IJRSET 2018 Volume 5, Special-Issue-2 Pages: 21-25
ISSN 2394-739X
International Journal for Research in Science Engineering and Technology

AN ANALYSIS OF LEUKEMIA DATA SET FOR VARIOUS CLUSTERING ALGORITHMS

[1] Mrs. R.Kiruthika,
[1] Assistant Professor,
[1] Department of computer Science,
[1] Rathinam College of Arts & Science,
[1] Coimbatore, Tamil Nadu, India.

ABSTRACT: The data mining process is to extract information from large database, and it is non-trivial process of identifying valid, novel, potential useful and understandable pattern in data. The data mining task is using two major categories of predictive and descriptive tasks. Data mining involves the outlier detection, classification, clustering, regression and summarization. The clustering is the most important technique in data mining, which divides data into groups of similar object .Each groups (= cluster) consist of object that are similar among themselves. A wide range of clustering algorithms is available in literature and still an open area for researcher. Here in my paper i make analysis of clustering based algorithms namely k-means, k-means++ and x-means and Affinity propagation over gene leukemia dataset.

Keyword: [Data mining, clustering, k-means,x-means,k-means++,Affinity propagation**]**

1. INTRODUCTION

The Data mining is the process of discovering useful information (i.e. patterns) underlying the data. Powerful techniques are needed to extract patterns from large data because traditional statistical tools are not efficient enough anymore. Clustering is an important data mining technique that puts together similar objects into a collection in which the objects exhibit certain degree of similarities. Clustering also separates dissimilar objects into different groups. This has made clustering an important research topic of diverse fields such as pattern recognition, bioinformatics and data mining. It has been applied in many fields of study, from ancient Greek astronomy to present-day insurance industry and medical.

To classify the various types of cancer into its different subcategories, different data mining techniques have been used over gene expression data. A common aim is to use the gene expression profiles to identify groups of genes. A common aim is to use the gene expression profiles to identify groups of genes or samples in which the members behave in similar ways. One might want to partition the data set to find naturally occurring groups of genes with similar expression patterns. Golub et al (Golub, 1999), Alizadeh et al (Alizadeh,2000), Bittner et al (Bittner,2000) and Nielsen et al (Nielsen,2002) have considered the classification of cancer types using gene expression datasets. There are many instances of reportedly successful applications of both hierarchical clustering and partitioning clustering in gene expression

analyses. Yeung et al (Yeung,2001) compared *k*-means clustering. In this paper, we make a comparative analysis of various clusteringbased algorithms like x-means, x-means, k-means++, and Affinity propagation over gene- leukemia dataset. Comparison is made in respect of accuracy and convergence rate.

2. CLUSTERING

Clustering is the task of segmenting a heterogeneous population into a number of more homogeneous sub groups or clusters. Clustering and clusters are not synonymous. A clustering is an entire collection of clusters; a cluster on the other hand is just one part of the entire picture. Clustering is a division of data into group of similar objects. Each group, called cluster consist of objects that are similar amongst themselves and dissimilar compared to objects of other groups. (9)

3. CLUSTER ANALYSIS

The process of grouping a set of physical or abstract object into classes of similar objects arecalled clustering. A cluster is a collection of data object that are similar to one another within the same cluster and are dissimilar to the object in other clusters.Cluster analysis is an important human activity. One learns how to distinguish between cats and dogs, or between animals or plants, by continuously improving subconscious clustering schemes. Cluster analysis has been widely used in numerous applications, that including pattern recognition, data analysis, image processing, and market research by clustering.

4. CLUSTERING ALGORITHMS

There are many types of algorithms are used in clustering. That is given below.

4.1 K-MEANS Algorithms:

The k-means algorithm (MacQueen, 1967) is one of a group of algorithms called Partitioning methods. The k -means algorithm is very simple and can be easily implemented in solving many practical problems. The k-means algorithm is the best-known squared error-based clustering algorithm. Consider the data set with 'n' objects,

i.e.,$S = \{xi : 1 \le i \le n\}$.

1) Initialize a k-partition randomly or based on some prior knowledge.

i.e. { C1, C2 , C3 ,…….., Ck }.

