International Journal of Computational Intelligence in Control

Using K-Means, LOF, and CBLOF as Prediction Tools

Irfan Ullah[1], Hameed Hussain[2], Shahid Rahman[2], Afzal Rahman[2], Muhammad Shabir[2], Niamat Ullah[2], Kifayat Ullah[3]

[1]Virtual University Lahore, Pakistan, ms140400343@vu.edu.pk

[2]University of Buner, KP, Pakistan, dr.hameed@ubuner.edu.pk, rahmanshahid059@gmail.com, afzal85@uop.edu.pk, shabir_adam@yahoo.com, niamatnaz@gmail.com,

[3]University of Swat, KP, Pakistan, kifayat@uswat.edu.pk

Date of Submission: 25th January 2021 Revised: 7th April 2021 Accepted: 20th June 2021

***Abstract* – Prediction about some disastrous situation occurring can minimize the final loss and very helpful to take some remedial actions in advance. Breast cancer diseases pose a great challenge in which some cells of the body grow abnormally. These cells then destroy alternative surrounding cells and their normal functions. Breast cancer has become the risky type of cancers in the world among women. Detection early of breast cancer is integral for reducing life losses. Similarly, churn prediction also helps in identifying those customers who are probable to stop a subscription, products or services, and is therefore very essential for any business. Loss of customers can be very costly as it is very expensive to obtain new customers in this age of competition.**

There are many breast cancer detection and churn prediction techniques, however; K-Means, Local-Outlier Factors (LOF) and Cluster-Based Local Outlier Factors (CBLOF) have not been used so far for these purposes. This paper aims to apply the aforementioned techniques for the said purposes. The results are evaluated and analyzed via Precision (Pr), Recall (Re) and F1-Measures to justify the efficiency and effectiveness of this research.

Index Terms - Breast Cancer, Churn Prediction, Outlier Detection, Local-Outlier Factors, Cluster-Based Local Outlier Factors.

INTRODUCTION

Breast cancer is one of the most frequent cancers amongst women. Also, the leading cause of death in developing countries as per, International Agency for Research on Cancer (IARC) via GLOBOCAN, 2012, described that the identification of cancer internationally in 2012. They determine a sort of cancer that contributed to the second highest death rate are breast cancer via a proportion of 11.9 % or 1.7 million women’s by [21]. The most effective way to reduce breast cancer deaths is its detection on earlier stages. Early stage breast cancer detection is important in reducing life losses [22]. Early identification needs an accurate and a reliable diagnosis procedure that enables physicians to differentiate caring breast tumors from malignant ones while not going for surgical biopsy. According to [23] the objective of these earlier stage detections is to assign patients to either a” Benign” cluster that is non-cancerous or a” Malignant” cluster that is cancerous. The automated breast cancer diagnosis is a crucial, real world medical problem. Therefore, discovery an effective and precise diagnosis technique is very necessary. In recent years strategies of machine learning, are used in prediction widely, particularly in diagnosis of medical by [24]. In today's 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 of banking systems in 2012, 50% of customer, totally change their bank or were planning to switch 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, given the significance of customer as the most appreciated resources for the banking system, customer retentions look 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. According to [3], in the 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. 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 breast cancer and churn prediction. A detailed comparison of the mentioned techniques (K-Means, LOF, and CBLOF) is carried-out to identify the most effective technique.

The performance of outlier detection / breast cancer and churn prediction is generally evaluated by standard measures such as Precision (Pr), Recall (Re) and F1-Measures [18].

The paper is organized as; related work and existing techniques are discussed in section two. Experimental results of the existing techniques (K-Means, LOF and CBLOF) are depicted in section three. Section four is dedicated to evaluation metrics. Comparison and evaluation are done in section five. Finally, section six concludes the paper.

RELATED WORK

The issue with the medical image detection procedure is that it could not show all the patterns and knowledge for a selected type of cancer or subtypes of cancer. Furthermore, breast cancer / outlier (abnormality) detection in medical terminology helps us to detect the real abnormality (outlier) cancerous area in the breast as well as the whole human body. Similarly, customer churn prediction / outlier detection in banking sector tries to indicate the switching of the 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 the 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 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.

