Journal of Theoretical and Applied Information Technology
20th December 2014. Vol.70 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

AN ALGORITHM TO CONSTRAINTS BASED MULTI-DIMENSIONAL DATA CLUSTERING AIDED WITH ASSOCIATIVE CLUSTERING

[1]B.KRANTHI KIRAN, [2]Dr. A VINAYA BABU

[1]Assistant Professor, Department of Computer Science and Engineering,
JNTUHCEJ, Karimnagar, Telangana, India

[2]Professor, Department of Computer Science and Engineering
JNTUniversity Hyderabad, Telangana, India

[1]kranthikiran9@gmail.com, [2]avb1222@jntuh.ac.in

ABSTRACT

To address the clustering problem related to multi-dimensional data clustering, a number of techniques have been implemented. A constraint based multi-dimensional data-clustering algorithm is proposed in this paper which helped with associative clustering can find out the number of clusters optimally present in a multi-dimensional data set. Now, by bays factor computation process associative constraint based clustering process is executed. Moreover, genetic algorithm is applied to optimization process to discover the optimal cluster results. The constraints based proposed algorithm assists in recognizing the right data to be clustered and the knowledge considering the data regarded as a constraint which enhances the precision of clustering. The data constraints furthermore assist in indicating the data related to the clustering task. The result of the proposed optimal associative clustering algorithm is compared with an existing algorithm on two multi dimensional datasets. Experimental result demonstrates that the proposed method is able to achieve a better clustering solution when compared with one existing algorithm.

Keywords: *Associative Clustering, Genetic Algorithm, Multi-dimensional Data, Bays Factor, Contingency Table*

1. INTRODUCTION

In finding out knowledge unseen in databases, Data mining develops as a promising solution. Data Mining has been properly termed as "the non-trivial extraction of implicit, formerly unidentified and potentially constructive information from data in databases" [1], [2]. Data mining has been exploited for multiple needs both in the private and public sectors. Accurate usage of data mining contain market segmentation, fraud detection, direct marketing, interactive marketing, market basket analysis, trend analysis and more [3, 4,5,7]. In several pervasive allocated computing environments, advances in computing and communication over wired and wireless networks have resulted. These environments frequently come with dissimilar distributed sources of data and computation. Mining in such environments obviously calls for correct utilization of these allocated resources. Most off-the-shelf data mining systems are planned to work as a monolithic centralized application on the other hand. They usually download the related data to a centralized location and next execute the data mining operations [1-7]. This centralized approach does not effort well in many of the emerging allocated, ubiquitous, probably privacy-sensitive data mining applications. In order to address this problem of mining data, Distributed Data Mining (DDM) proposes an alternate approach by distributed resources [6].

For above forty years, Clustering [16, 26] has been studied widely in data mining field and across several disciplines due to its broad applications. Clustering is the process of allocating data objects into a set of disjoint groups called clusters so that objects in each cluster are more related to each other than objects from dissimilar clusters. For competent clustering of data, the literature offers with a vast number of algorithms. These algorithms can be classified into nearest-neighbor clustering, fuzzy clustering, partitional clustering, hierarchical clustering, artificial neural networks for clustering, statistical clustering algorithms, density-based clustering algorithm and

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

so on. In these techniques, hierarchical and partitional clustering algorithms are two principal approaches of increasing interest in research communities. Hierarchical clustering algorithms can generally find satisfable clustering results. Even though the hierarchical clustering method is frequently portrayed as a better quality clustering approach, this method does not have any provision for the rearrangement of entities, which may have been badly classified at the early stage. In addition, most of the hierarchical algorithms are very computationally intensive and need much memory space [25].

Lately, big data clustering has been widely studied in many areas, together with statistics, machine learning, pattern recognition, and image processing [13-15]. In the regions, the scalability of clustering techniques and the methods for big data clustering much vigorous research has been dedicated. Different techniques has been introduced to overcome the problems happened in large database clustering, including initialization by clustering a model of the data and by means of an initial crude partitioning of the complete data set [7]. On the other hand, the most well-known representatives are partitioning clustering techniques such as CLARANS [11]; hierarchical clustering techniques such as BIRCH [10]; grid clustering techniques such as STING [8] and WAVECLUSTER [9]. Each technique has its benefits and shortcomings. For processing very large databases they are not appropriate. It is hard to obtain both high precision and competence in a clustering algorithm of large data. The two targets the entire time clash with each other. The power of a single computer is not sufficient in order to process massive data sets. Parallel and allocated clustering is the key method. In a distributed environment, it will be extremely scalable and low cost to do clustering.

In this paper, we propose an associative constraint based data clustering using multi-dimensional data. Bayes factor is employed to constraint based data clustering process in this paper. Additionally, genetic algorithm is used to obtain optimal clustering results. We evaluate the proposed algorithm on two real-world multi-dimensional data provided by UCI Machine Learning Repository. The remainder of this paper is organized as follows. Section 2 provides a review of related works. In section 3 explains basic concept of associative clustering. In section 4 focus on efficient implementation of proposed Associative Constraints Based Optimal Clustering algorithm. We devote section 5 to the experimental evaluation of our algorithm. Finally, we conclude in section 6.

