Journal of Computer Science 7 (12): 1875-1880, 2011
ISSN 1549-3636

A Cluster Feature-Based Incremental Clustering Approach to Mixed Data

A.M. Sowjanya and M. Shashi
Department of Computer Science and Systems Engineering,
College of Engineering Andhra University, Visakhapatnam, India

Abstract: Problem statement: The main objective of this study is to develop an incremental clustering algorithm that can handle numerical as well as categorical attributes in a given dataset. The authors have previously reported a cluster feature-based algorithm, CFICA that can handle only numerical data. **Appraoch:** Since many of the real life data mining applications work with datasets that contain both numeric and categorical attributes, there is a need for modifying the earlier algorithm to handle such mixed datasets. The core idea is to propose a new distance measure based on the weight age which is automatically generated and apply it to incremental clustering algorithms. The incremental data points are handled in two phases. In the first phase, k-means clustering algorithm is employed for initial clustering of the static databse.In the second phase, the designed distance measure is used to generate the appropriate cluster for the incremental data points. The combination of the two has proved to be more effective in handling mixed datasets. Clustering accuracy, clustering error and the computational time of the proposed approach have been evaluated with different k values and the thresholds. Variation of threshold values showed better results in terms of accuracy for different datasets. **Results:** The clustering error in this approach reduced considerably with different k values and thresholds. **Conclusion:** The results ensure the efficiency of the proposed approach in handling real mixed datasets composed of numerical and categorical attributes only.

Key words: Data mining, cluster feature, centroid, farthest neighbor points, mixed attributes, numerical attributes, categorical attributes, incremental clustering, k-means

INTRODUCTION

The process of categorizing a group of objects into various subsets such that objects of the same cluster are highly identical to each other is known as Clustering (Seokkyung and McLeod, 2005). Clustering is an important process for condensing and summarizing information because it is capable of providing a synopsis of the stored data. Image segmentation, information retrieval, web pages grouping, market segmentation and scientific and engineering analysis are some of the applications of clustering (Pham and Afify, 2007). Data clustering is both an essential data mining technique for knowledge discovery as well as a competent method for data analysis. Designing a new clustering algorithm is essential for handling the problem of continuous dumping raw data sets into available huge databases as this requires data clustering from scrape whenever there is an addition of new data instances into the database. This drawback can be avoided by designing a clustering algorithm that functions incrementally (Chien-Yu et al., 2002).

At present, huge size dynamically changing databases exist in almost every organization (Bo and McKay, 2006). It is desirable to perform updates incrementally by considering only the old clusters and the data inserted or deleted rather than employing the clustering algorithm to the (huge) updated database (Martin et al., 1998). An incremental clustering algorithm is capable of competently handling the ever-increasing existing databases

In this study, we have presented an approach, M-CFICA that efficiently extends the CFICA approach (Sowjanya and Shashi, 2010) to mixed data. First, initial clustering is performed on the static database using a partitional clustering technique along with a devised distance measure. In initial clustering, the distance measure is separately computed for numerical attributes as well as categorical attributes and it is combined with the automated weightage method. Then, cluster features that contain centroid and farthest neighbor points are computed for every cluster identified by the initial clustering process. Then, we identify the appropriate cluster for the incremental data points using an effective distance measure that is composed of the three distance values in between the

Corresponding Author: A.M. Sowjanya, Department of Computer Science and Systems Engineering,
College of Engineering Andhra University, Visakhapatnam, India

points namely, centroid, farthest neighbor and incoming data. With the help of computed distance values, we can find whether the incoming data point is to be added to existing cluster or to be formed as a new cluster. Subsequently, the cluster feature is updated with respect to the incremented cluster and the merging procedure can be done once we process a set of incremental data points. This procedure is iteratively done for each data points presented in the incremental database so as to obtain the resultant cluster.

Challenges: Incremental clustering of mixed data: Majority of the clustering methods are based on the use of a distance measure defined either on numerical attributes or on categorical attributes. However, in various applications, databases are composed of numerical as well as categorical attributes. Hence, the designing of an algorithm adaptable for mixed data is an indispensable one. Along with, the handling of mixed data for clustering is a challenging task in obtaining the better clustering accuracy. Existing solutions for clustering mixed numeric and categorical data fall into the subsequent classes (Chien-Yu et al., 2002): (1) Apply the distance measures utilized in numeric clustering for calculating the closeness between object pairs by encoding nominal attribute values as numeric integer values, (2) Apply categorical clustering algorithms along with discretizing the numeric attributes and (3) Generalize the criterion functions devised for one type of features to handle numeric and non-numeric feature values. However, when designing an algorithm appropriate for mixed data, a new distance measure should be proposed so that a new distance measure is designed based on the weightage which is automatically generated.

