IJCTA, 9(40), 2016, pp. 119-127

Analysing the Performance in Cluster Formation for Micro Financial Data using Frequent Itemset

Antony S. Alexander* C. Jothivenkateswaran and A. Clementking*****

Abstract : Data mining plays the major role in innumerable analysis for building applications. This work is initiated with the intervention of such data mining technique for processing the micro financial data for clustering, with frequent itemset after normalizing the data. The technique also isolates the noises on the data with missing values and outlier removal. The data preprocessing is done with the available data and then they are nurtured for the development of clusters. The Density Based Spatial Clustering of Application with Noise (DBSCAN) technique is used to determine the optimal clusters with the novel data preset as the frequent item of the given dataset. This clustered data centers can be used in the future classification process.

Keyword: Clustering, Normalization, Frequent itemset, DBscan, outlier.

1. INTRODUCTION

Clustering is a data mining technique to assemble set of data objects into multiple groups or clusters, so that objects with in the cluster will exist with high similarity, but will have dissimilar properties when compared to other clusters. The dissimilarities and similarities of and among the cluster are assessed based on the attribute value that describes the data objects.

Clustering techniques are widely used for data organization, categorization, compression and model construction etc. Many algorithms are designed and developed to perform clustering for various data sets which are of numeric or categorical, those algorithms are also categorized on several aspects such as partitioning methods, density-based methods, grid-based methods and hierarchical methods[6]. Further data set can be Cluster analysis are widely used as standalone data mining tool to gain comprehension into the data distribution, or as a preprocessing method for other data mining algorithms that are operating on detected clusters[7].

This study is on micro financial fund flow, to derive a sustainable growth within the optimal set of item. This optimal set is analyzed with Density Based Clustering technique by using the frequent itemset(s). A frequent itemset in this will be high dimensionality data set that takes the form of numerical value after preprocessing which occur together frequently and are good candidates for clusters. This paper is organized as follows: Section 2 about the related work. Section 3 briefly describes methodology, work flow and details our approaches relevant techniques and Section 4 presents the results and discussion. Finally we conclude this research in Section 5.

* Research Scholar, P.G. and Research Department of Computer Science and Computer Applications, Presidency College, Chennai, India. *Email: antonysalex@gmail.com*

** Associate Professor and Head, P.G. and Research Department of Computer Science and Computer Applications, Presidency College, Chennai, India. *Email: jothivenkateswaran@yahoo.co.in*

*** Associate Professor, Dept of Computer Science, College of Computer Science, King Khalid University, Abha, KSA *Email: clementking1975@gmail.com*

2. METHODOLOGY

The clustering method in this work is used to form some meaningful data clusters pairing with the frequent items occurring between the data. The clustering techniques are applied and utilized for the analysis of diversified varieties of domain data such as psychology and other social sciences, biology, statistics, pertaining to pattern recognition, information retrieval, machine learning, and data mining.

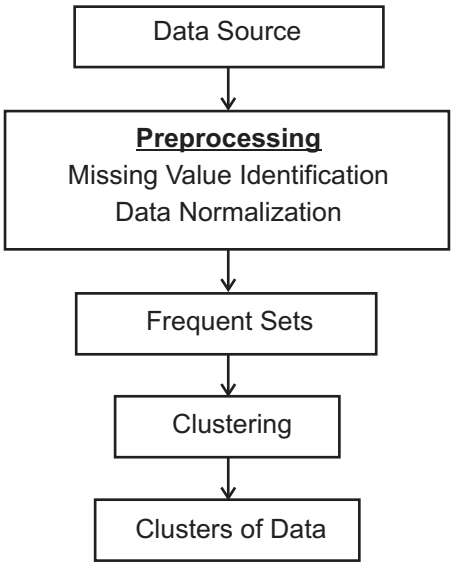


Figure 1: Methodology

2.1. Preprocessing

This phase deliberates the preprocessing of data for the easy utilization over the application build with data mining algorithms and for the improvement of the effectiveness or performance of the same on the micro financial datasets for optimal fund flow. During this process the slips like missing value identification, outlier detection and minimum and maximum normalization were incurred[5].

Missing data is one of the issues which are to be fathomed for given financial data. Improper data analysis leads to predisposition on the result. Missing values are handled with mean substitution method where the missing values for each attribute are substituted with the mean estimation of the observed values. The mean substitution method ensures for no missing values in the dataset.

Outliers are observation points that are distant from other observations. Two graphical techniques such as scatter plot and box plot are used for identifying the outliers. The box plot can be considered for detecting the outliers in datasets, which identify observations that are deemed improbable based on mean and standard deviation[13].

