A Novel Hybrid Clustering Algorithm: Integrated Partitional and Hierarchical Clustering Algorithm for Categorical Data

Rishi Syal,
Prof and Head, Department of Computer Science and Engineering,
Guru Nanak Engineering College, AP (India).
rishi_vps@yahoo.com

Dr G.V.S.R. Prasad
Prof , Department of Computer Science & Engineering, Gudlavalleru Engg. College, Gudlavalleru, A.P,India
gutta.prasad1@gmail.com

Dr V.Vijaya Kumar
Dean and Professor, Department of CSE, IT and MCA, GIET Rajahmundry, A.P., INDIA
vijayvakula@yahoo.com

Abstract **Data clustering became increasingly important in the field of computational statistics and data mining. Many algorithms have been developed in the literature for clustering where k-means clustering and hierarchical clustering are two well-known algorithms to partition the numerical data into groups. Due to the disadvantages of both categories of algorithms, recent researches have focused on hybrid clustering that combines the features of hierarchical and partitional clustering. The present paper developed a novel hybrid clustering algorithm called Integrated Partitional and Hierarchical clustering algorithm (IPHC) for categorical data. The proposed IPHC used modified k-mode clustering algorithm which is our previous proposed work for sub-clustering of categorical data and representative points (top-q points) are chosen from the sub-cluster. These representative points are then applied to agglomerative hierarchical clustering algorithm for constructing the hierarchical tree, called dendrogram. The proposed IPHC approach is validated with the aid of real categorical dataset available in the UCI machine learning repository. The experimental results demonstrate the effectiveness of the proposed IPHC algorithm.**

Keywords:- *Clustering, Partitional clustering, k-means clustering, Hierarchical clustering, Dendrogram, Categorical data.*

1. Introduction

Mining information and knowledge from the large volume of data has attracted in an increasing research attention because of its extensive applicability to enhance marketing strategies, business management, and user profile analysis and many more [1]. The term data mining, in other words, Knowledge Discovery in Databases (KDD) is commonly defined as the process of nontrivial extraction of implicit, previously unknown and potentially useful information from the data stored in a database [2]. In recent years, data mining [3], [4], [5], [6] as a multidisciplinary joint effort from databases, machine learning, and statistics, is significant in converting huge volume of data into chunks. Descriptive mining and Predictive mining are the two sub-divisions of data mining tasks. The former typifies the general properties of the data whereas the later generates a model to make predictions on unnoticed future events. Clustering, Association Rule mining and Sequential mining are few of the descriptive mining techniques [6, 7]. The predictive mining techniques comprise tasks such as Classification, Regression and Deviation detection [6], [8]. Clustering is one among the predictive mining techniques which is popularly studied in many research communities, e.g., statistical pattern recognition, machine learning, information retrieval, and data mining [1]. Clustering can be recognized as the unsupervised classification of patterns (observations, data items, or feature vectors) into groups (clusters) [10]. Clustering approaches can be classified as partitioning [11-13], hierarchical [14] [15], density based [16] [17], fuzzy clustering [15], artificial neural clustering [18], statistical clustering, grid based, mixed and more [19]. In these approaches, partitional and hierarchical clustering algorithms are two key approaches in research communities [10]. The most extensively employed partitional algorithm is the iterative k-means

approach. The k-means algorithm begins with k centroids (initial values are randomly chosen or derived from a priori information). Then, each pattern in the data set is allocated to the closest cluster (closest centroid). To conclude, the centroids are computed again as per the associated patterns. This procedure is done again until convergence is obtained [21]. Although K-means [13] was first introduced over 50 years ago, it is still regarded as one of the most extensively utilized algorithms for clustering. It is widely popular due to the ease of implementation, simplicity, efficiency, and empirical success [9].

On the whole, hierarchical clustering is regarded as a method of cluster analysis which tries to set up a hierarchy of clusters. Hierarchical clustering analyzes all the database items one at a time and deals each of them as separate clusters. The method recursively joins clusters by updating the inter-cluster distances. A number of algorithms employ this technique and they differ from each other in the way how they produce the sets. Usually, a tree data structure (dendrogram) is utilized to signify the hierarchical clustering. Two strategies exist in hierarchical clustering: Agglomerative and Divisive Algorithms.[23] An agglomerative approach (Bottom-Up approach) commences with each pattern in a distinct (singleton) cluster, and successively combines clusters jointly until a stopping criterion is satisfied. A divisive method (Top-Down approach) starts with all patterns in a single cluster and executes splitting until a stopping condition is achieved [10]. Hierarchical models produce good versatility because they do not need a priori definition of the number of clusters to be found out [5]. In comparison to the hierarchical and partitioned clustering, majority of the hierarchical algorithms are very computationally complex and consume high memory space. Conversely, the majority partitional clustering algorithms function in linear time. With better effectiveness, the clustering quality of a partitional algorithm is not as better as that of hierarchical algorithms [23].

