Journal of Theoretical and Applied Information Technology
10th October 2013. Vol. 56 No.1

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

EFFICIENT CENTRIODS BASED CLUSTERING ALGORITHM WITH DATA INTELLIGENCE

D.JOHN ARAVINDER[1] AND DR.E.R.NAGANATHAN[2]
[1]Assistant Professor, Department of CSE, Hindustan University, Chennai, Emil :
[2]Professor and Head, Department of CSE, Hindustan University, Chennai,
Emil : jaravindhar@hindustanuniv.ac.in, ernindia@gmail.com

ABSTRACT
Cluster analysis is important technique to find the similar and dissimilar group in data mining. From two decade of data mining process, most of technique extracts irrelevant knowledge to domain. This is the main aim of this paper. This paper proposes a new centroid based clustering algorithm. And also this paper includes some additional intelligence or measures with clustering process. This measure supported to find the relationship between data objects and clusters apart from distances. This algorithm tests with some synthetic datasets. Experimental results shows domain related clusters and needs to test with real time datasets.
Keywords: *Data Mining, Clustering, K-Means Algorithm, Actionable Clusters*

1. INTRODUCTION

Data mining makes use of ideas, tools, and methods from other areas, especially computational area such as database technology and machine learning. It is not much concerned with all areas in which statisticians are interested. Mining essentially assumes that the data have already been collected, and is concerned with how to discover its secrets. It is not a one short activity, but rather an iterative and interactive process. There are clear overlaps between Statistics and Data mining. Data mining should be *the nontrivial process of identifying valid, novel, potentially useful, and ultimately comprehensible knowledge* from databases such knowledge should be useful in making crucial decisions [1].

They are more generally significant issues of existing and future KDD. They hinder the shift from data mining to knowledge discovery, in particular, blocking the shift from hidden pattern mining to actionable knowledge discovery. The wide acceptance and deployment of data mining in solving complex enterprise applications is thus further restrained. Moreover, they are closely related and to some extent create a cause-effect relation, which is the involvement of domain intelligence contributing to actionable knowledge delivery. This paper explores the challenges and issues from the following aspects:

- Organizational and social factors surrounding data mining applications;
- Human involvement and preferences in the data mining process;
- Domain knowledge and intelligence making data mining close to business needs;
- Actionable knowledge discovery supporting decision-making actions;
- Decision-support knowledge delivery facilitating corresponding decision-making,
- Consolidation of the relevant aspects for decision-support.

There are many issues related to clustering technique. The rest of the paper is organized as follows: Section 2 and 3 provides the domain driven data mining and ubiquitous intelligence. Section 4 explains clustering technique. The related works are given in detail in section 5. Section 6 describes the proposed technique and algorithms. Section 7 presents the experimental results and comparative study. Section 8 consists of conclusion and future enhancement.

2. DOMAIN DRIVEN DATA MINING

The basic idea of domain driven data mining (DDDM) [4] is as follows. On top of the data-centered framework, it aims to develop proper methodologies and techniques for integrating *domain knowledge, human role and interaction, organizational and social factors,* as well as *capabilities and deliverables* toward delivering actionable knowledge and supporting business decision-making action-taking in the KDD process. *DDDM* targets the discovery of actionable knowledge in the real business environment. Such research and development is very important for developing the next generation

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

data mining methodologies and infrastructures. Most importantly, *DDDM* highlights the crucial roles of ubiquitous intelligence, including in-depth data intelligence, domain intelligence and human intelligence, and their consolidation, by working together to tell hidden stories in businesses, exposing actionable and operationalizable knowledge to satisfy real user needs and business operation decision making. End users hold the right to say "good" or "bad" to the mined results.

3. UBIQUITOUS INTELLIGENCE

This section have stated the importance of involving and consolidating relevant ubiquitous intelligence surrounding data mining applications for actionable knowledge discovery and delivery. Ubiquitous intelligence surrounds a real-world data mining problem. *DDDM* identifies and categories ubiquitous intelligence into the following types.

- Data intelligence reveals interesting stories and/or indicators hidden in data about a business problem. The intelligence of data emerges in the form of interesting patterns and actionable knowledge. There are two levels of data intelligence:
  - *General level of data intelligence*: refers to the patterns identified from explicit data, presenting general knowledge about a business problem, and
  - *In-depth level of data intelligence*: refers to the patterns identified in more complex data, using more advanced techniques, disclosing much deeper information and knowledge about a problem.

Taking association rule mining as an example, a general level of data intelligence is frequent patterns identified in basket transactions, while *associative classifiers* reflect deeper levels of data intelligence.