In addition to, by taking the distance measure into the incremental clustering problem is a core idea of the proposed approach. According to it, we redefine the distance measure and apply it to incremental clustering algorithms. In incremental database, mixed data points xΔ are added over time. These changes should be reflected in the resultant cluster C_R without extensively affecting the current clusters, C_i By considering, there are three possibilities that should be taken into account of incremental clustering of mixed data.

Case 1: Adding with the existing cluster (shown in Fig. 1.a).
Case 2: Making a new cluster (shown in Fig. 1.b).
Case 3: Possibility to merge the existing clusters when updated points are in between the existing two clusters (shown in Fig. 1.c). Cluster 'X' and 'Y' are the initial clusters Z is the new incoming data points

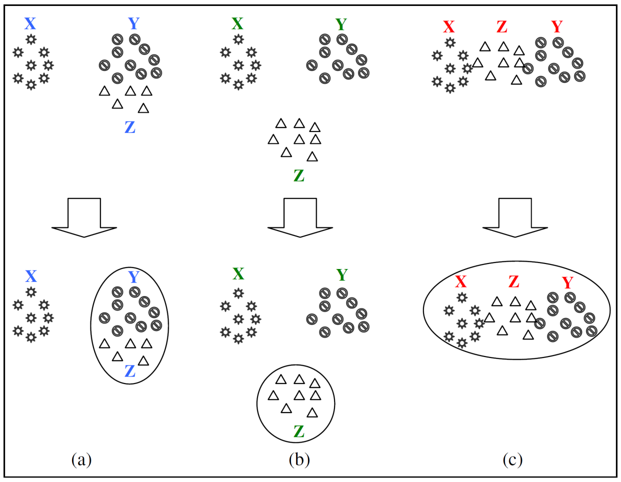

Fig. 1: (a) Adding to the existing cluster, (b) Formation of a new cluster, (c) Merging of closest cluster pair

M-cfica: There are plenty of clustering algorithms in the literature due to its wide applications. But, most of the clustering algorithms are used for clustering the static database containing only the numerical attributes. In recent times, data mining community turned their focus on incremental clustering of dynamically updated data points that contains numerical as well as categorical attributes. So, we have presented an incremental clustering approach to mixed datasets.

Here, we have designed a distance measure suitable for numerical and categorical attributes for handling of incremental data points in clustering. At first, we apply clustering algorithm that employs the technique of K-means clustering algorithm for initial grouping. Then, the devised approach capable of clustering mixed data is applied to the incremental data points. Let D and ΔD be the static and incremental database containing mixed data. The goal of the proposed approach is to find the clusters in the database D +ΔD, in which the data points from the incremental database ΔD are updated over time. The procedure employed for the proposed incremental clustering approach to mixed data is discussed in the following sub-sections.

Initial clustering with static database: At first, the clustering is performed on the static Database (D) where the data points do not change over time. Here, for clustering the static Database (D) we make use of the k-means clustering algorithm so that, k number of clusters are obtained. Let the static data base, Dcontaining of m data points with mixed attributes $z_{1,}$ $z_2,\ldots, z_m$ so that each data point is in R, the problem of finding the minimum variance clustering of the database into k clusters is none other than identifying k centroids $\{c_i\}$ (I = 1,2,…,k) in R whereas:

$$\frac{1}{m}\sum_{j=1}^{m}\left[\min_{i} d^{2}(z_{j},c_{i})\right]$$

Is minimized, where d (z_j, c_i) denotes the distance between z_j and c_i. The points {c_i} (i =1,2,…,k) are cluster centroids. We compute k cluster centroids, in which the mean squared error, MSE between a data point and its nearest cluster centroid is minimized.
Steps:

- Initialize k centroids, one for each cluster
- Compute the distance d (z_j,c_i) of each k centroid with data points z_j in D d (z_j,c_i) = (a*E_D (z_j,c_i)) + (β*dS (z_j,c_i))
- Assign data point z_j to clusterC_i whose distance is less
- Update the k centroids based on the memberships of new clusters. (Here, the centroid for the numerical attributes is obtained by taking the average of the relevant attribute value within a cluster. For categorical attribute, the relative frequency of every category within the cluster is computed and the category with high relative frequency of 'n' categorical attributes is chosen for the new centroid)
- Repeat Step 2-4, until there is no movement of the data points in between the clusters

