International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 17 (2017) pp. 7019-7028
© Research India Publications. http://www.ripublication.com

Robust Feature Selection Model for Outlier Detection Using Fuzzy Clustering and Rule Mining

Karthikeyan G. [1,2]
[1]Ph. D. (PT) Research Scholar, Department of Computer Science and Engineering,
Kongu Engineering College, Perundurai, Erode-638 052, Tamil Nadu, India.
[2]Sr.System Analyst, UST Global, Chennai-600096, Tamil Nadu, India.
Orcid Id: 0000-0002-3461-126X

Balasubramanie P. [3]
[3]Professor, Department of Computer Science and Engineering,
Kongu Engineering College, Perundurai, Erode-638 052, Tamil Nadu, India.
Orcid Id: 0000-0002-9990-2586

Abstract

Managing outliers in a robust SPARK assisted big data environment gains much of the importance with the dimensionality involved in it. On increasing number of elements meant for processing, the similar amount of features associated with those elements gets surged. Former mechanisms utilized in recognizing outliers gets confines to a classification methodology computed on the basis of distance assessed between them. This sort of outlier prediction based on distance assessment mechanism leads to a fallacy. In order to accomplish a proper decision regarding outliers involved, the relationship between each and every data instance should get revealed. Improper feature abstraction leads to misclassification in accordance with overlapping clusters formed. With the aim of overcoming these issues, this paper proposes a robust Feature Selection based Rule Mining (FSRM) methodology for detecting outliers from normal instances of multivariate data considered. The dataset taken in as input is initially preprocessed to alleviate the noisy and unfilled entries indulged in it. Furthermore, the endorsed work deploys Genetic Algorithm (GA) based feature selection mechanism on those preprocessed information. Robust features completely capable of attaining flawless decision regarding outlier emerges. Feature abstraction acquires a proper objective function to attain confined set of features. Clustering those obtained features by exploiting pre-defined fuzzy rules adapts a completely associated clusters suitable for further processing. Finally, support and confidence rule formulates the base for framing association rules as a means of detecting outliers. The robustness regarding detection of outlier get evaluated against prevalent outlier detection mechanisms in terms of Area Under Curve (AUC), sensitivity, specificity, memory consumption and detection accuracy.

Keywords: Association Rule Mining, Feature Selection, Fuzzy Clustering, Genetic Algorithm, Outlier Detection, Wisconsin Diagnostic Breast Cancer.

INTRODUCTION

Swiftly progressing information technology with abruptly altering inventive methodologies makes the capability of data to grow. Consistent growth in data drives researchers to make a move towards information extraction methodologies by utilizing some effortless way [1]. A completely refined and proficient form of data mining methodology is necessitated in order to accomplish information in a robust manner. The availability of enormous amount of data makes the process of realizing useful information from the entire dataset available through selection, projection and finally aggregation gets more complicated in accordance with the growth of data size that certainly reached an extreme extent [2]. As a means of resolving this criterion, the entire setup of data processing scenario is migrated into a big data environment. Such an environment is capable of handling a huge sized data that measures even several terabytes of memory with an enhanced proficiency in a real-time scenario too. Such a large sized information handling approaches together with robust data mining methodologies are usually applicable in many sort of recent and trendy developments such as video surveillance, financial applications, e-commerce, counterfeit credit card recognition, weather forecast and intrusion detecting activity in completely connected networks.

The key concept found applicable behind all these applications are termed as outlier detection mechanism. In general, a reading or a metric that is observed from a pre-defined set of distributions are recorded and some kind of metrics that abruptly gets diverged from usual set of values are confined as outliers [3] [4]. Specific features indulged within every metric is analyzed to recognize the characteristic of data [5]. In particular a multivariate dataset infers a frequent range of alterations in which each and every parameter varies with respect to the time period being analyzed. A predefined set of standards are utilized to resemble various occurrences monitored by the system. On the basis of standards defined, the normal values from dataset inferred are collected with respect to the features assessed. The features are those components that enable to enhance the accuracy of normal from those abnormal ones by mining through historical information archived. The huge amount of data tend to be processed in a cloud based spark assisted environment suffers from much of the practical issues such as, irregular entries, discontinuous boundaries if meant for some three dimensional structures, noisy values irrelevant to the actual values, unorganized and sparse representations of complex structures mess up the identification strategy and complicates it to an extreme extent. The task of segregating the outliers to a distinct set of clusters and isolating those inliners to another category of outlier cluster alone can enhance accuracy of entire processing [6].