2. REVIEW OF RELATED WORKS

Literature presents several techniques for distributed clustering. Here, we review some of the techniques presented for literature. Method on associative clustering for exploring dependencies among functional genomics datasets has been suggested by Samuel Kaski *et al.* [23]. High-throughput genomic measurements, construed as co-occurring data samples from multiple sources, expose a fresh problem for machine learning: What is in general in the dissimilar data sets, i.e., what kind of statistical dependencies are there among the paired samples from the dissimilar sets. They launch a clustering algorithm for investigating the dependencies. Samples inside each data set are grouped such that the dependencies among groups of dissimilar sets incarcerate as much of pair wise dependencies between the samples as feasible. In a new probabilistic way they have formalized this problem, as optimization of a Bayes factor. The technique is used to expose commonalities and exceptions in gene expression among organisms and to propose regulatory interactions in the form of dependencies between gene expression profiles and regulator binding patterns.

Using Associative Clustering Neural Network (ACNN), an approach to the study of gene expression data has been explained byYao yuhui *et al.* [24]. ACNN vigorously assesses similarity among any two gene samples through the interactions of a group of gene samples. It has possibility to more robust presentation than those similarities assessed by direct distances. The clustering presentation of ACNN has been checked on the Leukemias data set. In high dimensional data (7129 genes), the experimental results show that ACNN can attain superior presentation. The presentation can be further improved when some constructive feature selection methodologies are included. The study has illustrated ACNN can attain 98.61% precision on clustering the Leukemias data set with correlation study.

Based on the message passing model, Inderjit S *et al.* [12] offered a parallel execution of the k-means clustering algorithm. Their algorithm utilizes the intrinsic data-parallelism in the k-means algorithm. They analytically illustrated that the speedup and the scale up of their algorithm approach the optimal as the number of data points

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

raises. Wen-Yen Chen et al, [17] have explored representative techniques of approximating the dense similarity matrix because of the disadvantage of spectral clustering in large database. One of the methods was compared by sparsifying the matrix and the other by the Nyström method. Now, they planned a parallel execution and its scalability was assessed. They elevate the approach of sparsifying the matrix by holding nearest neighbors and exploring its parallelization. Now, both the memory applied and computation on distributed computers was parallelized by them. The experimental effect on different dataset illustrated that the planned algorithm can successfully handle big problems.

Eshref Januzaj *et al*. [18] have planned a scalable density-based distributed clustering algorithm. Now, a user-defined trade-off among clustering quality and the number of conveyed objects were permitted by the designed clustering algorithm. The procedure contained in the planned method were, according to a quality criterion reflecting their appropriateness to provide as local representatives they commanded all objects situated at a local site. The most excellent among these representatives was conveyed to a server site. It was next clustered with a slightly improved density-based clustering algorithm. Their experimental result illustrated that their planned algorithm outperformed in high quality clustering with scalable transmission cost. In distributed data clustering, Josenildo Costa da Silva and Matthias Klusch [19] have concentrated on the confidentiality problems, particularly the interference problem. Now, for distributed data clustering they planned an algorithm which was called as KDEC-S. It was to offer mining results when the confidentiality of original data was protected. The confidentiality level of KDEC-S method was declared only with the offered confidentiality framework. The fundamental plan of the planned method was not to rebuild the original data to the specified extent.

Ruoming Jin *et al*. [20] have planned a technique called Fast and Exact K-means Clustering (FEKM). Only one or a small number of passes on the complete dataset was necessary by the planned method and provably generated the similar cluster centers as reported by the original k-means algorithm. Now, the cluster centers were regulated by taking one or more passes over the complete datasets before this the planned algorithm created initial cluster centers by sampling. Moreover a theoretical study was offered by them to illustrate that the cluster centers were equal as the ones calculated by the original k-means algorithm. The experimental effect of real and synthetic datasets illustrated that the planned algorithm was executed better compared to K-means. In addition, here they explained and assessed a distributed version of FEKM which was called as DFEKM. It was most excellent for examining data that was allocated across loosely coupled machines. The DFEKM offered improved result than two other feasible options for correct clustering on distributed data, which were down-loading all data and running sequential k-means, or running parallel k-means on a loosely coupled configuration. The planned method outperformed parallel k-means if there was an important load imbalance.

Genlin Ji and Xiaohan Ling [21] have planned a distributed clustering technique based on ensemble learning to discover global clustering patterns. By the planned method, the distributed data sources were examined and mined. The two stages of the allocated clustering were, initially doing clustering in local sites and next in global site. The local clustering results were passed on to server site form an ensemble and joining plans of ensemble learning employed the ensemble to produce global clustering results. The produced global pattern from ensemble was mathematically changed to be a combinatorial optimization problem. A distributed clustering algorithm called DK-means was launched here as an execution for the model. The experimental effects demonstrated that the DK-means reached similar results to K-means which clusters centralized data set at a time. It was scalable to data distribution varied in local sites, and furthermore illustrated validity of the model.