- Human intelligence refers to (1) explicit or direct involvement of human knowledge or a human as a problem-solving constituent, etc., and (2) implicit or indirect involvement of human knowledge or a human as a system component.
- Domain intelligence refers to the intelligence that emerges from the involvement of domain factors and resources in pattern mining, which wrap not only a problem but its target data and environment. The intelligence of domain is embodied through the involvement into KDD process, modeling and systems.
- Network and web intelligence refers to the intelligence that emerges from both web and broad-based network information, facilities, services and processing surrounding a data mining problem and system.
- Organizational Intelligence refers to the intelligence that emerges from involving organization-oriented factors and resources into pattern mining. The organizational intelligence is embodied through its involvement in the KDD process, modeling and systems.

4. RELATED WORKS

The clusters are classified to five categories as follows: Well- Separated clusters – Each point is closer to all the points in its clusters than to any point in another cluster [5] [6]. Center based clusters – Each point is closed to the center of its cluster than to the center of any other clusters. Contiguity based cluster – Each point is closer to at least one point in its cluster than to any point in another cluster. Density based cluster – Cluster are regions of high density separated by regions of low density. Conceptual clusters – Points in a cluster share some general property that derives from the entire set of points.

4.1 Partitioning Clustering Algorithm

K-means clustering algorithm [1] divides the set of vertices of a graph into K clusters by first choosing randomly K seeds or candidate cetroids. It then assigns each vertex to the cluster whose centroid is the closest. K-means iteratively re-computes the position of the exact centroid based on the current members of each cluster, and reassigns vertices to the cluster with the closest centroid until a halting criterion is met (e.g. centroids no longer move). The number of clusters K is defined by user a priori and does not change.

4.2 Hierarchical Clustering Algorithms

Hierarchical algorithms [2] can be categorized into agglomerative and divisive ones. Agglomerative algorithms treat each vertex as a separate cluster, and iteratively merge clusters that have the greatest similarity from each other until all the clusters are grouped into one. The objective function of hierarchical clustering is intra-cluster similarity; i.e. greatest similarity at each merger.

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

Divisive algorithms start with all vertices in one cluster, and subdivide it into smaller clusters.

5. PROPOSED ALGORITHM

The K-means algorithm [3] is very commonly used for clustering data. To handle a large data set, a number of different parallel implementations of the K-means have also been developed. In partition clustering, a set D of N patterns $\{x_1, x_2, \ldots, x_N\}$ of dimension d is partitioned into K clusters denoted by $\{C_1, C_2, \ldots, C_K\}$ such that the sum of within cluster dispersions, i.e., the Squared Error (SSE) , as given in (1), becomes the minimum. Here, $M = \{M_1, M_2, \ldots, M_k\}$ is the set of cluster mean. Our proposed algorithm uses K-means to cluster data objects when clusters are of different size and different density. We use K-means as a guide to find the optimal solution to assign data objects to the correct cluster. Thus proposed algorithm is classified as partition clustering algorithm. The proposed algorithm uses the average mean distance between each cluster mean and each data object as a metric to take decision of merging and data intelligence. The average mean distance acts as a measure of density of objects; the following formula was used to determine the Average Mean Distance (AMD) Where is the cluster, is the cluster mean, is data object of X, n is the total number of data objects, is the total number of data objects that belong to cluster and is the Euclidian distance between the centroids.

In this proposed algorithm we try to reach the global optimal as possible as we can through multiple splitting using K-means and merging with respect to average mean distance. Initially proposed algorithm runs K-means with one additional centroid, and then we calculate the Average Mean Distance from each cluster mean. The two clusters with least AMD are merged into one cluster. It is important to exclude data objects that have shared in previous merging process, so they will not share in merging process next time. The proposed algorithm continues in this manner for K times, where K is the cluster number. We will see in the next section that the proposed algorithm does better than K-means in assignment data objects to clusters. The proposed algorithm makes the assignment more accurate and efficient.

Initially, the algorithm takes the number of cluster K as an input, after that we add one additional centriod, So the number of clusters NK= number of clusters (K) + 1, After running K-means with NK we gets NK clusters with a mean M_j for each cluster. Now we compute the Average Mean Distance from each data object to each cluster mean M_j according to equation 2. The least and closest two clusters distance are determined. Merging decision depends on the density of cluster, by comparing which of the two clusters has the largest Average Mean Distance, we will merge cluster of less density (lowest Average Mean Distance) with the other of higher density(largest Average Mean Distance). At this stage we have a new cluster; this new cluster will not share again in splitting and merging process. It is important to specify which data objects should not share in merging process this can be done by putting the merge flag to one for each merging data objects. The algorithm makes feedback of those data objects, which do not share in merging process, and again runs K-means with these objects to get the best clustering. We check the number of resulting clusters K after running the algorithm for K+1 times, it is possible to get one additional cluster. Additional cluster is due to the same Average Mean Distance of different clusters. In the case of additional cluster, simply get the average mean distance from each data object to each cluster mean M_j according to equation 2, the least and closest two cluster distance are determined. Now we compare which of them is more density (largest AMD) than other, merging cluster of less density (lowest AMD) with the other of higher density (largest AMD) so we get one cluster at the end. The algorithm repeats the previous steps for K iterations.

