IJCTA, 9(27), 2016, pp. 495-502 ISSN: 0974-5572
© International Science Press

# Linear Regressive Percentage Split Distribution Clustering

A. Joy Christy[1] and S. Hari Ganesh[2]

**ABSTRACT**

Data Mining (DM) refers to the extraction of meaningful data models from large data sources. Clustering is one of the descriptive techniques of DM that group's data objects based on their similarity. Though there are quite good numbers of algorithms available in clustering, the complexity, the time taken to build clusters and cluster compactness are still remain as issues. Hence, the improvisation of clustering algorithms is always considered as a thrust area of research. The core objective of this paper is to simplify the task of clustering by proposing a novel Linear Regressive Percentage Split Distribution clustering (LRPSDC) which maximizes the cluster performance with minimum time duration. The method works on the transformation of non-linear data into a linear form with respect to class label using linear regression model. A comprehensive experimental study is conducted to assess the performance of LRPSDC and evaluated against the existing clustering methods in terms of cluster compactness and time. The results clearly indicate that LRPSDC have built meaningful and compacted clusters with the reduced time than the existing ones.

***Keywords***: Correlation, Linear Regression, Percentage Split Distribution, Silhouette Distance.

## 1. INTRODUCTION

Clustering is categorized as one of the data descriptive analysis techniques that builds clusters with data objects in such a way that objects in a cluster are closer to each other than the objects of other clusters. Clustering is often said to be an unsupervised learning of data as it does not rely upon a labeled classifier. Clustering techniques reveals the divergence of data objects which is indeed helpful for taking future decisions by analyzing the distribution of data. Clustering plays a vital role in the areas like education, industries, medical diagnosis, business, weather forecasting and so on. Clustering techniques are majorly classified into hierarchical or partitional. Hierarchical clustering splits the data objects into a dendrogram like structure where the bottom-up segregation of hierarchical clustering is called agglomerative and top-down segregation of hierarchical clustering is called divisive. Unlike hierarchical, partitional clustering splits the data into k disjoint units through centroid or statistical measures. Though, the emergence of clustering seems to be superior, the existing clustering techniques suffer from the following reasons:

- Lack of global optimum solution
- Based only on the distance factor
- Evolves more iterations
- Bad initiliazation
- Generation of empty clusters
- Poor prediction of outliers

Moreover, the conceptualization of existing clustering algorithms is more critical, highly mathematical and time consuming. Hence, the improvisation of clustering algorithms is still considered as a thrust area in research. The

[1] Research Scholar, Department of Computer Science, Bishop Heber College, Trichy - 620 017

[2] Assistant Professor, Department of Computer Science, H.H. The Rajah's College,Pudukottai – 622 001.

process of clustering is not just a single task rather is accomplished by carefully executing a series of steps from data collection to cluster validation.

Data Collection: The data may either be collected manually or through web. There are plenty of web sources available for data collection within which the popular sources are said to be data repositories, data bases, web forums, and social networks. The choice of data collection is purely user oriented.

Data Cleaning: The data acquired by the data sources cannot be directly clustered as it may be in semi or unstructured form and may contain noisy, irrelevant and missing information. The process of data cleaning first changes the semi or unstructured data into a proper structure that outfits the clustering algorithm. The removal of irrelevant features and the data objects that contain missing values are also the primary concern of data cleaning process.

Clustering strategy: The method that constitutes clusters is called clustering strategy. Clustering strategy varies between the techniques of clustering. However, the success of clustering strategy depends upon the maximization of intra-cluster similarity and inter-cluster dissimilarity.

Cluster validation: The analysis on the effectiveness of cluster compactness is called cluster validation. Cluster validation is achieved through several internal and external validation metrics such as silhouette co-efficient, dunn index and davies-Bouldin index, confusion matrix and so on.