Researchers in [27] apply and compare 3 prediction models on SEER data for breast cancer detection. They have observed that algorithm C4.5 award the best performance concerning the accuracy up to 86.7 %. Furthermore, [28] pre-processed the SEER data to redundancies removing or missed evidence. And also, they compare accuracies of predictive model to the SEER data over 3 modeled of prediction that indicated the decision tree (C5) is the best predicting with accuracy up to 93.6 % on the holdout model.

The authors of [26] concerning prediction apply 3 AI techniques of survival in patients with carcinoma exploitation. And they claim that improvement of medical knowledge impacts on great amounts of information connected with health progressively. The exploitation prediction processing of data became a crucial gadget for the hospital management and medical analysis. Breast cancer data set in their analysis was got from a central Taiwan, regional teaching hospital in between 2002-2009.

According to [25] the AdaBoost algorithm and their analysis used for breast cancer survivability. In the data pre-processing approach to get information on improve medical checkup result, related medical problem, reduce treatment cost, and prioritize clinical studies of patient health. RELIEF algorithm was used in the pre-processing to pick the important attributes, whereas to extract-knowledge from information about patient survivability AdaBoost algorithm breast cancer was used.

Authors of [1] in their research proposed a novel algorithm called Relative Outlier Cluster Factors (ROCF) without top-n parameter, which can automatically figure out the outliers' rates of a dataset 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 γ is used to control the amounts of outlier, where k is the preferred amount of the cluster, while n_0 is the maximum amount of outlier, and γ parameter are used to categorize normal point and outliers. In general, whenever the values of γ increase, the number of outliers will be decreased.

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 and breast cancer detection. According to [9], the outlier detection is used in numerous applications areas such as, industrial damaged detection, image process, loan application process, weather prediction, marketing and customer segmentation. A-lots of work 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 breast cancer and 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 Figure 1.

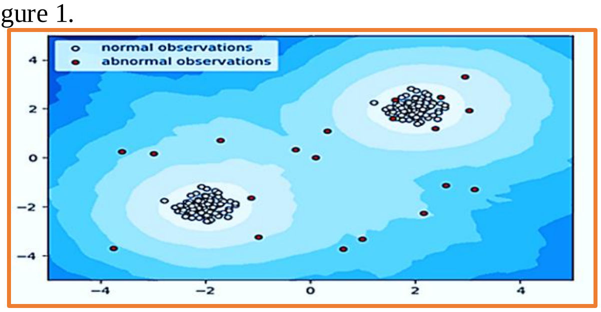

FIGURE 1
LOCAL OUTLIERS FACTORS

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 objects belongs to its clusters. For small cluster the distances to the closest' large clusters is 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] Figure 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.

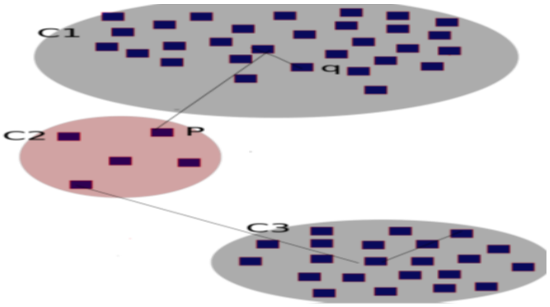

FIGURE 2
CLUSTER BASED LOCAL OUTLIER FACTOR (CBLOF)

RESULT AND DISCUSSIONS

In the first phase of the experiential analysis, we apply the aforementioned techniques for breast cancer detection. Breast cancer with evil cancers have high temperature than healthy breast and even breast with benign cancers. In this research, authors detecting the hottest area of abnormal breast which are the suspected areas. Main symptoms of breast cancers, contains increase in length or change in form of the breast, breast pain, swelling of all or a part of the breast, differences inside the color of the breast skin, a lump within the underneath arm vicinity etc. In the second phase of results and analysis we 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 phase 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 as outlier customer or churner.

