JCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Volume-2, Issue-11 E-ISSN: 2347-2693

A Probabilistic Estimation of Cluster Region Prone to Inter Cluster Data Movement

A.M.Rajee[1*] and F. Sagayaraj Francis[2]

[1*,2]*Department of CSE, Pondicherry Engineering College, India*
www.ijcseonline.org

Received: 08 Nov 2014 Revised: 16 Nov 2014 Accepted: 26 Nov 2014 Published: 06 Dec 2014

Abstract— Data clustering is an unsupervised learning methodology. We consider the problem of dealing with unclustered information to the already existing clustering setup. The new entrée may cause movement of data points between clusters, thereby altering the dynamics of the clustering system. With this scenario, this paper attempts to predict the region in the cluster which will facilitate such inter cluster data movement. A probabilistic model was built which will estimate the region, which has higher chance for enabling the data objects to move in and out of the cluster. Experimental studies were made with multiple instances of synthetic two dimensional data sets. The observed values were compared with the predicted values and the results displayed improved accuracy of the probabilistic model.

Keywords—Data Clustering; Inter Cluster Data Movement; Probabilistic Model; Un-Clustered Information

I. INTRODUCTION

Cluster analysis is an important data mining method which aims to classify a sample of subjects (or objects) into a number of different groups such that similar subjects are placed in the same group [1]. Cluster analysis finds its place in the field of psychiatry, where the characterization of patients on the basis of clusters of symptoms can be useful in the identification of an appropropriate form of therapy. In marketing, it may be useful to identify distinct groups of potential customers so that, for example, advertising can be appropriately targeted.

A clustering system is a group of similar objects (or synonymously data points, instances, observations, elements) which is grouped to K clusters N: $\{C_1, C_2,\dots, C_K\}$ with corresponding cluster centers labeled as $c_1,c_2,\dots., c_K$. Each individual data point in the cluster is labeled as x_i, $1 \le i \le n$, where n is the total number of objects in the clustering system. The Euclidean distance between the two objects x_i and x_j is denoted by $d(x_i,y_i)=\left(\sum_{i=1}^{m}(x_i-y_i)^2\right)^{\frac{1}{2}}$ where m is the number of dimensions. The Euclidean distance between the centers of two clusters C_i and C_j is denoted as $d\left(C_i,C_j\right)$. 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.

This paper presents the problem of bringing a new unclustered data object to the existing clustering model and developing a probabilistic approach to predict the clustering region facilitating such data movement.

This paper is organized as follows. Section 2 highlights the related literature works. Section 3 presents the estimation model problem scenario. Section 4 presents the experimental results and section 5 compares the approach with observed experimental results. Section 6 concludes the paper.

II. RELATED WORKS

Very few research works are progressing to handle new entrée to the existing clustering system. Each approach devised its own methodology to handle the incoming data point. In recent years, studies were made on continuous flow of data, also known as data streams. Traditional data mining clustering algorithms will not work on dynamic data sets. The clustering model must be restructured to confront the arrival of data points. As data stream arrive continuously, the decision of where to locate the next point to the existing clusters, becomes an issue. Two actions are possible. Re-cluster the entire set of points seen so far, including the last n points that prompted the decision to re-cluster (or) discard the old clusters, and produce a new set of clusters, by considering only the last n processed points [2].

Unclassified information is assigned to the K Means clusters by measuring the distance to the closest cluster or by setting up threshold limit. A. Campari and G. Serban proposed core based adaptive k-means (CBAk) [3, 4] and Hierarchical Adaptive Clustering (HAC) approach [4] for re-clustering an object set in the previously clustered system, where the attribute information is newly added. The dynamics of the system by inducing un-clustered

Corresponding Author: *RAJEE AM ,*
Email id: rajee.am@gmail.com

information to the K Clusters was studied.