Olivier Beaumont *et al*. [22] have planned the resource clustering problem in large scale distributed platforms, such as BOINC and WCG. Now, they planned to eliminate the single computing resource constraint by performing the task on a set of resources. Their objective was to plan a distributed method for a large set of resources which facilitates to build clusters. They described about a generic 2-phases method which was based on resource augmentation and whose approximation ratio was 1/3. In addition, a distributed version of the above technique was planned when the metric space was for a small value of D and the L∞ norm was applied to name distances. It obtains O rounds and messages both in expectation and with high possibility, where n was the total number of hosts.

3. ASSOCIATIVE CLUSTERING

Journal of Theoretical and Applied Information Technology
20th December 2014. Vol.70 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Associative clustering (AC) is a technique for separately clustering two` dataset when one-to-one associations among the sets, implying statistical dependency accessible. Assume that cluster two sets of data as said above with samples a and b, each discretely, such that (i) the clustering would capture as much as possible of the dependencies inside the data points (a, b), and (ii) the clusters would enclose (relatively) related data points.

Let us consider a phenomenon for example: If we memorize the pattern *umbrella*, the pattern *raining* will be associatively memorized. Once upon we think of *raining*, the pattern wet may be associated correspondingly. Thus the pattern *raining* helps to link the patterns *umbrella* and wet and make them associated. If all those three patterns are associated with each other, anyone of the three patterns can help to link the other two patterns and hence make-the association between them more and more explicit through time. Furthermore, other patterns that relate to {*raining, umbrella, wet*} would additionally interfere the associations of those three patterns. In the same time such related patterns would also interact the associations among them with each other. Finally all related patterns would form a mental lexicon. After forming the lexicon, any pattern in the lexicon can automatically recall the others like "*umbrella* → *raining* → *wet*"

4. ASSOCIATIVE CONSTRAINTS BASED OPTIMAL CLUSTERING ALGORITHM

In different domains, the current advancement in digital world generates very large data to do their related process. Within the small sets clustering played most important role to partition into a small sets to do related processes due to this unmanageable growth of data. However, once more, the additional challenge of cluster identification problem is how to deal with large data as most of algorithms are appropriate only for small data. The normal way of handling multi-dimensional data in clustering is to work out clustering problem with parallel algorithm. The significant hypothesis here is that the parallel algorithm can do better in terms of time consumption but the efficiency should be moreover satisfactory. It represents that the current cluster techniques should be applicable to do with multi-dimensional databases and presentation should reduce linearly with data size increase. In addition, when improving a large data clustering, the additional challenges like, dissimilar data format and data handling should be taken into account without much computational difficulty. Some of the existing clustering algorithms of large data either can handle both data kinds however are not competent when clustering large data sets or can handle large data sets competently but are restricted to numeric features. As a result lately, the parallel clustering offered important contribution in the large data clustering. Hence a scalable parallel clustering can assist in handling the above furnished problems. A number of methods have been executed in order to address the clustering problem associated to multi-dimensional data clustering. In this paper, to address the multi-dimensional data clustering, we propose an associative constraint based data clustering by multi-dimensional data.

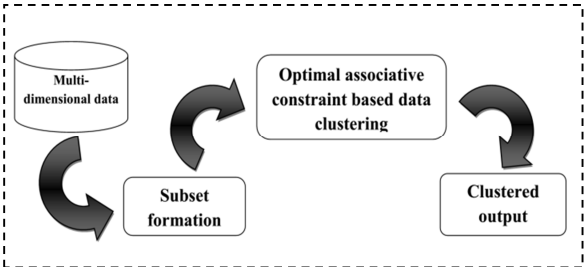

Figure 1: Illustration of proposed algorithm

Fig.1 shows our proposed algorithm for associative constraints based optimal clustering algorithm. The proposed approach, mainly concentrate on the constraints based multi-dimensional data clustering. The constraints helps in identifying the right data to be clustered and the knowledge regarding the data also considered as a constraint which improves the accuracy of clustering. The data constraints also help in specifying the data relevant to the clustering task. The dimensional or level constraints confine the dimension of data to be examined in a database. Here, we incorporate associative clustering method to the constraints based multi-dimensional data. The associative clustering method help to identify relationship between the two clusters based on the constraint values. The detailed multi-dimensional data clustering is explained following section.

4.1 Discretization

At first, the dataset $D = \{d_{ij}; 0 \le i \le m \, and \, 0 \le j \le n\}$ having n number of attributes is given to the discretization function to transmit the input data into discretized one. Discretization is an important step in the data processing to change the data into particular

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

interval means that the range of values is incarcerated into a particular interval. Now, we employ one discretization function based on the conventional way. The maximum and minimum value of every feature is recognized and the I interval is tracked by taking the ratio between the deviated value and the I value.

