EFB Grid based Structure for Discovering Quality Clusters in Density based Clustering

V. Kumutha
Assistant Professor
Department of Computer Science
D.J. Academy for Managerial Excellence
Coimbatore, Tamil nadu, India

S. Palaniammal
Professor & Head
Department of Science & Humanities
V.L.B. Janakiammal College of Engg. & Tech.,
Coimbatore, Tamil nadu, India

ABSTRACT
Clustering is one of the important data mining techniques which discover clusters in many real-world data sets. Recent algorithms attempt to find clusters in subspaces of high dimensional data. Density based clustering algorithms uses grid structure for partitioning each dimensions into intervals (bins) which yields good computation and quality results on large databases. In this paper, we propose equal-frequency based (EFB) grid structure for efficient computation of clusters for high dimensional data sets. The computation is reduced by partitioning the bins with equal frequency bin method. The performance evaluation is done with data sets taken from UCI ML Repository. The result gives better quality clusters compared with other grid structures.

Keywords
Cluster, Dimensionality, dense unit, equal-frequency, high dimensional.

1. INTRODUCTION
Clustering is one of the valuable data mining method applicable to various fields [6][7][8]. Traditional clustering algorithm considers all of the dimensions from the entire dataset. Accurate clusters cannot be formed, if there is irrelevant dimension or noise. Tradition methods lacks in finding out accurate and quality clusters in the presence of noise.

Density-based approaches have shown to successfully mine clusters even in the presence of noise. The idea is to define clusters as dense areas separated by sparsely populated areas. Density of an object is measured either by mere counting of objects or by complex functions on the number and location of objects in the neighborhood [13]**.** An object is considered dense if its density is above some threshold.

Grid-based approaches can be used for finding the dense regions in the subspace. The great advantage of grid-based clustering is its significant reduction of the computational complexity and the result of quality clusters, especially for clustering very large datasets. In grid structure, the data space is first partitioned into a number of units. The units with more data points and the units with closely packed datasets are considered as dense regions.

Many approaches use equi-sized units and variable sized units for unit cell size (bins). In equi-sized units (bins), the width of the interval will be the same and the data points are spread along the units. The units with more points are considered as dense region and the dense regions having their densities greater than the threshold forms the clusters.

The objective of this paper is to use Equal Frequency Based (EFB) method for partitioning the dimensions into intervals (bins) where the width of the intervals may vary. Each unit has the same number of data points. The number of intervals is fixed by the user parameter. Also, density of an object in each unit is measured by the location of objects in the neighborhood within that same unit, instead of counting the number of data points in the unit.

Taking the count of data points in each unit may include an outlier in the cluster. So in our approach, the distance between the data points in each unit are calculated and the units having minimum mean distance values greater than the threshold are taken as dense regions to form a cluster. The threshold is defined to be a fraction of the total number of records present in the unit.

This paper is organized as follows. Section 2 gives the literature review on the recent work on Density based and grid based approaches. Section 3 deals with the proposed method. Section 4 gives the experimental evaluation and section 5 concludes with the future work.

2. LITERATURE REVIEW
DBSCAN [1] discovers arbitrary shaped clusters. For each object of a cluster, the neighborhood of a given radius has to contain atleast a minimum number of points. The definition of a cluster is based on the notion of density reachability. It can find clusters completely surrounded by a different cluster. Due to the minpts parameter single-link effect is reduced. For each point of a cluster the density of data points in the neighborhood has to exceed some threshold. It requires two parameters and is mostly insensitive to the ordering of the points in the database.

OPTICS [2] Ordering Points To Identify the Clustering Structure, is an algorithm for finding density-based clusters in spatial data. It requires two parameters: epsilon, which describes the maximum distance (radius) to consider, and minpts, describing the number of points required to form a cluster.

It also considers points that are part of a more densely packed cluster, so each point is assigned a core distance that basically describes the distance to its minptsth point.

Wang et al [3] proposed a Statistical Information grid-based clustering method to cluster spatial database STING. Each cell at a high level is partitioned into a number of smaller cells in the next lower level. Statistical information of each cell is calculated and for each cell in the current level computes the confidence interval.

In MAFIA [10], higher-dimensional Candidate Dense Units (CDU) are formed by first identifying the variable-sized units (bins) in each dimensional and then grouping them into units in higher subspaces. From these CDUs dense units are

extracted with the thresholds defined for each variable-sized bin in each dimension. Initially, the histogram is constructed and the contiguous bins with similar histogram values are combined to form larger bins. The bins (cells) having low density of data are pruned and the other units forms the cluster.

