Journal of Modern Applied Statistical Methods

Volume 5 | Issue 1 Article 21

5-1-2006

A Combined Standard Deviation Based Data Clustering Algorithm

Kuttiannan Thangavel
Gandhigram Rural Institute, Deemed University, ktvel@rediffmail.com

Durairaj Ashok Kumar
Government Arts College, akudaiyar@rediffmail.com

Journal of Modern Applied Statistical Methods
May, 2006, Vol. 5, No. 1, 258-265

1538 – 9472/06/$95.00

A Combined Standard Deviation Based Data Clustering Algorithm

Kuttiannan Thangavel
Department of Mathematics
Gandhigram Rural Institute, Deemed University

Durairaj Ashok Kumar
Department of Computer Science
Government Arts College

The clustering problem has been widely studied because it arises in many knowledge management oriented applications. It aims at identifying the distribution of patterns and intrinsic correlations in data sets by partitioning the data points into similarity clusters. Traditional clustering algorithms use distance functions to measure similarity centroid, which subside the influences of data points. Hence, in this article a novel non-distance based clustering algorithm is proposed which uses Combined Standard Deviation (CSD) as measure of similarity. The performance of CSD based K-means approach, called K-CSD clustering algorithm, is tested on synthetic data sets. It compared favorably to widely used K-means clustering algorithm.

Key words: Clustering algorithm; combined standard deviation.

Introduction

A fundamental problem that frequently arises in a great variety of fields, such as pattern recognition, image processing, machine learning and statistics in the clustering problem (Narasimha, Jain, & Flyinn, 1999). In its basic form, the clustering problem is defined as the problem of finding homogenous groups of data points in a given data set. Each of these groups is called a cluster and can be defined as a region in which the density of objects is locally higher than in other regions.

K. Thangavel is Head of the Department of Mathematics, Gandhigram Rural Institute–Deemed University, Gandhigram–624 302, Tamilnadu, India. His research includes optimization algorithms, pattern searching and recognition algorithms, and neural networks. E–mail: ktvel@rediffmail.com. D. Ashok Kumar, a doctoral candidate, Lecturer, and Head in the Department of Computer Science, Government Arts College, Udumalpet–642 126, Tamilnadu, India. Email: akudaiyar@rediffmail.com. His research interest includes pattern recognition, neural networks, and genetic algorithms

Clustering methods can be classified into two categories: Hierarchical and Non-Hierarchical. The hierarchical methods can be further divided into agglomerative methods is viewed as a cluster and at each level, some clusters are divided into smaller clusters. There are also many non-hierarchical methods, which divide the set into clusters. These methods are further divided into two: the partitioning method, in which the clusters are mutually exclusive and the clumping method, in which overlap is allowed.

The simplest form of clustering is partitional clustering which aims at partitioning a data set into disjoint subsets (clusters) so that specific clustering criteria are optimized. The most widely used criteria in this clustering is the error criterion, which for each point computes its squared distance from the corresponding cluster center and then takes the sum of these distances for all points in the data set. A popular clustering method that minimizes the clustering error is the K-means clustering algorithm. However, the k-means clustering algorithm is a local search procedure and it is well known that its performance heavily depends on the initial starting conditions and centroid computed based on that (Pena & Larranaga, 1999). To treat this problem, several other techniques have been developed that are based on stochastic global optimization methods (eg. Genetic algorithm

simulated annealing). However, it must be noted that these techniques have not gained wide acceptance and in many practical applications the clustering method that is used in the K-means clustering algorithm with multiple restarts (Maulik & Bandyopadhyay, 2000).

The K-CSD clustering algorithm is proposed, which constitutes an effective clustering for minimization of the clustering error. The basic idea underlying the proposed method is that an optimal solution for a clustering problem with K clusters can be obtained using combined standard deviation. At each step, instead of placing the data point by minimum distance between centroid and the data point, the minimum combined standard deviation is used which leads to optimal clusters. In addition to effectiveness, the method is deterministic and does not depend on centroid. These are significant advantages over all clustering approaches mentioned above.

Clustering

Clustering has been always a key task in the process of acquiring knowledge. The complexity and especially the diversity of phenomena have forced society to organize the things based on their similarities (Spath, 1989). One can say that the objective of the cluster analysis is to sort out the observations into groups such that the degree of natural association is high among members of the same group and low between members of different groups. And clustering is a technique, which is used to find groups of clusters that are somehow similar in characteristic from the given data set for which the real structure is unknown.