Thus, the proposal of new clustering algorithm should address all the steps that are discussed above. This paper presents a linear regressive percentage split distribution clustering that utilizes the labeled classifier for building quality clusters. The main objectives of this study are:

1. To transform the non-linear data into a linear progression that fits into a linear equation using multiple linear regression
2. To find out and separate the outliers in the data objects
3. To overcome bad initialization issue
4. To split the linearly regressive data objects using percentage split distribution method

The rest of this paper is organized as follows: section 2 describes the review of literature, section 3 discusses the methodology of the proposed work, the illustration of the proposed work is deliberated in section 4, section 5 denotes experiment and results and finally section 5 elucidates the conclusion of this study.

## 2. REVIEW OF LITERATURE

This section presents a brief description on the limitations of existing methods and the recent works carried out.

### 2.1 Connectivity Models

Bouguettaya et al.[1] have stated that the connectivity based clustering suffers from high computational cost and built an hybrid clustering approach(KnA) that integrates the traditional K-Means with agglomerative clustering. Jeon et al. [2] have specified that the connectivity based clustering suffers from limited scalability and proposed a parallelization scheme for multi-threaded shared-memory machines. Almeida et al. [3] have indicated certain issues related to connectivity based clustering such as sensitivity to outliers, fluctuations of data points and automatic clustering and have proposed an improved Single Linkage Hierarchical Clustering (SLHCA). Pourjabbar et al. [4] have denoted that the traditional hierarchical clustering algorithms do not build meaningful clusters and have developed two fuzzy based models called Fuzzy Divisive Hierarchical Clustering (FDHC) and Fuzzy Hierarchical Cross-Clustering (FHCC).

## 2.2 Centroid Clustering Models

Wu et al. [5] have stated that the classic centroid models are sensitive to the selection of initial centroids and often converge with locally optimized solutions. Hence, the authors have presented a Hybrid Fuzzy K-Harmonic Means (HFKHM) clustering that combines an Improved Possibilistic C-Means (IPCM) with K-Harmonic Means (KHM). Prabha et al. [6] have proclaimed that the centroid models have noise and selection of initial clusters sensitivity and often terminate at a local optimum and have collaborated the concept of k-means with particle swarm optimization. Chawla et al. [7] have indicated that the centroid models are extremely sensitive to outliers and presented a naïve approach to discover the outliers and have proposed k-means—algorithm. Anand [8] have denoted that the centroid models are computationally complex and build clusters with poor quality and have developed a modified k-means algorithm.

## 2.3 Distributive Models

Liu et al. [9] have specified that the distributive models are highly rely upon the initialization and have presented a solution through an improved Expectation Maximization (EM) algorithm based on the Multivariate Elliptical Contoured Mixture Models (MECMM's). Volkovich et al. [10] have declared that the iterative clustering algorithms do not provide optimal cluster solutions as the cluster quality highly rely upon the good selection of initial partitions and have introduced a Cross-Entropy method. Chen et al. [11] have indicated that the distributive models are deficient in computational time and accuracy and have provided solutions with two expectation maximization models called FixEM and ModalEM. Yu et al. [12] have denoted that the distributive models are prone to outliers and initial selection and have presented a solution by proposing a Spatial-EM algorithm for finite elliptical mixture learning.

## 2.4 Density Models

Ghanbarpour et al. [13] have proclaimed that the density models have some inabilities in identifying clusters with different densities and sensitivity to noise and have introduced ExDBSCAN (Extended Density-Based Spatial Clustering of Applications with Noise) method. Zhang et al. [14] have specified that the density models are computationally expensive and the quality of the clusters depends on the selection of parameters and have come up with the concept of CGDBSCAN (Contribution-Grid based DBSCAN) algorithm

