Imperial Journal of Interdisciplinary Research (IJIR)
Vol-3, Issue-3, 2017
ISSN: 2454-1362, http://www.onlinejournal.in

An Efficient Clustering of Cancer Data Sets Using Modified Affinity Propagation

J.Vinitha[1] & M.Praneesh[2]
[1]MPhil Research Scholar , [2]Assistant Professor
[1,2]Department of Computer Science, Sankara College of science and commerce

Abstract:- *Clustering the data by identifying a subset of representative examples is important for processing sensory signals and detecting patterns in data. Such "exemplars" can be found by randomly choosing an initial subset of data points and then iteratively refining it. But this works well, only if that initial choice is close to a good solution. The proposed work have devised a method called Modified Affinity Propagation, which takes the similarity measures between pairs of data points as input. The experimental evaluation is conducted using affinity propagation to cluster data sets of leukemia and colon. It also identifies representative attribute with its resultant accuracy and convergence rate compared to that of k-means, k-means++, global k-means, affinity propagation and modified affinity propagation found clusters with much lower error compared with other methods, and it did so in less than one-hundredth the amount of time.*

Keywords— *Data Mining, Clustering, K-Means, colon, Leukaemia, Affinity propagation.*

1. INTRODUCTION

Data mining is a system of searching large amounts of data for patterns. It is a relatively new concept which is directly related to computer science. Despite this, it can be used with a number of older computer techniques such as pattern recognition and statistics. The goal of data mining is to extract important information from data that was not previously known. Data mining is a technique that has a large number of applications. However, it is commonly used by businesses or organizations that need to recognize certain patterns or trends. Generally, data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives.

Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases. Data mining is becoming increasingly common in both the private and public sectors. Industries such as banking, insurance, medicine, and retailing commonly use data mining to reduce costs, enhance research, and increase sales. In the public sector, data mining applications initially were used as a means to detect fraud and waste, but have grown to also be used for purposes such as measuring and improving program performance.

As with other aspects of data mining, while technological capabilities are important, there are other implementation and oversight issues that can influence the success of a project's outcome. One issue is data quality, which refers to the accuracy and completeness of the data being analyzed. A second issue is the interoperability of the data mining software and databases being used by different agencies. A third issue is mission creep, or the use of data for purposes other than for which the data were originally collected.

2. REVIEW OF LITERATURE

K-means and its variants (Larsen and Aone, 1999; Kaufman and Rousseeuw, 1990; Cutting, Karger, Pedersen, and Tukey, 1992) represent the category of partitioning clustering algorithms that create a flat, non-hierarchical clustering consisting of k clusters. The k-means algorithm iteratively refines a randomly chosen set of k initial centroids, minimizing the average distance (i.e., maximizing the similarity) of documents to their closest (most similar) centroids. The bisecting k-means algorithm first selects a cluster to split, and then employs basic k-means to create two sub-clusters, repeating these two steps until the desired number k of clusters is reached. Steinbach (2000) shows that the bisecting k-means algorithm outperforms basic k means as well as agglomerative hierarchical clustering in terms of accuracy and efficiency (Zhao and Karypis, 2002). Both the basic and the bisecting k-means algorithms are relatively efficient and scalable, and their complexity is linear to the number of documents. As they are easy to implement, they are widely used in different clustering applications. A major disadvantage of k-means, however, is that an incorrect estimation of the input parameter, the number of clusters, may lead to poor clustering accuracy. Also, the k-means algorithm is not suitable for discovering clusters of largely varying sizes, a common scenario in document clustering.

Furthermore, it is sensitive to noise that may have a significant influence on the cluster centroids, which in turn lowers the clustering accuracy. The k-medoids algorithm (Kaufman and Rousseeuw, 1990; Krishnapuram, Joshi, and Yi, 1999) was proposed to address the noise problem, but this algorithm is computationally much more expensive and does not scale well to large document sets.

3. METHODOLOGY

The objective of the thesis is to provide a comparative evaluation of K-family clustering algorithm for its accuracy and convergence rate with real data set colon and leukemia dataset.. The K-family presented are K-means, global k-means, k-means++ and k-family of affinity propagation and modified affinity propagation. Two different cancer datasets to make a study of k-family algorithms, the leukemia data set is a collection of gene expression measurements from 72 leukemia (composed of 62 bone marrow and 10 peripheral bloods) samples report by Golub. it contains an initial training set composed of 47 samples of acute lymphoblastic leukemia (ALL) and 25 samples of acute myeloblastic leukemia (AML).here two variants of leukemia dataset one with 50-genes and another one with 3859-Genes. The colon dataset is a collection of gene expression measurements from 62 colon biopsy sample reported by alon. It contains 22 normal and 40 colon cancer samples. The dataset consists of 2000 genes. The dataset has taken from uci repository.