WaveCluster[11] is a density and grid based approach which applied wavelet transforms to the feature space. It is computational efficient but is applicable to only low dimensional data.

CLIQUE [12] is another density and grid based approach for high dimensional data set, which also detects clusters in the highest dimensional subspaces. Density based algorithms which uses the size of the grid and a global density threshold for clusters, as input parameters. The units are obtained by partitioning every dimension into equal width interval.

3. PROPOSED METHOD

MAFIA, WaveCluster and CLIQUE uses different gird structure, our proposed method uses Equal Frequency Based (EFB) method to form the grid structure.

3.1 Preliminary

- Let D be the entire data set with N instances and let A={A_1, A_2,….A_d) be the set of d dimensions of the data set, and S={A_1 X A_2 X…….X A_d} be the corresponding d-dimensional data space.
- To find out the clusters, equal number of data points are spread in equal frequency based grid structure. The rectangular units are derived by partitioning each attribute into equal-frequency intervals such that each unit has same number of points.

Let x=[3,6,10,2,12,7,12,10,9,5,15,2,7,18,2,10,13,9,6, 5,11,5] be the sample data set. Using Clique grid based structure, the number of units formed are 5. The number of data points in each units are 9,7,4,2,0. The number of clusters formed is 2. But, the data points may be an outlier as we have taken only the count of the data points in each unit. Using EFB structure, the number of units is 5. The dense value in each unit are 0.5,0.5,1.5,1. This algorithm also selects 2 clusters, but as we take the location of the data points in the cluster, we can reduce the noise of outlier.

3.2 Proposed Algorithm

Drawbacks of existing methods

- In data sets with large number of dimensions and varying data distribution, a uniform grid size leads to enormous amount of computation for fine grids and very poor cluster quality for coarse grids.
- Finding the density using number of points may lead to noise. An outlier may be grouped wrongly within a cluster if the count of the data set is taken. But, the location of data points close to the neighborhood data points reduces noise by the outlier detection.

In this paper, we use equal-frequency units (bins) to find out the highly dense region. Density is calculated by using Euclidean distance measure.

The proposed algorithm consists of 2 phases. In the first phase, the data sets are sorted in ascending order and the units are formed by taking equal number of data count by applying equal-frequency interval algorithm. Then the distance between the data points in each unit is calculated using Euclidean distance measure. In the second phase, dense units are identified. The nearest two units having minimum mean distance satisfying the threshold are taken and the units are merged. Merging stopped, if no units have their distance greater than threshold. The threshold is taken as 20% of the total number of data points.

3.3 Pseudo-code

Input : D, the entire dataset
Output: C, number of clusters
  Input N data points and the number of bins *k*
  Calculate count=N/*k* (data in each unit), with the remaining data in last unit
  For each unit U_i
    Find the distance between the data points using Euclidean distance measure
    Find the mean distance in U_i
  Take the minimum of mean distance from all units
  Merge the 2 neighborhood units until greater than a given threshold.

This approach discards noise as it finds the distance between the data points. It is applicable to large datasets.

Table 1. Datasets with number of Instances and features

Dataset	Number of Instances	Number of Features
Liver	345	7
Diabetes	768	9

4. EXPERIMENTAL EVALUATION

4.1 Dataset

2 real datasets from UCI ML repository [4] are used to evaluate the clustering results. Table 1 shows the details of the data sets.

4.2 Empirical Evaluation

In this section, we compare the algorithms based on the performance of the existing algorithms with the proposed approach. The proposed algorithm yields a better computation and the number of quality clusters formed is 2. Figure 1 & 2 shows the comparison of results. Figure 1 shows the number of clusters selected using EWB method and figure 2 shows the number of quality cluster selected using EFB methods. The program was implemented using MatLab 7.0.1. The comparison shows that the proposed approach is efficient in computation, noise deduction and quality clusters.