# 3. METHODOLOGY

The proposed framework of LRPSDC is comprised of five major steps which is described in this section. The methodology starts by inputting the raw dataset into feature selection process for the purpose of cleaning the dataset by removing the irrelevant features and missing values. The output of feature selection is a reduced dataset that consists only of meaningful features that are to be passed on to the linear regression segment. The linear regression segment converts the non-linear data into a linear form with regard to the labeled classifier. The addition of linearly transformed data is then performed to yield a single representation of data object. As the presence of outliers bias to an irregular distribution of data objects, an extreme value analysis method is employed to identify and exclude the outlier data objects from clustering. The next segment of the proposed methodology computes the upper and lower boundaries of each cluster through the Percentage Split Distribution (PSD) method. Once when the cluster boundaries are set, an if-then association is performed to place all data objects within the boundaries they fall under. The segment, cluster visualization concentrates on the visual alignment of data objects over the geographical space. The quality of clusters generated by the proposed methodology is then assessed by a novel SR silhouette cluster validation technique and finally the performance reports of the proposed methodology are presented. The elaborated description of the methodology is presented in the subsequent section.

## 3.1 Feature Selection

Not all features of the datasets are useful for the construction of a knowledge descriptive model, as some of them have very low impact with decision making. The identification and negation of those less impacted features are the primary concern of feature selection algorithms. In spite with the existence of various feature selection techniques this paper manipulates Pearson correlation method as the earlier findings of the author's works have proven that the percentage split distribution method is found to be effective with correlated features for obtaining enhanced clustering results. Hence, in this paper, the standard Pearson Correlation method is employed.

Pearson correlation is a method for examining the relationship between two variables R and S which would usually be a positive, negative or no correlation ranging from the values -1 to 1. The value 1 represents a positive correlation (when an increase in R lets an increase in S), -1 represents a negative correlation (increase in R lets a decrease in S) and 0 represents no correlation (when there is no such relationship) [15]. The Pearson correlation formula is denoted in Equ.1.

$$\mathrm{PCC} = \frac{\Sigma(\mathrm{R}-\mathrm{R})(\mathrm{S}-\mathrm{S})}{\sqrt{\Sigma(\mathrm{R}-\mathrm{R})^2.\Sigma(\mathrm{S}-\mathrm{S})^2}} \tag{1}$$

Where R' and S' are the means of variables R and S. Thus, this segment of methodology extracts only the most correlating features from the original set.

### 3.2 Linear Regression (LR)

Certain experiments have proven that the percentage split distribution method creates meaningful clusters with linear data. It is not obvious that the correlated data often forms a linear relationship. Hence, in this segment the correlated features of class label are transformed into a linear form through linear regression method [16]. A linear regression model that is formed with single dependent (scalar) and independent (explanatory) variable is called simple linear regression model which is denoted in equ.2.

$$y = mx + b \tag{2}$$

Where '*m*' is a slope and '*b*' is an intercept on y-axis. The computation of slope '*m*' is represented in equ.3 and line intercept on y-axis '*b*' is denoted in equ.4.

$$m = c \times \frac{S_e}{S_s} \tag{3}$$

'$S_e$' is the standard deviation of explanatory variable, '$S_s$' is the standard deviation of scalar variable and '*c*' is the correlation value of explanatory and scalar variables.

$$c = mean_y - m \times mean_x \tag{4}$$

A linear regression that is formed with single dependent and multiple independent variables is called a multiple linear regression which can be denoted with a notation presented in equ.5 for any object i.

$$y_i = c + m_1x_{1i} + m_2x_{2i} + m_3x_{3i} + \ldots + m_px_{pi} \tag{5}$$

Where '*c*' is the line intercept of y-axis, '*m*' is slope of line. The accumulation of RHS of the notations forms a Single Representation (SR) of data objects.