The process of cluster formation is found resistant with noise only when its features are analyzed and it is clustered proficiently. If not clustered properly, the clusters formed gets overlapped with each and in case obtaining these clusters in a spark assisted big data environment becomes more complex and enhances the cost incurred for computing it [7]. Many of those prevailing methodologies utilize much of the distance based mechanisms to identify outliers and alleviate it but these overlapping clusters induce in making false predictions for outliers. The overlapping clusters with respect to sparseness or density of cluster induces the errors in processing those clusters. Many prevailing methodologies does not suffice for managing a large sized dataset available in a big data environment. Inter Quartile Range (IQR) mechanism that is proficient in clustering with an enhanced accuracy, is not capable of mitigating classifying instances in as means of boosting the accuracy in detecting outlier [8]. On realizing K-means algorithm for accomplishing perfect clustering, the robustness of accuracy is completely diminished in accordance with the surging size of data involved [9]. Though K-medoids algorithm is found robust for identifying outliers lying in a sparse set of data points, the result obtained is incurring a highly sophisticated computation procedure [10]. Henceforth, to revoke these issues a novel Feature Selection based Rule Mining (FSRM) assisting in outlier detection methodology is framed with incorporation of feature selection methodology and a finest rule mining approach for segregating outliers from those normal data instances. The big data scenario utilized in this paper deployed a SPARK architecture in which a huge-sized data are handled and managed with an enhanced proficiency. The vital contributions projected in this paper are,

- Detecting outliers in the Wisconsin Diagnostic Breast Cancer (WDBC) dataset through a novel GA based feature selection methodology the completely mitigates the overall number of features covered for a single data instance. Subsequently, the overall dimensionality gets diminished.
- Employ fuzzy clustering for merging those similarly characterized features into single clusters and hence, capable of handling multivariate data adapted from multi-dimensional platform.
- To realize association rule mining on those clustered features on the basis of support and confidence measures acquired for every single feature of a data instance.

The remaining sections of this paper is organized as follows: Section II reviews some of the existing works related to outlier detection and practical issues unveiled in prevailing systems. Section III presents the detailed description of the proposed FSRM system. Section IV presents the performance results of both existing and proposed techniques. Finally, this paper is concluded and the future work to be carried out is stated in Section V.

RELATED WORK

This section exemplifies the prevalent techniques available for analysing and segregating outliers from those normal entries within big data and multivariate data.*Romero and Ventura* [11] endorsed a the proficiency of data mining methodology in exploring and excavating some useful information from those information existing in varied granular levels. *Veiga, et al.*[12] evaluated the shortcoming inferred in managing that large sized information residing in a big data environment. The necessity for a consistent rise in overall size of the information in accordance with growth happened in practical scenario, the computing algorithms was also induced to grow inevitably by utilizing an expensive and complex computational infrastructure. *Li, et al* [13] designed a specifically proficient unifying structure for recognizing outliers in the repository that comprised information regarding Process Trace Data (PTD) through an exploitation of CUmulative SUM (CUSUM) approach. A complete domain knowledge was created by means of analyzing that information through a deployment of Fast Greedy Algorithm (FGA). Though the existence of outliers were revealed in a robust manner, the threshold necessitated was to get fixed in prior to the process initialization that consequently surged the overall computational complexity.

Albanese, et al.[14] devised a rough set analysis based approach for identifying outliers with a utilization of Rough Outlier Set Extraction (ROSE) by means of deploying upper as well as lower approximations. However, the formulated ROSE approach certainly outperformed all sorts of prevailing state-of-art methodologies in terms of mitigating computational time, it is incapable of managing those fine granularities with respect to the increase in size of information being processed.

Zhao, et al.[15] exploited weighted density functionality for analyzing those outliers available within the categorical information being processed. Though the overall time complexity was mitigated on exploiting this approach, the accuracy in those identified outliers consistently degrades on realizing the similar approach on a multivariate information platform or for a data abstracted from a dynamic environment.

Bouguessa [16] exploited weighted density functionality for analyzing those outliers available within the categorical information being processed. Though the overall time complexity was mitigated on exploiting this approach, the accuracy in those identified outliers consistently degrades on realizing the similar approach on a multivariate information platform or for a data abstracted from a dynamic environment recommended a proficient mechanism constructed on the basis of some other pre-defined set of principles that completely assists segregating the inliners from those outliers of a spatio-temporal data. Furthermore, a mixed set of elements were deliberated for processing and the outliers were identified with an enhanced proficiency devoid of utilizing any sort of transformation to be realized on features.

Gupta, et al. [17] surveyed on various outlier mechanisms that are capable of processing information residing in varied data types such as temporal data prevailing in temporal networks, spatio-temporal data, information residing in distributed data streams, data abstracted from the time-series based information platform etc. The lack of availability of mechanisms for processing point based information model through online web based applications was revealed. Consequently, this aspect inhibited the overall proficiency in detection of outliers existed in temporal data in a comprehensive manner.

Schubert, et al.[18] deliberated a proficient outlier identification methodology for informative objects residing within a dataset that was static and comprised only low-density areas. The static information is analyzed within a Euclidean space and also possessed a set of static vectors. Though it was capable of figuring out noisy information along with alleviation of redundancy, its scalability was highly restricted to act upon a large set of information exposed in varied data types given in textual format, time series etc. Its compatibility was mitigated to a single domain. *Taha and Hadi* [19] recommended a proficient approach to analyze and segregate the outliers after recognizing it from those frequent itemsets explored by means of deploying a minimum support measure in categorical information taken on a multivariate basis. A distinct score was stipulated for every single object acquired through the accomplishment of decision attributes in an automatic manner and it incurred an enriched computational complexity. The frequent itemsets that were explored comprised of sensitive outliers and it was incapable of handling it. Hence, these criteria ultimately lead to plant model discrepancy owing to the realization of attribute appropriation in a biased manner. In order to resolve this issue, this approach necessitated an additional attribute appropriation model for assessing noise dissemination in itemsets and it gained an added complexity for getting deployed on data-driven models.