Hierarchical and partitional algorithms have their own advantages and disadvantages. The integration of the two clustering algorithms called hybrid clustering enables good computational and conceptual simplicity [22]. Overall, these algorithms initially partition the input data set into *m* sub-clusters. Then, these algorithms set up a hierarchical structure on the basis of these *m* sub-clusters. Majority of the clustering algorithms in the literature show better results for numerical data that are ordered values, namely the height of a person and the speed of a moving vehicle [6]. In this research, we developed a novel IPHC algorithm that combines the two popular methods, k-means clustering and hierarchical clustering for categorical data. Categorical data are unordered values, for an instance, the kind of a drink and the brand of a car. Initially, the categorical data is partitioned into blocks and modified k-means clustering algorithm is applied to each block for sub-clustering. The categorical objects corresponds to each sub-cluster is sorted in accordance with the dissimilarity measure and the sorted data is divided into partitions which are used for finding the representative points. Then, the representative points are used to construct the dendrogram by applying the hierarchical clustering algorithm. Based on the constructed dendrogram, the categorical objects in the partitions are given as a resulted cluster.

The rest of the paper is organized as follows: The review of related researches is given in section 2. The two popular clustering algorithms are presented in section 3. The proposed novel hybrid clustering algorithm is presented in section 4. The experimental results of the proposed hybrid clustering algorithm are given in section 5. Conclusion is summed up in section 6.

2. Review of Related Researches

A handful of researches are available in the literature for hybrid clustering of numerical data. It looks more challengeable for designing the hybrid clustering algorithm for categorical data. Here, some of the recent researches are presented for hybrid clustering of numerical data along with the clustering algorithms for categorical data.

Cheng-Ru Lin and Ming-Syan Chen [23] have presented a similarity measure, recognized as cohesion, to determine the inter-cluster distances. By employing the cohesion measure, they had developed a two-phase clustering algorithm, known as cohesion-based self-merging (CSM), which functions in time linear to the size of input data set. Joining the features of partitional and hierarchical clustering methods, algorithm CSM divided the input data set into a number of small sub-clusters in the first phase and continuously combined the sub-clusters on the basis of cohesion in a hierarchical manner in the second phase.. Ickjai Lee and Jianhua Yang [24] studied a topology on the basis of merge technique for merging partitioning and hierarchical clustering. The presented merge technique improved the efficiency of partitioning clustering while maintaining the efficiency of hierarchical clustering. It balanced the clustering's drawbacks and then amplified their positive features.

Chen *et al*. [25] introduced a hybrid approach that appeared diverse from the available method. Initially it executed hierarchical clustering to choose a location and number of clusters in the first round and run the K-means clustering in next round. They cluster around half of the data by hierarchical clustering and accomplish it by K-means for the remaining half in one single round. The approach enabled a mechanism to handle outliers. In comparison to the earlier hybrid clustering approach and K-means clustering in 2 diverse distance measure on Eisen's yeast microarray data, that method always

produces much better quality clusters. P.A. Vijaya *et al*. [26] presented efficient bottom-up hybrid hierarchical clustering (BHHC) techniques for the use of prototype selection for protein sequence classification. In the first stage, an incremental partitional clustering technique namely leader algorithm (ordered leader no update (OLNU) method) which needed only one database (db) scan is employed to determine a set of sub-cluster representatives. In the second stage, either a hierarchical agglomerative clustering (HAC) scheme or a partitional clustering Algorithm—'K-medians' was employed on these sub-cluster representatives to acquire a requisite number of clusters. Therefore, the hybrid scheme scalable and so would be appropriate for clustering large data sets. They also have a hierarchical structure comprising of clusters and sub-clusters and the representatives of which are employed for pattern classification. The proposed methods are noticed to be computationally efficient with comparatively good Clustering Accuracy.

Zengyou He *et al*. [28] analyzed clustering algorithms for categorical data on the basis of cross-fertilization between the two disjoint research fields. It is defined that the CDC (Categorical Data Clustering) problem is an optimization problem from the perspective of CE (Cluster Ensemble), and employed CE approach for clustering categorical data. The experimental results on real datasets demonstrated that CE based clustering method is competitive with available CDC algorithms with regard to clustering accuracy. Fuyuan Cao *et al*. [29] have described an initialization method for categorical data. It employed the initialization method to k-modes algorithm and fuzzy k-modes algorithm. Experimental results showed that the presented initialization method is better than random initialization method and can be employed to large data sets for its linear time complexity with regard to the number of data objects. Dae-Won Kim *et al*. [30] extended k-modes-type algorithms for clustering categorical data by denoting the clusters of categorical data with k-populations as an alternative of the hard-type centroids employed in the conventional algorithms. Use of a population-based centroid representation enables it to maintain the uncertainty inherent in data sets as far as possible before actual decisions is finalized. The k-populations algorithm is noticed to provide better clustering results through various experiments.