### 3.3 Outlier Detection

Outliers are the remarkable variation of data objects from the distribution of other objects. Detection of outliers is useful for the prediction of abnormality in the data which may be corrected or deleted. Detection of outliers is helpful in identifying the fraudulent activities in credit cards, military surveillance and network intrusion [17]. The performance of PSD is highly affected by outliers as the clusters are built as per the distribution of SR of data objects. Hence, the detection of outliers is executed as a process of normalizing the data objects. In the proposed methodology an extreme value analysis method is used as an outlier detection method which computes upper and lower quartile medians of the distribution so as to define the inter-quartile (RQ) median of data objects. The inter-quartile median is further used for computing the inner and outer fences of outlier data objects. The value that crosses the inner fence is called mild outliers and outer fence is called an extreme outliers. The data objects that falls under the inner and outer fences are identified and excluded from clustering. Fig.1. depicts the pseudo code for outlier detection.

```
1. Calculate the median R2 from SR values
2. Compute the lower quartile R1 from the median
3. Compute the upper quartile  R3 from the median
4. Compute the range of inter-quartile RQ by
   subtracting R3-R1
5. Compute lower inner fence as R1-1.5*RQ
6. Compute upper inner fence as R3+1.5*RQ
7. Compute lower outer fence as R1-3*RQ
8. Compute upper outer fence as R3+3*RQ
```

**Fig.1. Pseudo Code – Outlier detection**

### 3.4 Percentage Split Distribution

One of the primary objectives of this paper is to propose a simplified clustering technique that helps the user to conceptualize the task of clustering in a better way. Hence, the strategy of the proposed clustering consists of two simplified steps such as accumulation percentage split distribution and if-then association.

Percentage Split is a crucial step of the proposed method, where the data objects are assumed that they are well distributed between the range 0 to100% in which the min (SR) represents 0% and the max (SR) represents 100%. The clustering portions are then defined by dividing the 100% with number of inputted cluster. This step is an iterative process where each of the iterations calculates the lower and upper limits of each cluster$C_1 \ldots C_t$ using percentage split distribution formula denoted in Equ.6.

$$PSD = ((\max(SR) - \min(SR)) * percentage) + \min(SR) \tag{6}$$

Where max (SR) and min (SR) denote maximum and minimum values of SR and percentage represents the split value for each cluster.

If-then Association step gets the lower and upper limits of each cluster from the previous percentage split distribution step and formulates an if-then structure to assign the data objects in to its suitable clusters. SR value of a data object is compared with the boundary limits of each cluster and assigns the data object into the cluster that it falls under. This step is repeated until all data objects are assigned in a cluster. Fig.2. shows the step-by-step process of LRPSD.

### 3.5 Cluster Validation

The term cluster validation is concerned with the assessment of quality of clusters generated by any algorithm. Cluster compactness is a validation measure deals with the analysis of closeness of data objects within and between the cluster members. Compactness of the cluster members are often measured with the unit called "variance" [18] where the variance of a data object within the cluster should be minimized than the members of other clusters. In this paper, a novel SR silhouette cluster compactness measure is taken to analyze the compactness of cluster.

SR Silhouette coefficient metric measures the fitness of data object with its associating cluster. The SR silhouette coefficient of any individual object 'n' is computed using the formula denoted in equ.7.

$$s(n) = \frac{y(n) - x(n)}{\max\{x(n),\ y(n)\}} \tag{7}$$

where ‘x’ is the average distance of ‘n’ with the objects in the same cluster, ‘y’ is the minimum of average distance of ‘m’ to the objects with other clusters. The small value in x (n) and large value in y (n) denotes the best fit of data object ‘n’ with its associating cluster$C_i$. As like correlation the value closer to 1is expressed as the best results in SR silhouette co-efficient. The distance between the objects for analyzing the cluster compactness is computed using a new measure called SR Difference (SRD) measure which computes only the absolute SR difference of an object with all other objects using equ.8.