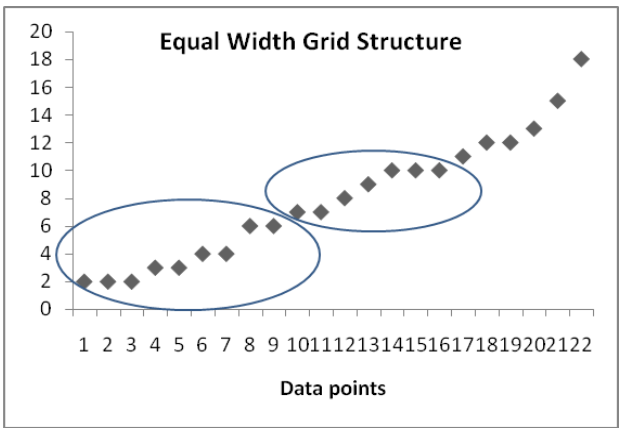

Fig 1: EWB structure with number of clusters

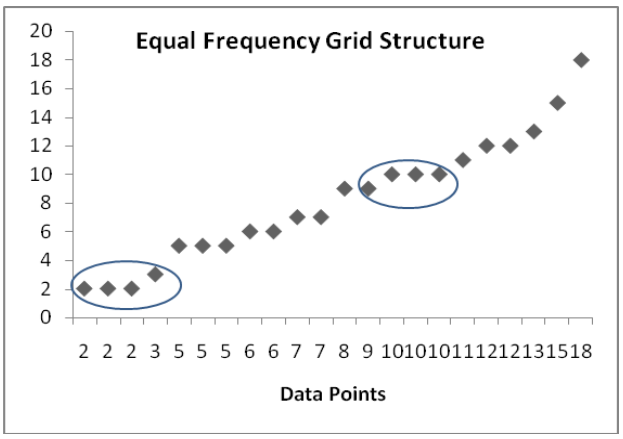

Fig 2: EFB structure with number of clusters

5. CONCLUSION

Clustering algorithms are attractive in various fields. High dimensional data is increasingly common which suffer from the curse of dimensionality, degrading the quality of the results. In this paper, we devised a novel approach to discover clusters with equal-frequency model. This discovers the regions which have relatively high densities compared to the average region density in the grid. As shown by the experimental results, this approach can discover the clusters with high quality and efficiency better than the previous works.

6. REFERENCES

[1] M.Ester, H.P. Kriegel, J.Sander, and X.Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. Data Mining and Knowledge Discovery, pages 226-231, 1996.

[2] M. Ankerst, M.M. Breunig, H.P. Kriegel, J. Sander,OPTICS: Ordering Points to Identify the Clustering Structure, ACM SIGMOD International Conference on Management of Data, ACM Press, pp. 49-60, 1999.

[3] Wang J. Yang, R. Muntz, STING : A Statistical Information Grid Approach to Spatial Data Mining, International Conference on Very large database, 1997.

[4] UCI repository of machine learning databases [http:/mlaern.ics.uci.edu/MLRepository.html]. 1998.

[5] C.C. Agarwal, A.Hinnerburg, and D.Keim, On the surprising behavior of Distance Metrics in High Dimensional space. ICDT, 2001.

[6] M.S.Chen, J.Han, P.S. Yu. Data mining: An Overview from Database Perspective, TKDE, 1996

[7] U. M. Fayyad, G. Piatetsky-Shapiro, P.Smyth, and R. Uthurasarmy, Advances in Knowledge Discovery and Data Mining, MIT Press, Cambridge,, MA, 1996.

[8] J.Han and M.Kamber, Data Mining: Concepts and Techniques. Morgan Kaufmann, 2000.

[9] A.K. Jain, M.N. Mutry, and P.J. Flynn, Data Clustering: A Review, ACM Computing Surveys, vol. 31, no. 3, pp. 264-323, 1999.

[10] S. Goil, H. Nagesh, and A. Choudhary, MAFIA: Efficient and Scalable Subspace Clustering for Very Large Data Sets, Technical Report CPDC-TR-9906-010, Northwestern Univ., 1999.

[11] G.Sheikholeslami, S.Chatterjee, A.Zhang, WaveCluster: A Multi-Resolution Clustering Approach for Very Large Spatial Databases, Proceedings of the International Conference on Very Large Data Bases (1998) Volume: M, Issue: 24, Publisher: Citeseer, Pages: 428-439.

[12] R.Agrawal, J. Gehrke, D.Gunopulos, P.Raghavan, Automatic Subspace Clustering of High Dimensional Data for Data Mining Applications, Data Mining and Knowledge Discovery, vol. 11, no. 1, pp. 5-33, 2005.

[13] Ira Assent, Ralph Krieger, Emmanuel Muller, Thomas Seidl, DUSC: Dimensionality Unbiased Subspace Clustering, Proceedings IEEE Conference on Data Mining (ICDM 2007)