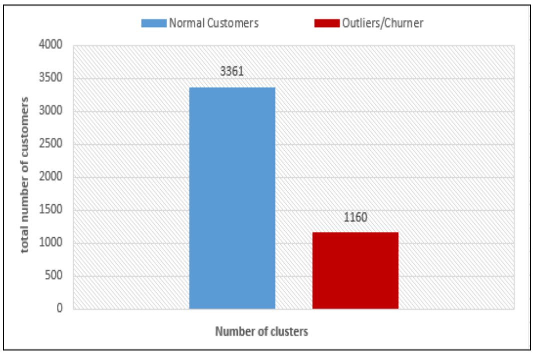

Fig: 4 K-Means detect the outlier and normal customers

In above Fig: 3, we have 02 clusters, the red color indicates true outlier customers, while the blue color indicates normal customer in a particular cluster. Furthermore, in the distance-based approach, similarity between two objects is measured with the help of distance between the two objects in data space, if this distance exceeds a particular threshold, then the data object will be called as the outlier. Finally, we run our K-Means distance-based algorithm and it detects those customers which has long period since no transactions turned as outliers. The K-Means detect the churner/outlier customer on the basis of large distance to its nearest neighbors and fail on those customer detections which has not large distance to his neighbors but in reality, it's churner/outliers, so this problem is well solves through LOF.

B. Local Outlier Factor (LOF) Algorithms:

The local density outliers are a measure of Local Outlier Factor (LOF), which captures the degree of outlier-ness of every object in the data set, to pick up local outliers. Density-based approach is performed by calculating the local-densities of the points beings investigate and the local-density of its nearest neighbors. Therefore, Densities base approaches are usually more effective than the distance-based approaches but it suffers more execution times. Furthermore, authors apply the LOF technique on the same datasets and again obtained results are in two clusters.

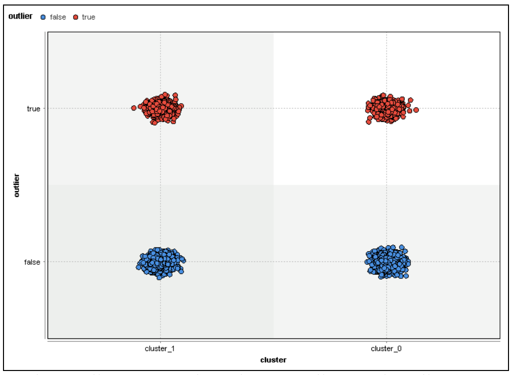

Fig: 5 False outliers (normal) and true outlier (real outliers) customers using LOF.

In above Fig: 5, we have 02 clusters, the blue color indicates the normal customers, while the red color indicates the outlier/churner customer in a particular cluster. Furthermore, in densities-based outlier’s detections, the objects O is an outlier’s if its density is comparatively much low than of its neighbors. Also, the densities-based clusters locate region of high densities that are separate from one another by region of low densities. As a result, if we compare the LOF with K-Means then we can find that the density based is better than the distance-based outlier detection because in distance-based we miss such outlier which is same density like a normal customer see Fig: 5, while LOF detects it very efficiently.

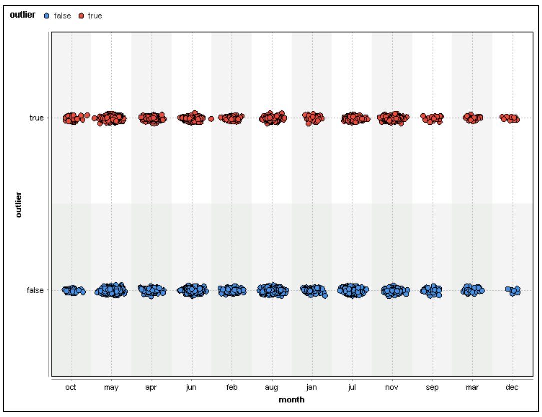

Fig: 6 LOF detect outlier in every month

In Fig: 6, LOF divides banks data set into two clusters e.g. normal (false) and outliers/churner (true). The blue color indicates normal customers in particular period e.g. months. While, red colors indicate outliers/churners in given months. Also, the number of true outliers is 1414 and normal customers are 3107 see Fig: 7.