I. K-Means Algorithm:

For outlier's detection authors applied the K-Means clustering technique on both datasets. Furthermore, the K-Means algorithm clusters the whole data into two clusters for breast cancer detection as shown below in Figure 3.

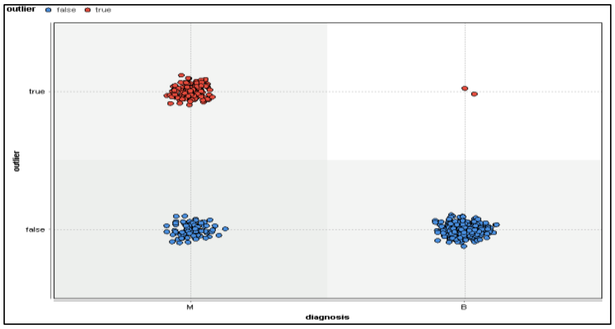

FIGURE 3
BREAST CANCER ANOMALY DETECT BY K-MEANS

In addition, the K-Means technique is also applied for churn prediction on bank dataset. As results shown in Figure 4 have 02 clusters, the red color indicates true outlier customers, while the blue color indicates normal customer in a particular cluster. In the distance-based approach, between two object similarity is measured with the help of distance between the two objects. If this distance surpasses a specific threshold, then the data objects will be termed as the outlier

.

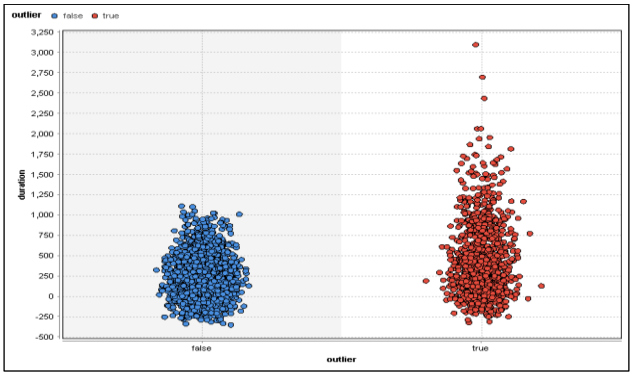

FIGURE 4
K-MEANS IMPLIES INTO TWO CLUSTERS FOR OUTLIERS

In Figure 4 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.

K-Means algorithm splits breast cancer dataset into 2 clusters. Authors get normal (438) and outliers/cancerous patients (131). While, in bank dataset the number of true outliers is 1160 and normal customer (false outlier) is 3361 see Figure 5.

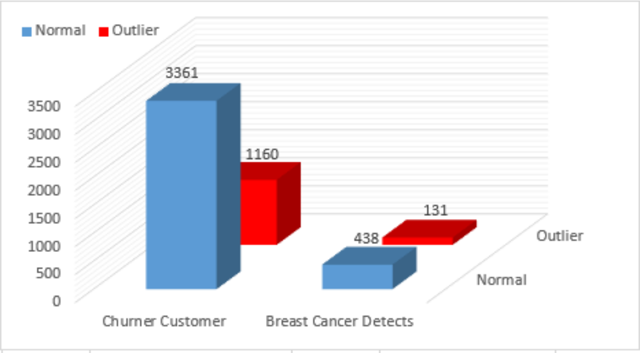

FIGURE 5
K-MEANS APPLY FOR OUTLIER AND NORMAL DETECTION ON BOTH DATASET

Authors run K-Means distance-based algorithm and it detect those patients which have suffer from the daises of breast cancer and also detects those customers which has long period since no transactions.

II. 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 effectives than the distance based approached but it suffers more execution times. Authors apply the LOF technique on the mention two datasets and again the technique divide breast cancer dataset into two clusters as shown in below Figure 6.