Charu C. Aggarwal, Philip S. Yu presented an online approach for clustering massive text and categorical data streams. A time-sensitive weightage was assigned to each data point. An entry was maintained to keep track of the last time; a point was added to the cluster [5]. Angie King suggested a discounted center updating rule as a modification of the updating rule proposed by Lloyd. This proposed exponential smoothing heuristic algorithm works for the naturally clusterable data and for the cluster centers, which are moving over time [6]. 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 [7]. Based on these inputs, this paper attempts to identify a specific location within each cluster enabling inter cluster data movement [8].

III. PROABABILISTC ESTIMATION OF CLUSTER LOCATION PRONE TO INTER CLUSTER MOVEMENT

When a new data object comes to the clustering setup, it may either become member of any its nearest cluster or become a member of the cluster, facilitating movement of data points between clusters. This process is known as Inter Cluster Migration [8].

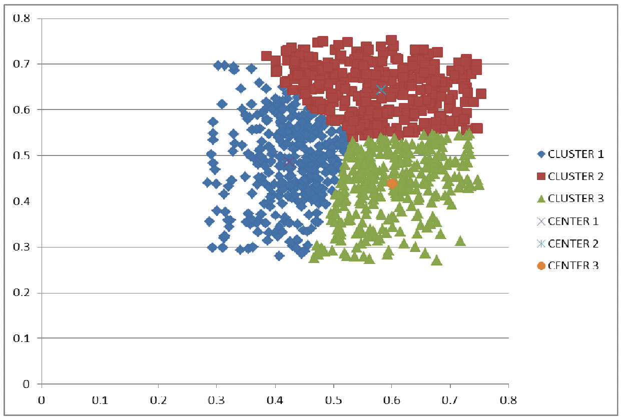

Fig.1. Clustered system with 1000 x 2 instances (K=3)

The problem scenario is based on the new point initiating inter cluster movement, disturbing the existing cluster setup. A new unclustered data object arrives at a distance d from its closest cluster center[9,10]. Let the closest cluster to the new point be C_i. Assuming, the new point triggers inter cluster movement of data objects between C_i and other clusters in the system. We need to estimate the region in C_i, where there is higher probability of inter cluster data movement.

Let the cluster C_i be the closest cluster to the new data point. Let the first closest cluster of C_i is C_j and the second closest cluster of C_i is C_k. A set of data points in C_i is closest to C_j if the Euclidean distance between each of the data points in C_i and the center c_j is minimum among other data members of C_i. A set of data points in C_i is farthest to C_j if the Euclidean distance between each of the data points in C_i and the center c_j is maximum among other data members of C_i.

Consider a three clustered system C_1, C_2 and C_3 formed from two dimensional synthetic data set consisting of 1000 instances. This clustered setup is shown in Fig.1. The first and second nearest clusters of each cluster is given in Table 1. It is conspicuous from Table 2 that if the first nearest cluster of C_2 is C_3, then the reverse may not be true.

TABLE I. First and second nearest clusters

Clusters	C_1	C_2	C_3
First nearest to	C_3	C_3	C_1
Second nearest to	C_2	C_1	C_2

Let C_i be the closest cluster to the new entrée causing inter cluster movement.

The function near returns a set of member data points of C_i, whose distance from the center of nearest cluster of C_i is minimum. The function far returns a set of data points of C_i, whose distance from the center of farthest cluster of C_i is maximum. Table 2 gives a brief introduction of the near and far functions used throughout this paper.

TABLE II. First and second nearest clusters

Terms	***Definition***
Near(C_i,C_j)	Set of data points in C_i, which is nearest to c_j
Far(C_i,C_j)	Set of data points in C_i, which is farthest to c_j.
Near(C_i,C_k)	Set of data points in C_i, which is nearest to c_k
Far(C_i,C_k)	Set of data points of C_i, which is farthest to c_k

When a new entree arrives to C_i, the member data point(s) of C_i which are farthest from c_i and nearest to c_j are very likely to move between clusters. The region enclosing these data points will be more prone to facilitate such inter cluster data movement. The region with the data points farthest from c_i and nearest to c_k will receive the next lower probability of data movement. The probability of those data points, which are very nearest to c_i and likely to move between clusters, is zero. Algorithm 1 sets the probability for the regions in the cluster C_i based on these assumptions.