3. Clustering Algorithms

3.1 Partitional Clustering Algorithm

The k-means algorithm is a popular partitional algorithm for data clustering. K-means clustering is a method of cluster analysis which aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean. When we apply k-means algorithm to the categorical data, this has counteracted by two problems such as, (1) Formation of cluster center (2) Finding the similarity measure in between the cluster center and the categorical objects. By handling these two issues, we have used the k-mode algorithm [32] based on the k-means algorithm. The algorithmic procedure of the k-means algorithm for categorical data (k-mode algorithm) is discussed below.

K-means algorithm for categorical data (k-mode algorithm)

Let us consider a categorical data set $D = \{X_1, \ldots, X_n\}$ of categorical objects to be clustered, where each object $X_i = (x_{i,1}, \ldots, x_{i,m}), 1 \le i \le n$ is defined by m categorical attributes. Then, this problem can be mathematically formulated as follows:

Minimize

$$P(W,Q) = \sum_{l=1}^{k}\sum_{i=1}^{n} w_{i,l} d(X_i, Q_l)$$

Subject to, $\sum_{l=1}^{k} w_{i,l} = 1,\ 1 \le i \le n,$

$$w_{i,l} \in \{0,1\},\ 1 \le i \le n, 1 \le l \le k$$

Where $W = [w_{i,l}]_{n \times k}$ is a partition matrix, $Q = \{Q_1, Q_2, \ldots, Q_k\}$ is a set of representatives, and $d(X_i, Q_l)$ is the dissimilarity between object X_i and representative Q_l.

Basic steps of k-mode clustering algorithm:

1) Initialize k representatives, one for each cluster.
2) Compute the dissimilarity $d(X_i, Q_l),\ l = 1,2,\ldots,k$ of each k representative with categorical objects X_i in D.
3) Assign categorical object X_i to cluster C_l whose dissimilarity measure is less.
4) Update the k representatives based on definition 2.
5) Repeat Step 2 to step 4, until there is no movement of the objects between the clusters.

Definition 1: (Dissimilarity Measure)

The dissimilarity of categorical object X_i with the representative Q_l is computed based on the following equations.

$$d(X_i, Q_l) = \sum_{j=1}^{m} \delta(x_{i,j}, q_{l,j})$$

$$\delta(x_{i,j}, q_{l,j}) = \begin{cases} 0; & if\ x_{i,j} = q_{l,j} \\ 1; & otherwise \end{cases}$$

Definition 2: (Updating of k-representatives)

Initially, the categorical object X_i related with cluster C_l, $l = 1,2,\ldots,k$ are obtained and then, we compute the relative frequency $f_{x_{i,j}}$ of every category $x_{i,j}$ within the cluster C_l. The categories of categorical attributes within the cluster C_l are arranged in accordance with their relative frequency $f_{x_{i,j}}$. The category $x_{i,j}$ with high relative frequency of 'm' categorical attributes is chosen for the new representative. For example, gender is a categorical attribute having two categories (male and female) and hair color is also a categorical attribute having a number of categories (blonde, brown, brunette, red and more).

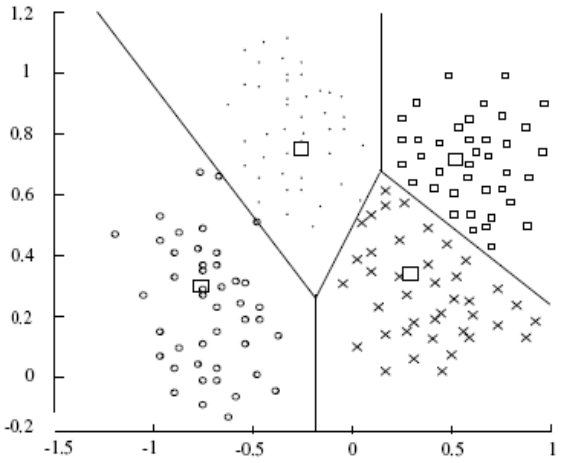

Figure 1: Sample result of K-means clustering algorithm

3.2 Hierarchical clustering

Hierarchical clustering creates a hierarchy of clusters which may be represented in a tree structure called as dendrogram. There are two types of methods (1) Agglomerative method (2) divisive method. Agglomerative hierarchical clustering is a bottom-up clustering method, where clusters have sub-clusters which in turn have sub-clusters, etc. Divisive hierarchical clustering is top-down clustering method which separates n objects successively into finer groupings. Differences between methods arise because of the different ways of defining distance (or similarity) between clusters like, (1) single-linkage (2) complete linkage.