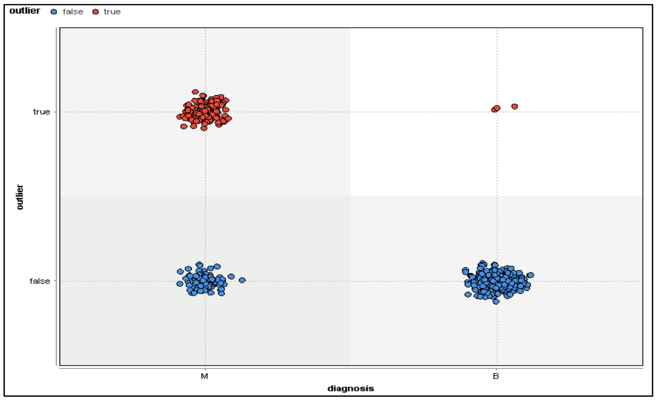

FIGURE 6
BREAST CANCER ABNORMALITY DETECT VIA LOF

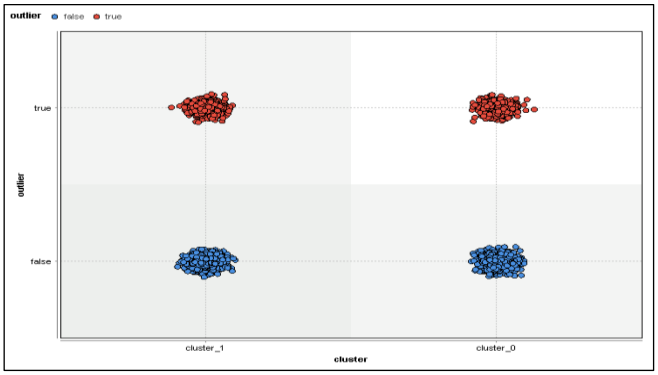

FIGURE 7
FALSE OUTLIERS (NORMAL) AND TRUE OUTLIER (REAL OUTLIERS) CUSTOMERS VIA LOF

In above two Figure 6 and Figure 7, authors obtained 02 clusters, the blue color indicates the normal objects, while the red color indicates the outliers in a particular cluster. Furthermore, in densities-based outlier's detections, the object O is an outlier's if its density is comparatively much low than of its neighbors. Also, the densities bases 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 entity see Figure 6 and Figure 7, while LOF detects it very efficiently.

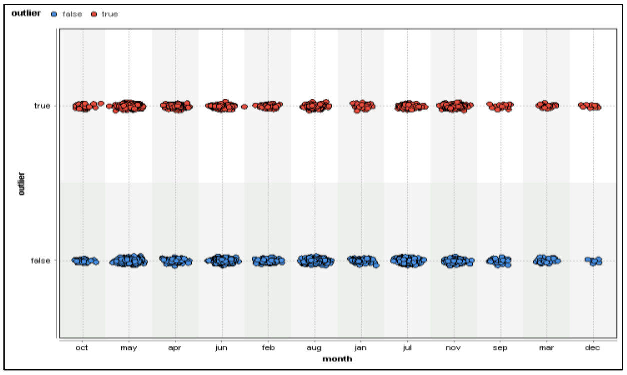

FIGURE 8
LOF DETECT OUTLIER IN EVERY MONTH

In Figure 8, LOF divides banks data sets into two clusters e.g. normal (false) and outliers (true). The blue color indicates normal customers in particular period e.g. months. While, red colors indicate outliers/churners in given months.

Also, the amount of true predicted outlier in breast cancer dataset are 123 while the normal patients are 446 furthermore, true outliers in bank dataset re 1414 and normal customer are 3107 see Figure 9.

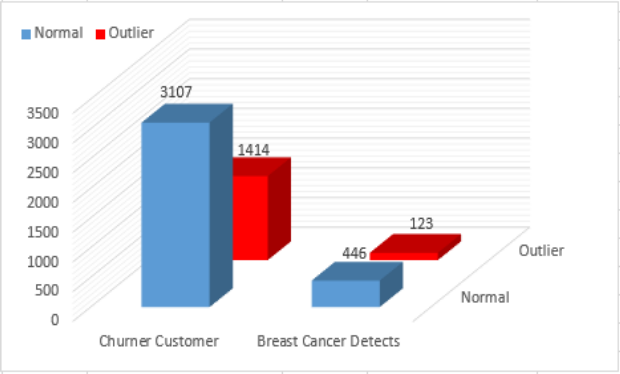

FIGURE 9
THE LOF DETECT NORMAL AND OUTLIER

According to [15] The local reachable densities are signs of the densities of the region's arounds a data point. 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 clusters point by, [19]. Finally, this problem is well solved through CBLOF.

