Asian Journal of Information Technology 13 (10): 599-605, 2014
ISSN: 1682-3915

Multi Cluster Dimensional Projection on Quantum Distributed Interference Data

[1]L.V. Arun Shalin and [2]K. Prasadh
[1]Department of Computer Science and Engineering,
Manonanium Sundarnar University, Tirunelveli, Tamil Nadu, India
[2]Department of Computer Science and Engineering,
Mookambika Technical Campus, Ernakulam, Kerala, India

Abstract: Clustering intends to divide the subset and group the similar objects with respect to the given similarity measure. Clustering includes number of techniques which includes statistics, pattern recognition, data mining and other fields. Projected clustering screens the data set into numerous disjoint clusters with the outliers so that each cluster exists in a subspace. The majority of the clustering techniques are considerably incompetent in providing the objective function. To overcome the attribute relevancy and the redundancy by providing the objective function, we are going to implement a new technique termed Multi cluster Dimensional Projection on Quantum Distribution (MDPQD). This technique evolved a discrete dimensional projection clusters using the Quantum Distribution Model. Multi cluster dimension on quantum distribution considers the problem of relevancy of attribute and redundancy. An analytical and empirical result offers a multi cluster formation based on objective function and to evolve a dimensional projection clusters. Performance of the multi cluster dimensional projection on quantum distribution is measured in terms of an efficient multi cluster formation with respect to the data set dimensionality, comparison of accuracy with all other algorithms and scalability of quantum distribution.

Key words: Dimensional projection, discrete, interference data, Quantum Distribution Model, redundancy, relevancy attributes, multi cluster

INTRODUCTION

Clustering is an accepted data mining technique for a diversity of applications. One of the incentives for its recognition is the capability to work on datasets with minimum or no a prior knowledge. This builds clustering realistic for real world applications. In recent times, discrete dimensional data has awakened the interest of database researchers due to its innovative challenges brought to the community. In discrete dimensional space, the space from a report to its adjacent neighbor can approach its space to the outermost reports. In the circumstance of clustering, the problem causes the space between two reports of the same cluster to move toward the space between two reports of different clusters. Traditional clustering methods may not succeed to distinguish the precise clusters using Quantum Distribution Model.

Fuzzy techniques have been used for handling vague boundaries of arbitrarily oriented clusters. However, traditional clustering algorithms tend to break down in high dimensional spaces due to inherent sparsity of data. Puri and Kumar (2011) propose a modification in the function of Gustafson-Kessel Clustering algorithm for projected clustering and prove the convergence of the resulting algorithm. It present the results of applying the proposed projected Gustafson-Kessel Clustering algorithm to synthetic and UCI data sets and also suggest a way of extending it to a rough set based algorithm.

As in the case of traditional clustering, the purpose of Discrete Dimensional Projected Clustering algorithms is to form clusters with most encouraging quality. However, the traditional functions used in estimate the cluster quality may not be appropriate in the predictable case. The algorithms will consequently be likely to select few attributes values for each cluster which might be inadequate for clustering the reports correctly. In some previous researches on projected clustering (Bouguessa and Wang, 2009), the clusters are evaluated by using one or more of the following criteria:

- Space between the values of the attribute to produce the relevance result
- Size of the selected attribute value in the cluster
- Size of the associate reports in the cluster

Corresponding Author: L.V. Arun Shalin, Department of Computer Science and Engineering, Manonanium Sundarnar University, Tirunelveli, Tamil Nadu, India

Spontaneously, a small standard space between attribute values in a cluster indicates that the associate reports agree on a small range of values which can make the reports easily restricted. A large number of selected attributes value towards the reports are analogous at a discrete dimensional, so they are very credible to belong to the same real cluster. Finally, a large number of reports in the cluster point out there are a high support for the selected attributes value and it is improbable that the small distances are merely by chance.

All these are indicators for a high-quality multiple clusters but there is actually a tradeoff between them. Assume a given set of reports, it selects only attributes value that tend to create the common space among reports, fewer attributes will be selected. Similarly, for a space obligation, locating more reports into a cluster will probably increase the average number of attribute value chosen.

It's important to point out that in this research; we focus on Multi cluster Dimensional Projection on Quantum Distribution (MDPQD) under the situation of Quantum Distribution Model to determine a discrete dimensional clustering. This form of multi cluster is unusual from all other clustering schemes. Thus, we compare the multi cluster dimensional projection on quantum distribution with Partitioned Distance-Based Projected Clustering algorithm and enhanced approach for projecting clustering in terms of efficient multi cluster formation with respect to the data set dimensionality, comparison of accuracy with all other algorithms. An analytical and empirical result offers a multi cluster formation based on objective function and to evolve a dimensional projection clusters.