Ienco, et al. [20] endorsed an inventive anomaly recognition technique for processing categorical information in a semi-supervised manner that utilized a distinctive prototype for analyzing anomalous occurrences. The discrimination is acquired by means of evaluating both anomalous and normal occurrences of categorical information involved. The time complexity experienced in recognizing anomalies was much higher when processed on categorical data. Also, the outliers was not segregated on the basis of specific features inferred from those outliers and it was differentiated in a general structure.

Liu, et al. [21] recommended an innovative approach for recognizing outliers in uncertain data no deploying a learned classifier using Support Vector Data Description (SVDD) mechanism. It was carried out in a dual-fold manner given as employing confidence score for the creation of pseudo training set of features and further prolonging the same towards implication of SVDD technique. However, sensitivity measured with respect to noise recognized in the input data was assessed in a robust manner, the entire recognition capability for unveiling outliers was mitigated to a considerable extent when employed on an uncertain data. *Zimek, et al.*[22] presented a Feature Bagging (FB) approach for prompting outliers after recognizing their diversity through a subsampling mechanism. The outliers were identified and segregated by means of exploiting a classification paradigm formulated depending upon ensembles formed on the basis of subsamples acquired in prior. Though it was found sufficient for handling large sized data, it was not suitable for processing a small amount of information. Hence the features realized from those data are highly similar and subsets of objects were left without any sort of recognition for outliers.

Sharma and Panigrahi [23] illustrated a fraud detection strategy for getting deployed in financial accounting procedure in order to empower forensic accounting through a robust neural network and Bayesian belief networks methodology. Regression analysis was proficient enough to segregate fraudulent accounting activities. However, the misclassifying tendency results in diminished sensitivity of the overall detection procedure and the cost inferred was too high.

Vadgasiya and Jagani [24] devised an inventive algorithm for creating a completely balanced trade-off between clusters and the actually available outliers. Outliers were ultimately mitigated in a circumstance at which, those clusters were found more than the overall number of outliers. The crucial drawback in realizing a clustering approach for identifying and segregating outliers was its instable outcome for every distinct instance it was employed on the itemsets. Carton, et al. [25] projected a proficient Eyes Wide Open (EWO) strategy for accomplishing an automated analysis of outliers by means of utilizing Evidence Accumulation Clustering (EAC) for rule based systems implied on a varied corpus. However, the overall proficiency inferred was very moderate than other manually inferred approaches.

PROPOSED METHOD

This section describes proposed Feature Selection and Rule Mining (FSRM) based Outlier Detection system. The multi-variate data obtained from WDBC dataset is handled with assistance of SPARK architecture.

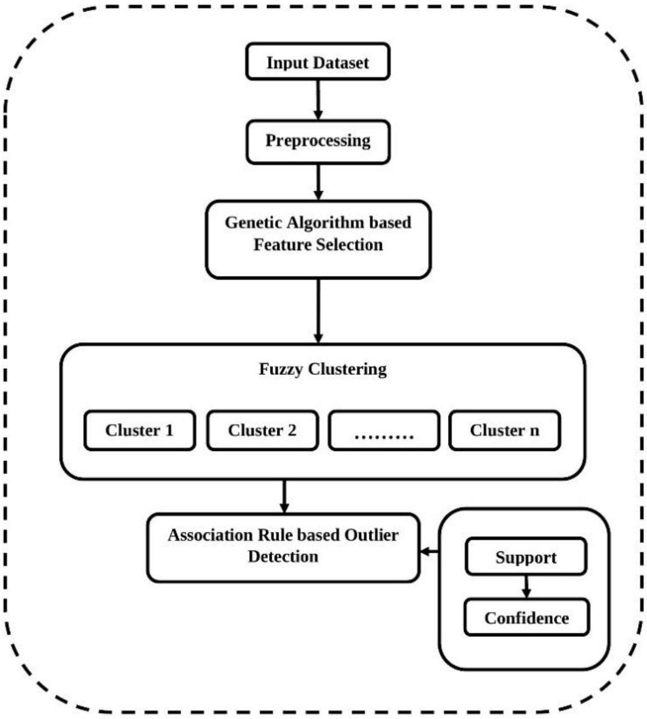

Figure 1: Overall flow of the proposed system

The overall flow of the proposed system is shown in Fig 2 and. Initially, that information in that dataset are preprocessed and the noisy elements are alleviated. After preprocessing that information, the procedure further opts for feature selection procedure. Genetic Algorithm (GA) is employed to assess the features of data instance through computation of an optimal fitness functionality. On realizing this procedure on the preprocessed dataset, a particular feature that is capable of differentiating that specific instance to be normal or an outlier opts. On the basis of those features selected, a clustering methodology constructed on the basis of the fuzzy technique is employed to cluster those associated features into distinct clusters. These clusters are then processed to acquire the support and confidence measure by means of utilizing a rule mining mechanism that formerly frames the support and confidence rules in a comparative way.