A. **k-means++ Algorithm**

k-means++ is another variation of k-means, a new approach to select initial cluster centers by random starting centers with specific probabilities is used. The steps used in this algorithm are described below,

Step 1, Choose first initial cluster center c_1 randomly from the given dataset X. Step 2, choose next cluster center $c_i = x_j \in X$ with probability p_i where, denote the shortest distance from x to the closest center already choosen. Step 3, Repeat step2 until k cluster centers are chosen. Step 4, after initial selection of k cluster centers, Apply k-means algorithm to get final k clusters.

B. Global K-Means Clustering Algorithm

Global K-means Algorithm is an improved version of k-means which can avoid getting into locally optimal solution in some degree, and reduce the probability of dividing one big cluster into two or more ones owing to the adoption of squared-error criterion.

Algorithm: Improved K-means(S, k), $S=\{x_1,x_2,\ldots,x_n\}$

Input, The number of clusters k^1 ($k^1 > k$) and a dataset containing n objects(X_i)

Output: A set of k clusters (C_j) that minimize the squared-error criterion)

Step 1: Draw multiple sub-samples {SI, S2, . . . , Sj} from the orginal dataset,

Step 2: Repeat step 3 for m=l to j

Step 3: Apply K-means algorithm for subsample S_m for k^1 cluster

Step 4: Choose minimum of as the refined initial points Z_j , j ϵ [1,k^1] Step 5: Now apply k-means algorithm again on dataset S for k^1 clusters.

Step 6: Combine two nearest clusters into one cluster and recalculate the new cluster center for the combined cluster until the number of clusters reduces into k.

The tested result of global k-means on leukemia and colon dataset (cancer). Global *k*-means deploys the *k*-means algorithm to find locally optimal solutions by trying to keep the clustering error to a minimum. The *k*-means algorithm starts by placing the cluster center arbitrarily and at each step moves the cluster center with the aim to minimize the clustering error. The down side to this algorithm is that it is sensitive to the initial position of the cluster centers. To overcome this, *k*-means can be scheduled to run several times and each time with a different starting point.

C. K-means++ Algorithm

The algorithm consists of mainly two steps which are repeated until completion. Step1, (Improve-Params) in this step, we apply k-means algorithm initially for k clusters till convergence. Where k is equal to lower bound supplied by the user . Step2, (Improve -Structure) this structure improvement step begins by splitting the each cluster center into two children in opposite directions along a randomly chosen vector. After that we run k-means locally within each cluster for two clusters. The decision between the children of each center and itself is done comparing the BIC-values of the two structures. Step 3, if k> =k_{max} (upper bound) stop and report to best scoring model found during search otherwise go to step 1.

D. Affinity Propagation and Modified Affinity Propagation

The exemplar for data point i. When the goal is to minimize squared error, each similarity is set to a negative squared error (Euclidean distance): For points xi and xk, s(i,k) = $-\|xi - xk\|2$. Indeed, the method described here can be applied when the optimization criterion is much more general. When an exemplar-dependent probability model is

available, s(i,k) can be set to the log-likelihood of data point i given that its exemplar is point k. Alternatively, when appropriate, similarities may be set by hand.

Rather than requiring that the number of clusters be pre-specified, affinity propagation takes as input a real number s(k,k) for each data point k so that data points with larger values of s(k,k) are more likely to be chosen as exemplars. These values are referred to as preferences. The number of identified exemplars (number of clusters) is influenced by the values of the input preferences, but also emerges from the message-passing procedure. If a priori, all data points are equally suitable as exemplars, the preferences should be set to a common value this value can be varied to produce different numbers of clusters. The shared value could be the median of the input similarities (resulting in a moderate number of clusters) or their minimum (resulting in a small number of clusters).