Literature review: Most existing clustering algorithms become substantially inefficient if the required similarity measure is computed between data points in the full-dimensional space. To address this problem, Bouguessa and Wang (2009) a number of projected clustering algorithms have been proposed. However, most of them encounter difficulties when clusters hide in subspaces with very low dimensionality.

Gajawada and Toshniwal (2012) propose VINAYAKA, a Semi-Supervised Projected Clustering Method based on DE. In this method, DE optimizes a hybrid cluster validation index. Subspace Clustering Quality Estimate index (SCQE index) is used for internal cluster validation and Gini index gain is used for external cluster validation in the proposed hybrid cluster validation index. Proposed method is applied on Wisconsin breast cancer dataset.

Hierarchical clustering is one of the most important tasks in data mining. However, the existing hierarchical clustering algorithms are time-consuming and have low clustering quality because of ignoring the constraints. In this study, Hang *et al*. (2009), a Hierarchical Clustering algorithm based on K-Means with Constraints (HCAKC) is proposed.

Shanmugapriya and Punithavalli (2012), an algorithm called Modified Projected K-Means Clustering algorithm with effective distance measure is designed to generalize K-Means algorithm with the objective of managing the high dimensional data. The experimental results confirm that the proposed algorithm is an efficient algorithm with better clustering accuracy and very less execution time than the Standard K-Means and General K-Means algorithms.

Survey by Kriegel *et al*. (2009), tries to clarify: the different problem definitions related to subspace clustering in general, the specific difficulties encountered in this field of research, the varying assumptions, heuristics and intuitions forming the basis of different approaches and how several prominent solutions tackle different problems. Sembiring *et al*. (2010), PROCLUS performs better in terms of time of calculation and produced the least number of un-clustered data while STATPC outperforms PROCLUS and P3C in the accuracy of both cluster points and relevant attributes found.

Inspired from the recent developments on manifold learning and L1-regularized models for subset selection, Cai *et al*. (2010) propose in a new approach, called Multi-Cluster Feature Selection (MCFS) for unsupervised feature selection. Specifically, we select those features such that the multi-cluster structure of the data can be best preserved. The corresponding optimization problem can be efficiently solved, since it only involves a sparse eigen-problem and a L1-regularized least squares problem.

Yang and Chen (2011), proposed Weighted Clustering Ensemble algorithm provides an effective enabling technique for the joint use of different representations which cuts the information loss in a single representation and exploits various information sources underlying temporal data but does not contain the extracted feature. Jiang *et al*. (2011) have one extracted feature for each cluster. The extracted feature, corresponding to a cluster is a weighted combination of the words contained in the cluster. By this algorithm, the derived membership functions match closely with and describe properly the real distribution of the training data.

Nie *et al*. (2012) employs a Probabilistic algorithm to estimate the most likely location and containment for each object. By performing such online inference, it enables online compression that recognizes and removes redundant information from the output stream of this substrate. We have implemented a prototype of the

inference and compression substrate and evaluated it using both real traces from a laboratory warehouse setup and synthetic traces emulating enterprise supply chains. To evolve a discrete dimensional projection clusters, a new technique named Multi cluster Dimensional Projection on Quantum Distribution (MDPQD) scheme is presented.

MATERIALS AND METHODS

Proposed multi cluster dimensional projection using Quantum Distribution Model: The proposed research is efficiently designed for projecting the clusters in discrete dimensional by adapting the Multi cluster Dimensional Projection on Quantum Distribution (MDPQD). The proposed Multi cluster Dimensional Projection on Quantum Distribution (MDPQD) is processed under different input, intermediate and output processes:

- The input unit takes the Habitats of Human data based on Socio Democratic Cultures (HHSD) dataset
- The objective functions are choosing to obtain the precise goal by eliminating the relevancy and redundancy of attributes
- The quantum distribution technique is used to discover a model without the relevancy analysis and redundancy of attributes
- The activity formed in the proposed model is the multi cluster formations using the different set if attribute values
- The output unit contains the discrete dimensional projection clusters

The architecture diagram of the proposed Multi cluster Dimensional Projection on Quantum Distribution (MDPQD) is shown in Fig. 1.