Data Preprocessing

The multivariate information residing in the dataset constitutes a pre-defined set of real-value attributes involved in resembling the physical features of the cell nuclei prevailing in the digitized imagery of a cytological diagnosis. Numerous relevant features recognized from those imageries are further processed to acquire a useful model that is capable of assisting in the formulation of a basic model that serves in identifying outliers with trailing procedure. After the accomplishment of preprocessing procedure, the entire information residing within the dataset becomes complete without including any sort noise or unfilled entries in it. Subsequently, the proficiency of data mining procedure gets surged in a stable manner. The dataset entirely gets fragmented up into benign or malignant medical imageries. Each and every imagery is assessed for its individual features abstracted from it in the form of numerical values. These processed input entries aids in proceeding with the further procedure.

Genetic Algorithm based Feature Selection

A concept of survival of the fittest is utilized in the searching conception utilized in this procedure. Search attributes are deliberated as the numerical values that certainly indicates the area and other features of nuclei. Distinct elements of those characteristics are recognized as genes and separate features are denoted as chromosomes. Vital operators involved in GA are stipulated as follows.

Population Generation

All kinds of individuals involved in the procedure of generating a separate group of characteristics are termed as population and formulation of those generations at the initial stage is likely called as population generation. The overall complexity of the problem typically defines the population to be formed. Logically in GA the initial population is almost randomly opted and is formulated in the form of as if defined in Table 1.

Table 1: Arbitrarily generated initial population

Population	Chromosome n	[a;b;c;d]

Selection

A choosing procedure is accomplished at this stage by means of opting for two feasible set of parent individuals. The ultimate aim of opting for an appropriate parental is ensued on accomplishing this procedure. The chromosomes with an extreme amount of fitness value are opted by GA mechanism chosen here. The fitness value possessed by every single chromosome or a feature of an instance gets suggested through the objective function deliberated in an arbitrary manner.

Crossover

The fittest set of parental features are suggested through selection procedure implied in order to obtain a best of those features opted from those overall set of features opted.

Complete procedure trailed in this crossover is consolidated as,

- Opt for an optimal chromosome pair from the pool of selected features in an arbitrary manner.
- The length of population framed is assessed and cross site is allocated accordingly
- Swapping happens in between those opted cross sites with respect to their original positions

Algorithm for genetic algorithm based feature selection

Input: Preprocessed Dataset (Pd)

Output: Best of the selected features (Sbf)

```
for i=0: I do
    Sf ← select random features from dataset
    for j=0: Sf do
        Tc ← load chromosomes from sf
        end for
        for j=0: Tc do
        S ← selection
        end for
        for j=0: S do
            $N_t = ch(j) + Sf(j)$
        P ← $1/N_t$
        end for
        for j=0: Tc do
        SCF ← select crossover features
        end for
        for j=0: Scf do
        Cac ← chromosomes after crossover
        end for
        for j=0: Cac do
        M ← mutation
        end for
        Sbf ← selected best features
end for
```

Features obtained for every data instance is arbitrarily chosen from the overall set of data instances in the actual dataset. All those selected features are confined as chromosomes for further processing. A probability for those selected set of features is computed and hence, the probability is stipulated for each and every feature of every single data instance being chosen and processed. On the basis of this fitness function, those features with utmost fitness criterion opt for crossover operations. Afterward, mutation procedure is implied on those chromosomes in order to obtain the best set of features that certainly enhances the accuracy of detecting outliers.

Table 2: Parameters used in this work

D	Load Dataset
Pd	Preprocessed dataset
I	Iteration count
P	Probability
ch	Chromosome
sf	Selected features
Sbf	Best of the selected features
Tc	Total chromosome count
Of	Objective function
Noc	Total number of clusters formed
Cd	Clustered Data
A	Dataset attributes
Su	Support
Occ	Occurrence
To	Total number of occurrence

Fuzzy Clustering

The selected set of best features is obtained as an input from previous feature selection procedure. A fuzzy approach based clustering is utilized for accomplishing the clustered data from those optimal set of best features. The feature set formulated is assessed on the basis of distance inferred between them with respect to the data space observed. The necessitated clusters are articulated by means of allocating specific data points for every distinct feature being acquired. This procedure of fixing the distance between a data point and cluster center is reiterated for every feature obtained. The cluster center (v_j) are periodically updated at the end of each and every iteration (μ_{ij}). The objective function of every single population accomplished is assessed and thus the cluster center is fixed. The objective function gets articulated as,

$$J(U,V) = \sum_{i=1}^{n} \sum_{j=1}^{c} (\mu_{ij})^m \,||x_i - v_j||^2 \qquad (1)$$

Algorithm for forming clusters with fuzzy rules

```
Input: Best of the selected features ($Sbf$)
Output: clustered data ($Cd$)

$\mu_{ij} = 1\Big/\sum_{k=1}^{c}\left(d_{ij}/d_{ik}\right)^{(2/m-1)}$
$v_j = \left(\sum_{i=1}^{n}\left(\mu_{ij}\right)^m x_i\right)\Big/\left(\sum_{i=1}^{n}\left(\mu_{ij}\right)^m\right)$
Compute $Of$ by (1)
```

On implementing the fuzzy clustering approach on those selected set of best features, like features are clustered into a distinct set of clusters for every iteration fixed for a specific data instance. The clustering membership is fixed for distinct set of specifications and the clusters are formulated on the basis of objective function computed.