III. 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,

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

Lastly, according [20] outlier's scores are computing by the distances of an occurrence to its clusters 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 Figure 10.

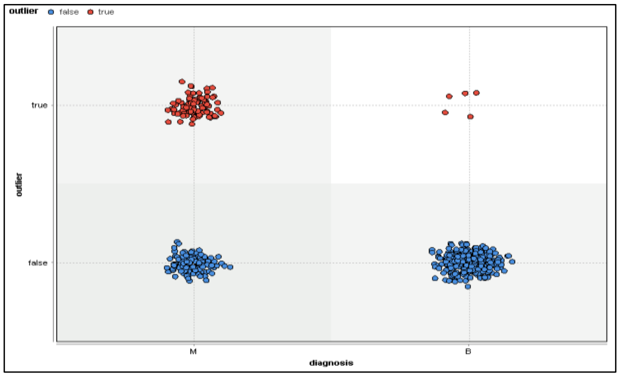

FIGURE 10
BREAST CANCER ABNORMALITY DETECT VIA CBLOF

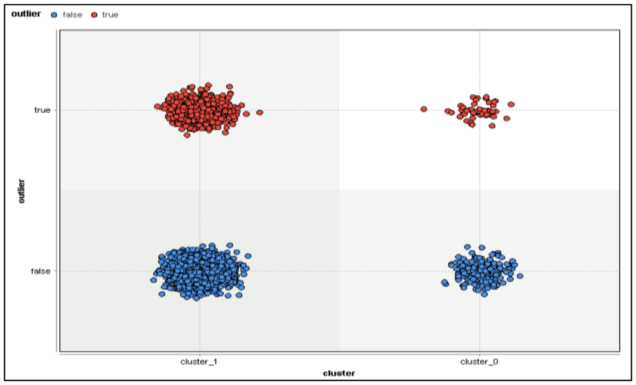

FIGURE 11
FALSE OUTLIERS (NORMAL) AND TRUE OUTLIER (REAL OUTLIERS) CUSTOMERS VIA CBLOF.

In the above two Figure 10 and Figure 11, authors obtained 02 clusters normal (false) and outlier (true), the red color indicates the outlier instances, while the blue color indicates the normal entities in a particular cluster. Finally, The CBLOF technique detects the accurate numbers of outlier/abnormality in each cluster against their medical treatments and accounts since transactions usage.

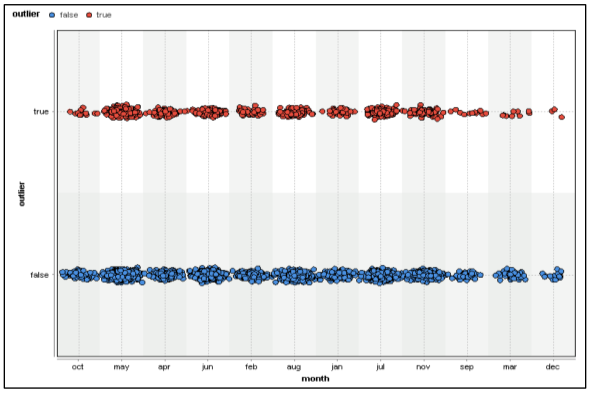

FIGURE 12
CBLOF DETECT OUTLIER IN EVERY MONTH

In Figure 12, 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.

The CBLOF algorithm split breast cancer dataset into 2 clusters and gives outlier (91) and normal (478) respectively. While on bank dataset the number of true outliers is 888 and normal customer are 3633 see Figure 13.