In the Fig. 1 while generating a HHSD dataset, the size of each cluster and the domain of each attribute were first determined arbitrarily. This dataset is used to list the different types of attributes (i.e.,) food habits, culture, weather conditions and business conditions. The values are assigned to the attributes depending on the environment. The similar attribute values are grouped together to form the clusters. Each cluster then precise in selecting the attribute value. The multiple clusters are formed for the different types of attributes. For each attribute value of a cluster a confined mean was chosen precisely from the domain. Each report in the multi cluster determines whether to follow the precise attribute values according to the data error rate.

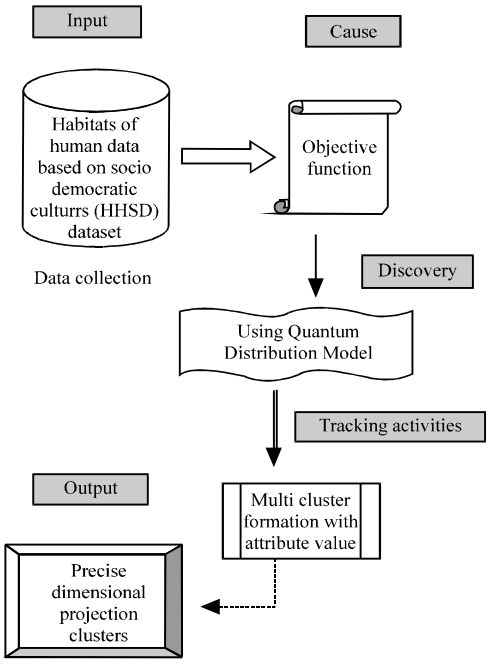

Fig. 1: Proposed Multi cluster Dimensional Projection on Quantum Distribution (MDPQD) process

Quantum distribution is optimized with a given constraints and with variables that need to be minimized or maximized using programming techniques. An objective function can be the result of an effort to communicate a business goal in mathematical terms for use in assessment analysis and optimization studies. Applications are frequently necessitate a QD Model by simpler form so that an available computational objective function approach can be used. An a priori bound is resultant on the quantity of error which such an rough calculation can encourage. This leads to a natural criterion for selecting the most excellent precise attribute value is chosen.

Objective function using QDM: We assume that the Habitats of Human data based on Socio Democratic Cultures (HHSD) dataset consists of a set of incoming reports which are denoted by $\overline{Y}1 \ldots \overline{Y}i$. It is assumed that the data point $\overline{Y}i$ is received at the time stamp Si. It is assumed that the discrete dimensionality of the HHSD data set 'h'. The 'h' dimensions of the report $\overline{Y}i$ are denoted by ($y^1i \ldots y^di$). In addition, each data point has an error associated with the dissimilar dimensions. The error associated with the kth dimension for data point Yi is denoted by $\psi k(\overline{Y}i)$.

We remind that many indecisive data mining algorithms use the probability density function in order to characterize the underlying behavior. Consequently,

we construct the more modest assumption that only error variances are available. Since, the diverse dimensions of the data may replicate diverse quantities, they may correspond to very different scales. In order to take the precise behavior of the different discrete dimensions into account, we need to perform quantum distribution across different discrete dimensions. This is done by maintaining global statistics. This statistics is used to compute global variances. These variances are used to scale the data over time with values.

In order to include the greater importance of recent data points in a developing stream, we use the concept of an objective function f(s) which quantifies the relative importance of the different data points over time. The objective function is drawn from the range (0, 1) and serves as a quantize factor for the relative importance of a given data point. This function is a decreasing function and represents the objective of importance of a data point over time.

A commonly used objective function is the exponential objective function. This function is defined as follows. The exponential objective function f(s) with parameter λ is defined as follows as a function of $f(s) = 2\text{-}\lambda.s(1)$. We note that the value of f(s) reduces by a factor of 2 every $1/\lambda$ time units. This corresponds to the half-life of the function f(s).

Dimension projection clustering process: By the concept, it notify that the discrete dimensional projection clusters do not depend on consumer limitation in determining the attribute value of each cluster. It strives to maximize the quantum distribution technique for the each selected attribute value and the number of selected attributes of each cluster at the same time. As confer previously when it indicates the superiority of a dimensional projected cluster it maximizes the accuracy by eliminating the redundancy. It is the inclusion process of multi clustering that allows us to implement a dynamic threshold adjustment scheme that maximizes the criteria simultaneously.