The distribution is examined with Gaussian distribution which is a statistical method used to find the fitness of the financial datasets. The dataset is scaled to fit into a specific range for analyzing the data. This work is done using min-max normalization. The features are transformed from x_1 to y_1 which fits in the range $[x_2, y_2]$. It is given by the following formula

$$y_1 = \left(\frac{\left(x_1 - \text{Minimum value of } x_1\right)}{\text{Maximum value of } x_1 - \text{Minimum value of } x_1} \right) * \left(y_2 - x_2\right) + x_2$$

2.2. Frequent Sets

Data mining is an important technique to extract interesting knowledge in large databases and to find frequent itemset for clustering from massive volume of data[14]. Frequent sets are essential for many Data Mining tasks such as finding interesting patterns from databases, association rules, sequences, episodes, correlations, clusters and classifiers[10]. Identifying frequent itemset is one of the important techniques,

since it supports in the identification of sets of paired items, like products, identical symptoms and related characteristics, which often occur together in the given huge databases[12].

The goal for searching frequent itemset of this data set is to analyze the financial transaction data, for example customer (benefactor) along with their money transaction in terms of the deposit or credit for being utilized for various purposes[5].

Frequent transaction of money along with the nature of investments and the outcome in terms of profit out of the expenditure/investment can be often described as frequent item in this micro financial data set[15].

Formally, let **I** denote the set of items, **D** for dataset and **T** to mean a transaction.

1. A transaction over **I** is a couple **T** = (*tid*, **I**) where *tid* is the transaction identifier and **I** is the set of items from **I**.
2. Dataset D over **I** is a set of transactions done over **I** such that, each transaction has a unique identifier. **I** is left, whenever it is clear from the context.
3. A transaction **T** = (*tid*, **I**) is said to support a set **X**, if **X C I**. The cover of a set **X** in **D** consists of the set of transaction identifiers of transactions in **D** that support **X**, The support of a set **X** in **D** is the number of transactions in the cover of **X** in **D**. The frequency of a set **X** in **D** is the probability that **X** occurs in a transaction, or in other words, the support of **X** divided by the total number of transactions in the database. D is eliminated whenever it is clear from the context

2.3. Clustering Techniques

The clustering techniques can be mainly categorized in to partition and hierarchical. As discussed in introduction many other methods are used for cluster analysis. This research had used partition techniques, in Density Based Spatial Clustering of Application with Noise (DBSCAN) classic techniques for clustering. DBSCAN can be used to finds all types of clusters properly, sovereign of the size, shape, and location to each other. It produces clusters by considering the density of nearest neighborhood objects. It is based on "density reachability" and "density connectability", both depends on two input parameters[11].

There are

1. input parameter- size of epsilon neighborhood ε and
2, minimum terms of local distribution of nearest neighbors *m*.

Here epsilon (ε) parameter controls size of neighborhood and clusters. It starts with an initial arbitrary point that has not been visited[1]. The point's ε -neighborhood is retrieved, and if it contains suitable points, a cluster is initiated. Otherwise the point is labelled as noise. The number of point parameter impacts detection of outliers. DBSCAN will target on low-dimensional spatial data[2]. The implementation of this idea for partitioning of a finite set of points requires concepts of density, connectivity and boundary data. The approach to cluster data objects is as follows:

1. The influence of an object to its neighborhood is given by an influence function.
2. Overall density is modeled as the sum of the influence functions of all objects.
3. Clusters are determined by density attractors, where density attractors are local maximum of the overall density function.

The influence function can be an arbitrary one, as long as it is determined by the distance $d(\hat{x},\hat{y})$ between the objects. Examples are the square wave influence function:

$$f_{Square}(\hat{x},\hat{y}) = \begin{cases} 0 & \text{if } d(\hat{x},\hat{y}) > \sigma \\ 1 & \text{otherwise} \end{cases}$$

where σ is a threshold, or a Gaussian:

$$f_{\text{Gaussian}}(\hat{x},\hat{y}) = e^{-\frac{d(\hat{x},\hat{y})^2}{2\sigma^2}}$$

This work, has taken *Ester et. al's*. employ a spatial index to help in discovery the neighbors of a data point. Thus, the complexity is improved to O($nlog_n$), as opposed to O(n^2) without the index. Finally, if Euclidean distance is used to measure proximity of objects, its performance degrades for high dimensional data[4].

Let X = {x_1, x_2, x_3, ..., x_n} be the set of points of the normalized data. DBSCAN is processed with two parameters: ε – epsilon and the minimum points required to form a cluster[4].

Pseudo code for DBScan Clustring

Step 1: Start with the initial arbitrary starting point that has not been visited yet.