$$D_{ij} = \forall_{i=1}^{n} \forall_{j=1}^{n} \left| SR_i - SR_j \right| \tag{8}$$

where i and j denote the data objects. Moreover, the SR difference measure is simple and able to obtain the same result as Euclidean distance with less computational cost by evading the square, and root operations.

```
1.  Input dataset
2.  Obtain the most correlated subset of features using Pearson Correlation
3.  Store the subset of features in a dataset D
4.  Perform multiple linear regression to correlate the data objects with class label
        a. $y_i = c + m_1 x_{1i} + m_2 x_{2i} + m_3 x_{3i} + ... + m_p x_{pi}$
5.  Determine the SR value for each object
6.  Detect outliers using extreme value analysis method
7.  Get the number of cluster from the user
8.  Compute the Cluster Percentage (CP) as 100/ number of cluster
9.  percentage = CP
10. for each cluster C1.. Ct compute the percentage split distribution as
    PSD=((max(SR)-min(SR) )*percentage/100)+min(SR)
    percentage + = CP
11. Set the upper and lower limits for each cluster with the values obtained in step 6.
12. Assign the data object i in the cluster where the value of SRi falls under  through if-then association
13. Repeat step 8 until all data objects are clustered
14. Validate cluster with SR silhouette distance metric
```

**Fig. 2. Pseudo Code- LRPSDC**

## 4. EXPERIMENTATION

The performance of LRPSD clustering is analyzed by developing a JAVA program. The objective of the experiment is to cluster the cognitive and competency skills of the students in terms of their intelligent, attitude and hands on training where the intellectuals of the students represents the cognitive skills and attitude and hands on training refers to competency skills. The dataset is taken from StatCrunch data repository which contains four attributes namely IQ score, GPA, Gender and interview performance with 78 instances out of which the two most correlating features GPA and IQ score have been selected for the experiment. The goal of the experimentation is to group similar students with

- High cognitive and High competency skills - 26 students
- High cognitive and Low competency skills - 19 students
- Low cognitive and High competency skills - 15 students
- Low cognitive and Low competency skills - 18 students

## 5. RESULTS AND DISCUSSION

The results yielded from the experimentation clearly shows that the proposed algorithms work effectively in analyzing the cognitive and competency skills of students in higher education by correctly clustering the student data. Fig. 3.a. shows the cluster assignments of LRPSD clustering and fig.3.b. shows the SR Silhouette distance measure of LRPSD Clustering.