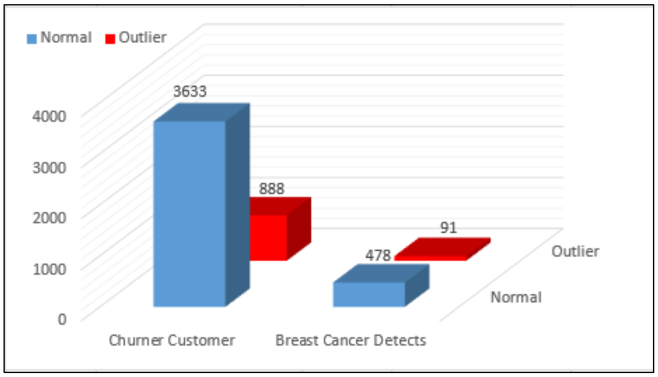

FIGURE 13
CBLOF DETECT THE NORMAL AND OUTLIER OBJECTS IN EACH CLUSTER

METRICS FOR MEASUREMENT

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

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

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

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

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

From above discussion we have following definitions:

- True-Positive (TP): Predicts the positive numbers of instances properly.
- True-Negative (TN): Predicts the numbers of negative instances correctly.
- False-Positive (FP): Predicts the numbers of negative instances incorrectly as positive.
- False-Negative (FN): Predicts the numbers of positive instances incorrectly as negative.

The Pr or Re solely can’t describes the competence of a outliers meanwhile descent performance in one of those directories doesn’t certainly implies excellent performances on-the others. For this aims, F-1 Measures, a prevalent mixture is usually use as a singleton metric for evaluation outliers’ performance. F-1 Measures, is define as the harmonic-means of Pr and Re such as:

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

A value closed to ones suggests that a well combine Pr and Re is achieved by the outliers, as per [18].

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 check Table: 1 Better the values of Pr, Re and F-1 Measures means better is the technique for breast cancer and churn prediction.

TABLE I
THE PERFORMANCE OF ALGORITHMS

Algorithm	Pr	Re	F1-M	Based on
K-Means	76.95	100	95.86	Breast Cancer
	74.36	100	85.30	Banking
LOF	78.36	100	87.82	Brest Cancer
	68	100	81.48	Banking
CBLOF	84.01	100	91.30	Brest Cancer
	80.38	100	89.12	Banking

In above Table: 1 the value of recall is same for all outliers because as we know that, Re is the detail of the outlier entity that are applicable to the queries which are correctly retrieved i.e. in our datasets we have no missing values. According to [14], K-Means algorithms is sensitive for outliers' detections in breast cancer detection it only detects those patients who has very deeply attacked by cancer and does not detects those who has in early stages and also the K-Means is unable to detect correct churner customer because its only detect those customer who has longest distance from the rest of the customers. As, authors mention above the detection in early stages saves life of somebody and also it will be recoverable in early stage furthermore, churner customer not only have longest distance, but it has also a shortest period like somebody open new account for specific purposes when he gets it then close their account immediately. The LOF is much time consuming instead of K-Means and CBLOF. Furthermore, the LOF is not detecting accurate results of outliers like sometimes it detects raw data as outliers and also does not detect those instances which have the same densities like a normal instance, but she/he is true outliers, so it's also not good for breast cancer and customer churn prediction. The CBLOF is better in detecting real outliers clustering, because it gets accurate results instead of LOF and K-Means for breast cancer and customer's churner/outliers. The CBLOF detect those outliers/abnormality who have the same densities like normal patient, but abnormal tissues are not very different from normal, in reality it is an outlier or cancerous area. The detection of this early stage is necessary. Furthermore, it also detect those customers churner/outliers who have the same density 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 can be accurately detected trough CBLOF.

CONCLUSION AND FUTURE WORK

In this research study, authors convert two raw datasets of breast cancer and banks in to suitable data and then converted this data facts to valuable information through Data mining mechanisms. Authors have taken out the datasets for selected attribute from UCI repository. Authors use K-Means, LOF, and CBLOF techniques to distinguish significant cancerous areas and customers appearances to predicts churner. The detection of cancer in early stages can saves life of somebody and also churner customer is more important than normal customers because when banks know in advance that this customer is near to switch then banks can offer him/her some extra bonus for retention.