Distance measure to find the nearest cluster of incremental data points: Developing an effective clustering algorithm needs a very effective distance measure to identify the most relevant cluster. So, we make use of the distance strategy that computes the distance values for both numerical and categorical data separately to find the suitable cluster for an incoming data point. Here, the distance measure employed depends on the centroid c, farthest neighbor point Q and the incoming data point x_Δ rather than the mean point. The numerical distance measure is computed for the following three set of points: mean and incoming data point ($c_{i,}$ x_Δ), farthest neighbor point and incoming data point ($Q_{i,}$ x_Δ), mean and farthest neighbor point ($c_{i,}$ Q_i), Susing Euclidean distance measure:

D1= E_D (ci, xΔ); d_2 = E_D ($Q_{i,}$ x_Δ); d_3 = E_D ($c_{i,}$ Q_i)

Where:
E_D ($c_{i,}$$x_\Delta$) = Euclidean distance between the points c_i and x_Δ
E_D ($Q_{i,}$ x_Δ) = Euclidean distance between the points Q_i and x_Δ
E_D ($c_{i,}$ Q_i) = Euclidean distance between the points c_i and Q_i

Then, the categorical dissimilarity is computed for the same set of data points:

dS_1 = dS (c_i, x_Δ); dS_2 = dS ($Q_{i,}$ x_Δ); dS_3 = dS ($c_{i,}$ Q_i)

Where:
dS (c_i, x_Δ) = Dissimilarity measure between the points c_i and x_Δ
dS ($Q_{i,}$ x_Δ) = Dissimilarity measure between the points Q_i and x_Δ
dS ($c_{i,}$ Q_i) = Dissimilarity measure between the points c_i and Q_i

Then, the weightage is automatically found based on the number of numerical as well as categorical attribute in the dataset:

$$\alpha = \frac{l}{l+n}\beta = \frac{n}{l+n}$$

Where:
L = Total number of numerical attributes in the dataset
n = Total number of categorical attributes in the dataset

Hence, the distance $D^{(i)}_{x_\Delta}$ is calculated using the following equation by making use of numerical distance, categorical dissimilarity and weightage. The cluster quality of the proposed approach is enhanced extensively if these three measures are included into the distance value for finding the suitable cluster:

$D^{(i)}_{x\Delta}$ = (a*d_1+ β* d_{S1}) + ((a*d_2+ β* dS_2)
× (a*$_{d3}$+ β* dS_3))

MATERIALS AND METHODS

Applying designed distance measure to handle incremental data points: The incremental database ΔD, consists of t data points $x_{\Delta 1,}$ $x_{\Delta 2\ldots}$ $x_{\Delta t;}$$1 \leq j \leq t$; Initially, we attain a set of cluster given as, C = {C_1,C_2…C_k}; $1\leq$ i $\leq$k from the initial clustering algorithm by inputting the static database D. Then, for each initial cluster computed by the initial clustering algorithm, we compute the cluster feature ($^{CF}{}_M$) of the mixed data. The cluster feature of the initial clusters for the mixed data contains centroid c_i and p-farthest neighbor points. The cluster feature $^{CF}{}_M$ is represented; $^{CF}{}_M{}^{(i)}$ = {c_i, $q_i{}^{(j)}$}; $I \leq j \leq p$, where C_i is the centroid of the cluster C_i and $q_i{}^{(j)} \in D$ gives the p-farthest neighbor points of cluster C_i. The p-farthest neighbor points of the cluster C_i are computed as follows: The distance in between every point presented in the cluster C_i with the centroid of the relevant cluster c_i is computed using the distance measure described as follows:

d(z_j, c_i) = (a* E_D (z_j, c_i)) + (β*dS (z_j, c_i)

Subsequently, we sort the data points in descending order based on the measured distance value. Then, from every cluster, we select the top p-farthest neighbor points that are known as p-farthest neighbor points of the cluster C_i ($q_i^{(j)}$; $i \leq j \leq p$) with respect to the centroid value c_i.

Case 1: Adding to the existing cluster: After computing the cluster feature of every cluster, we start to group the dynamically updating data points presented in the incremental database, ΔD. For the incoming data point x_Δ, we find the farthest neighbor point Q_i that is nearer to the incoming data point x_Δ using p-farthest neighbor points. To identify the farthest neighbor point Q_i, we find the distance in between the p-farthest neighbor points and the incoming data point using the distance measure given as follows, $d\ (q_i, x_\Delta) = (a{*}E_D\ (q_i, x_\Delta)) + (\beta{*}dS\ (q_i, x_\Delta))$.