There are two thresholds in the dimensional projections as $|B_{min}|$ and Q_{min} that restricts the smallest amount of selected attributes for each cluster and the minimum quantum attribute values of them. An attribute value is selected by a cluster if and only if its quantum index with respect to the cluster is not less than Q_{min}. Under this MDPQD scheme, if an attribute value is not selected by either of two clusters, it will be selected by the new cluster formed by merging them. However, if an attribute value is selected by only one cluster, so that we can obtain a precise result to the query depending on the variance of the mixed set of values at the attribute. Two clusters are allowed to merge if and only if the resulting cluster has utmost $|B_{min}|$ selected attributes.

Initially, both thresholds are set at the highest possible values so that all allowed merges are very feasible to engage reports from the analogous real cluster. At some point, there will be no more talented merges with the current attribute values. This signals the algorithm to release the values and establish a new round of integration. The process repeats until no more merging is possible or a target number of clusters are reached with the attribute value. By vigorously adjusting the threshold values in response to the merging process, the number and relevance of the selected attributes are both exploit.

Algorithm for Multi cluster Dimensional Projection on Quantum Distribution (MDPQD): There below describes the steps to be performed:

h: Habitats of Human data based on Socio Democratic Cultures (HHSD) dataset dimensionality
$|B_{min}|$: Minimum number of selected attributes value per cluster
Q_{min}: Minimum number of Quantum index of a selected attribute kth dimensions of dataset
Begin
Step 1: Each report in a cluster
Step 2: For pace: = 0 to d-1 do
{
Step 3: $|B_{min}|$: = d-pace
Step 4: Rmin: = 1-pace = (d-1)
Step 5: Foreach cluster C
Step 6: SelectAttrisVal(C, Q_{min})
Step 7: BuildQDmodel ($|B_{min}|$, Q_{min})
Step 8: While Quantizeresult
{
Step 9: MC1 and MC2 are multiple clusters formed with objfunc
Step 10: Various attribute value which forms the new cluster Cn
Step 11: Cn: = MC1, MC2……MCn
Step 12: SelectAttrisVal (Cn, Q_{min})
Step 13: Update Quantize result
Step 14: If clusters remained = k
Step 15: Goto 16
}
}
Step 16: Output result
End

To form the multiple clusters, each cluster keeps an objective function that place the attribute value between the each clusters and the best achieve is propagated. After calculating the attribute value of all other clusters, the information of the best cluster will be extracted from the dataset using the Quantum Distribution Model. The entries involving the two clusters with the same attribute value will be removed and the value between the clusters will be inserted into the clusters. The process repeats until no more possible merges exist and a new clustering step will begin by achieving the multi clustering concept.

Experimental evaluation: In Habitats of Human data based on Social and Demographic culture (HHSD)

dataset, clusters can be formed in discrete dimensions. Only a discrete dimension of attributes is precise to each cluster and each cluster can have a different set of precise attribute value. An attribute is precise to a cluster if it helps identify the member reports of it. This means the values at the precise attributes are distributed around some specific values in the cluster while the reports of other clusters are less likely to have such values. Determining the multi clusters and their precise attribute value from a HHSD dataset is known as the discrete dimensional projected clusters.

For each cluster, a discrete dimensional projected clustering algorithm determines a set of attributes. It assumed to be more precise to the users. Discrete Dimensional Projected clustering is potentially useful in grouping the clusters and forms the multiple clusters using attribute value. In these datasets, the habitat levels of different human being is taken as samples and recorded. We can view the cultural level of different peoples as attributes of the samples or it is also taken as the samples as the attributes of different culture people.

Clustering can be performed on attribute value of each sample. A set of precise attribute value might co express simultaneously in only some samples. Alternatively, a set of precise attribute value samples may have only some of the culture habits are co expressed simultaneously. Identifying the precise attributes using the objective function may help to improve the multi clustering objective. The selected attributes may also suggest a smaller set of habitat for researchers to focus on, possibly reduces the efforts spent on expensive natural experiments. In this study, we develop a progression of experiments considered to estimate the correctness of the proposed algorithm in terms of:

- Comparison of accuracy
- Execution time
- Multi cluster formation efficiency

RESULTS AND DISCUSSION

In this research, we have seen how the clusters have been projected in discrete dimensional projected spaces. Figure 2 describes the performance of the proposed Multi cluster Dimensional Projection on Quantum Distribution (MDPQD). In the consequence, we compared MDPQD against Partitioned Distance-Based Projected Clustering algorithm (PCKA) and Enhanced Approach for Projecting Clustering (EAPC) in terms of accuracy.