One of the simplest agglomerative hierarchical clustering methods is single linkage, also known as the nearest neighbor technique. In single-linkage, the dissimilarity between two clusters is computed as the dissimilarity measure in between the two closest objects in the two clusters.

$$S_d(r,s) = Min\left\{ S_d(i,j) : where\ object\ i\ is\ in\ cluster\ r \right\}$$
$$and\ object\ \text{j}\ is\ in\ cluster\ s$$

In the complete linkage, also called as the farthest neighbor, clustering method is the opposite of single linkage. Dissimilarity between groups is defined as the dissimilarity between the most distant pair of objects, one from each group.

$$S_d(r,s) = Max\left\{ S_d(i,j) : where\ object\ i\ is\ in\ cluster\ r \right\}$$
$$and\ object\ \text{j}\ is\ in\ cluster\ s$$

Here the dissimilarity between every possible object pair (i, j) is computed, where object i is in cluster r and object j is in cluster s.

4. Proposed Algorithm : A Novel IPHC Algorithm for Efficent Clustering of Categorical Data

This section describes the proposed IPHC algorithm for efficient clustering of the categorical data. The proposed hybrid algorithm combines two popular and well known categories of clustering algorithms (Partitional and Hierarchical algorithm) and it provides better clustering results since the advantages of the hierarchical algorithms are the disadvantages of the partitional algorithms, and vice versa [31]. This research paper uses the early proposed modified k_Mode algorithm as in[] for the first step of partitional algorithm followed by implementation of hierarchical algorithm for efficient working of hybrid clustering algorithm. Several clustering algorithms [23-27] have been proposed to combine the features of these two types of clustering algorithms. In general, these algorithms first partition the input data set into sub-clusters using partitional clustering algorithm. Then, they construct a hierarchical tree called dendrogram based on these

sub-clusters. This hybrid technique of clustering is first proposed in [35]. Based on this hybrid technique, we develop a novel IPHC algorithm that uses the *top-* q points of sub-clusters as a representative to construct the dendrogram. The working procedure of the proposed IPHC algorithm is shown in Figure 2 and the steps involved in the proposed hybrid algorithm is given as,

(1) Partition the large categorical dataset D into p blocks.

(2) Perform sub-clustering on each block using modified *k*-means clustering algorithm.

(3) Select top q points of each sub-cluster, on the basis of the proposed hybrid clustering algorithm, from all p blocks to obtain the compressed dataset C.

(4) Perform clustering on the compressed dataset C using single linkage – agglomerative hierarchical clustering algorithm.

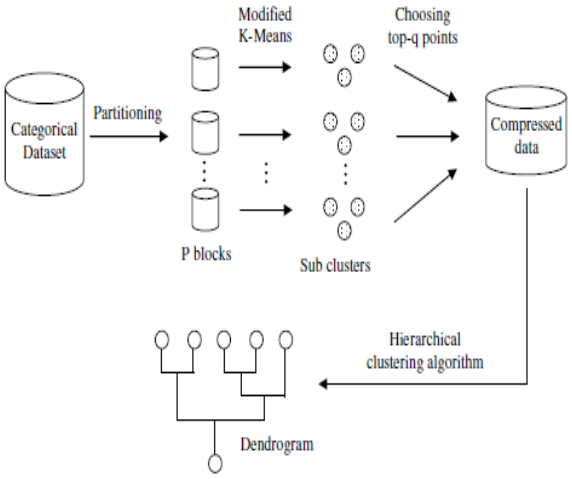

Figure 2: Block diagram of the proposed IPHC algorithm

4.1 Partitioning the Categorical Dataset into Blocks

Partitioning is the first step of the IPHC algorithm. It is the process of partitioning the categorical dataset D into p blocks with equal size ($D_1, D_2, \cdots, D_p$). This partitioning process can be done randomly or systematically. Clustering across extremely large categorical data set, typically many million categorical objects, is a complex and time consuming process. Partitioning the categorical dataset can lead to better performance through parallel operations. p blocks are generated here.

4.2 Sub Clustering Each Block Using Modified K-Mode Algorithm

This step provides k number of sub clusters in each block so that we can obtain sub-clusters of size $p * k$ [37]

4.3 Obtaining the Representative Points for Compressed Data

The proposed IPHC take the *top-* q representative points from every sub-cluster obtained from the modified k-means algorithm for compressed data C . This technique effectively reduces the computation time since it processes the *top-* q representative points only. The hybrid clustering algorithm presented by N. M. Murty *et al.* [35] made use of the centroid of sub-clusters for constructing a dendrogram and they achieved satisfactory clustering results. However, using only one point to signify the sub-cluster may easily lose some potential information about the distributions of sub- clusters. It looks more advantage if we used the *top-* q representative points for representing the sub cluster. The procedure used for choosing *top-* q representative points is that by sorting the categorical objects X_i in a cluster C_l based on their dissimilarity measure $d(X_i, Q_l)$ and divide it into q partitions systematically. Afterwards, the relative frequency is computed for the categories of categorical attributes within the q partitions. For every q partition, the high relative frequency categories are chosen from m categorical attributes. In this way, the *top-* q representative points of cluster C_l are obtained and form the compressed data of size $p * k * q$.

