Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 46 (2015) 210 – 215

Procedia Computer Science

International Conference on Information and Communication Technologies (ICICT 2014)

Building Inter Cluster Movement Estimation (ICME) model-A step by step approach

Rajee A M[a,]*, Sagayaraj Francis F[a]

[a]*Department of CSE, Pondicherry Engineering College, Puducherry 605014, India*

Abstract

One of the most well-known techniques in data mining is clustering. This paper presents a scenario of introducing new unclustered information to the already clustered system. Consequently, there is an occurrence of movement of data points between clusters, to accommodate the new entrée. This paper attempts to develop an Inter Cluster Movement Estimation (ICME) model to predict this behaviour of the data points in the clustering system. Better prediction will result in the reduction in the number of times, re-clustering is done on the data sets. Experiments were conducted on datasets with multiple instances and attributes. On analysis, the study revealed that ICME model was in concurrence with observed values with a lower error rate. Real data sets from UCI Data Repository were used for comparative analysis of ICME with similar partitional clustering algorithms including adaptive K Mean and Fuzzy C Means. Reports prove that ICME was found to converge faster consuming lesser number of iterations than adaptive K means and Fuzzy C Means.

Peer-review under responsibility of organizing committee of the International Conference on Information and Communication Technologies (ICICT 2014)

Keywords: Data clustering; Cluster parameters; Inter cluster movement; Estimation model; Multi-dimensional data sets

1. Introduction

Clustering is the most important un-supervised learning problem which divides data into groups of similar objects. Clustering dichotomizes data without the need for labelled samples. A cluster is a group of data objects that possess similar characteristics.

* Corresponding author. Tel.: +91-8438487063.
E-mail address: rajee.am@gmail.com

Peer-review under responsibility of organizing committee of the International Conference on Information and Communication Technologies (ICICT 2014)
doi:10.1016/j.procs.2015.02.013

In the process of clustering, we tend to minimize the intra cluster distance while maximizing the inter cluster distance[2]. Cluster analysis is widely used in many areas in the current era especially in market research, data analysis, pattern recognition and image processing.[1, 10, 15]

A partition or clustering in *X* is a set of disjoint clusters that partitions *X* into *K* groups from the data set $N: C = \{C_1, C_2, ..., C_K\}$. A dataset is defined as a set of *N* objects (or synonymously, data points, instances, cases, patterns, tuples, transactions) represented as vectors in an F-dimensional space: $X = \{x_1, x_2, ..., x_N\} \subseteq \Re^F$. The Euclidean distance between object x_i and x_j is denoted as $d(x_i, x_j)$ and the Euclidean distance between the clusters C_i and C_j is denoted as $d(C_i, C_j)$. A cluster C_i is closest (or synonymously nearest) to cluster C_j if the Euclidean distance between their centers is smallest among each of the cluster pairs in the clustering system. We therefore refer C_i is closest to C_j and vice versa.

The clustering system was built with a set of K Clusters $\forall$ K $\geq$ 1. When new unclustered information is fed to this system, the clusters regroup themselves to accommodate the new entrée. The representative data points of each cluster tend to move between the clusters, changing the composition of clustering system. This concept is known as inter cluster movement (or synonymously inter cluster migration) of data points.

With this context, an Inter Cluster Movement Estimation (ICME) model was built which predicts the movement of data points between clusters (initiated by the new entrée) thereby avoiding repeated re-execution of the clustering algorithm.

The remaining section of this paper is organized as follows. Section 2 discusses the recent and related research works on clustering. Section 3 presents the formulation of ICME model. Section 4 builds the experimental setup and performs analysis of the model in varying instances of multi-dimensional data sets. A Comparative study of ICME over similar clustering approaches is made in section 5. Section 6 draws the conclusion from the paper.

2. Related Works

Research works were progressing in adaptive and other variants of clustering. A few of them are discussed in the remaining lines. Unclassified information is assigned to the K means clusters by measuring the distance to the closest cluster or by setting up threshold limit. A. Campan and G. Serban proposed Core Based Adaptive k-means (CBAk) [16] and Hierarchical Adaptive Clustering (HAC) approach[17] for re-clustering an object set in the previously clustered system, when the attribute information is newly added. Similar cluster models were also seen in adaptive clustering where the existing clustering solutions are changed from the queries and user feedback[13]. Input data stream was assigned to the already known data structure by O. Georgieva, F. Klawonn and parallel proposed hard and fuzzy algorithm variants[6]. Dynamic weighting of data by interactively updating centers was introduced by Si-Bao Chen, Hai-Xian Wang, Bin Luo[9].

Seokkyung Chung and Dennis McLeod performed incremental clustering from web articles (documents) that change over time. The proposed algorithm incrementally clusters documents based on neighborhood search and computes their similarity. The re-clustering was effected by merging the documents to a singleton cluster[4, 11].

A novel method for clustering incoming data point by finding farthest neightbor point was developed by A.M.Sowjanya and M. Shashi. The efficiency of this method was tested on mixed data sets[5].