Step 2: Extract the nearest neighborhood of this point using ε -epsilon

Step 3: If there are satisfactory neighborhoods around this point then clustering process is initiated and point is marked as visited otherwise it is labeled as noise

Step 4: If a point is found to be a part of the cluster then its ε -epsilon neighborhood is also the part of the cluster and procedure of **step 2** is repeated for all ε -epsilon neighborhood points.

Step 5: A new unvisited point is retrieved and processed, leading to the discovery of a further clusters or noise.

Step 6: This process continues until all points are marked as visited

3. RESULTS AND DISCUSSION

In this section shields with data normalization, finding of frequent itemset and clustering results respectively.

3.1. Data Normalization

The Micro Financial dataset which is taken as example has fifteen different fields which are of both numeric and string. Using mean deviation, data normalization is done for identifying the missing values and outlier remova[l4,2]. The data are replaced with the nearest neighborhood technique. The resultant value is transformed to be numerical by using the Min-Max normalization technique[11]. The below Table -1 contains the sample of few fields of the normalized data.

Table 1

Normalized Data

F1	*F2*	*F3*	*F4*	*F5*	*F6*	*F7*
0.11659	0.030303	0.057175	0.057175	0.001121	0.11765	0.73318
0.057175	0.030303	0.13173	0.089686	0.000561	0.33333	0.73318
0.057175	0.030303	0.13173	0.089686	0.000561	0.64706	0.26682
0.036996	0.060606	0.03139	0.03139	0.001121	0.84314	0.73318

F8	*F9*	*F10*	*F11*	*F12*	*F13*	*F14*	*F15*
0.10818	0.032511	0.20291	0.066143	0.031447	0.002242	0.14583	0.025408
0.10818	0.032511	0.20291	0.066143	0.34591	0.24271	0.375	0.065336
0.86715	0.03083	0.71469	0.06222	0.13208	0.24271	0.16667	0.029038
0.86715	0.1093	0.20291	0.81783	0.19497	0.24271	0.125	0.021779

3.2. Frequent itemset

The normalized dataset is subjected with the frequent itemset process. Where, each transaction T is described as set of items of the micro financial data. Each transaction is entitled and identified with the Transaction ID (TID). An Itemset will contain a set of items. K-itemset will have K item of frequent items. The support count or frequency count are identified in the frequent item, which will define the number of times an itemset occurs in the process. If the support count of the itemset satisfies given the support threshold, then the Itemset I is considered to be frequent itemset[4].