Outlier Detection based on Association Rules

The clustered data with the selected set of best features employs the Association Rule Mining (ARM) concept to understand the behavior of those features for every data instance inferred. Let the attributes from the previous module are represented as $A = (a_1, a_2, \dots, a_n)$ and D be the dataset that contains a set of transactions. The rule formulation depends on the two measures such as support and confidence. The pseudo code to implement the FSRM for the banking services as follows:

Algorithm for segregating outliers by association rules

```
Input: clustered data ($Cd$)
Output: detected outliers

for $i$=0: $Noc$ do
    for $j$=0: $A$ do
        for $k$=0: $A$ do
            $Su = (\sum_1^n Occ)/\sum_i^n To$
        end for
    end for
end for
$Co \leftarrow Confidence$
for $i$=0: $Noc$ do
    for $j$=0: $A$ do
        for $k$=0: $A$ do
            $Co = (Su_{jk})/Su_j$
        end for
    end for
end for
```

The set of attributes is denoted as (Soa) is used to measure those measures as follows:

Support: The measure of the frequency of rule within the transactions refers to support and such rule ($A \Rightarrow B$) involve the great part of the dataset for high support values.

$$supp(A \Rightarrow B) = p(A \cup B) \tag{2}$$

In this paper, the support value for the normal and outlier attributes is formulated as

$$S = \frac{supp(Soa)}{N} \tag{3}$$

Confidence: The measure of the percentage of transactions containing A which contain also B refers the confidence value. The mathematical formulation for the confidence estimation is conditional probability estimation is represented as

$$Confidence\ (C) = P\left(\frac{B}{A}\right) = supp(A,B)/supp(A) \tag{4}$$

This formulation is modified with normal attributes as follows:

$$C = \frac{supp(Soa_A \cup Soa_B)}{Soa_B} \tag{5}$$

The threshold value to predict the normal data corresponding to the normal attribute is computed from the average value of confidence as follows:

$$T_C = \frac{\left(\sum_{i=1}^{C} c_i\right)}{C.size} \tag{6}$$

The comparison between the confidence value of each transaction with the threshold value (T_C) decides the outlier.

Performance Analysis

This section articulates the performance results of both existing and proposed techniques in terms of Average AUC, run time-training, run time-testing, detection accuracy, and memory consumption. The existing techniques considered in this work are Semi-supervised Anomaly Detection for Categorical Data (SAnDCat), Local Outlier Factor (LOF), unconstrained Least Square importance Fitting (uLSiF), one-class Support Vector Machine (OSVM) and Feature Regression and Classification (FRaC) techniques. Moreover, the performance of the proposed approach is realized for its proficiency in varied data dimensionality against prevalent methodologies are also compared.

Dataset Description

The Wisconsin Diagnostic Breast Cancer (WDBC) [26] contains various attributes namely, diagnosis, ID number and real valued features. There are ten real valued features namely, radius, area, perimeter, smoothness, texture, compactness, concave points, concavity, symmetry and fractal dimension computed from digitized image of the breast mass. The WDBC dataset is multi-variant dataset that comprises 569 instances and 32 attributes. WDBC describes the characteristics of cell-nuclei in an image. These entire set of information is realized in the SPARK architecture that is capable of handling large-sized information in a big data environment.

Average AUC

The overall proficiency of the proposed FCRM technique is typically assessed on the basis of detection rate inferred for the dataset in which it is implied [20]. The process trailed for assessing the detection of outliers is accomplished by computing aggregated count of irregular occurrences correctly classified and the regular occurrences misclassified as irregular. In order to deliberate both these metrics at a single instance, the AUC is worked out for the existing SAnDCat, LOF, uLSIF, OSVM, FRaC and proposed FSRM techniques.

$$AUC = \frac{S_0 - {n_0(n_0+1)}/{2}}{n_0 n_1} \qquad (7)$$

Where, total number of instances appropriated into the normal class (n_0) and those instances appropriated for abnormal or irregular instances g(n_1) are computed for obtaining the rank (r_i) in the test set (S_0) and is assessed as, $S_0 = \sum_{i=1}^{n_0} r_i$.

From this analysis, it is observed that the proposed FSRM based outlier detection technique provides the highest average AUC when compared with prevalent techniques as depicted in Fig 2. It is due to the enhancement inferred with clustering accuracy in accordance with the support and confidence measures assessed. Since the relationship between those features are revealed for every instance of a test set the AUC gets enhanced obviously.

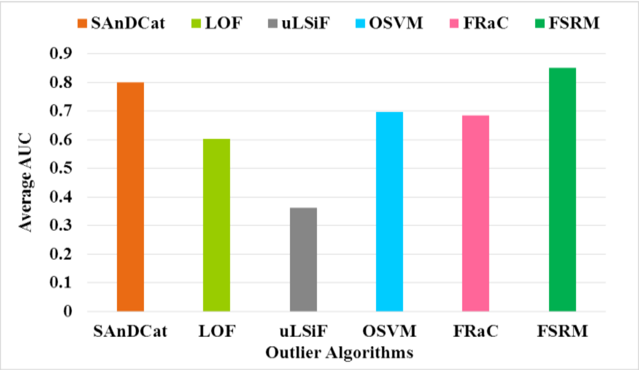

Figure 2: Average AUC analysis

In existing methods, the SAnDCat offers maximum average AUC (0.8) compared to previous methods. But, the proposed work further improves the average AUC into 0.85 which is 5 % improvement compared to existing approach.