Figure 2 describes the average data accuracy based on the number of clusters formed with respect to the HHSD dataset. The dimensionality of the cluster of the proposed Multi cluster Dimensional Projection on Quantum Distribution (MDPQD) is compared with an existing partitioned distance-based projected clustering algorithm (PCKA) and Enhanced Approach for Projecting Clustering (EAPC).

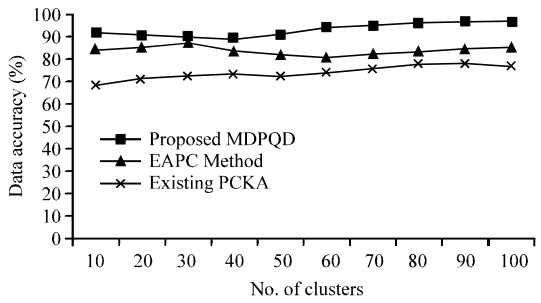

Fig. 2: No. of clusters vs. data accuracy

Figure 2 describes the average data accuracy based on the number of clusters partitioned with respect to the dataset. The set of experiments was used here to examine the impact of accuracy in the Multi cluster Dimensional Projection on Quantum Distribution algorithm. MDPQD is capable to accomplish vastly precise results and its performance is normally reliable. As we can see from Fig. 2, MDPQD is more scalable and accuracy in cluster formation than the existing EAPC and PCKA algorithm. If the average clusters accuracy is very low, only in the EAPC and PACKA by providing unsatisfactory results. Experiments showed that the proposed MDPQD algorithm efficiently identifies the clusters using the objective function and its dimensions precisely in a variety of situations.

MDPQD eradicates the choice of inappropriate dimensions in all the data sets used for experiments. This can be achieved by the fact that MDPQD initiates its process by detecting all the regions and their positions in every dimension, facilitating it to control the calculation of the discrete dimensions. Compared to an existing PCKA and EAPC, the proposed MDPQD achieved better accuracy and the variance is 30-40% high.

Figure 3 describes the presence of time taken to execute based on the discrete data dimensionality partitioned with respect to the HHSD dataset. The execution time of the cluster of the proposed Multi cluster Dimensional Projection on Quantum Distribution (MDPQD) is compared with an existing partitioned distance-based projected clustering algorithm (PCKA) and Enhanced Approach for Projecting Clustering (EAPC).

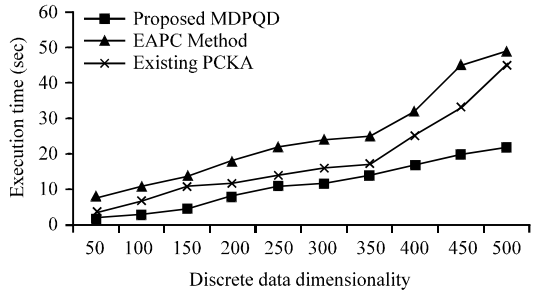

Fig. 3: Discrete data dimensionality vs. execution time

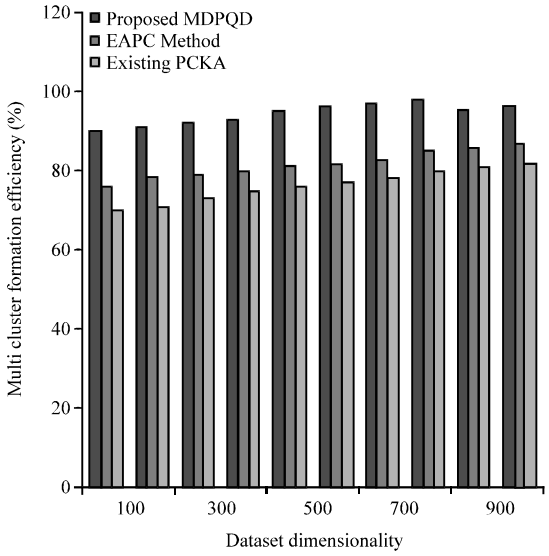

Fig. 4: Dataset dimensionality vs. multi cluster formation efficiency

Figure 3 describes the presence of time to execute based on the discrete data dimensionality with respect to the HHSD dataset. As observed from the Fig. 3, MDPQD exhibit reliable performance from the first set of experiments on data sets with lesser execution time taken. In tricky cases, MDPQD presents much improved results than the existing EAPC and PCKA Method. The results stated in Fig. 3 recommend that the proposed MDPQD is more interested to the proportion of data sets in execution time parameter.