Similarly, for the same new arrival, data points in C_j and C_k may also shift between clusters, but will follow the similar near and far patterns in their inter cluster movement. In that case, the reference cluster is C_i, the first and second nearest cluster of C_i are C_j and C_k respectively.

The cluster C_i is divided into four regions, R_1,R_2,R_3 and R_4, with all the member data points x_i of C_i, , $1 \leq i \leq |C_i|$ is enclosed within $R_{xi} \forall$ $1 \leq x \leq 4$ as shown in fig.2. Let $p(R_{xi})$ be the probability of region R_{xi} of cluster C_i facilitating movement of data points between clusters.

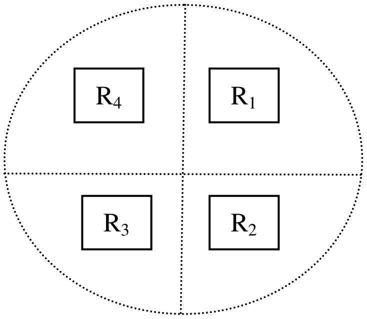

Fig.2. Cluster C_i with four regions

```
Algorithm 1

Assumption: Let $C_i$ be the nearest cluster to the new point.

Input: Set of K clusters $\{C_1, C_2,\ldots, C_K\}$, Region $R_{xi}$ of $C_i$
Output: $p(R_{xi})$ probability (maximum likelihood) of the region

Steps:
For each region $R_{xi}$ in $C_i$,
    1. Let $R_{1i}= \{x_i \forall near(C_i,C_j) \subseteq R_{1i}, near(C_i,C_j) \cap R_{1i} \geq | near(C_i,C_j)| -2\}$, $p(R_{1i})=0.6$
    2. Let $R_{2i}= \{x_i \forall far(C_i,C_j), 1\leq i\leq|C_i|, far(C_i,C_j) \subseteq R_{2i}, far(C_i,C_j) \cap R_{2i} \geq | far(C_i,C_j)| -2\}$, $p(R_{2i})=0.06$
    3. Let $R_{3i}= \{x_i \forall near(C_i,C_k), 1\leq i\leq|C_k|, near(C_i,C_k) \subseteq R_{3i}, near(C_i,C_k) \cap R_{3i} \geq | near(C_i,C_k)| -2\}$, $p(R_{3i})=0.3$
    4. Let $R_{4i}= \{x_i \forall far(C_i,C_k), 1\leq i\leq|C_k|, far(C_i,C_k) \subseteq R_{4i}, far(C_i,C_k) \cap R_{4i} \geq | far(C_i,C_k)| -2\}$, $p(R_{4i})=0.04$

End for
```

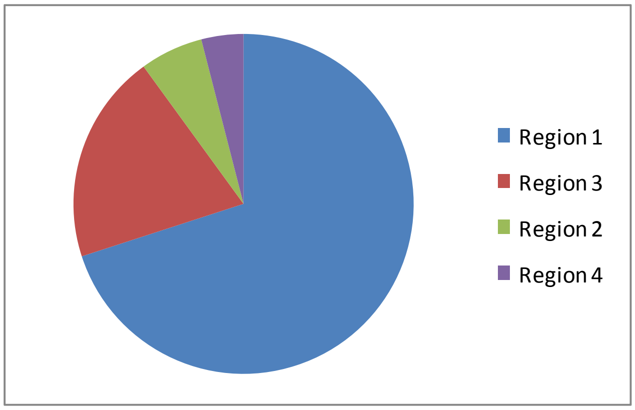

Fig.3.Probability of the region facilitating inter cluster data movement for a cluster

IV. EXPERIMENTAL RESULTS AND ANALYSIS

A clustering system with 6 clusters was built with varied 2 dimensional instances. A new point is fed to the clustering setup. The cluster which is nearest to the new point is assumed to be the reference cluster. The regions in the reference cluster C_i were labeled as R_{1i}, R_{2i}, R_{3i} and R_{4i}. The algorithm 1 is executed to estimate the probability of the region facilitating inter cluster movement. The predicted probability values of the regions are compared with the region supporting data movement, with the arrival of new point. Table 1 gives the results. The value of the region where there happens inter cluster movement is assumed to be 1. Similar experiments were also performed with other clusters, and similar results were obtained.