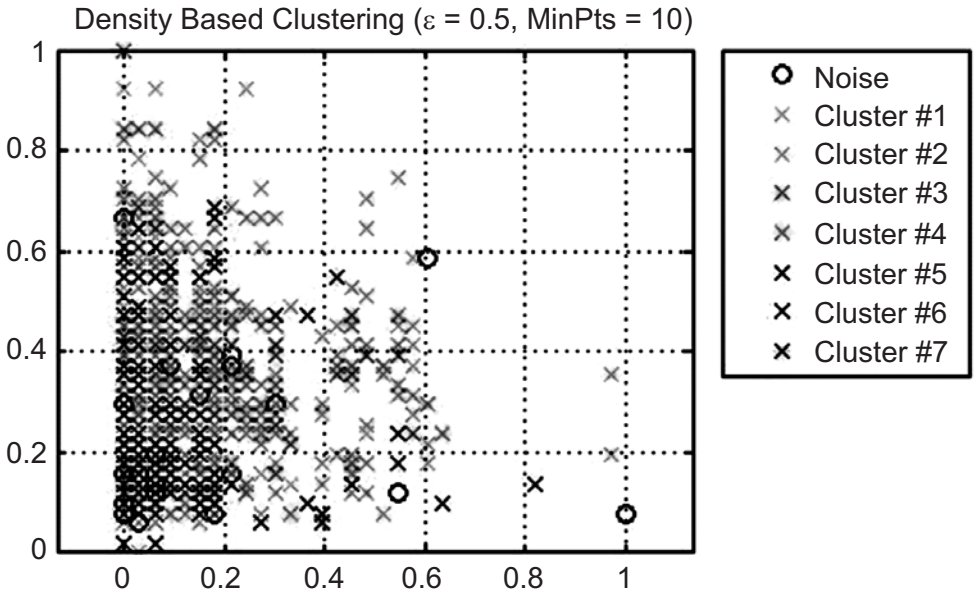
Table 2
Frequent Itemsets

Transaction	*No. of Itemsets*
T1	9
T2	36
T3	84
T4	126
T5	126
T6	84
T7	36
T8	9

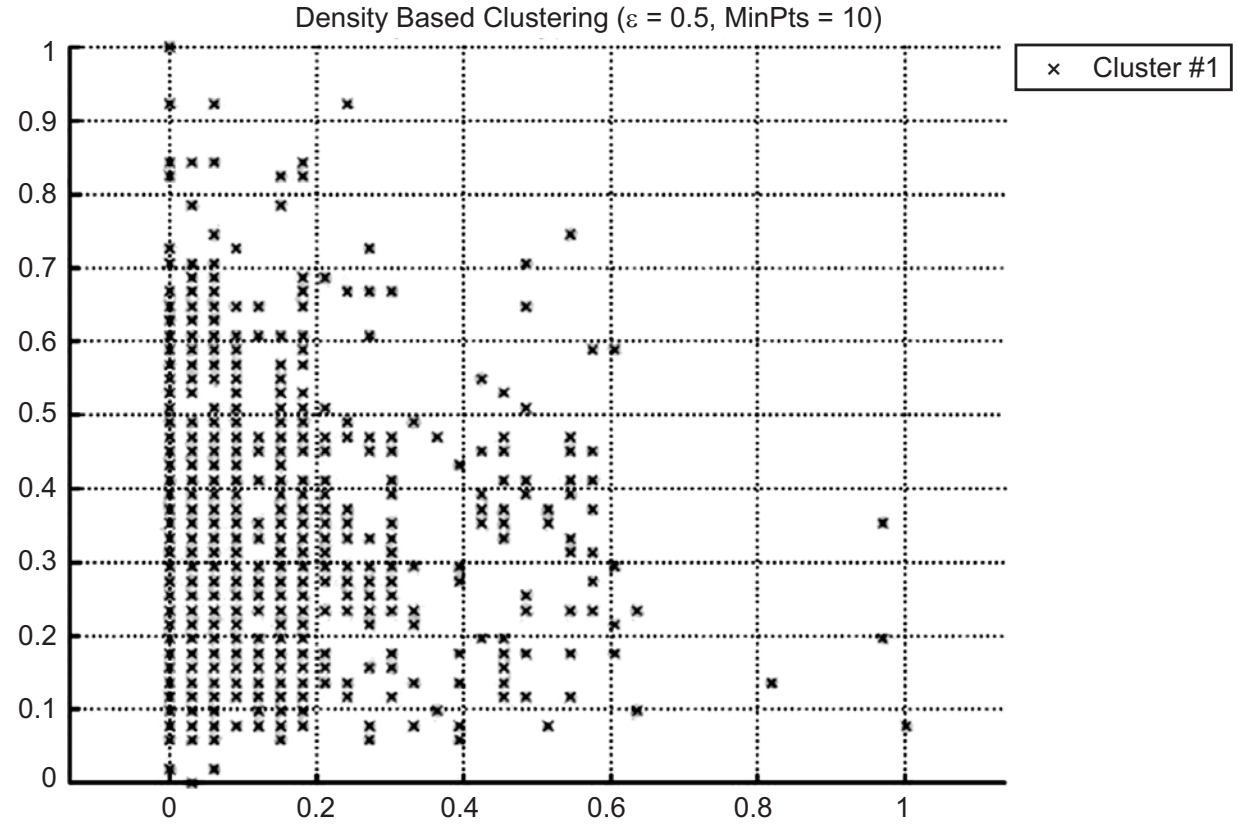
The table-2 is represents an example of eight transactions of frequent itemset derived from five hundred and two of financial dataset. Each transaction represents the different item which varies with diverse purpose money transaction in the financial dataset.