2) Calculate the cluster prototype matrix M (distance matrix of distances between k-clusters and data objects).

M = { m1 , m2 , m3, …….. , mk } where mi is a column matrix 1× n.

3) Assign each object in the data set to the nearest cluster - Cm i.e.

$xj \in Cm$ if $\| x j - Cm \| \le \| x j - Ci \| \; \forall \; 1 \le j \le k$, $j \ne m$ where j=1,2,3,…….n.

4) Calculate the average of each cluster and change the k-cluster centers by their averages.

5) Again calculate the cluster prototype matrix M.

6) Repeat steps 3, 4 and 5 until there is no change for each cluster.

4.2 X_MEANS ALGORITHM:

X-means algorithm (Dan Pelleg and Andre Moore, 2000) searches the space of cluster locations and number of clusters efficiently to optimize the Bayesian Information Criterion (BIC) or The Akaike Information Criterion (AIC) measure. The kd-tree technique is used to improve the speed for the algorithm. In this algorithm, numbers of clusters are computed dynamically using lower and upper bound supplied by the user.

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.

Step3: if k> =kmax(upper bound) stop and report to best scoring model found during search otherwise go to step 1.

4.3 K-MEANS++ ALGORITHM:

k-means++ (David Arthur et. Al., 2007) 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 c1 randomly from the given dataset X .

Step 2: choose next cluster center ci =xj ∈ X with probability pi where pj $= \frac{D(Xj)^2}{\sum_{x\epsilon X} D(X)^2}$,

D(X) denote the shortest distance from x to the closest center already chosen.

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.

4.4. AFFINITYPROPOGATION ALGORITHM

Affinity propagation (AP) can be viewed as a method that searches for minima of an energy function

$E(C)= -\sum^{N} S(I,cj)\ s(I,cj) \leq 0\ I=1$

Each label ci indicates the exemplar of the data point i, while s(i,ci) is the similarity between data point i and its exemplar ci.

For ci = i, s(i, ci) is the input preference for data point i indicating how suitable data point i can be the exemplar. In most cases, the statistical and geometrical structure of a data set is unknown so that it is reasonable to set all the preference value the same. The bigger this shared value is, the larger the number of clusters is. Throughout the following of this paper, the preferences are set to the same value if not mentioned. The process of AP can be viewed as a message communication process with two kinds of messages exchanged among data points, named responsibility and availability. The algorithmic is stated below:[8]

Input: s(i, k): the similarity of point i to point k.

p(j): the preferences array which indicates the preference that data point j is chosen as a cluster center.

Output:

idx(j): the index of the cluster center for data point j.

dpsim: the sum of the similarities of the data points to theirclustercenters.

netsim: the net similarity (sum of the data point similarities

and preferences).

expref: the sum of the preferences of the identified cluster centers netsim: the net similarity (sum of the data point data point similarities and preference)

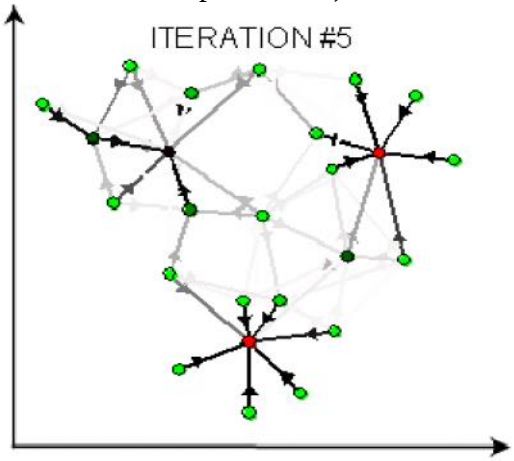

Figure.3 Iteration affinity propagation

Step1: Initialization the availability a(i.k) to zero a(i,k)=(1)

Step2: Update the responsibility using rule

r(i,k)← s(i ,k) – max {a(i, k'). s(i, k')}. (2)k's.t. k' ≠ k

Step3: Update the availability using the rule

a (i , k) ←min{0, r (k ,k)∑ max{0,r(i',k)}} i' s.t. i'≠ i ,k (3)

The self-availability is updated differently

a (k , k) → ∑ max{0, r (i' , k)}. (4)i' s.t. i'≠ k

Step 4: The message-passing procedure may be terminated after a fixed number of iterations, after changes in the messages fall below a threshold or after the local decisions stay constant for some number of iterations.