TABLE III. Comparison of estimated probability with observed results

Number of Instances	**Region where migration actually happened- Experimental values**	**Probability values for the region- from trails**	**Estimated probability of the region enabling migration**	**Accuracy (in %)**
1204	R1	1	0.6	60
1500	R3	1	0.3	30
1073	R1	1	0.6	60
866	R1	1	0.6	60
512	R1	1	0.6	60
2982	R3	1	0.3	30
197	R1	1	0.6	60

V. CONCLUSION

This paper attempts to handle the incoming new data point that tends to disturb the existing cluster dynamics. Due to this new entrée, the data objects may move between clusters, disturbing the clustering system. This paper devised a prediction mechanism to estimate the region in the cluster, facilitating such data movement. Experimental studies were conducted with data set with multiple instances and varied number of clusters. The results show that the probabilistic model was in concurrence with observed values with relatively higher accuracy.

REFERENCES

[1] J. Han and M.Kamber, "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, **2001**.

[2] S.Lloyd, "Least squares quantization in PCM", IEEE Transactions on Information Theory, **1982, pp.129-136**.

[3] A. Campan and G. Serban, "Adaptive Clustering algorithms", Advances in Artificial Intelligence, Springer, **2006**.

[4] G.Serban and A.Campan, "Adaptive Clustering using a Core-based Approach", Informatica, Volume L, Number 2, **2005**.

[5] Charu C. Aggarwal, Philip S. Yu, "A Framework for Clustering Massive Text and Categorical Data Streams", ACM SIAM Data Mining Conference, **2006**

[6] Angie King, "Online k-Means Clustering of Non-stationary Data", Prediction Project Report, **2012**

[7] Seokkyung Chung and Dennis McLeod, "Dynamic Pattern Mining: An Incremental Data Clustering Approach", Journal on Data Semantics, Volume 2, **2005**

[8] A.M.Rajee and F.Sagayaraj Francis, "Inter Cluster Movement Estimation model based on cluster parameters", in Proc. IEEE International Conference on Computational Intelligence and Computing Research", **2013, pp.369-372**.

[9] Jain A. K, "Data Clustering: 50 Years Beyond K-means", Pattern Recognition Letters 31(8), **2010, pp.651–666**.

[10] Jain A. K, Murty M. N and Flynn, P. J, "Data Clustering: A Review. ACM Computing Surveys", 31(3), **1999, pp. 264–323**.

AUTHORS PROFILE

Rajee.A.M received B.E degree in Computer science and engineering from Bharathidasan University, Tamilnadu, India in 2002 and M.Tech degree in Computer engineering from Manonmaniam Sundaranar University, Tamil nadu, India in 2004. She has got 8 years of academic experience in various engineering college. Currently she is a Full time doctoral student in computer science and engineering at Pondicherry Engineering College, Puducherry, India. Her research areas are Data clustering analysis and techniques. She has published 3 research articles in various conferences and won 1 Best Paper Award.

Sagayaraj Francis.F holds a PhD in Computer Science and Engineering and M.Tech in Computer Engineering, both from Pondicherry University. He is currently working as Professor in the Department of Computer Science and Engineering at Pondicherry Engineering College, Puducherry, India. His areas of interest include Data Analysis and Knowledge discovery, Information systems, Business Intelligence. He is currently guiding 8 PhD Scholars. He is an active life member of ISTE and International Association of Computer Science and Information Technology.