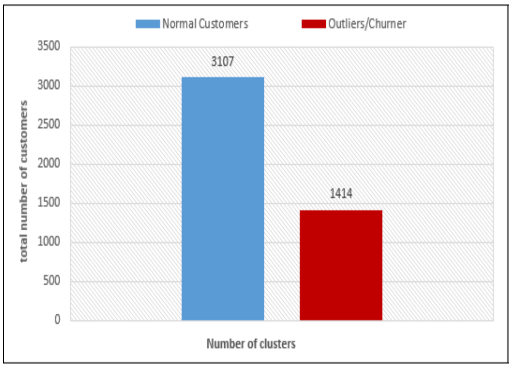

Fig:7 The LOF detect normal and outlier customer

According to [15] The local reachable densities are signs of the densities of the region’s arounds a data point. Furthermore, LOF contest MnPntsdst earlier detentions in term of, larger MnPntsdst corresponding to a sparse-regions (outlier) a small MnPntsdst corresponding to a dense-regions (normal). Also, occasionally outliers object may-be relatively closed to each other’s in the data spaces, starting small group of outliers’ objects. Meanwhile, MnPnts discloses the minimum number of points to-be considers as clusters, if the MnPnts is set too-low, the group of outlier’s objects will be wrongly recognizing as cluster. While, MnPnts is also use to computes the densities of each point so if MnPnts is set too-high, some outlier nearby dense-cluster may be mis-identified as cluster's point by, [19]. Finally, this problem is well solved through CBLOF.

C. Clusters-Based-Local-Outliers Factors (CBLOF):

The clusters-based-local-outliers-factors, detect outlier as point that don’t lies in or positioned faraway separately from any cluster and outlier is a noise of clustered implicit-lies. The clusters-based-local-outliers detections algorithm can detect the outliers clustered an object is an CBLOF if,

- it doesn’t belong to any cluster.
- there is a huge distance amongst the objects and its neighbor cluster.
- it goes to a minor or thin cluster.

Lastly, according [20] outlier’s scores are computing by the distances of an occurrence to its cluster's center multiply by the occurrences belongs to its clusters. For small cluster, the distances to the neighboring big cluster is uses. The process of using the amounts of clustered member as a scaling factor should approximate the locals-densities of the cluster. As a result, CBLOF, cluster the same data set in to following two clusters, see Fig: 8.

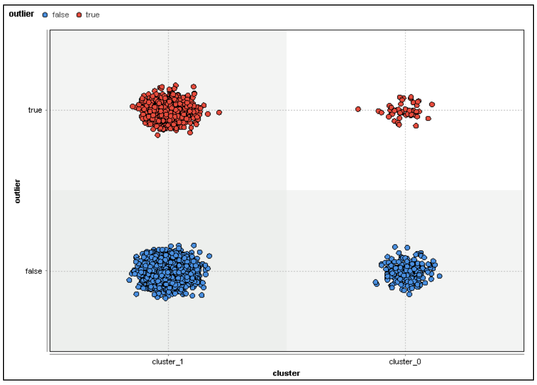

Fig: 8 False outliers (normal) and true outlier (real outliers) customers using CBLOF.

In the above Fig: 8, we have 02 clusters normal (false) and outlier (true), the red color indicates the outlier customers, while the blue color indicates the normal customer in a particular cluster. Finally, The CBLOF technique detects the accurate numbers of churn customer in each cluster against their account since transactions usage. Regarding to less duration or no transaction is taken turning it to true outlier/churn customers.

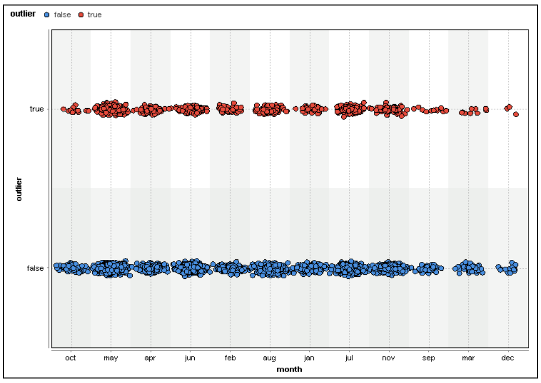

Fig: 9 CBLOF detect outlier in every month

In Fig: 9, bank data set is divided into two clusters e.g. normal(false) and outliers (true). The blue color indicates normal customers in particular period e.g. months. While, red color indicates outliers/churners in given months.
Also, the number of true outliers is 888 and normal customers are 3633 see Fig: 10.