3.3. Clustered development

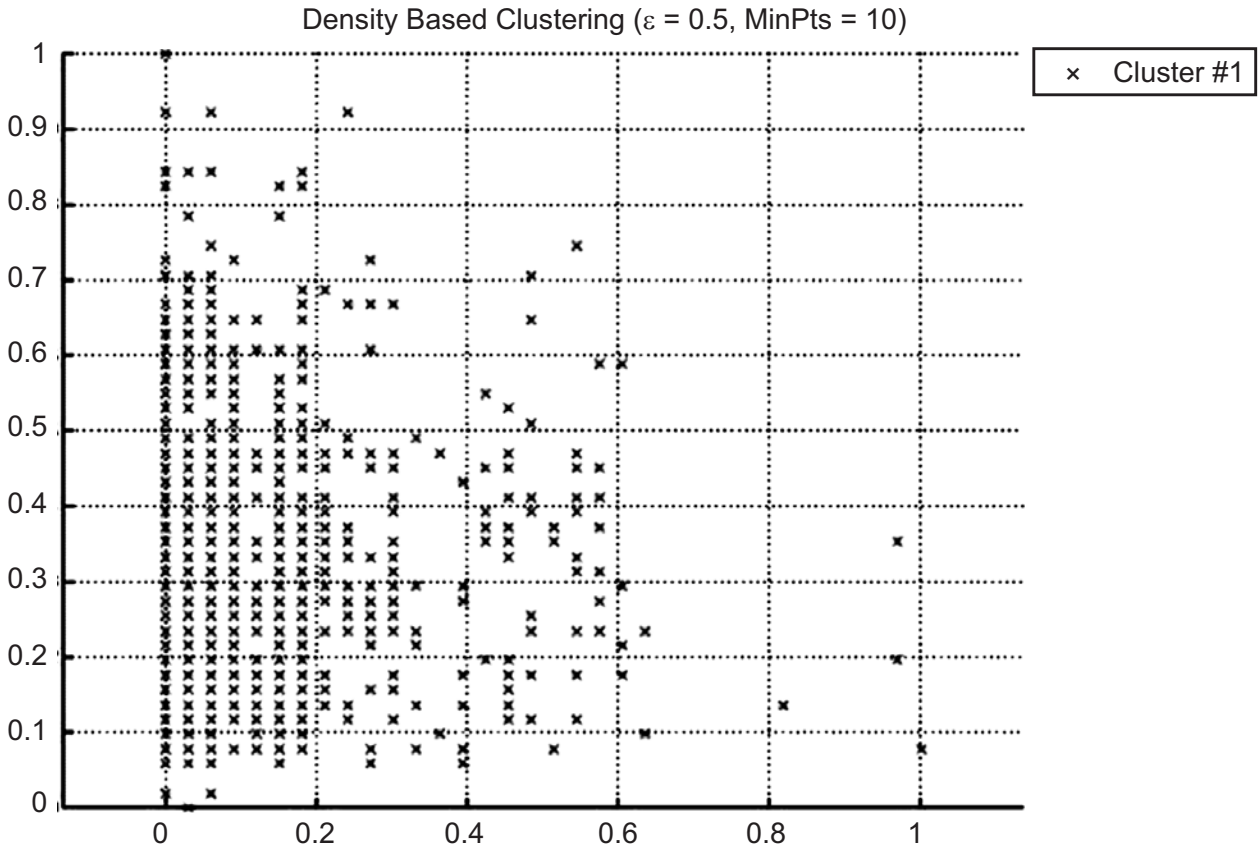
The result of the various transaction itemset clusters are imagined below. The result is derived with the value of epsilon- ε = 0.5 and density points (minimum points) = 10.