- For instance, initially, deviation is calculated as every k value

$$Dev(k) = \frac{Max(d_k) - Min(d_k)}{2} \quad (3)$$

- After calculate deviation for each row values, values are converted to the following condition:

$$\left.\begin{array}{ll} 0, & input < 1(Dev(k)) \\ 1, & input < 2(Dev(k)) \\ 2, & input < 3(Dev(k)) \\ 4, & input < 4(Dev(k)) \end{array}\right\} \quad (4)$$

Then, every value that comes under within the range is replaced with the interval value so that the input data is transformed to the discretized data D_d.

4.2 Subset Formation

In this section, we aim to partition data space D_d into subspaces to separate the dataset into equal subset. Using distance formula; the input dataset is partitioned into small and equal subsets in order to get equal subset. Now, Euclidean distance function is applied to find the nearest neighbor data point to partition the subset formation. If we consider data partitioning as a mapping P from an N-dimensional data space to j subspaces of dimensionality, n_j

$$p: \quad R^N \rightarrow \bigcup_j R^{n_j}, \qquad n_j \leq N$$

Given an N-dimensional dataset D_d, m[th] subset is formed as follows:

- Initially, a point is randomly selected from the high dimensional space and distance estimated between that data point and other data points. Next, first l-number of shortest distances is selected and that data points are chosen as m[th] subset.
- After that, novel data point is chosen (except m[th] subset data points) from the high dimensional space and distance computed between that novel data point and other data points. Next, first l-number of shortest distances is selected and that data points are chosen as novel subset. The process is replicated till j numbers of subspaces are produced.

4.3 Optimal Associative Constraint Based Clustering Using Genetic Algorithm

In this section, we propose an optimal associative clustering algorithm for generate better clustering accuracy and also to obtain optimal cluster or partition set. Now, genetic algorithm is applied to attain the optimal associative clustering solution from the high dimensional data. Genetic algorithm is an evolutionary computing method that can be employed to work out problems with a huge solution space [27].

4.3.1 Chromosome representation and population initialization

In genetic algorithm, a chromosome or solution representation is used to explain each chromosome in the population. Each chromosome is made up of a sequence of genes from a particular alphabet. An alphabet can contain binary digits, floating-point numbers, integers, symbols (i.e., A, B, C, D), etc. therefore; binary value representation is exploited to explain the chromosome in this document. In this depiction, two task of associative clustering process is programmed by the chromosome. Particularly, each chromosome is explained by a sequence of $M = 1 \times [n + k]$ binary digit numbers, where n is the dimension of the data space. k is the number of initial subsets.

In the proposed GS algorithm, an initial population of p contain binary digits can be arbitrarily generated. Fig. 2 demonstrates the solution encoding process. The length of the particular chromosome c is L_c, it contains two main functions (i) first n value signifies the dimension whether need or not and (ii) $[n + 1]$ to $[n + k]$ signifies that data points of subset have to go whether cluster 1 or cluster 2.

Journal of Theoretical and Applied Information Technology
20th December 2014. Vol.70 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

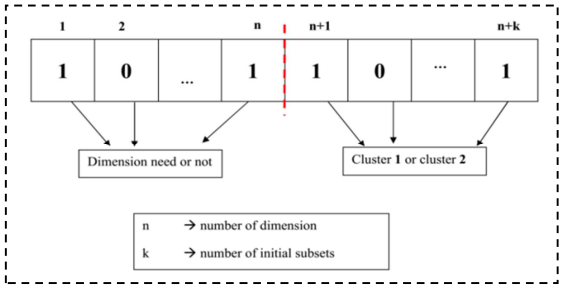

Figure 2: Solution encoding process of proposed GA algorithm

4.3.2 Fitness computation using Associative clustering

The fitness function is applied to describe a fitness value to each candidate solution. The fitness computation is executed for each chromosome. The fitness of a chromosome points out the degree of goodness of the solution it represents. This is compiled of three steps. Initially, novel data set is generated based on the initial solution. Secondly, subset is created to the chosen dataset and lastly, bayes factor is calculated to measure the dependency two between cluster sets.

In fig. 3, the fitness computation process is demonstrated. Following population initialization, the novel dataset is produced based on the binary '1' digit, which are present in the initial solution in fig. 2 and chosen attribute '1' column from original dataset D_d. After that, subset is formed to equal group generation by means of Euclidean distance. Next, a contingency table is produced for each combination of subset in the multi-dimensional data. After a bayes factor is introduced to constraint based clustering process and it has the benefit of properly taking into account the finiteness of the data while still being asymptotically related to mutual information.

Contingency Table: A contingency table is basically a display format employed to examine and record the relationship between two or more categorical variables. In addition, it is the categorical equivalent of the scatter plot applied to examine the relationship between two incessant variables. For instance, consider a $r \times s$ contingency table $(n_{i,j})(i = 1,2,...,r; j = 1,2,...,s)$ where, $(n_{i,j})$ indicates the frequency in cell (i, j).