Clustering is often confused with classification, but there are some differences between the two. In classification, the data are assigned to predefined classes or clusters, whereas in clustering the classes or clusters are also to be defined and also when the only data available are unlabelled. The classification problems are, sometimes, referred to as unsupervised classification. Cluster analysis can be defined as a wide variety of procedures that can be used to create a classification. These procedures empirically form clusters of groups of highly similar entities. In other words, it can be said that cluster analysis defines group of cases through a number of procedures, which are more similar among them than all the others.

The clustering methods can be basically classified into two categories: Hierarchical and Nonhierarchical. The hierarchical methods can be further divided into the agglomerative methods and the divisive methods. The agglomerative methods merge together the most similar clusters at each level and the merged clusters will remain in the same cluster at all higher levels. In the divisive methods, initially, the set of all object is viewed as a cluster and at each level, some clusters are divided into smaller clusters. There are also many nonhierarchical methods which divide the dataset into clusters. These methods are further divided into two: the partitioning method, in which the clusters are mutually exclusive and the clumping method, in which overlap is allowed.

For years, many clustering techniques were proposed in partitional clustering and are now available in the literature (Narasimha, Jain, & Flyinn, 1999). The methods are Forgy's algorithm, Kmeans algorithm, ISODATA and its variants. The extensive studies (Tseng & Yang, 1999; Narashinha & Sridhar, 1991; Maulik & Bandyopadhyay, 2000) dealing with comparative analysis of different clustering methods suggests that there is no general strategy, which works equally well in the different problems domain. However, it has been found that it is usually beneficial to run schemes that are simpler, and execute them several times, rather than using schemes that are very complex but need to be run only once.

K-Means Clustering Algorithm

The aim of this study is a clustering technique that will not assume any particular underlying distribution of the data set being considered. As well, it should be conceptually simple like the K-means algorithm (Duda & Hart, 1973; Macqueen, 1967). The searching through algorithm is explored in order to search for appropriate cluster centers in the feature space such that a similarity metric of the resulting cluster is optimized.

In fact, to compare the performance or to check the optimality, one does not have the sufficient information regarding the structure of the data set. Thus, to determine the best clusters,

a better algorithm is devised which is more valid. It can be established by ranking the utility of clustering results obtained from different clusters algorithms, with respect to certain application domains, where utility can be measured. As the cluster centers are updated in the K-means and proposed algorithms, the distance between the cluster centers and each of its points can be treated as a unique measure. Mathematically, the clustering metric μ for K clusters $C_1, C_2, \ldots, C_K$

$$\mu(C_1, C_2, \ldots, C_K) = \sum_{i=1}^{K} \sum_{xj \in Ci} \|xj - zi\|$$

where Ci are clusters and zi are cluster centers.

The clustering algorithm searches for the appropriate cluster centers $z_1, z_2, \ldots, z_K$ such that the clustering metric μ is minimized. The K-means algorithm is briefly described below in the sequel:

Input: Set of sample patterns $\{x_1, x_2, \ldots, x_m\}$, $x_i \in R^n$

Output: Set of Clusters $\{ C_1, C_2, \ldots, C_K \}$.

Step 1: Choose K initial cluster centers $z_1, z_2, \ldots, z_K$ randomly from the m patterns $\{ x_1, x_2, \ldots, x_m \}$ where $K < m$.

Step 2: Assign pattern x_i to cluster C_j, where $i = 1, 2, \ldots, m$ and $j \in \{1, 2, \ldots, K\}$, if and only if $\|x_j - z_j\| < \|x_j - z_p\|$, $p = 1, 2, \ldots, K$ and $j \neq p$. Ties are resolved arbitrarily. Compute cluster centers for each point x_i as follows,
$z_i = (1/n_i) \sum x_j$, $i = 1, 2 , \ldots , K$. $x_j \in C_i$
Where n_i is the number of elements belongs to cluster Ci.

Step 3: Assign each pattern x_i to cluster Cj, where $i = 1, 2, \ldots, m$ and $j \in \{1, 2, \ldots, K\}$ if and only if $\|x_j - z_j\| < \|x_j - z_p\|$, $p = 1, 2, \ldots, K$ and $j \neq p$, where $\| \bullet \|$ is an Euclidean metric norm. Ties are resolved arbitrarily, without changing the cluster centers z_j, $j = 1, 2, \ldots, K$

Step 4: Stop.

K-CSD Clustering Algorithm

In a nutshell, the clustering capability of proposed clustering technique using combined standard deviation (Gupta, 2001) is stated in the following steps:

Input: Set of sample patterns $\{x_1, x_2, \ldots, x_m\}$, $x_i \in R^n$

Output: Set of clusters $\{ C_1, C_2, \ldots, C_K \}$.