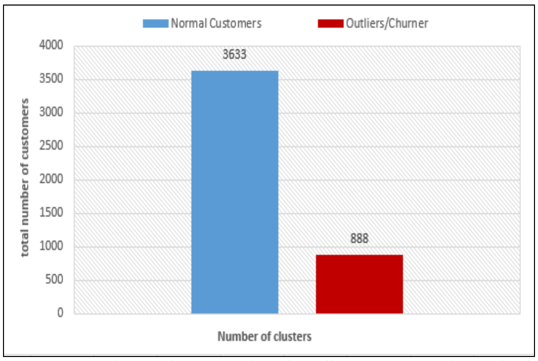

Fig: 10 CBLOF detect the normal and outlier customers in each cluster

IV. Metrics for Measurement

According to, [16] for the evaluation and interpretation steps, authors use Precision (Pr) and Recall (Re) for outlier/churner detections.

Precision (Pr): Precision (Pr) is the element of retrieve churn/outlier customers that are applicable to the query.

$$Pr = \frac{TP}{TP+FP}$$

Recall (Re): Recall (Re) is the element of the churn/outlier customers that are applicable to the query that are successfully retrieve.

$$Re = \frac{TP}{TP+FN}$$

From above confusion matrix we have following definitions:

- True-Positive (TP): Numbers of positive instances predicts correctly.
- False-Negative (FN): Numbers of positive instances incorrectly predicts as negative.
- False-Positives (FP): Numbers of negative instances incorrectly predicts as positive.
- True-Negative (TN): Numbers of negative instances correctly predicted.

Generally, Precision and Recall are opposite of each other. If Recall is high, then Precision will be low and vice versa. F1-Measure is used to combine the result of both Precision and Recall. F1-Measure is the weighted average of Precision and Recall. For this aims, F-1 Measures, a prevalent mixture is usually use as a singleton metric for evaluation churner performance. F-1 Measures, is define as the harmonic-means of Pr and Re such as:

$$F-1\ \text{Measures} = \frac{2\times \Pr\times \text{Re}}{\Pr+\text{Re}}$$

A value closed to one's suggests that a well combine Pr and Re is achieved by the churner customers, as per [18].

V. Comparisons of the K-Means, LOF and CBLOF

The detail comparisons of K-Means, LOF and CBLOF techniques are evaluated by the value of Pr, Re and F-1 Measures see Table 1. Better the values of Pr, Re and F-1 Measures means better is the technique for churn prediction.

Table 1 The Performance of Algorithms (Highest value in bold)

Algorithm	Pr (%)	Re (%)	F-1 M (%)
K-Means	74.36	100	85.30
LOF	68.75	100	81.48
CBOF	**80.38**	100	**89.12**

According to [14], K-Means algorithms is sensitive for outliers' detections also the K-Means is unable to detect correct churner customer because its only detect those customers who has longest distance from the rest of customers. As, authors mention above the churner customer not only have longest distance but it has also a shortest period like somebody open new account for specific purpose when he gets it then close their account immediately. The LOF is also fail due to much time consuming instead of K-Means and CBLOF. Furthermore, the LOF is not detecting accurate results of customers like sometimes it detects raw data as churner/outliers and also does not detect those customers which have same densities like normal customers but she/he is true churner/outlier customers. So, it is also not better for customer churn detection. The CBLOF is good in detecting real outliers, because it gets accurate results instead of LOF and K-Means. Also, for customer churned detection it detects those customers who have same densities like normal customers but change their behavior/switching to churner/outliers (e.g. LOF) and also long distance from its neighbors or long time (e.g. K-Means) since no transaction then it will be very accurately detects trough CBLOF.

VI. Conclusion and Future work

In this research study, authors convert raw customer's data of a bank in to suitable data and then converted this data facts valuable information through Datamining Mechanisms. Authors have taken out the dataset for selected attribute from raw customers bank datasets for an elected set of 4521 customers. We use K-Means, LOF, and CBLOF technique to distinguish significant customers appearances to predicts churner. However, the prediction of churner customer is more important than normal customers because when banks know in advance that this customer is near to switching then banks can offer him some extra bonus for retention.