Bayes factor computation: Bayes factor is applied to clustering process after generated contingency table for each combination of data space in subset formation. Commonly, by comparing a model of dependent margins to another model for independent margins [28], bayes factors have been applied as dependency measures for contingency tables. The clustering process is performed by following bayes factor form:

$$BF = \frac{\prod_{i,j} \Gamma\left(n_{i,j} + n^{(d)}\right)}{\prod_i \Gamma\left(n_{i\bullet} + n^{(x)}\right) \prod_j \Gamma\left(n_{\bullet i} + n^{(y)}\right)}$$

Where, $n_i = \sum_j n_{ij}$ and $n_j = \sum_i n_{ij}$ express the margins. The hyper-parameters $n^{(d)}$, $n^{(x)}$, and $n^{(y)}$, arise from Dirichlet priors. We have set all three hyper-parameters to unity, which makes the BF equivalent to the hyper-geometric probability classically used as a dependency measure of contingency tables. The fitness function of chromosome t is defined as maximized baye factor, i.e., the maximized bayes factor is computed for each chromosome. The objective of the genetic algorithm is to maximize this fitness function.

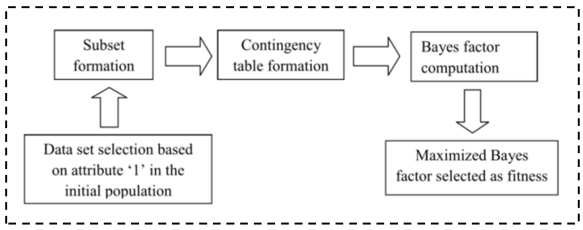

Figure 3: Fitness computation process

4.3.3 Selection operator

The aim of the selection operator is to remove the poor solutions. Thus, at the beginning of the each iteration ascending or descending order formation is applied to select the suitable solution for crossover process.

4.3.4 Crossover operator

The most important aim of the cross over operator is to form expanded and potentially promising novel chromosomes. It joins the features two parent chromosomes to form two offspring by swapping related segments of the parents. The perception behind the applicability of the crossover operator is data exchange between different potential solutions. In this piece of writing, we applied one point cross-over algorithm. The cross-over point is arbitrarily chosen and after that the

Journal of Theoretical and Applied Information Technology
20th December 2014. Vol.70 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

two segments of parents are replaced to form offsprings, as shown in fig. 4.

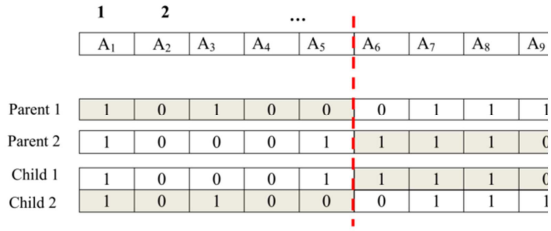

Figure 4: Sketch map of single point crossover

4.3.5 Mutation operator

Mutation provides genetic diversity and enables the genetic algorithm to search a wider space. It introduces random changes in structures in the population, and it may occasionally have beneficial results: escaping from a local optimum. Now, a mutation operator is executed as illustrated in fig. 5 and it arbitrarily chosen the data point based on the fixed mutation rate, in order to direct the search to get of a local optimum.

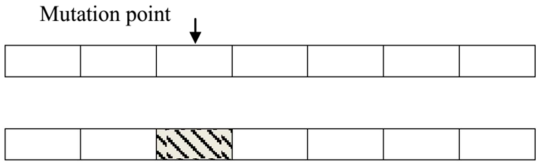

Figure 5: Mutation process

4.3.6 Termination

In this paper, we have executed the algorithm for a fixed number of iterations. The process terminates after some number of generations either by the user or dynamically by the program itself, where the best chromosome obtained will be received as the best solution.

4.4 Description of The Algorithm

In our GA based optimal associative clustering algorithm, a chromosome representing the associative clustering task or process by two set of functions (dimensional selection and optimal cluster generation) is used and each chromosome is individually evaluated by using the fitness function explained in section 4.3.2. In the evolutionary loop, a set of individuals is chosen for evolutionary cross over and mutation. The possibility of evolutionary operator is chosen adaptively. The crossover operator converts two individuals (parents) into two offspring by joining parts from each parent. Now, single point crossover is used to convert two individuals. The mutation operator works on a single individual and forms an offspring by mutating that individual. On the basis of the fitness function and form the novel generation the recently generated individuals are assessed. The chromosome with the best fitness value is selected in every generation. The process ends after some number of generations either by the user or vigorously by the program itself, where the best chromosome acquired will be taken as the best solution. The best string of the last generation provides the solution to our clustering problem.

5. RESULT AND DISCUSSION

The proposed clustering algorithm is executed in a windows machine containing configurations Intel (R) Core i5 processor, 3.20 GHz, 4 GB RAM, and the operation system platform is Microsoft Wnidow7 Professional. Also, we have employed mat lab latest version (7.12) for implementation.

5.1 Dataset Description

For the experimental results, two real-world datasets namely adult and census downloaded from the UCI Repository of Machine Learning Databases [29].