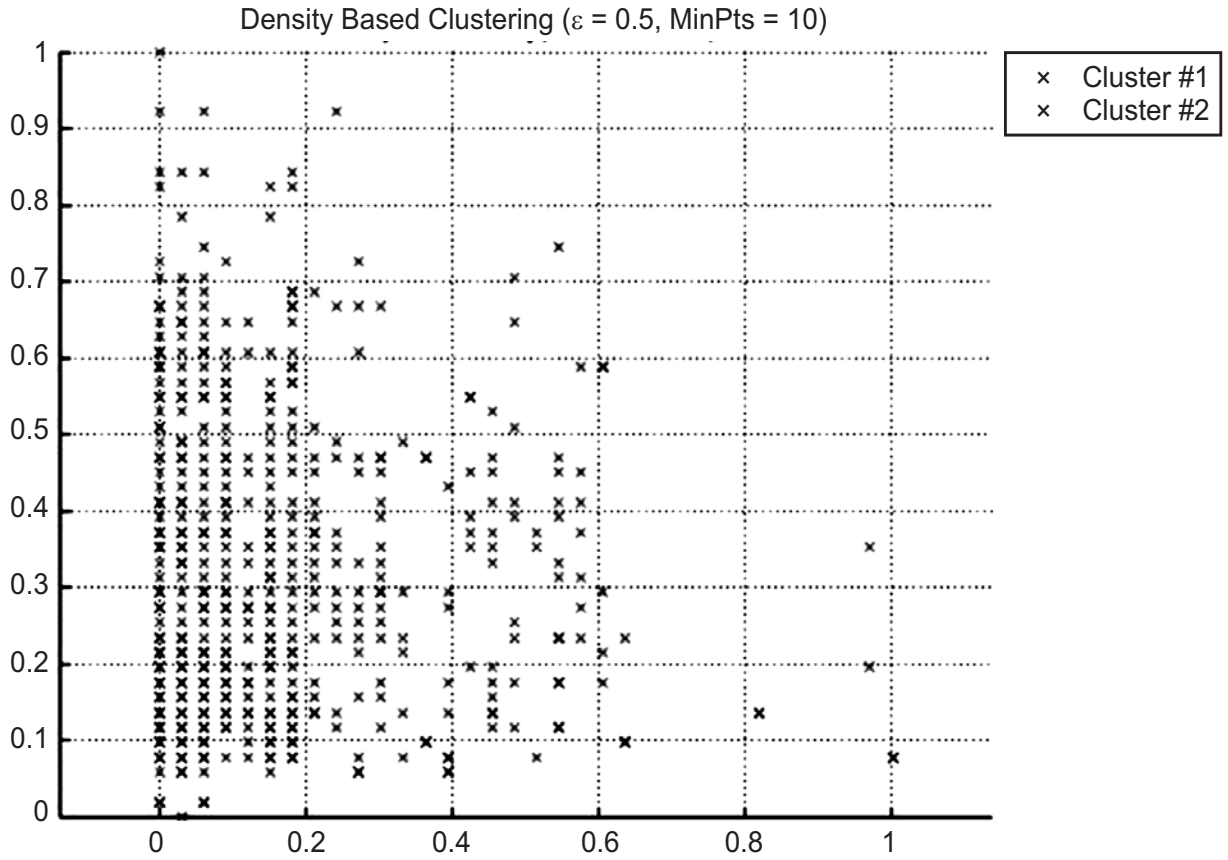
(*a*) T1 {2,6,7,8,10,11,12,14,15}



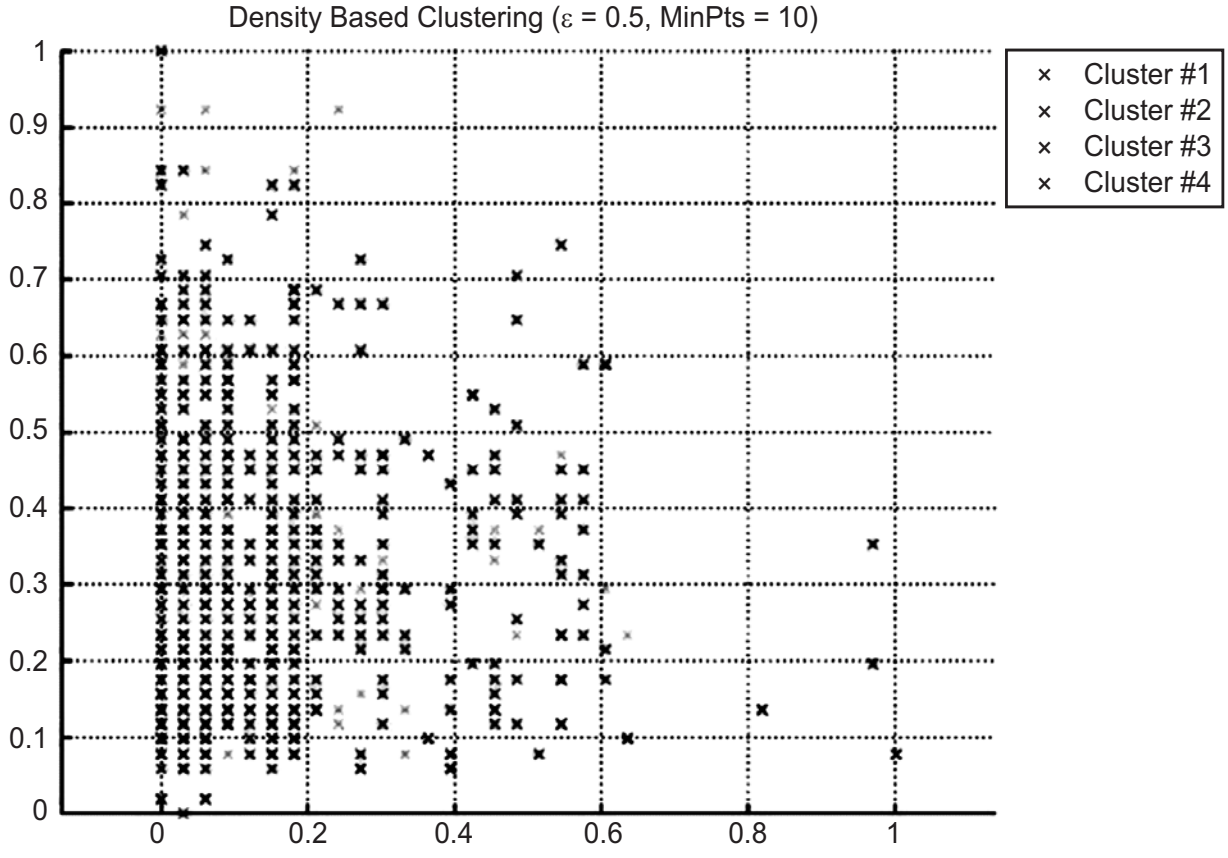
(*b*) T2 {2,6}



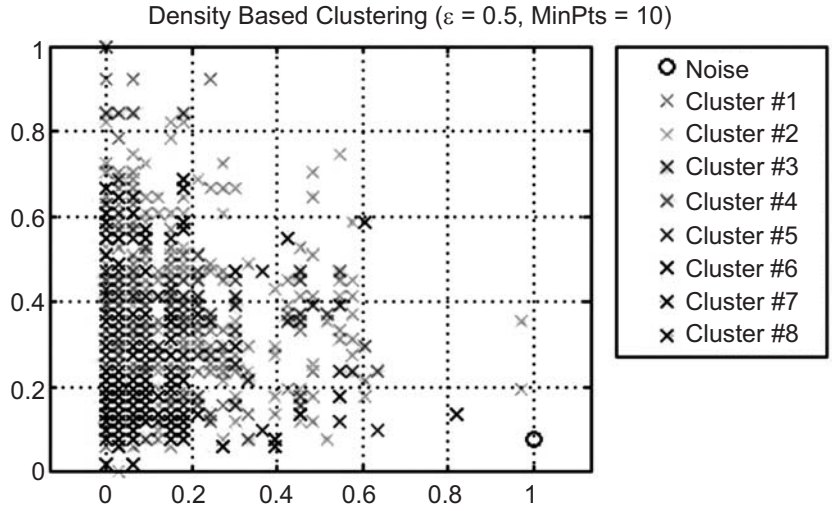
(*c*) T3 {2,6,8}



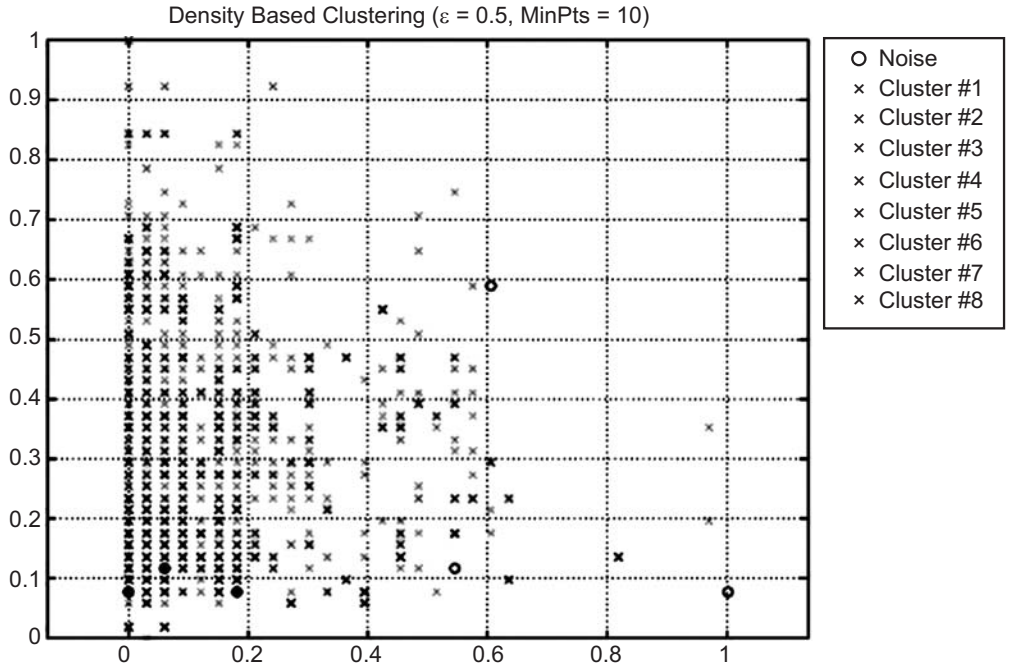
(*d*) T4 {2,6,7,11}



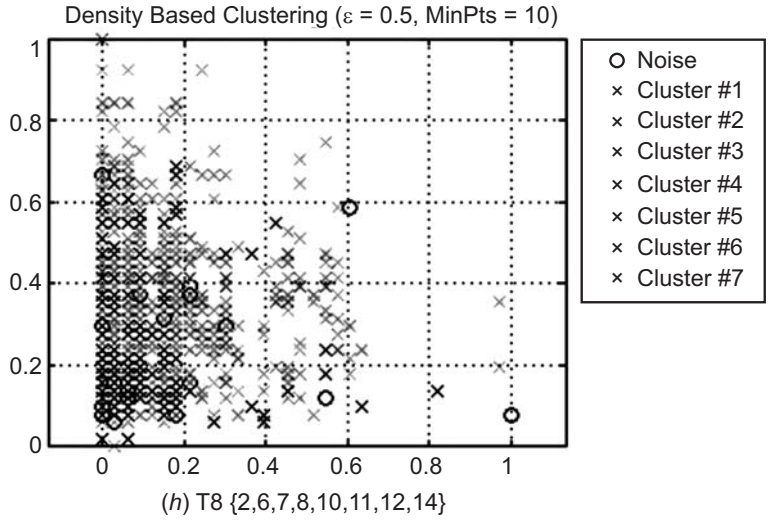
(*e*) T5 {2,6,8,10,14}



(*f*) T6 {2,7,8,10,11,12}



(*g*) T7 {2,8,10,11,12,14,15}



(*h*) T8 {2,6,7,8,10,11,12,14}

Figure 2: Different clusters after various transactional dataset

In the above figure 2-(*a*) has nine items which is derived into seven different clusters centres with few data as noise, (*b*) and (*c*) has two and three itemset respectively which is derived into single cluster without noise data, (*d*) has four itemset which is derived into two cluster centres which doesn't have noise data, (*e*) has five itemset which is derived into four cluster centres which doesn't have noise data, (*f*) has six itemset which is derived into eight cluster centres which noise, (*g*) seven itemset had been derived into eight cluster centres which noise data, (*e*) With eight itemset it has derived seven cluster centres which negligible amount noise data,

4. CONCLUSION