Step 1: Choose K initial cluster points $z_1, z_2, \ldots, z_K$ randomly from the m patterns $\{x_1, x_2, \ldots, x_m\}$ (where $K < m$) for each cluster.

Step 2: Assign pattern x_i to cluster C_j, where $i = 1, 2, \ldots, m$ and $j \in \{1, 2, \ldots, K\}$, if and only if $CSD(x_j , C_j) < CSD(x_j, C_p)$, $p = 1, 2, \ldots, K$ and $j \neq p$. Ties are resolved arbitrarily. The $CSD(x_j , C_j)$ is obtained by including point x_i into Cluster C_j and find the Combined Standard Deviation of new cluster C_j .

Step 3: Compute cluster centers for each point x_i as follows, $z_i = (1/n_i)\sum x_j$, $i = 1, 2 , \ldots , K$. $x_j \in C_i$ Where n_i is the number of elements belongs to cluster Ci.

Step 4: Assuming z_i are the new initial points to each cluster C_j. Assign each pattern x_i to cluster C_j, where $i = 1, 2, \ldots, m$ and $j \in \{1, 2, \ldots, K\}$ if and only if $CSD(x_j , C_j) < CSD(x_j, Cp)$, $p = 1, 2, \ldots, K$ and $j \neq p$. Ties are resolved arbitrarily, without changing the cluster centers z_j, $j = 1, 2, \ldots, K$

Step 5: Stop

Experimental Results

The experimental results are carried out to compare the Proposed Algorithm clustering algorithm with the K-means clustering algorithm using two synthetic data sets: Data1 and Data2. These are described below:

Data1: This is a non-overlapping two dimensional data set where the number of classes is three. It has several

patterns which are selected from those classes by giving equal probabilities. The value of K is chosen to be 3 for this data set.

Class 1: [0, 20] X [40, 60]
Class 2: [40, 60] X [0, 20]
Class 3: [80,100] X [60, 80]

The results of K-means clustering algorithm and Proposed Algorithm clustering algorithm are shown in the following Tables: Table 1, Table 2, Table 3, and Table 4 for 30, 60, 90, and 120 patterns of Data 1 respectively for different configurations of data sets generated.

Table 1 : 30 patterns

Configu-ration	K-means		K-CSD	
	Number of Clusters	μ – Euclidean metric	Number of Clusters	μ - Euclidean metric
1	3	186.17	3	115.69
2	3	145.12	3	131.74
3	3	156.12	3	130.42
4	3	186.05	3	235.82
5	3	77.52	3	129.23
Total	15	750.98	15	742.90
Average	3	150.196	3	148.58

Table 2 : 60 patterns

Configu-ration	K-means		K-CSD	
	Number of Clusters	μ – Euclidean metric	Number of Clusters	μ - Euclidean metric
1	3	282.32	3	320.43
2	3	214.27	3	187.92
3	3	274.54	3	201.53
4	3	102.26	3	187.97
5	3	224.85	3	179.29
Total	15	1098.24	14	1077.14
Average	3	219.648	2.8	215.428

Table 3 : 90 patterns

Configu-ration	K-means		K-CSD	
	Number of Clusters	μ – Euclidean metric	Number of Clusters	μ - Euclidean metric
1	3	264.46	3	216.52
2	3	282.80	3	250.27
3	3	187.65	3	140.41
4	3	338.13	3	344.81
5	3	128.46	3	128.94
Total	15	1201.50	15	1080.95
Average	3	240.30	3	216.19

Table 4 : 120 patterns

Configu-ration	K-means		K-CSD	
	Number of Clusters	μ – Euclidean metric	Number of Clusters	μ - Euclidean metric
1	3	252.87	3	272.63
2	3	326.26	3	278.94
3	3	371.83	3	272.04
4	3	323.89	3	277.12
5	3	276.22	3	248.57
Total	15	1551.07	15	1349.30
Average	3	310.214	3	269.86

Data2: This is an overlapping two dimensional data set where the number of classes is three. It has several patterns which are selected from those classes by giving equal probabilities. In the K-means algorithms, the value of K is chosen to be 3 for this data set.

Class 1: [-3.3,-0.7] X [0.7, 3.3]
Class 2: [-1.3, 1.3] X [0.7, 3.3]
Class 3: [-3.3,-0.7] X [-1.3, 1.3]

The results of K-means clustering algorithm and the Proposed Algorithm clustering algorithm are shown in the following Tables: Table 5, Table 6, Table 7 and Table 8 for 30, 60, 90 and 120 patterns of Data 2 respectively for different configurations of data sets generated.