4.4 Cluster the Compressed Data Using Hierarchical Clustering Algorithm

This sub section details about the hierarchical clustering of compressed data C . The compressed data C discussed in section 4.3 is fed to the hierarchical clustering algorithm that forms a tree-like structure, called as dendrogram. Here, we have used the agglomerative hierarchical clustering based on the method, single-linkage.

Steps used for agglomerative hierarchical clustering algorithm:

(1) Consider each categorical object in the compressed data C as an individual cluster.

(2) Compute the dissimilarity matrix D_m that gives the dissimilarity between all possible pairs in the compressed data C.

(3) Find the minimum dissimilarity in the matrix D_m. This minimum dissimilarity pair is merged and it forms a new cluster.

(4) Compute the dissimilarity measure for the newly formed cluster with other clusters and insert it into the dissimilarity matrix D_m.

(5) Repeat the step 3 and step 4 until every point are merged into the single cluster.

4.5 Dissimilarity measure used in the proposed:

IPHC method:

We have used Goodall4 measure [33, 34] for dissimilarity calculation of categorical data in agglomerative hierarchical clustering. The dissimilarity of categorical objects X and Y is calculated as follows.

$$S(x,y) = 1 - \left(\frac{\sum_{i=1}^{m} S(x_i, y_i)}{m} \right)$$

$$S(x_i, y_i) = \begin{cases} P^2(x_i); if \ x_i = y_i \\ 0 \quad ; \text{otherwise} \end{cases}$$

$$P^2(x_i) = \frac{f(x_i)(f(x_i)-1)}{N(N-1)}$$

Where, x_i and y_i are the categorical attributes of the categorical object X and Y. $f(x_i)$ is the relative frequency of the categorical attribute x_i in compressed data C. N is the number of categorical objects in the compressed data C.

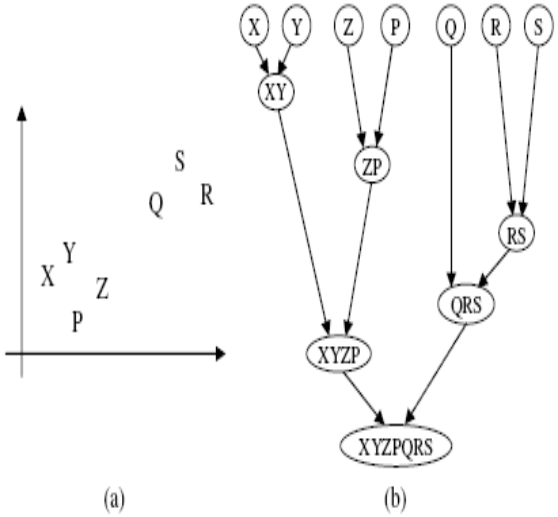

Figure 3: Sample result of agglomerative hierarchical clustering (a) Input data points (b) Dendrogram of the input data.

The resultant cluster is generated from the constructed dendrogram which is built only with the help of *top-q* representative points of every cluster. The representative points signify the partitions discussed in sub section 3.3. The partitions of the corresponding representative points are retrieved and it is given to the final cluster based on the dendrogram.

5. Experimental Results

The experimental results of the proposed IPHC

Table 1: Categorical attributes in nursery data

Categorical Attributes	Categories of Categorical Attributes
Parents	usual, pretentious, great_pret
Has_nurs	proper, less_proper, improper, critical, very_crit
Form	complete, completed, incomplete, foster
Children	1, 2, 3, more
Housing	convenient, less_conv, critical
Finance	convenient, inconv
Social	non-prob, slightly_prob, problematic
Health	recommended, priority, not_recom

algorithm are presented in this section. We have implemented the proposed IPHC system using Java (jdk 1.6). The proposed IPHC algorithm has been evaluated with the aid of Nursery Database available in the UCI machine learning repository. Nursery Database was derived from a hierarchical decision model originally developed to rank applications for nursery schools. The dataset consists of 12960 categorical objects with 8 categorical attributes. To apply the proposed IPHC algorithm on nursery database, we discarded the class label attributes. The categories of the categorical attribute are shown in table 1.

In the experiment, we partition the 12960 categorical objects into 4 blocks and each block we apply the modified k-means algorithm with k=5. We obtain 20 sub-cluster and find the top-q (q=10) representative points from every sub-cluster in such a way that the compressed data of 200 representative points are obtained. This compressed data is used for constructing the dendrogram from which, we obtain different clustering results. Initially, the proposed IPHC assigns 200 representative points presented in the compressed data as a separate cluster ($C_{0,}C_1,\cdots,C_{199}$). Subsequently, they are merged on the basis of dissimilarity measure. The merging pairs and its dissimilarity measure are shown in table 2 for the first 20 level. The dendrogram of the proposed IPHC algorithm for the first 20 level is shown in figure 4.