There are two kinds of message exchanged between data points, and each takes into account a different kind of competition. Messages can be combined at any stage to decide which points are exemplars and, for every other point, which exemplar it belongs to. The responsibility r(i,k), sent from data point i to candidate exemplar point k, reflects the accumulated evidence for how well-suited point k is to serve as the exemplar for point i, taking into account other potential exemplars for point i (Fig). The "availability" a(i,k), sent from candidate exemplar point k to point i, reflects the accumulated evidence for how appropriate it would be for point i to choose point k as its exemplar, taking into account the support from other points that point k should be an exemplar (Fig). r(i,k) and a(i,k) can be viewed as log-probability ratios. To begin with, the availabilities are initialized to zero: a(i,k) = 0. Then, the responsibilities are computed using the rule

In the first iteration, because the availabilities are zero, r(i,k) is set to the input similarity between point i and point k as its exemplar, minus the largest of the similarities between point i and other candidate exemplars. This competitive update is data-driven and does not take into account how many other points favor each candidate exemplar. In later iterations, when some points are effectively assigned to other exemplars, their availabilities will drop below zero as prescribed by the update rule below. These negative availabilities will decrease the effective values of some of the input similarities s(i,k′) in the above rule, removing the corresponding candidate exemplars from competition. For k = i, the responsibility r(k,k) is set to the input preference that point k be chosen as an exemplar, s(k,k), minus the largest of the similarities between point i and all other candidate exemplars. This "self-responsibility" reflects accumulated evidence that point k is an exemplar, based on its input preference tempered by how ill-suited it is to be assigned to another exemplar..

4. RESULT AND DISCUSSION

The result of K-means, global k-means, k++ and affinity propagation algorithm and modified affinity propagation for clustering data. To group genes with similar functionalities based on colon and leukemia dataset There are two different cancer datasets to make a study of various k-mean based algorithms. The Leukemia data set is collections of gene expression measurements from 72 leukemia (composed of 62 bone marrow and 10 peripheral bloods) samples reported by Golub. It contains an initial initial training set composed of 47 samples of acute lymphoblastic leukemia (ALL) and 25 samples of acute myeloblastic leukemia (AML).

Here we take two variants of leukemia dataset one with 50-genes and another one with 3859-genes. The Colon dataset is a collection of gene expression measurements from 62 Colon biopsy samples reported by Alon. It contains 22 normal and 40 Colon cancer samples .The Colon dataset consists of 2000 genes.

This dataset is similar to the yeast gene expression dataset: it contains expression levels of 2000 genes taken in 62 different samples. For each sample it is indicated whether it came from a tumor biopsy or not. Numbers and descriptions for the different genes are also given.

This dataset is used in many different research papers on gene expression data. It can be used in two ways: you can treat the 62 samples as records in a high-dimensional space and can treat the genes as records with 62 attributes. Leukemia dataset is of the same type as the colon cancer dataset and can therefore be used for the same kind of experiments. In fact, most of the that use the colon cancer data also use the leukemia data

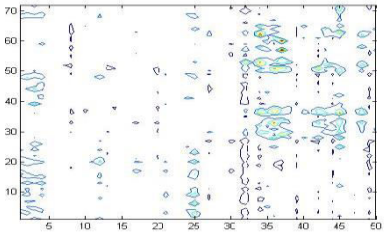
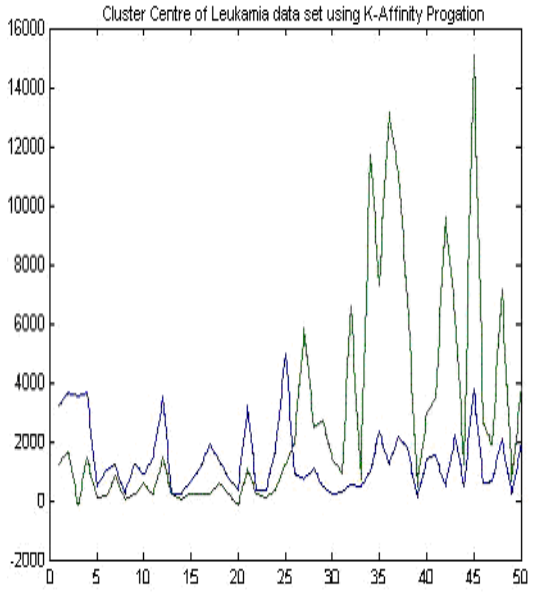


Fig 1. Contour mapping of Leukemia Dataset with 50-gene

Table 1. Result Over 50 Gene Leukemia

Result over different variations of k-means algorithm using 50-gene leukemia		
(Total number of record present in dataset=72)		
Clustering Algorithm	Correctly Classified	Average Accuracy
k-means	68	94.88
Global k-means	66	91.67
k-means++	66	95.83
Affinity propagation	69	96.43
Modified Affinity Propagation	71	98.23



Graph 1 Modified Affinity Propagation Algorithm For Leukemia Data Set

The way to initialize the means was not specified. One popular way to start is to randomly choose k of the samples.