Table 5 : 30 patterns

Configu-ration	K-means	K-CSD		
	Number of Clusters	μ Euclidean metric	Number of Clusters	μ Euclidean metric
1	3	10.22	3	14.33
2	3	13.55	3	9.40
3	3	8.17	3	9.82
4	3	14.27	3	14.21
5	3	16.22	3	9.88
Total	15	62.43	15	57.64
Average	3	12.486	3	11.528

Table 6 : 60 patterns

Configu-ration	K-means	K-CSD		
	Number of Clusters	μ Euclidean metric	Number of Clusters	μ Euclidean metric
1	3	13.65	3	10.07
2	3	13.54	3	12.92
3	3	14.03	3	16.64
4	3	13.25	3	17.64
5	3	17.79	3	13.10
Total	15	72.26	15	70.37
Average	3	14.452	3	14.074

Table 7 : 90 patterns

Configu-ration	K-means	K-CSD		
	Number of Clusters	μ Euclidean metric	Number of Clusters	μ Euclidean metric
1	3	26.38	3	15.29
2	3	21.22	3	27.18
3	3	23.83	3	17.03
4	3	20.83	3	16.55
5	3	17.19	3	16.63
Total	15	109.45	15	92.68
Average	3	21.88	3	18.536

Table 8 : 120 patterns

Configu-ration	K-means	K-CSD		
	Number of Clusters	μ Euclidean metric	Number of Clusters	μ Euclidean metric
1	3	28.63	3	24.74
2	3	30.44	3	19.80
3	3	18.56	3	18.37
4	3	19.22	3	21.87
5	3	20.13	3	20.72
Total	15	116.98	15	105.5
Average	3	23.396	3	21.10

Table 9

Data	No. of Patterns	K-means		K-CSD	
		Number of Clusters	Average Euclidean metric - μ	Number of Clusters	Average Euclidean metric - μ
1	30	3	150.196	3	148.580
	60	3	219.648	3	215.428
	90	3	240.30	3	216.190
	120	3	310.214	3	269.860
2	30	3	12.486	3	11.528
	60	3	14.452	3	14.074
	90	3	21.88	3	18.536
	120	3	23.396	3	21.100
Total		24	992.572	24	915.296
Average		3	124.072	3	114.412

Conclusion

The implemented K-means and proposed K-CSD clustering algorithm is tested with two different synthetic datasets to optimize the clustering metric μ. The tested average metric measures of the Data 1 and Data 2 are tabulated in Table 9.

From the Table 9, it could be seen that the average metric is reduced in the proposed algorithm. Future work is planned to design and implement algorithms to cluster data sets with large amount of objects. Such algorithms are required in a number of data mining applications, such as partitioning very large heterogeneous sets of objects into a number of

smaller and more manageable homogeneous subsets that can be more easily modeled and analyzed and detecting underrepresented concepts, e.g., fraud in a very large number of insurance claims.

References

Anderberg, M. R. (1973). *Cluster analysis for applications*. Academic Press.

Duda, R. O., & Hart, P. E. (1973). *Pattern classification and scene analysis.* New York, N.Y.: Wiley.

Gupta S.P., *Statistical methods*, Sultan Chand & Sons, 2001.

Lin, Y. T. & Shiueng, B.Y (2001). A genetic apporach to automatic clustering problem. *Pattern Recognition, 34*(2), 415-424.

Macqueen, J. (1967). Some methods for classification and analysis of multivariate data. Proc. 5th Berkeley Symposium on probability and statistics. Berkeley, C.A.: University of California Press.

Narashinha, M. M. & Sridhar, V. (1991). A knowledge based clustering algorithm. *Pattern Recognition letters, 12*, 511-517.

Narasimha, M. M., Jain, A. K., & Flyinn, P. J. (1999). Data clustering : A review. ACM Computing Surveys, ,*31*(3), 264-323.

Pena, J. A. L. J. M. & Larranaga, P. (1999). An empirical comparison of four initialization methods for the K-means algorithm. *Pattern Recognition Letters, 20*, 1027-1040.

Spath, H. (1989). *Cluster analysis algorithms*. Chichester, U.K.: Ellis Horwod.

Ujiwal, M. & Sanghamitra, B. (2000). Genetic algorithm based clustering technique. *Pattern Recognition Letters, 33*, 1455-1465.