UCI Adult data: This is the annual income data consisting of 48842 instances (mix of continuous and discrete) or 45222 instances (if instances with unknown values are removed). Also, it contains 6 continuous, 8 nominal attributes and 1 class attribute. This is extracted from the census dataset.

UCI Census data: The census data has 2,458,284 records with 68 categorical attributes, about 352 Mbytes in total. It was derived from the USCensus1990raw data set which was obtained from the (U.S. Department of Commerce) Census Bureau website using the Data Extraction System.

5.2 Evaluation Criteria

To evaluate the effectiveness of clustering algorithms, we will introduce one evaluation indices, i.e., accuracy, which is defined by follows:

Accuracy: For example, if we obtain 3 clusters are obtained as shown in fig. 6 through clustering algorithm, the accuracy of this assignment is measured by counting the number of correctly assigned data points and dividing by N.

Journal of Theoretical and Applied Information Technology
20th December 2014. Vol.70 No.2

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

$$Clustering\ accuracy(\Omega, C) = \frac{1}{N}\sum_{k}\max_{j}\left|\omega_k \cap c_j\right|$$

Where, $\Omega = \{\omega_1,\ldots,\omega_k\}$ is the set of clusters

$C = \{c_1,\ldots,c_j\}$ is the set of classes.

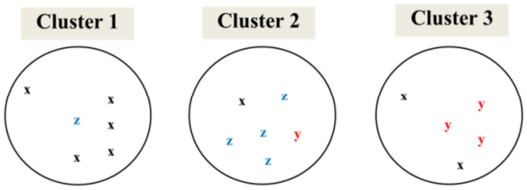

Figure 6: Accuracy computation process.

Majority class and number of members of the majority class for the threeclusters are, x, 5 (cluster 1); y, 4 (cluster 2); 3 (cluster 3). Clustering accuracy is $\frac{1}{17} * (5 + 4 + 3) = 0.71$.

5.3 Parameter Setting

The parameter settings and stopping conditions are listed in Table 1. In the experiments, the population size is taken as 10. The crossover and mutation rate is fixed as $p_c = 0.5$ and $p_m = 0.2$. The number of cluster is set as 2 in our paper.

Table 1:Parameter settings and termination conditions for proposed algorithm

Stopping iteration	≥10
Size of the initial population	10
Cross over rate p_c	0.5
Mutation rate p_m	0.2
Number of cluster	2
Length of the chromosome	$n+k$
Min\Max Gene limit	(0-1)

5.4 Comparative Analysis

The determination of number of clusters is important in clustering problem. Many methods have been proposed for identifying the number of clusters automatically in recent years. But associative data clustering algorithm has been developed only few. In this paper, we propose an associative constraint based data clustering algorithm discussed in section 4. Now, we have fixed the cluster set as 2 and these clusters are grouped optimally by using genetic algorithm. In this paper, we compare the performance of the proposed method with one existing method [23]. The performance of the proposed and existing method is evaluated through accuracy and time.

The accuracy and time performance of USI adult data is illustrated in fig. 7. Fig. 7a shows the accuracy rates and time obtained varying the initial subsets applying to the associative clustering process. Analyzing the fig. 7a, the proposed method achieved the better performance in subset 20 having the accuracy of 77.1%, where existing algorithm [23] achieved only 71%. Fig. 7b shows the time performance for different iterations and it says that our proposed algorithm takes minimum time when compares with existing [23] for clustering process.

The accuracy and time performance of USI census data is illustrated in fig. 8. Fig. 8a shows the accuracy rates and time obtained varying the initial subsets applying to the associative clustering process. From 8a, it seen that the proposed method is outperformed having the accuracy of 76.4%, which is high compared with existing algorithm only achieved 71.33% in subset 20. Fig. 8b illustrates the time performance for different iteration in proposed and existing clustering algorithm. It seen that using UCI census data, the proposed algorithm takes minimum time when compares with existing [23] for clustering process.