Run-time Analysis

Runtime criterion of the algorithm is stated as the association between the overall length of the input and the amount of time incurred to process those input instances on the basis of detection and is prompted in seconds. Here, the runtime is assessed for prevalent technique SAnDCat available against proposed FSRM methodology in accordance with the varying percentage of instances in the dataset. From this analysis, it is observed that the proposed FSRM necessitates run search time when compared to the other techniques. Such a significant mitigation in runtime is achieved by the proposed FCRM methodology only because of its appropriate feature selection constraint prior to let instances into processing. Fig 3. And Fig. 4 Graphically illustrates the runtime training and testing inferences respectively.

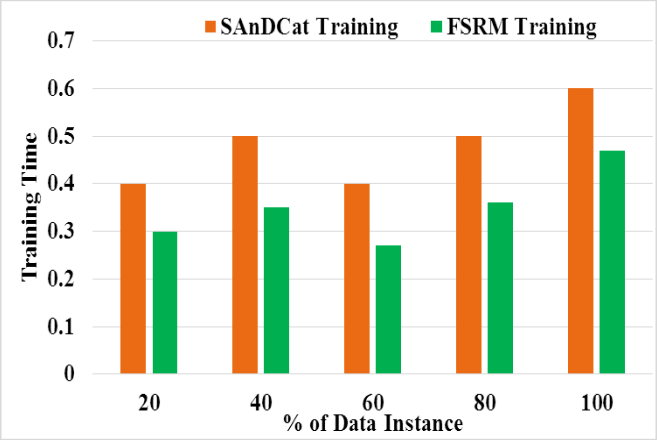

Figure 3: Runtime training Inference

In existing methods, SAnDCat technique consumed minimum time for training procedure. The SAnDCat technique exposes 0.4 secs for 20% of data instance and 0.6 Secs when there is 100% of data instance. Similarly, proposed FCRM technique exposes 0.3 secs for 20% of data instance and 0.47 Secs when there is 100% of data instance. The proposed work further reduces the time consumption to 0.1 and 0.13 secs which are 25 and 21.66 % reduction compared to SAnDCat respectively.

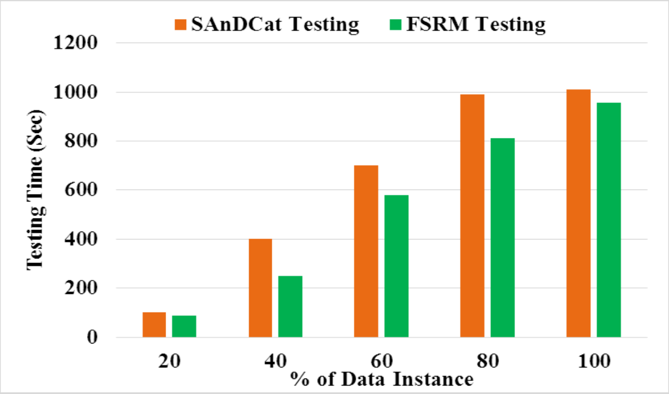

Figure 4: Runtime testing Inference

In existing methods, SAnDCat technique consumed minimum time for the testing procedure. The SAnDCat technique exposes 100 secs for 20% of data instance and 1010 Secs when there is 100% of data instance. Similarly, proposed FCRM technique exposes 87 secs for 20% of data instance and 956 Secs when there is 100% of data instance. The proposed work further reduces the time consumption to 3 and 54 secs which are 30 and 5.346 % reduction compared to SAnDCat respectively.

Memory Usage

The aggregate volume of memory necessitated for the proficient computation and processing of instances inferred from those datasets for an efficacious functioning of the system deployed. Here, the memory consumption is estimated with respect to varying number of dimensions in datasets for both existing Memory Efficient increment Local Outlier Factor detection algorithm for more flexible version (MiLOF_F) [27] and the proposed FSRM techniques. From this analysis, it is evaluated that the proposed FSRM requires least possible memory for processing when compared to the other techniques depicted in Fig 5.

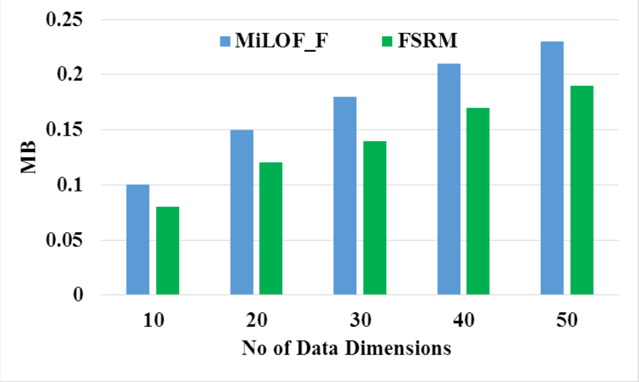

Figure 5: Memory usage analysis

In existing methods, MiLOF_F utilized the least amount of minimum memory to store and process the data. The MiLOF_F offers 0.1 and 0.23 MB for the data dimension of 10 and 50 respectively. Instead, the proposed work additional mitigates the utilization of memory to 0.08 and 0.19 MB for the data dimension of 10 and 50 which are 0.02 and 0.04 MB lesser than those existing methodologies. On the whole,20% and 17.39 % reduction compared to MiLOF_F evidently.