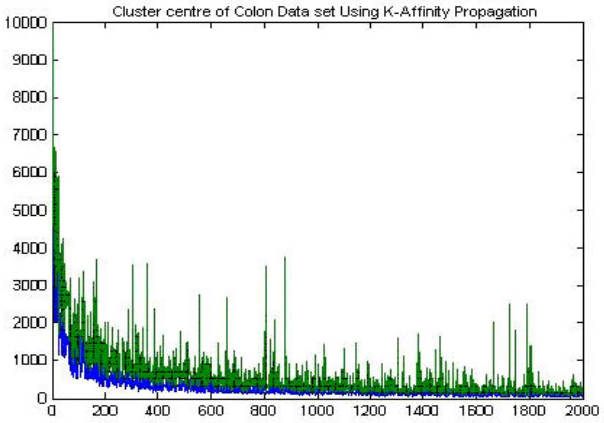
The results produced depend on the initial values for the means, and it frequently happens that suboptimal partitions are found. The standard solution is to try a number of different starting points.

It can happen that the set of samples closest to $\mathbf{m}_i$ is empty, so that $\mathbf{m}_i$ cannot be updated. This is an annoyance that must be handled in an implementation, but that we shall ignore.

The results depend on the metric used to measure $\| \mathbf{x} - \mathbf{m}_i \|$. A popular solution is to normalize each variable by its standard deviation, though this is not always desirable.

The results depend on the value of k.

This last problem is particularly troublesome, since the often have no way of knowing how many clusters exist. In the example shown above, the same algorithm applied to same data produces the 3-means clustering.



Graph 2. Modified Affinity Propagation For Colon Data

5. CONCLUSION

The k-means family of Cluster Algorithm's comparison performance evaluated in this thesis work shows that accuracy of these algorithms is good for the leukemia data set and slightly less for colon dataset. However performance of affinity propagation and modified affinity propagation is comparable. Modified affinity propagation has several advantages over related techniques. Methods that is, k-centers clustering, k-means clustering, and the global k-means algorithm store a relatively small set of estimated cluster centers at each step. These techniques are improved upon by methods that begin with a large number of clusters and then prune them, but they still rely on random sampling and make hard pruning decisions that cannot be recovered.

Clustering algorithm is to divide the dataset into disjoin clusters and modified affinity propagation algorithm to clustering data. As far as convergence rate is concerned, and observe that convergence rate of Affinity propagation is higher than all other variants of k-means. Modified affinity propagation is able to avoid many of the poor solutions caused by unlucky initializations and hard decisions. Resultant accuracy and convergence rate compared to that of k-means, k-means++, global k-means and affinity propagation algorithm. Modified affinity propagation algorithm found clusters with much lower error compared with other methods. The limitation of the work implemented and tested is given for future enhancement the future enhancement of the proposed algorithms i.e., modified fuzzy c means, cluster algorithms can be improved further with the help of fuzzy logic and rough set theory. To get better quality of clusters we can use these concepts. In case of k-means initial selection, cluster

centers play a very important role. So we will work on the possibility to improve these algorithms by using some good initial selection technique and fuzzy logics to achieve better results in tumor classification.

REFERENCE

[1] "Anjan Goswami. *Department of Computer Science and Engineering" Fast and Exact Out of-Core and Distributed K-Means Clustering 2001*

[2] A. Ben-Dor, R. Shamir, and Z. Yakhini, *"Clustering gene expression patterns" in J Comput Biol 6(3-4):281-97.*

[3] Alizadeh A., Eisen M.B, Davis R.E, et al. *Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. Nature. 2000; 403(6769):503–511*

[4] Arun. K. Pujari, *"Data Mining Techniques", Universities press (India) Limited 2001, ISBN81-7371-3804.*

[5] Bagirov, A.M.[Adil M.], *Modified global k-means algorithm for minimum sum-of-squares clustering problems, October 2008, pp. 3192-3199.*

[7] Bingham E, Mannila H: *Random projection in dimensionality reduction: applications to image and text data. Knowledge Discovery and Data Mining 2001:245-250*

[8] Bloisi, D.D.[Domenico Daniele], Iocchi, L.[Luca], Rek-Means: *A k-Means Based Clustering Algorithm,Springer DOI Link*

[9] Bouguessa, M.[Mohamed], Wang, S.R.[Sheng-Rui], Jiang, Q.S.[Qing-Shan], *A K-means-based Algorithm for Projective Clustering*

[10] Brendan J. Frey and Delbert Dueck. *Clustering by Passing Messages between Data Points. Science 315, 972 (2007).*