The distance $d\ (q_i, x_\Delta)$ is computed for all p-farthest neighbor point with incoming data point $x\Delta$ and we sort the p-farthest neighbor points with the help of computed distance value and thereby, we choose one point, named as farthest neighbor point Q_i that has the minimum distance. Later, the devised distance measure $D^{(i)}_{x\Delta}$ is computed based on the centroid, farthest neighbor point and incoming data point. The same procedure is repeated for the every cluster C_i obtained from the initial clustering process. After that, the incoming data point $x\Delta$ is assigned to the cluster C_i only if (1) the calculated distance $D^{(i)}_{x\Delta}$ is less than the predefined threshold level, N_T the calculated distance should be minimum distance for the cluster C_i.

Case 2: Formation of a new cluster: The new cluster from the existing one is formed only if the above two conditions is not satisfied with the incoming data point so that, the number of cluster is incremented with one. Then, if the incoming data point x_Δ is added to the existing cluster or formed as a new cluster, there should be update the cluster feature CF_M for further processing the incoming data points. The updating of cluster feature $CF_M\ ^{(I)}$ of the incremented cluster $C^{(I)}$ is carried out by: (1) computing the centroid of the incremented cluster $C^{(I)}$ (2) identifying the p-farthest neighbor points of the incremented cluster $C^{(I)}$ Here, the centroid of the incremental cluster for the numerical attributes is computed by taking the average of it. For categorical attributes, the relevant frequency is found out and chooses the highest relevant frequency category as centroid value. Then, the distance measure $d\ (c_i, z_j)$ is computed for every data points in the incremented cluster $C^{(I)}$ with the updated centroid ^{cI}. From the sorted list of data points, we take a new set of top p-farthest neighbor points in such a way that the new cluster feature is computed for the incremented cluster $C^{(I)}$ given as, $CF_M\ ^{(I)} = \{c_I, q_I^{(j)}\}$ $i \leq j \leq p$. The same procedure is repeated for all the data points in the incremental database and those points are added to existing or forming as a new cluster.

Case 3: Merging of closest cluster pair: The merging of cluster pair is carried out only when ,$t_{n,}$ number of data points are processed with the above two cases. The motivation behind this process is that a non-optimal clustering structure may be obtained during the incremental clustering process. In order tackle this challenge, the merging process is utilized here to lead the incremental clustering process with good clustering accuracy. In addition to, it can provide a reasonable number of clusters by reducing the clustering error. The merging of two clusters can be done utilizing the cluster feature CF_M of the identified clusters after processing 't_n' number of data points. The steps used for merging strategy are:

- Compute the distance d (ci, cj) in between the centroid of the cluster with other cluster
- Merge the cluster pair whether the distance d (ci, cj) is less than the merging threshold level, M_T
- Re-compute the centroid of the merged cluster. (Here, we take the average for numerical attribute and for categorical attribute; we select the category which has highest relevant frequency)
- Repeat step (i) to (iii) until no cluster pair is merged

Finally, the resultant cluster C_R is obtained after processing all data points in the incremental database, ΔD

RESULTS AND DISCUSSION

The performance of the proposed approach is evaluated on Zoo dataset and the Flags dataset using Clustering Accuracy (CA). The evaluation metric used in the proposed approach is given below,

$$\text{Clustering Accuracy, } CA = \frac{1}{N}\sum_{i=1}^{T} X_i$$

$$\text{Clustering Error, } CE = 1 - CA$$

Where:

N = Number of data points in the dataset
T = Number of resultant cluster
X_i = Number of data points in both cluster
i = its corresponding class

Zoo dataset : consists 101 instances of animals with 17 features and output classes of 7. The first attribute is name of the animal.