Table 2: Merging pairs and its dissimilarity measure

Merging pairs	Dissimilarity Measure
$C_{158}\rightarrow C_{189}$	0.986463567839196
$C_{159}\rightarrow C_{174}$	0.9867273869346733
$C_{187}\rightarrow C_{192}$	0.9867587939698492
$C_{153}\rightarrow C_{191}$	0.9868844221105527
$C_{197}\rightarrow (C_{174}, C_{159})$	0.9869095477386934
$C_{112}\rightarrow (C_{189}, C_{158})$	0.9870351758793969
$C_{169}\rightarrow C_{183}$	0.9871168341708543
$C_{110}\rightarrow C_{165}$	0.9871231155778895
$C_{179}\rightarrow (C_{189}, C_{158}, C_{112})$	0.9871231155778895
$(C_{165},C_{110})\rightarrow (C_{179}, C_{189}, C_{158}, C_{112})$	0.9871231155778895
$C_{199}\rightarrow (C_{187}, C_{192})$	0.9871482412060302
$C_{171}\rightarrow (C_{179}, C_{189}, C_{158}, C_{112}, C_{165},C_{110})$	0.9872236180904522
$C_{115}\rightarrow C_{117}$	0.9872298994974874
$C_{184}\rightarrow C_{186}$	0.9872298994974874
$C_{181}\rightarrow (C_{184}, C_{186})$	0.9872424623115578
$C_{156}\rightarrow C_{175}$	0.987286432160804
$C_{190}\rightarrow (C_{191}, C_{153})$	0.9872989949748744
$C_{152}\rightarrow C_{176}$	0.9873052763819096
$C_{163}\rightarrow C_{172}$	0.9873429648241207
$C_{198}\rightarrow (C_{192}, C_{187}, C_{199})$	0.9873743718592964

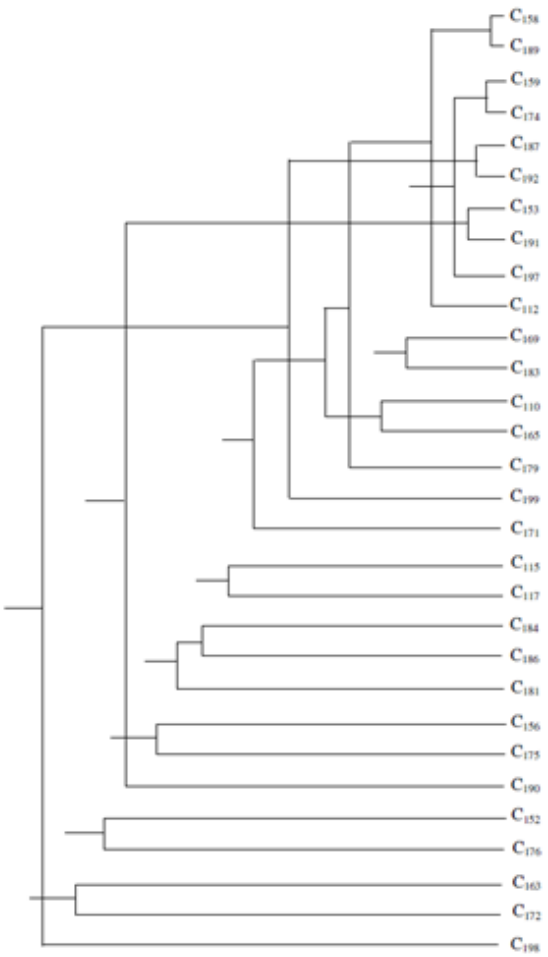

Figure 4: Dendrogram of the proposed hybrid clustering algorithm for the first 20 level merging

6. Conclusion

This paper uses our early proposed partitional algorithm along with hierarchical algorithm to develop a hybrid clustering algorithm that employs the two well-known methods namely, hierarchical and partitional clustering. The proposed IPHC algorithm partitions the categorical data set into several small sub-clusters in the first level and continuously merges the sub-clusters using hierarchical clustering in the second level. The

hierarchical clustering algorithm of IPHC used the representative points of sub-clusters to construct the dendogram. The proposed IPHC algorithm is validated with the real categorical dataset available in the UCI Machine Learning Repository and the results indicates that the proposed IPHC algorithm is efficient.

Acknowledgement

The authors would like to thank Dr.K.V.V. Satya Narayana Raju, MLC and chairman, K.Sasikiran Varma, Secretary GIET for providing advanced research lab facilities through Srinivasa Ramanujan Research Forum GIET.

The authors would like to thank Sardar Tavinder Singh Kohli Chairman, Sardar Gagandeep Singh Kohli Vice Chairman and Dr H. S. Saini Managing Director Guru Nanak Technical Institutions for their encouragement and support for this research. The authors would like to express their gratitude to the reviewers for their valuable suggestions and comments. The work is (partially) supported by research grants from the R&D of Guru Nanak Engineering College.