Detection Accuracy

The proportion between the summation of overall data instances correctly inferred against summation of a number of instances inferred for both of its normal and abnormal occurrences is stated as the detection accuracy of the entire system defined. The accuracy is calculated as follows,

$$Accuracy = \frac{(TN+TP)}{(TN+TP+FN+FP)}$$

$$= \frac{Number\ of\ true\ correct\ assessment}{Number\ of\ all\ assessment} \qquad (8)$$

Where,

TN -True Negative,

TP -True Positive,

FN -False Negative

FP-False Positive.

The overall detection accuracy is considerably surged owing to the exact relationships defined between those features of objects being selected in prior that subsequently mitigated the entire dimensionality criterion. The accuracy of both existing and proposed outlier detection techniques. In accordance with the performance of those prevailing techniques, MiLOF_F approach accounts for the maximized accuracy but the accuracy of the proposed FCRM methodology ranges to an enhanced accuracy of 2% for 10 number of data instances and 2.10%. The accuracy of both existing and proposed outlier detection techniques are shown in Fig 6.

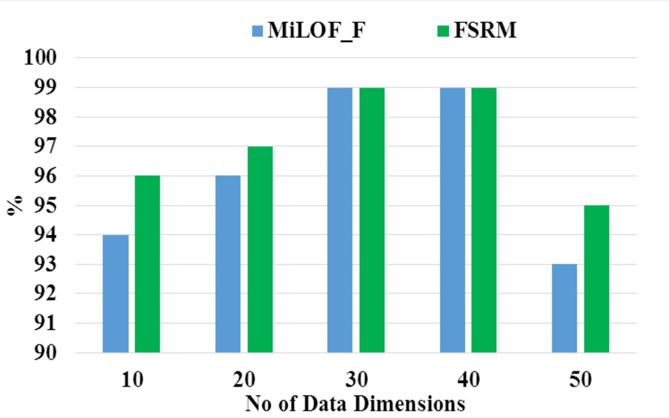

Figure 6: Accuracy Analysis

In existing methods, MiLO_F offers an accuracy of 94 for 10 data instances and 93 for 50 data instances. Similarly, the proposed FSRM enhances the accuracy into 96 for 10 instances and 95 for 50 data instances respectively.

CONCLUSION AND FUTURE WORK

The concept of outlier detection finds its application in various fields such as counterfeit credit card recognition, video surveillance etc, Features inferred for each and every data instance inferred from the information are consistently raising with respect to increase in the size of information in the present scenario. As a means of acquiring a proper outlier detection mechanism implied, perfect feature abstraction is necessitated for all kinds of information. Hence, a novel FSRM methodology is deployed to overcome the issues in detecting outliers within multivariate data. The dataset taken as input are initially preprocessed to alleviate the noisy and unfilled entries indulged in it. The preprocessed information obtained are likely to get processed by means of employing GA approach for opting the finest set of fittest features. Robust features that are capable of attaining flawless decision regarding outlier is abstracted. Feature abstraction acquires a proper objective function to attain a confined set of features. Clustering those obtained features by exploiting pre-defined fuzzy rules is adapted for obtaining a completely associated clusters that are suitable for further processing. At the end, support and confidence rule formulates the base for framing association rules as a means of detecting outliers. The robustness regarding detection of outlier gets evaluated against prevalent outlier detection mechanisms in terms of Area Under Curve (AUC), sensitivity, specificity, memory consumption and detection accuracy. As a sign of enhancement, the proposed FSRM approach tends to achieve a robust detection accuracy of about 2% and 2.10 % for a maximum and a minimum number of data instances. In future, the means of enhancing this work finds its way in enriching the clustering ability and the l oading cost of the overall outlier detection system.

REFERENCES

[1] X. Wu, X. Zhu, G.-Q. Wu, and W. Ding, "Data mining with big data," *ieee transactions on knowledge and data engineering,* vol. 26, pp. 97-107, 2014.

[2] E. Schubert, A. Zimek, and H.-P. Kriegel, "Fast and scalable outlier detection with approximate nearest neighbor ensembles," in *International Conference on Database Systems for Advanced Applications*, 2015, pp. 19-36.

[3] J. Huang, Q. Zhu, L. Yang, and J. Feng, "A non-parameter outlier detection algorithm based on Natural Neighbor," *Knowledge-Based Systems,* vol. 92, pp. 71-77, 2016.

[4] M. Radovanović, A. Nanopoulos, and M. Ivanović, "Reverse nearest neighbors in unsupervised distance-based outlier detection," *IEEE transactions on knowledge and data engineering,* vol. 27, pp. 1369-1382, 2015.