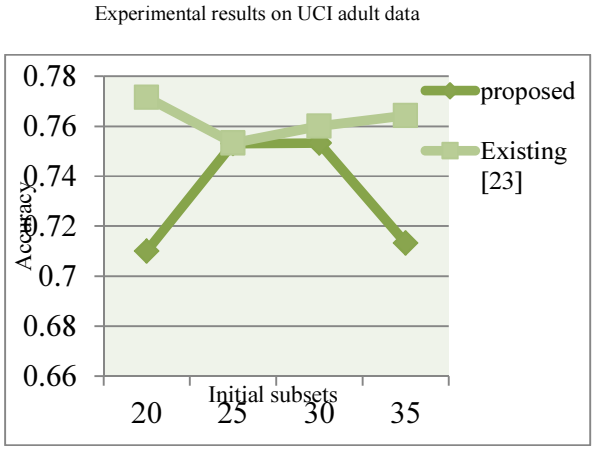

Figure 7a: Accuracy performance of adult dataset: Initial subset vs. accuracy

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

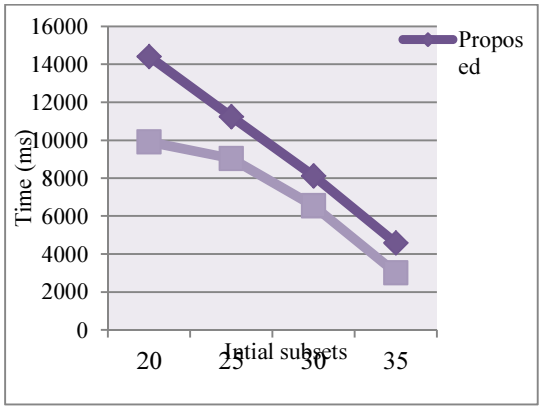

Figure 7b: Time performance of adult dataset: Time vs. initial subsets

Experimental results on UCI census data

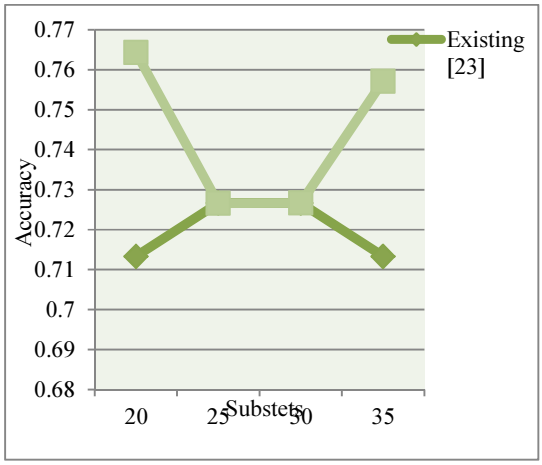

Figure 8a: Accuracy performance of census dataset: Accuracy vs. initial subsets

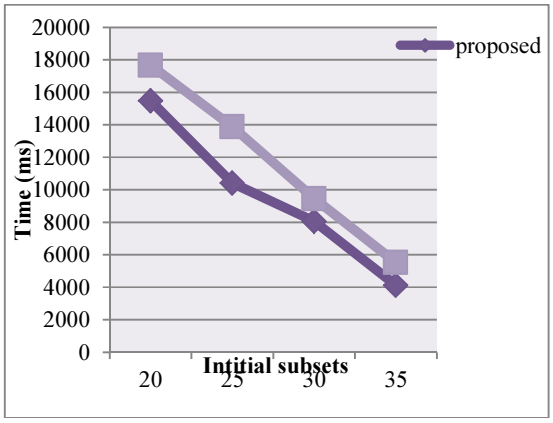

Figure 8b: Time performance of census dataset: Time vs. initial subsets

6. CONCLUSION

In this article, we propose an efficient approach to high-dimensional clustering using genetic algorithm. Then, by bays factor computation process associative constraint based clustering process was executed. Also, genetic algorithm is applied to optimization process to discover the optimal cluster results. The constraints based proposed algorithm assists in recognizing the right data to be clustered and the knowledge considering the data regarded as a constraint which enhances the precision of clustering. The data constraints furthermore assist in indicating the data related to the clustering task. Our experimental evaluation demonstrated that the proposed algorithm compares favorably to one existing algorithm on two multi dimensional dataset. Experimental results showed that the performance of this clustering algorithm is high, effective, and flexible.

REFERENCES:

[1] Osmar R. Z., "Introduction to Data Mining", In: Principles of Knowledge Discovery in Databases. CMPUT690, University of Alberta, Canada, 1999.

[2] Kantardzic, Mehmed. "Data Mining: Concepts, Models, Methods, and Algorithms", John Wiley and Sons, 2003.

[3] E. Wainright Martin, Carol V. Brown, Daniel W. DeHayes, Jeffrey A. Hoffer and William C. Perkins, "Managing information technology", Pearson Prentice-Hall 2005.

[4] Andrew Kusiak and Matthew Smith, "Data mining in design of products and production systems", in proceedings of Annual Reviews in control, vol. 31, no. 1, pp. 147- 156, 2007.

[5] Mahesh Motwani, J.L. Rana and R.C Jain, "Use of Domain Knowledge for Fast Mining of Association Rules", in Proceedings of the International Multi-Conference of Engineers and Computer Scientists, 2009.

[6]. Souptik Datta Kanishka Bhaduri Chris Giannella Ran Wolff Hillol Kargupta "Distributed Data Mining in Peer-to-Peer Networks", Journal of internet computing, vol.10, no.4, pp.18-26. 2006.

[7] Ron Wehrens and Lutgarde M.C. Buydens, "Model-Based Clustering for Image Segmentation and Large Datasets via Sampling", Journal of Classification, Vol. 21, pp.231-253, 2004.

[8] W. Wang, J. Yang, R. Muntz, STING,"A Statistical Information Grid Approach to Spatial Data Mining", VLDB, 1997.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

[9] G. Sheikholeslami, S. Chatterjee, and A. Zhang, "WaveCluster: A multi-resolution clustering approach for very large spatial databases", VLDB, pp. 428-439, 1998.