6. DATA INTELLIGENCE

Data Intelligence reveals interesting stories and/or indicators hidden in data about a business problem. The intelligence of data emerges in the form of interesting patterns and actionable knowledge. Even though mainstream data mining focuses on the substantial investigation of varying data for hidden interesting patterns or knowledge, the real-world data and its surroundings are usually much more complicated. Data intelligence (DI) show the relationship between two or more items when they occur together more often than expected, if they were statistically independent. A DI value greater than 1 indicates that the rule body and the rule head appear more often together than expected. This means that the occurrence of the rule body has a positive effect on the occurrence of the rule head. A DI smaller than 1

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

indicates that the rule body and the rule head appear less often together than expected. This means that the occurrence of the rule body has a negative effect on the occurrence of the rule head. A DI value near to 1 indicates that the rule body and the rule head appear almost as often together as expected. This means that the occurrence of the rule body has almost no effect on the occurrence of the rule head [7].

$$Data\,\text{int}\,elligence(xandy) = \frac{P(x \cup y)}{P(x)P(y)}$$

Input : Dataset (*D*), number of clusters (*K*), number of trails (*n*)
Output : *K* – Clusters

1. Initialize the list of clusters to contain the cluster consisting of all points from *D*.
2. Repeat
3. Remove a cluster from the list of clusters
4. For i=1 to *n* do
5. Split the selected cluster using basic *K*-means
6. End for
7. Select the two clusters from the bisection with the lowest total *SSE* and *data intelligence*.
8. Add these two clusters to the list of clusters
9. Until the list of clusters contains *K* clusters.

7. EXPERIMENTAL RESULTS

This work use two-dimensional dataset. Then, this paper uses *normrnd* function in Matlab to generate the random synthetic datasets. Two synthetics datasets are considered for these experimental results. The synthetics datasets are DS1, and
DS2.

S.NO	Datasets	Size	Dimension
1	DS1	400	2
2	DS2	220	2

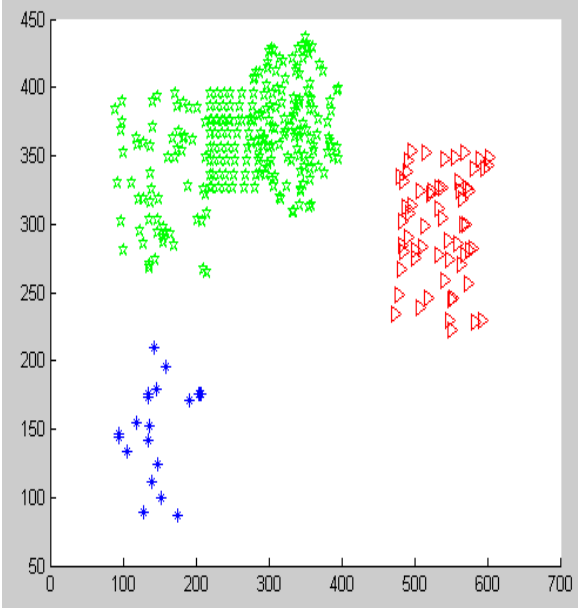

Figure 1: K Means Algorithms With DS1 Dataset

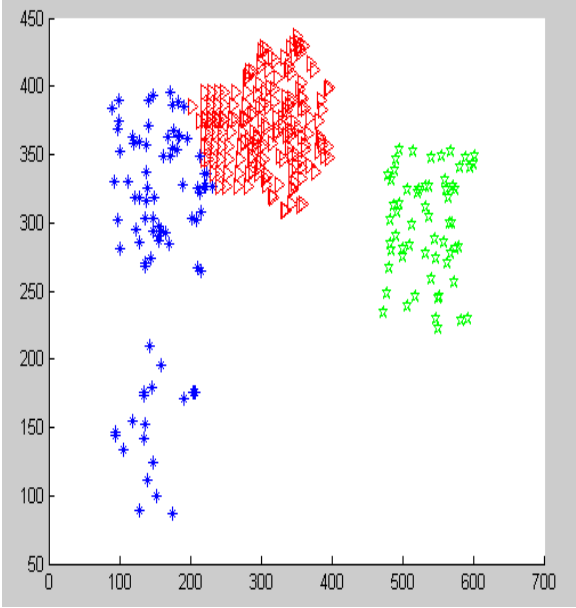

Figure 2: K Means Algorithms With DS2 Dataset

Figure 1 and 2 shows the mined clusters DS2 using k-means algorithm. There are three clusters which are green cluster, red cluster and blue cluster (denoted by points). Blue cluster and green cluster have different density, red cluster is denser than both cluster blue and cluster green.