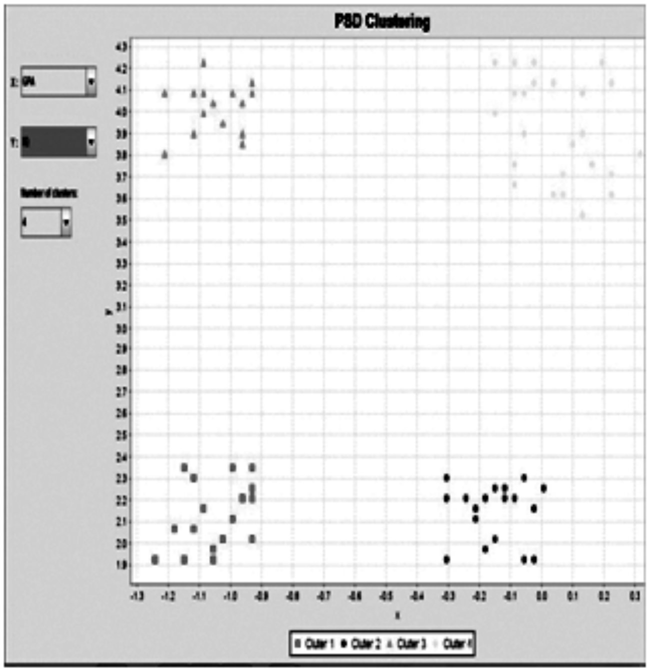


Fig.3.a. Cluster Assignments of LRPSD

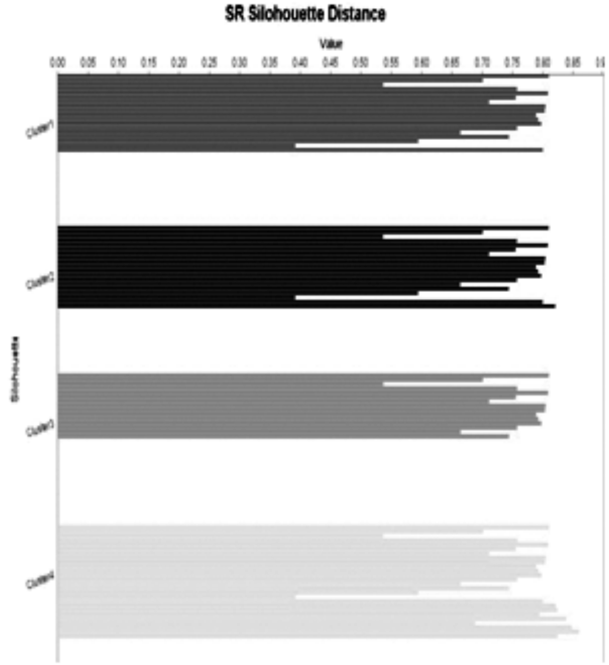
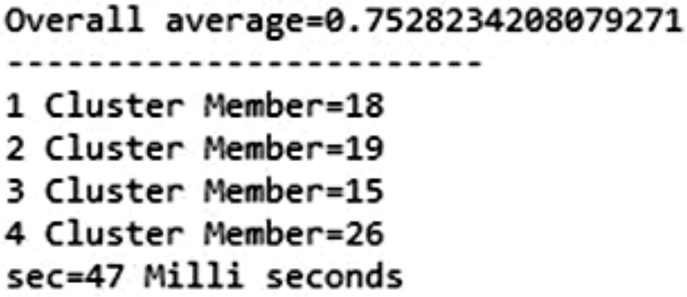


Fig.3.b. SR Silhouette Distance Measure of LRPSD

The average silhouette distance of LRPSD clustering with its cluster elements are given in fig.4.



```
Overall average=0.7528234208079271
-----------------------
1 Cluster Member=18
2 Cluster Member=19
3 Cluster Member=15
4 Cluster Member=26
sec=47 Milli seconds
```

Fig.4. Average Silhouette Distance with Cluster Elements

The results shown in fig.4. explicates that the cluster results of LRPSD is exactly matches with the original dataset with highest silhouette distance with 100% accuracy. Moreover the time taken to build clusters is also 47 milliseconds as the clustering is done in a single iteration.

The performance of LRPSDC is compared with traditional k-means algorithm with regard to time, cluster compactness and accuracy and the results are displayed in Table. 1.The results have clearly proven that LRPSDC is superior than k-means for all three measures.

Table 1

LRPSDC Performance Analysis

| *S. No.* | *Algorithm* | *Time (in milliseconds)* | *Cluster Compactness* | *Accuracy* |
|---|---|---|---|---|
| 1 | LRPSDC | 47 | 0.7528234 | 100% |
| 2 | K-means | 436 | 0.4646739 | 56% |

## 6. CONCLUSION

This paper presents a new procedure called Linear Regressive Percentage Split DistributionClustering which improves the efficiency of clustering process with minimized time and maximized accuracy. A novelty of this work is the implementation of multiple linear regression models with PSD method which normalizes the non-linear data into linear form. Unlike other clustering algorithms, LRPSDC clusters the data items with a single iteration, improves the efficiency of clustering by converging the local optimum solutions to global optimum. Moreover, LRPSD eliminates the concept of initialization and builds quality clusters.

## REFERENCES

[1] Bouguettaya, Athman, Qi Yu, Xumin Liu, Xiangmin Zhou, and Andy Song. "Efficient agglomerative hierarchical clustering,"*Expert Systems with Applications*,**42**, 2785-2797, 2015.