References

[1] Kun-Ta Chuang and Ming-Syan Chen, *"Clustering Categorical Data by Utilizing the Correlated- Force Ensemble"*, IEEE Transactions On Knowledge and Data Engineering, Vol. 17, No. 2, pp. 269-278, February 2005.

[2] Fayyad. U, *"Data Mining and Knowledge Discovery in Databases: Implications fro scientific databases",* In Proceedings of 9th International Conference on Scientific and Statistical Database Management, pp. 2-11, Olympia, Washington, USA, 1997.

[3] Daxin Jiang, Jian Pei and Aidong Zhang, *"DHC: A Density-based Hierarchical Clustering Method for Time"*, In proceedings of the 3rd IEEE Symposium on BioInformatics and BioEngineering pp. 393, 2003.

[4] Gokhan Silahtaroğlu, *"Clustering Categorical Data Using Hierarchies (CLUCDUH)"*, World Academy of Science, Engineering and Technology, Vol. 56, No.64, pp. 334-339, 2009.

[5] Pedro Pereira Rodrigues, Joao Gamaz and Joao Pedro Pedroso, *"ODAC: Hierarchical Clustering of Time Series Data Streams",* Vol. 20, No. 5, pp. 615-627, May 2008.

[6] Cheng-Ru Lin, Ken-Hao Liu and Ming-Syan Chen, *"Dual Clustering: Integrating Data Clustering over Optimization and Constraint Domains"*, IEEE transactions on knowledge and data engineering, Vol. 17, No. 5, may 2005.

[7] Charly K., *"Data Mining for the Enterprise",* In Proceedings of 31st Annual Hawaii International Conference on System Sciences, IEEE, Vol. 7, pp. 95-304, 1998.

[8] F. Coenen, Leng .P and Goulbourne. G, *"Tree Structures for Mining Association Rules",* Journal of Data Mining and Knowledge Discovery, Vol. 15, pp. 391-398, 2004.

[9] Anil K. Jain, *"Data clustering: 50 years beyond K-means"*, Lecture Notes in Computer Science, Springer, vol. 5211, pp. 3-4, 2008.

[10] A.K. Jain, M.N. Murty and P.J. Flynn, *"Data Clustering: A Review"*, ACM Computing Surveys, Vol. 31, No. 3, September 1999.

[11] R.C. Dubes, "How Many Clusters Are Best?—An Experiment", Pattern Recognition, Vol. 20, No. 6, pp. 645-663, 1987.

[12] C.-R. Lin and M.-S. Chen, *"On the Optimal Clustering of Sequential Data",* In Proceedings of Second International Conference on Data Mining, April 2002.

[13] J. McQueen, *"Some Methods for Classification and Analysis of Multivariate Observations",* In Proceedings of Fifth Berkeley Symposium on Math.Statistics and Probability, Vol. 1, pp. 281-297, 1967.

[14] P.H.A. Sneath and R.R. Sokal, *"Numerical Taxonomy. London: Freeman",* 1973.

[15] J.C. Bezdek, *"Pattern Recognition with Fuzzy Objective Function Algorithms",* New York: Plenum Press, 1981, ISBN:0306406713.

[16] M.M. Breunig, H.-P. Kriegel, P. Kro¨ger, and J. Sander, *"Data Bubbles: Quality Preserving Performance Boosting for Hierarchical Clustering",* In Proceedings of ACM SIGMOD, Vol. 30, No. 2, pp. 79-90, 2001.

[17] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, *"A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise",* In Proceedings Second International Conference on Knowledge Discovery and Data Mining, pp. 226-231, 1996.

[18] J. Hertz, A. Krogh, and R.G. Palmer, *"Introduction to the Theory of Neural Computation. Reading",* Mass Addison-Wesley, 1991.

[19] Cheng-Fa Tsai and Chia-Chen Yen, *"ANGEL: A New Effective and Efficient Hybrid Clustering Technique for Large Databases"*, Lecture Notes in Computer Science, Springer, pp. 817-824, 2007.

[20] Nizar Grira, Michel Crucianu, Nozha Boujemaa,*"Unsupervised and Semi-supervised Clustering:a Brief Survey"*, Springer, August 16, 2005.

[21] F. Samadzadegan and S. Saeedi, *"Clustering Of Lidar Data Using Particle Swarm Optimization Algorithm In Urban Area"*, 2009.

[22] Trevor Hastie, Robert Tibshirani and Jerome Friedman, *"The Elements of statistical learning"*, Springer, Second Edition, 2008.

[23] Cheng Ru,Lin and Ming-Syan Chen, *"Combining Partitional and Hierarchical Algorithms for Robust and Efficient Data Clustering with Cohesion Self-Merging"*, IEEE transactions on knowledge and data engineering, Vol. 17, No. 2, pp. 145-159, February 2005.