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

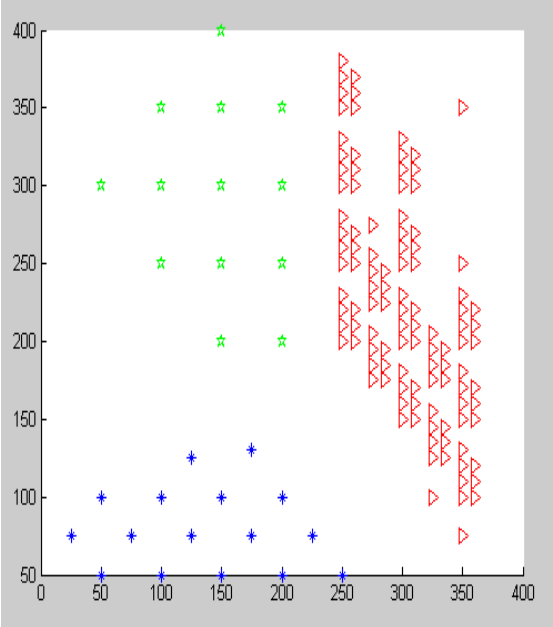

Figure 3: Proposed Algorithms With DS1 Dataset

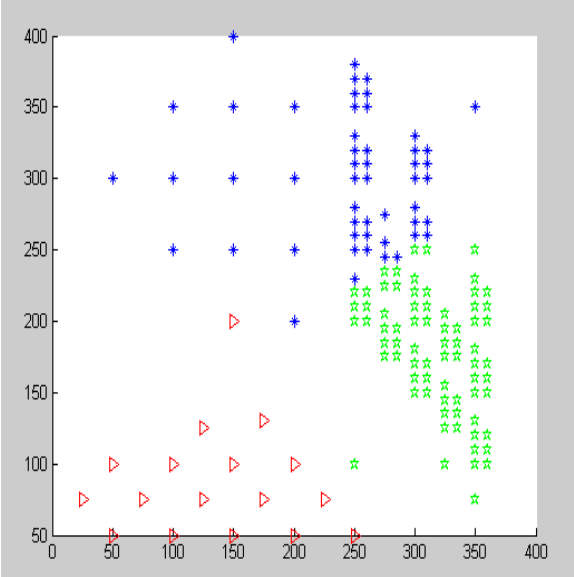

Figure 4: Proposed Algorithms With DS1 Dataset

Figure 3 and 4 shows the mined clusters DS2 using our proposed algorithm. There are three clusters which are green cluster, red cluster and blue cluster (denoted by points). Blue cluster and green cluster have different density, red cluster is denser than both cluster blue and cluster green.

8. CONCLUSION AND FUTURE ENHANCEMENT

This paper presented a novel algorithm for extracting actionable clusters using K-means based clustering algorithms. The clusters are of different size and different density. The proposed algorithm used one additional centroid, the distance measurement depends on the density of data objects from all clusters mean. Also this algorithm validated the item using data intelligence. These experimental results demonstrated that our scheme could do better than the traditional K-means algorithm. While our proposed algorithm solve the problems when clusters are of differing Sizes and Densities, the traditional K-means failed.

REFERENCE

[1]. MacQueen J. B., "Some Methods for classification and Analysis of Multivariate Observations. Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability". Berkeley, University of California Press, 1:281-297, 1967.

[2]. Johnson S. C., "Hierarchical Clustering Schemes". Psychometrika, 2:241-254, 1967.

[3]. Fayyad, U., Piatetsky-Shapiro, G. and Smyth, P. 1996. "From data mining to knowledge discovery: an overview." in U. Fayyad, G. Piatetsky-Shapiro, P. Smyth and R. Uthursamy, eds, 'Advances in Knowledge Discovery and Data Mining', AAAI-Press, pp.1–34.

[4]. Longbing Cao, 2012, Domain-Driven Data Mining: Challenges and Prospects, IEEE Transactions On Knowledge And Data Engineering, Vol. 22, No. 6,pp.755-769.

[5]. S. M. Savaresi and D. Boley. 2004, 'A comparative analysis on the bisecting K-means and the PDDP clustering algorithms'. Intelligent Data Analysis , 8(4), pp.345–362.

[6]. G. Hamerly and C. Elkan. Alternatives to the k-means algorithm that find better clusterings. In Proc. of the 11th Intl. Conf. on Information and Knowledge Management, pp:600–607, McLean, Virginia, 2002. ACM Press

[7]. Geng, L., Hamilton.H.J. 2006. "*Interestingness Measures for Data Mining: A Survey*", ACM Computing Surveys, Vol. 38, No. 3, Article.9, pp.1-32.