[2] Jeon, Yongkweon, and Sungroh Yoon. "Multi-threaded hierarchical clustering by parallel nearest-neighbor chaining."*Parallel and Distributed Systems, IEEE Transactions,***26**, 2534-2548, 2015.

[3] Almeida, J. A. S., L. M. S. Barbosa, A. A. C. C. Pais, and S. J. Formosinho. "Improving hierarchical cluster analysis: A new method with outlier detection and automatic clustering."*Chemometrics and Intelligent Laboratory Systems,***87**, 208-217, 2007.

[4] Pourjabbar, A., C. Sârbu, K. Kostarelos, J. W. Einax, and G. Büchel. " Fuzzy hierarchical cross-clustering of data from abandoned mine site contaminated with heavy metals."*Computers & Geosciences,***72,** 122-133, 2014.

[5] Wu, Xiaohong, Bin Wu, Jun Sun, ShengweiQiu, and Xiang Li. "A hybrid fuzzy K-harmonic means clustering algorithm."*Applied Mathematical Modelling*,**39**, 3398-3409, 2015.

[6] Prabha, K. Arun, and N. KarthikeyaniVisalakshi. "Improved Particle Swarm Optimization Based K-Means Clustering."*In Intelligent Computing Applications (ICICA)*, International Conference on IEEE, 59-63, 2014.

[7] Chawla, Sanjay, and Aristides Gionis. "k-means-: A Unified Approach to Clustering and Outlier Detection." *In SDM (2013)*, 189-197. 2013.

[8] Khandare, Anand D. "Modified K-means algorithm for emotional intelligence mining." *Computer Communication and Informatics (ICCCI), 2015 International Conference on IEEE*, 1-3,2015.

[9] Liu, Zhe, Yu-qing Song, Cong-huaXie, Feng Zhu, and Xiang Bao. "Clustering gene expression data analysis using an improved EM algorithm based on multivariate elliptical contoured mixture models."*Optik-International Journal for Light and Electron Optics***125**, 6388-6394, 2014.

[10] Volkovich, Z., R. Avros, and M. Golani. "On Initialization of the Expectation-Maximization Clustering Algorithm."*Global Journal of Technology and Optimization*,**2**, no. 117 (2011).

[11] Chen, Shu-Chuan Grace, and Bruce Lindsay. "Improving mixture tree construction using better EM algorithms."*Computational Statistics & Data Analysis,***74**, 17-25, 2014.

[12] Yu, Kai, Xin Dang, Henry Bart, and Yixin Chen. "Robust model-based learning via Spatial-EM algorithm."*Knowledge and Data Engineering*, *IEEE Transactions,***27**, no. 6 (2015), pp: 1670-1682.

[13] Ghanbarpour, Asieh, and BehroozMinaei. "EXDBSCAN: An extension of DBSCAN to detect clusters in multi-density datasets."*In Intelligent Systems (ICIS), 2014 Iranian Conference on IEEE*, 1-5,2014.

[14] Zhang, Linmeng, ZhigaoXu, and Fengqi Si. "CGDBSCAN: DBSCAN Algorithm Based on Contribution and Grid."*In Computational Intelligence and Design (ISCID)*, 368-371, 2013.

[15] Yu, Lei, and Huan Liu. "Feature selection for high-dimensional data: A fast correlation-based filter solution."*In ICML, vol. 3*, 856-863, 2003.

[16] Ari, Bertan, and H. Altay Güvenir. "Clustered linear regression."*Knowledge-Based Systems* **15**, 169-175, 2002.

[17] Singh, Karanjit, and ShuchitaUpadhyaya. "Outlier detection: applications and techniques."*International Journal of Computer Science Issues***9**, 307-323, 2012.

[18] Liu, Yanchi, Zhongmou Li, HuiXiong, Xuedong Gao, and Junjie Wu. "Understanding of internal clustering validation measures."*In Data Mining (ICDM), 2010 IEEE 10th International Conference on IEEE,* 911-916, 2010.