The optimal frequent item set subject to clusters for the Micro Financial Dataset has been executed with the DBSCAN (Density Based Spatial Clustering of Application with Noise) based on frequent itemset. Though it is not apparent at first sight, after execution the resulted cluster values are analyzed. This analysis as in the visualized figure show that the development of cluster using this technique on the transactional micro financial data with diversified data types has been executed faster with in the time complexity which has resulted with independent and optimal cluster. It has also ascertained that density-based clustering done on N number of itemset will have K item of frequent items among the selected frequent item occurrence. The clusters itemset can be feed forwarded to classification.

5. REFERENCES

1. Verma, Manish, et al. "A comparative study of various clustering algorithms in data mining", *International Journal of Engineering Research and Applications (IJERA)* 2.3 (2012): 1379-1384.
2. Han, Jiawei, Jian Pei, and Micheline Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.
3. Mining, Data. "Concepts and Techniques" *Jiawei Han and Micheline Kamber*(2001).
4. Ester, Martin, et al. "A density-based algorithm for discovering clusters in large spatial databases with noise", *Kdd*. Vol. 96. No. 34. 1996.
5. Han, Jiawei, Jian Pei, and Micheline Kamber. "*Data mining: concepts and techniques"*, Elsevier, 2011.
6. Velmurugan, T. and Naveen, A., "Analysing MRI Brain Images Using Fuzzy C–Means Algorithm", International Journal of Control Theory and Applications, vol. 9, no. 10, pp. 4661-4675, 2016
7. Naveen, A., and T. Velmurugan. "Identification of calcification in MRI brain images by k-means algorithm", *Indian Journal of Science and Technology* 8.29 (2015): 1.
8. Zhang, Wen, et al. "Text clustering using frequent itemsets", *Knowledge-Based Systems* 23.5 (2010): 379-388.
9. Malik, Hassan H. "*Efficient algorithms for clustering and classifying high dimensional text and discretized data using interesting patterns*" Diss. Columbia University, 2008.
10. Ghosh, Soumadip, et al. "Mining frequent itemsets using genetic algorithm", *arXiv preprint arXiv:1011.0328* (2010).
11. Huang, Zhexue. "A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining" *DMKD*. 1997.
12. Andritsos, Periklis. "Data clustering techniques" *Rapport technique, University of Toronto. Department of Computer Science* (2002).
13. Wagstaff, Kiri, et al. "Constrained k-means clustering with background knowledge" *ICML*. Vol. 1. 2001.
14. Beil, Florian, Martin Ester, and Xiaowei Xu. "Frequent term-based text clustering." *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2002.
15. Zimek, Arthur, Ira Assent, and Jilles Vreeken. "Frequent pattern mining algorithms for data clustering", *Frequent Pattern Mining*. Springer International Publishing, 2014. 403-423.