Comprehensive evaluation of 03 different unsupervised-outlier detection algorithms on bank dataset has been achieved for the first-times. Data Mining goals to passage information and insights through the investigation of larger quantity of data using K-Means, LOF and CBLOF technique. As our concern, the CBLOF algorithm shows on average good performance from other used technique, demonstrating a more outstanding and computes penetrating density estimations is not essentially requires. In term of computational complexities, CBLOF is faster than their nearest neighbors' competitor. Though in rehearsal, we recommend to restarts the fundamental K-Mean several time in orders for obtaining a stable clusters outcome. However, when process speeds are very important or a clustered modeled can be updates in a data flowing applications, a CBLOF might be used. Furthermore, K-Means algorithms is sensitive for outliers' detections also the K-Mean is unable to detect correct churner customer because its only detect those customers who has longest distance from the rest of customer. Also, the LOF does not detect accurate results of customer like sometimes it detects raw data is churner/outliers and also does not detect those customers which have same densities like normal customers but she/he is true churner/outlier customers,

In this research paper authors predict churner customer in bank data set, future study should emphasis on looking for new discerning structures, which might allow to describe the customer behaviors and distinct the 02-customer classed well i.e. one opportunity is to generate multi-months' attribute on short period of times in order to describe customer switching more precisely. Another possibility is to calculate values differences of each attributes over times, rather than using the static values etc.

References

[1]. Huang Jinlong, Zhu Qingsheng," A novel outlier cluster detection algorithm without top-n parameter", Elsevier Science Publishers, Knowledge-Based Systems, Vol. 121, Pages 32-40, 1 April 2017.

[2]. Guojun Gan a, Michael Kwok-Po Ng, "k-means clustering with outlier removal", Elsevier Science Publishers, Pattern Recognition Letters, Vol. 90, Pages 8-14, April 2017.

[3]. Ing. Massimo Ferrari, Dott. Riccardo Panizzolo "Machine learning techniques for customer churn prediction in banking environments", Universit`a degli Studi di Padova, everis Italia S.p.a. 2015-2016. 20, ISSN 0973-4562, 2015.

[4]. Amiri, H., and Daume III, H. Target-dependent churn classification in microblogs. In Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015. Bhasin, M.L. "Data Mining: A Competitive Tool in the Banking and Retail Industries"

[5]. Ernst & Young, NG Data. Predicting & Preventing Banking Customer Churn by Unlocking Big Data. NG Data. Retrieved from http://www.ngdata.com/ , October 15, 2014.

[6]. Aurélie Lemmens & Sunil Gupta, managing churn to maximize profits, 2013.

[7]. Nguyen, E. H. Customer Churn Prediction for the Icelandic Mobile Telephony Market. 60 ECTS Thesis, University of Iceland, Engineering & Natural Sciences, Sigillum 2011.

[8]. Nie G, Rowe W, Zhang L, Tian Y, Shi Y Credit card churn forecasting by logistic regression and decision tree. Expert Syst Appl 38:15273–15285, 2011.

[9]. Moh‟dBelal Al-Zoubi "An Effective Clustering-Based Approach for Outlier Detection" European Journal of Scientific Research, Vol. 28, No.2, 2009.

[10]. J.JamesManoharan and Dr.S.Hari Ganesh, "Initialization of optimized K-means Centroids using Divide-and-Conquer Method", ARPN Journal of Engineering and Applied Sciences", Vol. 11, No. 2, ISSN 1819-6608, January 2016.

[11]. J.JamesManoharan and Dr.S.Hari Ganesh ,"Improved K-means Clustering Algorithm using Linear Data Structure List to Enhance the Efficiency", International Journal of Applied Engineering Research, Vol. 10, No

[12]. He, Z., Xu, X., Deng, S.: Discovering cluster-based local outliers. Pattern Recogn. Lett. 24 (9–10), 1641–1650, 2003.

[13]. Zengyou He, XiaofeiXu, and Shengchun Deng. Discovering cluster-based local outliers. Pattern Recognition Letters, 24(9-10):1641 {1650, 2003.

[14]. Class 13. Unsupervised learning Clustering. Shimon Ullman + Tomaso Poggio Danny Harari + Daneil Zysman + Darren Seibert

[15]. Breunig, M.M., Kriegel, H.P., and Ng, R.T., "LOF: Identifying densitybased local outliers.", ACM Conference Proceedings, 2000, pp. 93-104.

[16]. Z.P. Zhang, Y.X. Liang, A data stream outlier detection algorithm based on reverse k nearest neighbors, Advanced Materials Research., Trans Tech Publ., 2011.

[17]. Charu C. Aggarwal. 2016. Recommender Systems – The Textbook. Springer.

[18]. T. Fawcett, An introduction to roc analysis, Pattern Recognition Letters27 (8) (2006) 861–874.

[19]. Mohammad Zaid Pasha & Nitin Umesh, "A Comparative Study on Outlier Detection Techniques" International Journal of Computer Applications (0975 – 8887) Volume 66– No.24, (March 2013).

[20]. Markus Goldstein and Seiichi Uchida, "A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data", doi: 10.1371/journal.pone.0152173, (Apr19, 2016).