[5] M. Hauskrecht, I. Batal, M. Valko, S. Visweswaran, G. F. Cooper, and G. Clermont, "Outlier detection for patient monitoring and alerting," *Journal of Biomedical Informatics,* vol. 46, pp. 47-55, 2013.

[6] A. Nurunnabi, G. West, and D. Belton, "Outlier detection and robust normal-curvature estimation in mobile laser scanning 3D point cloud data," *Pattern Recognition,* vol. 48, pp. 1404-1419, 2015.

[7] M. H. Bhuyan, D. Bhattacharyya, and J. Kalita, "A multi-step outlier-based anomaly detection approach to network-wide traffic," *Information Sciences,* vol. 348, pp. 243-271, 2016.

[8] T. Santhanam and M. Padmavathi, "Comparison of K-Means clustering and statistical outliers in reducing medical datasets," in *Science Engineering and Management Research (ICSEMR), 2014 International Conference on*, 2014, pp. 1-6.

[9] T. S. a. M. S. PAdmavathi, "An efficient model by applying genetic algorithms for outlier detection in classifying medical datasets " *Australian Journal of Basic and Applied Sciences,* pp. 583-591, August 2015.

[10] T. Velmurugan, "Efficiency of k-means and k-medoids algorithms for clustering arbitrary data points," *Int. J. Computer Technology & Applications,* vol. 3, pp. 1758-1764, 2012.

[11] C. Romero and S. Ventura, "Data mining in education," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery,* vol. 3, pp. 12-27, 2013.

[12] J. Veiga, R. R. Expósito, X. C. Pardo, G. L. Taboada, and J. Tourifio, "Performance evaluation of big data frameworks for large-scale data analytics," in *Big Data (Big Data), 2016 IEEE International Conference on*, 2016, pp. 424-431.

[13] Z. Li, R. J. Baseman, Y. Zhu, F. A. Tipu, N. Slonim, and L. Shpigelman, "A unified framework for outlier detection in trace data analysis," *IEEE Transactions on Semiconductor Manufacturing,* vol. 27, pp. 95-103, 2014.

[14] A. Albanese, S. K. Pal, and A. Petrosino, "Rough sets, kernel set, and spatiotemporal outlier detection," *IEEE Transactions on Knowledge and Data Engineering,* vol. 26, pp. 194-207, 2014.

[15] X. Zhao, J. Liang, and F. Cao, "A simple and effective

outlier detection algorithm for categorical data," *International Journal of Machine Learning and Cybernetics*, vol. 5, pp. 469-477, 2014.

[16] M. Bouguessa, "A practical outlier detection approach for mixed-attribute data," *Expert Systems with Applications*, vol. 42, pp. 8637-8649, 2015.

[17] M. Gupta, J. Gao, C. C. Aggarwal, and J. Han, "Outlier detection for temporal data: A survey," *IEEE Transactions on Knowledge and Data Engineering*, vol. 26, pp. 2250-2267, 2014.

[18] E. Schubert, M. Weiler, and A. Zimek, "Outlier Detection and Trend Detection: Two Sides of the Same Coin," in *Data Mining Workshop (ICDMW), 2015 IEEE International Conference on*, 2015, pp. 40-46.

[19] A. Taha and A. S. Hadi, "A general approach for automating outliers identification in categorical data," in *Computer Systems and Applications (AICCSA), 2013 ACS International Conference on*, 2013, pp. 1-8.

[20] D. Ienco, R. G. Pensa, and R. Meo, "A Semisupervised Approach to the Detection and Characterization of Outliers in Categorical Data," *IEEE transactions on neural networks and learning systems*, vol. 28, pp. 1017-1029, 2017.

[21] B. Liu, Y. Xiao, L. Cao, Z. Hao, and F. Deng, "SVDD-based outlier detection on uncertain data," *Knowledge and information systems*, vol. 34, pp. 597-618, 2013.

[22] A. Zimek, M. Gaudet, R. J. Campello, and J. Sander, "Subsampling for efficient and effective unsupervised outlier detection ensembles," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, 2013, pp. 428-436.

[23] A. Sharma and P. K. Panigrahi, "A review of financial accounting fraud detection based on data mining techniques," *arXiv preprint arXiv:1309.3944*, 2013.

[24] M. G. Vadgasiya and J. M. Jagani, "An enhanced algorithm for improved cluster generation to remove outlier's ratio for large datasets in data mining," *Development*, vol. 1, 2014.

[25] C. Carton, A. Lemaitre, and B. Coüasnon, "Eyes Wide Open: an interactive learning method for the design of rule-based systems," *International Journal on Document Analysis and Recognition (IJDAR)*, pp. 1-13, 2017.

[26] W. N. S. Dr. William H. Wolberg, Olvi L. Mangasarian. (1993). *Breast Cancer Wisconsin (Diagnostic) Data Set*. Available: https://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+(Diagnostic)

[27] M. Salehi, C. Leckie, J. C. Bezdek, T. Vaithianathan, and X. Zhang, "Fast Memory Efficient Local Outlier Detection in Data Streams," *IEEE Transactions on Knowledge and Data Engineering*, vol. 28, pp. 3246-3260, 2016.