This brief study on related work gives us a platform to unleash the working behaviour of the estimated model.

3. Building Inter Cluster Movement Estimation (ICME) Model

A system is set up with a set of K Clusters $\forall$ K $\geq$ 1. New unclustered information is fed to the existing clustering system. The new unclustered element will become member of its nearest cluster and will facilitate any of the following actions.

- Facilitate the movement of other data points between clusters, thus upsetting the clustering setup

- Does not influence inter cluster movement, thereby becoming a silent member of the cluster.

This paper is inclined to study the behaviour of the clustering system considering the former case. To reduce the number of times, re-clustering is done on the data set, an Inter Cluster Movement Estimation (ICME) model was built, which will predict the occurrence of inter cluster movement, thereby letting the user to decide on the re-execution of the clustering system. Various cluster parameter including cluster size, cluster separation, cluster cohesion were studied. These studies implied that the cluster parameters tend to influence the inter cluster movement[18,19]. The ICME model is finding the estimator *D* which is the minimum distance of the new point from its closest cluster, effecting inter cluster movement and is expressed as a function of Cluster Separation (CS), distance between the center of two closest clusters $d(C_i, C_j)$ and Sum of Squared Error (SSE). SSE or the within sum of squares is defined as the summation of the squared distances from each data point to its cluster center. An algorithm for finding D is discussed in the following lines.

Algorithm 1: Inter Cluster Movement Estimation

Input: a set of K Clusters
Output: an estimator for predicting the occurrence of inter cluster movement, due to a new point
Steps:

1. Introduce a new unclustered data point x_Y to the clustering system. Let C_{near} be the closest cluster to x_Y.

2. For the cluster C_{near}, find its cluster separation measure $CS_{near} = |C_{near}| * (m - m_{near})^2$ where $|C_{near}|$ is the number of data points in the cluster C_{near}, m is the overall center of the clustering system and m_{near} is the center of cluster C_{near}

3. Compute $d(C_{near}, C_j)$ which is Euclidean distance between the centers of C_{near} and its closest cluster C_j.

4. Inter Cluster Movement Distance estimator is formulated as $$D \propto \frac{CS_{near} * d(C_{near}, C_j)}{SSE} \quad (1)$$

5. Introducing 'migrator' as a proportionality constant, $$D = \text{migrator} * \frac{CS_{near} * d(C_{near}, C_j)}{SSE} \quad (2)$$

6. If $D \geq d(x_Y, C_{near})$, then there is a possibility of inter cluster movement due to x_Y.

7. Else, there will be no occurrence of inter cluster movement.

   7.1 Assign x_Y to C_{near} and $|C_{near}| = |C_{near}| + 1$.

Assigning a value to the 'migrator' constant depends on the inter dependence of cluster separation on SSE, $d(C_i, C_j)$ and cluster size.

The ICME model will estimate the minimum distance D, that is, whether a new point, when positioned at the distance D will facilitate the movement of data between the clusters or not. This distance D is the Euclidean distance between the new point and the center of its nearest cluster.

4. Experimental Analysis

Synthetic data sets were created with varied number of multi-dimensional instances. Table 1 and Table 2 shows two clusters built from different instances of two dimensional and three dimensional data sets respectively.

Table 1. Result summary of two clusters from two dimensional data sets

Observed experimental values	Inter Cluster Movement Estimator D	Relative Error (%)
191.2	186.7009	2.353068
182.5	178.2933	2.305024
174.4	170.3847	2.302362
167	163.0461	2.367602
160	156.3562	2.277404
153.5	150.3995	2.019899
148.1	145.2641	1.914832
143.3	141.0378	1.578656

The values in column one shows the observed distance of the new point which initiated movement of data points between the clusters. The second column displays the estimated D values. A lower error percentage in the third column describes the reliability of the estimator model, explaining the adequacy for its implementation in the clustered system.

Table 2. Result summary of two clusters from three dimensional data sets

Observed experimental values	Inter Cluster Movement Estimator D	Relative Error (%)
71.9	68.24211	5.087466
65.4	61.15259	6.494517
62.6	59.4003	5.111347
47.4	43.81566	7.561902
40.2	36.319587	9.652769
66.8	65.479974	1.976087
45	42.04955	6.556555
9.8	10.262842	4.722878

Similar experiments were repeated with other instance of multi - dimensional data sets and the results obtained revealed that, the estimated ICME values are in concurrence with the experimental values.

5. Comparison of ICME with similar clustering models

Some of the partitional clustering algorithms like adaptive K –Means[14], Fuzzy C Means are sensitive to the incoming data point[7, 8]. These algorithms possess their own methodology to handle the unclustered data into the existing clustered setup. Since the proposed Inter Cluster Movement Estimation model exhibit this similar feature, a run time assessment all the clustering algorithms would be a better analogous indicator for selecting ICME over these approaches.