Comprehensive evaluation of 03 different unsupervised-outlier detections algorithm on breast cancer and bank datasets has been achieved for the first-time. Data Mining goals to passage information and insights through the investigation of larger quantity of data using K-Means, LOF and CBLOF techniques. As our concern, the CBLOF algorithm shows on average better performance from other used techniques, 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.

In this research authors detects breast cancer patients and churn customers. Future study should be 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 breast cancer attacked and 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 J, Qingsheng. Z, Yang L et.al. "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] Gan G and Ng M. K, "k-means clustering with outlier removal", Elsevier Science Publishers, Pattern Recognition Letters, Vol. 90, Pages 8-14, April 2017.

[3] Valentino A, "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] Lemmens A. and Gupta S, 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] Al-Zoubi M. B "An Effective Clustering-Based Approach for Outlier Detection" European Journal of Scientific Research, Vol. 28, No.2, 2009.

[10] Manoharan J.J and Ganesh .S. H, "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] Manoharan J.J and Ganesh .S. H, "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 and Deng S, Discovering cluster-based local outliers. Pattern Recogn. Lett. 24 (9–10), 1641–1650, 2003.

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

[14] Ullman, J. D., & Rajaraman, A, Clustering, Mining of Massive Datasets, 241-280. 2012.

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

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

[17] Aggarwal C. C, Recommender Systems – The Textbook. Springer. Vol(1), 2016

[18] Fawcett T, An introduction to ROC analysis, Pattern Recognition Letters 27 (8) (2006) 861–874.

[19] Pasha M. Z and Umesh N, "A Comparative Study on Outlier Detection Techniques" International Journal of Computer Applications (0975 – 8887) Volume 66– No.24, (March 2013).

[20] Goldstein M and Uchida S, "A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data", doi: 10.1371/journal.pone.0152173, (Apr 19, 2016).

[21] Jalil, A. T, Dilfi, S. H, & Karevskiy A. Survey of Breast Cancer in Wasit Province, Iraq. Global Journal of Public Health Medicine, 1(2), 33-38. 2019

[22] American Cancer Society. Breast Cancer Facts & Figures 2005-2006. Atlanta: American Cancer Society, Inc. (http://www.cancer.org/).

[23] Gupta, S., Kumar, D., & Sharma, A. "Data Mining Classification Techniques Applied for Breast Cancer Diagnosis and Prognosis", Indian Journal of Computer Science and Engineering (IJCSE), ISSN: 0976-5166 Vol. 2 No. 2 Apr-May 2011.

[24] Kumar, G. R., Ramachandra, G. A., & Nagamani, K, "An Efficient Prediction of Breast Cancer Data using Data Mining Techniques", SSN: 2319-1058, Vol. 2 Issue 4 August 2013.

[25] Thongkam, J., Xu, G., Zhang, Y., & Huang, F., Breast Cancer Survivability via Ada Boost Algorithms, In: Health data and knowledge management: proceedings of the Second Australasian Workshop on Health Data and Knowledge Management (HDKM), Wollongong, NSW, Australia, Vol. 80, pp.55-64, 2008.

[26] Cheng T.Y., Cheng M. C., Bor W. C., Prediction of Survival in Patients with Breast Cancer using Three Artificial Intelligence Techniques, Journal of Theoretical and Applied Information technology, Vol.60, No.1, pp. 179-183, 2014.

[27] Bellachia A. and Guvan E.,"Predicting breast cancer survivability using data mining techniques", Scientific Data Mining Workshop, in conjunction with the 2006 SIAM Conference on Data Mining, 2006.

[28] Delen D., Walker G. and Kadam A. (2005), Predicting breast cancer survivability: a comparison of three data mining methods, Artificial Intelligence in Medicine, vol.34, pp.113-127.