Figure 4 describes the consumption of time to perform the dimensional projection of clustered based on the Quantum Distribution Model described in the dataset. The proposed MDPQD balances linearly with the increase in the data dimensionality. As specified in the scalability experiments with respect to the data set size, the execution time of MDPQD is generally provides improved results than that of EAPC and PCKA when the time required to project the clusters in discrete dimensionality employed for regular runs is also included.

The time consumption is measured in terms of seconds. Compared to the existing PCKA and EAPC, the proposed MDPQD consumes less time since it gives better cluster dimensionality result and the variance in time consumption is ~30-40% low in the proposed MDPQD.

Figure 4 describes the multi cluster formation efficiency with respect to the dataset dimensionality. The multi cluster formation efficiency for the proposed Multi cluster Dimensional Projection on Quantum Distribution (MDPQD) is compared with an existing Partitioned Distance-Based Projected Clustering algorithm (PCKA) and Enhanced Approach for Projecting Clustering (EAPC).

Figure 4 describes the multi cluster formation with the help of the objective function. This function forms the cluster with the precise attribute values. The different types of attributes with different set of values are used to achieve a multi cluster. Compared to an existing EAPC and PCKA, the proposed MDPQD provides the efficient multi cluster formation and the variance in is approximately 20-25% high in the proposed MDPQD.

CONCLUSION

In this research, we efficiently achieve the multi clustering concept in Habitats of Human data based on Socio Democratic Cultures (HHSD) dataset by professionally introducing the proposed Multi cluster Dimensional Projection on Quantum Distribution (MDPQD) Model. The proposed scheme describes the quantum distribution model by analyzing the data and rectify the redundancy occur on the attribute value in the dataset. We compared MDPQD with Partitioned Distance-Based Projected Clustering algorithm and Enhanced Approach for Projecting Clustering, in terms of accuracy and multi cluster formation efficiency. The experimental evaluations showed that dimensional projection clusters considerably outperforms objective function, especially in multi clustering. The experimental results showed that the proposed MDPQD scheme for the sensitive data attributes worked efficiently by improving 25-35% scalability and less execution time. We show that the use of objective function in the computations can significantly improve the quality of the underlying results.

REFERENCES

Bouguessa, M. and S. Wang, 2009. Mining projected clusters in high-dimensional spaces. IEEE Trans. Knowledge Data Eng., 21: 507-522.

Cai, D., C. Zhang and X. He, 2010. Unsupervised feature selection for multi-cluster data. Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 25-28, 2010, ACM, New York, USA., pp: 333-342.

Gajawada, S. and D. Toshniwal, 2012. Vinayaka: A semi-supervised projected clustering method using differential evolution. Int. J. Software Eng. Applic., 3: 77-85.

Hang, G.Y., D.M. Zhang, J.D. Ren and C.Z. Hu, 2009. A hierarchical clustering algorithm based on K-means with constraints. Proceedings of the 4th International Conference on Innovative Computing, Information and Control, December 7-9, 2009, Kaohsiung, pp: 1479-1482.

Jiang, J.Y., R.J. Liou and S.J. Lee, 2011. A fuzzy self-constructing feature clustering algorithm for text classification. IEEE Trans. Knowledge Data Eng., 23: 335-349.

Kriegel, H.P., P. Kroger and A. Zimek, 2009. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering and correlation clustering. ACM Trans. Knowledge Discovery Data, Vol. 3. 10.1145/1497577.1497578.

Nie, Y., R. Cocci, Z. Cao, Y.L. Diao and P. Shenoy, 2012. SPIRE: Efficient data inference and compression over RFID streams. IEEE Trans. Knowledge Data Eng., 24: 141-155.

Puri, C. and N. Kumar, 2011. Projected Gustafson-Kessel Clustering Algorithm and its Convergence. In: Transactions on Rough Sets XIV, Peters, J.F., A. Skowron, H. Sakai, M.K. Chakraborty, D. Slezak, A.E. Hassanien and W. Zhu (Eds.). Springer-Verlag, Berlin, Heidelberg, pp: 159-182.

Sembiring, R.W., J.M. Zain and A. Embong, 2010. Clustering high dimensional data using subspace and projected clustering algorithms. Int. J. Comput. Sci. Inform. Technol., 2: 162-170.

Shanmugapriya, B. and M. Punithavalli, 2012. A modified projected K-means clustering algorithm with effective distance measure. Int. J. Comput. Applic., 44: 32-36.

Yang, Y. and K. Chen, 2011. Temporal data clustering via weighted clustering ensemble with different representations. IEEE Trans. Knowledge Data Eng., 23: 307-320.