Fig.1. portrays different clustering solutions, built from the real data sets from the UCI machine learning repository as specified in Table 3. ICME was found to converge fastest spanning less number of iterations. Fuzzy C Means is seen to iterate more than adaptive K means. The ICME model integrated with K Means spends less spell on the new point, which illustrates a clear picture on its better efficiency.

Table 3. Real data sets from UCI repository

Data sets	Number of Instances	Number of Attributes
Breast Tissue	106	9
Ecoli	336	7
Iris	150	4
Transfusion	748	4
Vehicle	846	5
Yeast	1484	8

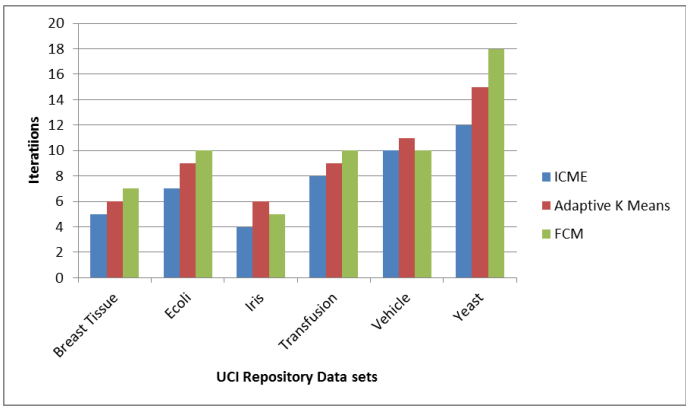

Fig. 1. Comparative analysis of ICME over similar clustering models.

6. Conclusion

We propose an approximation model to predict the occurrence of movement of data points between clusters, with a new input. The estimated value computes the distance of the new point that will cause a change in cluster dynamics. This model will be useful in letting the user to decide on the strenuous activity of redundant re-execution of the clustering procedures. Experiments were conducted on various instances of multi-dimensional data sets with different number of clusters. On analysis, the results show that the estimated model was in compatible with the observed experimental values with a significant lower error rate. Real data sets from UCI repository was also considered with ICME model spanning lesser number of iterations than adaptive K Means and Fuzzy C Means.

References

1. Daniel Barbard, Requirements for Clustering Data Streams. *SIGKDD Explorations* 2002; **3**-43.
2. Charu C Aggarwal , Philip S Yu. A Framework for Clustering Massive Text and Categorical Data Streams. *ACM SIAM Data Mining Conference* 2006.
3. Angie King.Online k-Means Clustering of Non-stationary Data. *Prediction Project Report*; 2012.
4. Seokkyung Chung and Dennis McLeod, Dynamic Pattern Mining: An Incremental Data Clustering Approach. *Journal on Data Semantics*, 2005; **2**-11.
5. Sowjanya A M and Shashi M, A Cluster Feature-Based Incremental Clustering Approach to Mixed Data. *Journal of Computer Science* 2011.
6. Georgieva O and Klawonn F, Dynamic data assigning assessment clustering of streaming data. *Journal of Applied Soft Computing* 2000: **8**-12.
7. Jain A K. Data Clustering: 50 Years Beyond K-means. *Pattern Recognition Letters* 2010; **31**-1, p.651–666.
8. Jain A K, Murty M N, Flynn P J. Data Clustering: A Review. *ACM Computing Surveys* 1999; **31**-3, p.264–323.
9. Si-Bao Chen, Hai-Xian Wang, Bin Luo, On Dynamic Weighting of Data in Clustering with K-Alpha Means, International Conference on Pattern Recognition 2010.
10. S. Lloyd, Least Squares Quantization in PCM. *IEEE Transactions on Information Theory*, 1982**, 28**-2, p.129–136.
11. Adil M.Bagirov, Julien Ugon, Dean Webb. Fast modified global k-means algorithm for incremental cluster construction. *The Journal of the Pattern Recognition Society* 2011.
12. Saka E, Nasraoui O. On dynamic data clustering and visualization using swarm intelligence. *IEEE 26th International Conference on Data Engineering Workshops (ICDEW)* 2010.
13. Moses Charikar et al, Incremental clustering and dynamic information retrieval. *SIAM Journal on Computing* 2004.
14. Campan A, Serban G. Adaptive Clustering algorithms. *Advances in Artificial Intelligence* 2006.
15. Han J, Kamber M. *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, 2001.
16. Gabriela Serban, Alina Campan. Adaptive Clustering using a Core-based Approach, *Informatica* 2008; **L**-2.
17. Gabriela Serban, Alina Campan. Hierarchical Adaptive Clustering, *Informatica* 2008; **19**-1, p.101–112.
18. Bhavani Raskutti and Christopher Leckie. An Evaluation of Criteria for Measuring the Quality of Clusters. *IJCAI*, 1999.
19. Rajee AM, Sagayaraj Francis F. Inter Cluster Movement Estimation model based on cluster parameters. *IEEE International Conference on Computational Intelligence and Computing Research* 2013, p.369-372.