[10] T. Zhang, R. Ramakrishnan, and M. Livny, "BIRCH: an efficient data clustering method for very large databases", In Proceedings of the ACM SIGMOD International Conference on Management of Data, pp.103-114, 1996.

[11] Ng R. T., Han J.: "Efficient and Effective Clustering Methods for Spatial Data Mining", Proceedings 20th International Conference on Very Large Data Bases, pp.144-155, 1994.

[12] Inderjit S. Dhillon and Dharmendra S. Modha, "A Data-Clustering Algorithm On Distributed Memory Multiprocessors", Proceedings of KDD Workshop High Performance Knowledge Discovery, pp. 245-260, 1999.

[13] M. Ester, H. Kriegel, J. Sander, and X. Xu, "A density-bsed algorithm for discovering clusters in large spatial databases with noise", In SIGKDD, pp. 226–231, 1996.

[14] R. T. Ng and J. Han, "Efficient and effective clustering methods for spatial data mining", In VLDB, 1994.

[15] T. Zhang, R. Ramakrishnan, and M. Livny. Birch, "An efficient data clustering method for very large databases", In SIGMOD, pp. 103–114, 1996.

[16] Jinchao Ji , Wei Pang, Chunguang Zhou, Xiao Han, Zhe Wang, "A fuzzy k-prototype clustering algorithm for mixed numeric and categorical data", journal of Knowledge-Based Systems, vol. 30, pp. 129-135, 2012.

[17] Chen L, Chen CL, Lu M., "A multiple-kernel fuzzy C-means algorithm for image segmentation", IEEE Transaction on System Man Cybernetics: Part B, vol. 41, no. 5, pp. 1263-74, 2011.

[17] Wen-Yen Chen, Yangqiu Song, Hongjie Bai, Chih-Jen Lin and Edward Y. Chang, "Parallel Spectral Clustering in Distributed Systems", IEEE Transaction on Pattern Analysis and Machine Intelligence, Vol.33, No.3, pp. 568 – 586, 2011.

[18] Eshref Januzaj, Hans-Peter Kriegel and Martin Pfeifle, "Scalable Density-Based Distributed Clustering", Proceedings of 8th European Conference on Principles and Practice of Knowledge Discovery in Databases (PKDD), pp. 231-244, 2004.

[19] Josenildo Costa da Silva and Matthias Klusch, "Inference in Distributed Data Clustering", Engineering Applications of Artificial Intelligence, Vol.19, No.4, pp.363-369, 2005.

[20] Ruoming Jin, Anjan Goswami and Gagan Agrawal, "Fast and Exact Out-of-Core and Distributed K-Means Clustering", Journal of Knowledge and Information System, Vol. 10, No.1, pp. 17-40, 2006.

[21] Genlin Ji and Xiaohan Ling, "Ensemble Learning Based Distributed Clustering", Emerging Technology in Knowledge Discovery and Data Mining, Vol. 4819, pp 312-321, 2007.

[22] Olivier Beaumont, Nicolas Bonichon, Philippe Duchon, Lionel Eyraud-Dubois and Hubert Larcheveque, "A Distributed Algorithm for Resource Clustering in Large Scale Platforms", Principles of Distributed Systems, Vol.5401, pp.564-567, 2008.

[23] Samuel Kaski, Janne Nikkila¨ , Janne Sinkkonen, Leo Lahti, Juha E.A. Knuuttila, and Christophe Roos," Associative Clustering for Exploring Dependencies between Functional Genomics Data Sets", IEEE/ACM Transactions On Computational Biology And Bioinformatics, Vol. 2, No. 3,pp: 203-216, 2005.

[24] Yao Yuhui, Chen Lihui, Andrew Goh, Ankey Wong, " Clustering Gene Data Via Associative Clustering Neural Network", Proceedings of the 9th International Conference on Neural Information Processing, Vol.5, pp: 2228- 2232, 2002.

[25] Hesam Izakian, Ajith Abraham, Vaclav Snasel, "Fuzzy Clustering Using Hybrid Fuzzy c-means and Fuzzy Particle Swarm Optimization", World Congress on Nature and Biologically Inspired Computing (NaBIC 2009), India, IEEE Press, pp. 1690-1694, 2009.

[26] Swagatam Das, Ajith Abraham, Amit Konar, "Automatic Clustering Using an Improved Differential Evolution Algorithm", IEEE Transactions on Systems, Man, and Cybernetics—Part A: Systems And Humans, Vol. 38, No. 1, 2008.

[27] J.H. Holland, Adaptation in Natural and Artificial Systems, MIT Press, Cambridge, MA, 1992.

[28] I.J. Good., "On the Application of Symmetric Dirichlet Distributions and Their Mixtures to Contingency Tables," Annals of Statistics, vol. 4, pp. 1159-1189, 1976.

[29] UCI Repository of Machine Learning databases, University of California, Irvine, Department of Information and Computer Science, http://www.ics.uci.ed/~mlearn/MLRepository.html