[24] Ickjai Lee and Jianhua Yang, *"Voronoi-based Topological Information for Combining Partitioning and Hierarchical Clustering"*, In Proceedings of the International Conference on Computational Intelligence for Modelling, cimca, Vol. 2, pp. 484-489 ,2005.

[25] Chen, B. Tai, P.C. Harrison and R. Yi Pan , *"Novel hybrid hierarchical-K-means clustering method (H-K-means) for microarray analysis"*, in Proceedings of the IEEE Conference on Computational Systems Bioinformatics, pp. 105-108, 21 November 2005.

[26] P.A. Vijaya, M. Narasimha Murty and D.K. Subramanian, *"Efficient bottom-up hybrid hierarchical clustering techniques for protein sequence classification"*, Pattern Recognition, Vol. 39, No. 12, pp. 2344-2355, January 2006.

[27] K.Thammi Reddy, M.Shashi and L.Pratap Reddy, *"Hybrid Clustering Approach for Concept Generation"*, International Journal of Computer Science and Network Security",Vol. 7, No. 4, pp. 62-69, 2005.

[28] Zengyou He, Xiaofei Xu and Shengchun Deng, *"A cluster ensemble method for clustering categorical data"*, Information Fusion, Vol. 6, No. 2 , pp 143-151, June 2005.

[29] Fuyuan Cao,Jiye Liang and Liang Bai, *"A new initialization method for categorical data clustering"*, Expert Systems with Applications, Vol. 36, No. 7, pp. 10223-102284, September 2009.

[30] Dae-Won Kim,KiYoung Lee,Doheon Lee, and Kwang H. Lee, *"A k-populations algorithm for clustering categorical data"*, Pattern recognition, Vol. 38, No. 7, pp.1131-1134, July 2005.

[31] Swagatam Das, Ajith Abraham and 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, January 2008.

[32] Zhexue Huang, *"Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values"*, Data Mining and Knowledge Discovery, vol. 2, pp. 283–304, 1998.

[33] Shyam Boriah, Varun Chandola and Vipin Kumar, *"Similarity Measures for Categorical Data: A Comparative Evaluation"*, In Proceedings of 2008 SIAM Data Mining Conference, Atlanta, GA, April 2008.

[34] D. W. Goodall, *"A New Similarity Index Based on Probability"*, in Biometrics, vol. 22, pp. 882-907. 1966.

[35] N. M. Murty and G. Krishna, *"A hybrid clustering procedure for concentric and chain-like clusters"*, International Journal of Computer and Information Sciences, vol. 10, no.6, pp.397-412, 1981.

[36] Moth'd Belal and Al-Daoud, *"A New Algorithm for Cluster Initialization,"* World Academy of Science, Engineering and Technology (WASET), Vol. 4, pp. 74 -76, 2005.

[37] Rishi syal, Dr V.Vijay kumar *" innovative modified k-mode clustering algorithm"* int'l journal of engg research and appln vol 2 july 2012

Prof. Rishi Sayal, Guru Nanak Engineering College Hyderabad, holds a BE (Computer Science) and M.Tech(IT) with 22 years of teaching, training and consultancy at various Engineering colleges, Software corporates and is currently pursuing Ph.D(CSE) from Mysore University in data Mining under the guidance of Dr. V. Vijaya Kumar. His main area of interests is databases, data mining, data warehouses ,computer networks and distributed systems. He is the life member of CSI and ISTE AND MEMBER OF IEEE.. He has published 6 papers in international conferences and 5 papers in international journals.

Dt G.V.S.N.R.V.Prasad obtained his Ph.D recently from JNTU Kakinada under the guidance of Dr V Vijaya Kumar in the area **of** data mining .He did his MS Software Systems, BITS Pilani and M.Tech in Computer Science and Technology in Andhra University .He has 15 years of teaching experience. Published 7 Research Papers in various National and International Conferences and 3 Research papers in National and International Journals. He is a member in various Professional Bodies . Presently working as Professor in CSE at Gudlavalleru Engineering College , Gudlavalleru ,A.P. His area of interest is Data Mining, Network Security and Image Processing.

Prof. Vijaya Kumar did his MS Engineering in Computer Science [USSR –TASHKENT STATE UNIVERSITY] and Ph.D in Computer Science . Worked as Associate Professor in Department of CSE and School of Information Technology (SIT) at Jawaharlal Nehru Technological university (JNTU) Hyderabad. Having a total of 13 years of experience. He Published 60 Research Papers in various National and International Conferences /Journals. Guiding 10 Research scholars . He is a Member for various National and Inter National Professional Bodies.Presently working as Dean for CSE & IT at GODAVARI INSTITUTE OF ENGINEERING AND TECHNOLOGY Rajamundry .