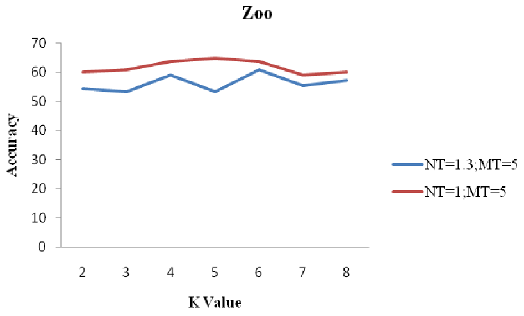

Fig. 2: Clustering accuracy with different k values in Zoo dataset

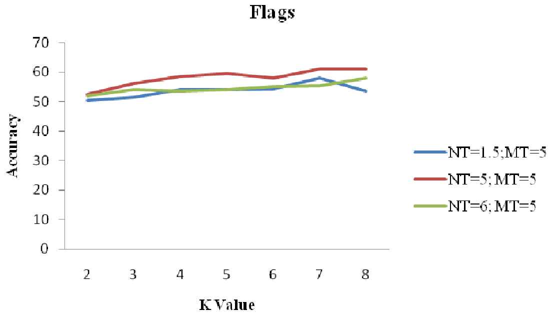

Fig. 3: Clustering accuracy with different k values in flags dataset

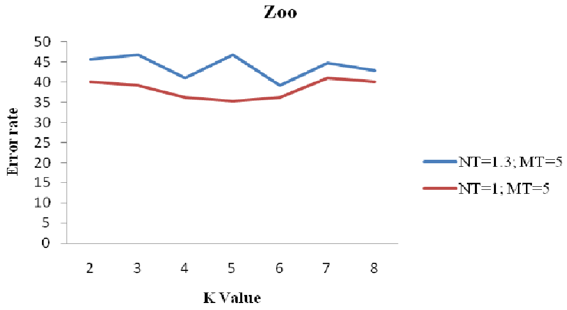

Fig. 4: Clustering error with different k values in Zoo dataset

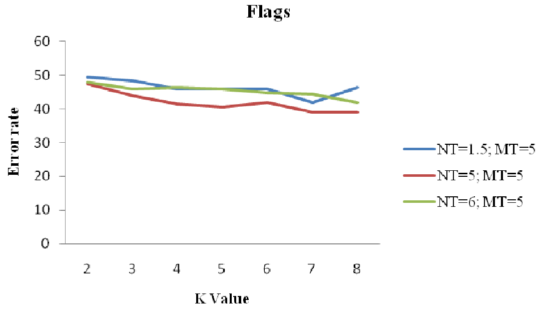

Fig. 5: Clustering Error with different k values in flags dataset

It possess 15 boolean characteristics representing the existence of hair, feathers, eggs, milk, backbone, fins, tail, whether airborne, aquatic, predator, toothed, breathes, venomous, domestic, cat size. The character attribute represents the amount of legs in the set {0, 2, 4, 5, 6, 8}.

Flags dataset: Possess the facts of 194 nations and their flags. There exist 30 attributes, like name, landmass, religion, color in the top-left corner of the flag, etc, among them 10 are numeric-valued and others are Boolean- or nominal-valued.

The evaluation results, clustering accuracy, clustering error and the computational time of the proposed approach are plotted as a graph and shown in Fig. 2-5. These results have been evaluated with different k values and the thresholds.

The accuracy of the proposed approach with different datasets produces better results by varying the threshold value and the merging threshold. While varying the k values for different thresholds, the accuracy of the proposed approach produces almost similar results.

The clustering error of the proposed approach gets comparatively reduced with different k values and the thresholds.

CONCLUSION

This study presents an efficient cluster feature-based approach to incremental clustering of mixed data. The proposed approach contains two phases - initial clustering and incremental clustering. K-means algorithm is used for initial clustering, in which the automatic weightage is assigned for numerical and categorical attributes in finding the distance value. Then, an effective distance measure comprising of three distance values is employed to find the relevant cluster of the incremental data points. Real mixed datasets like Zoo dataset and Flags dataset are used to analyze the performance of the proposed approach in terms of clustering accuracy and computation time. The results ensure the efficiency of the proposed approach.

REFERENCES

Bo, L., J. Pan, R.B. McKay, 2006. Incremental clustering based on swarm intelligence. Lectu. Notes, Comp. Sci., 4247: 189-196. DOI: 10.1007/11903697_25

Chien-Yu, C., S.C. Hwang and Y.J. Oyang, 2002. An incremental hierarchical data clustering algorithm based on gravity theory. Lec. Notes Comput. Sci., 2336: 237-250, DOI: 10.1007/3-540-47887-6_23

Martin, E., H. Kriegel, J. Sander, M. Wimmer and X. Xu, 1998. Incremental clustering for mining in a data warehousing environment. Proceedings of the 24th International Conference on Very Large Data Bases, New York, USA**,** pp: 323-333.

Pham, D.T. and A.A. Afify, 2007. Clustering techniques and their applications in engineering. J. Mechan. Engin. Sci., 221: 1445-1460. DOI: 10.1243/09544062JMES508

Seokkyung, C. and D. McLeod, 2005. Dynamic pattern mining: An Incremental data clustering approach. Lec. Notes Comput. Sci., 3360: 85-112, DOI: 10.1007/978-3-540-30567-5_4

Sowjanya, A.M. and M. Shashi, 2010. Cluster feature-based incremental clustering approach (CFICA) for numerical data. IJCSNS Int. J. Comp. Sci. Network Security, 10.