Availabilities and responsibilities can be combined to make the exemplar decisions. For point i, the value of k that maximizes a(i, k)+r(i, k) either identifies point i as an exemplar if k=i or identifies the data point that is the exemplar for point i. When updating the messages, numerical Oscillations must be taken into consideration. As a result, each message is set to λ times its value from the previous iteration plus 1- λ times its prescribed updated value. The λ should be larger than or equal to 0.5 and less than 1. If λ is very large, numerical oscillation may be avoided, but this is not guaranteed. Hence a maximal number of iterations are set to avoid infinite iteration in AP clustering.

5. LEUKEMIA DATASETS

We used the cancer data sets to make a study of various clustering based algorithms. The Leukemia data set is a collection of gene expression measurements from 72 leukemia (composed of 62 bone marrow and 10 peripheral blood) samples reported 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 we take two variants of leukemia dataset one with 50-genes.

6. RESULT OVER LEUKEMIA DATASET

The analysis of different variants of clustering algorithm is done with the help of the cancer dataset (Leukemia). Variants of clustering used in this study are k-means, x-means, k-means++ and affinity propagation. The Average accuracy rate of these variants of k-means are shown below in table.

Clustering Algorithm	Correctly Classified	Average Accuracy
K-Means	68	94.88
X-Means	66	91.67
Affinity Propagation	69	95.15
K-Means++	69	95.83

Table 1: Result over different variations of k-means algorithm using 50-gene leukemia.

(Total number of record present in dataset=72)

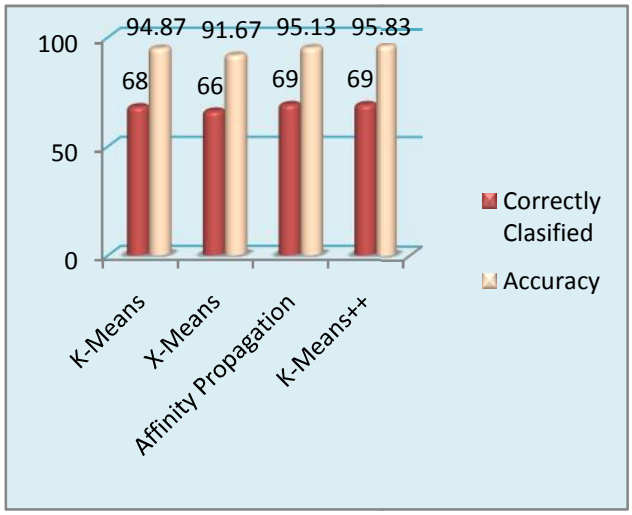

Figure: Comparison of Clustering Algorithms.

CONCLUSION AND FUTURE WORK

The analysis of Clustering Algorithms are done with the help of leukemia Data set. The average Accuracy is shown that the performance of K-means++ and Affinity propagation is slightly better in 50 gene leukemia dataset, on clustering execution time and convergence rate and found much low error when compare with k-means.

Performance of this algorithm can be improved with the help of variants 3859 gene leukemia using efficient k-means, fuzzy logic to get better quality of cluster. So these algorithm help to get Good Result.

REFERENCES

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

[2]. 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.

[3]. Dan Pelleg and Andrew Moore: X-means: Extending K-means with Efficient Estimation of the Number of Clusters, ICML 2000.

[4]. David Arthur and Sergei Vassilvitskii: k-means++:The advantages of careful seeding, Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms. pp. 1027—1035, 2007.
[5]. Gibbons F.D, Roth F.P. Judging the quality of gene expression-based clustering methods using gene annotation. Genome Res. 2002;12(10):1574–1581.
[6]. Golub T.R, Slonim D.K, and Tamayo P, et al. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. Science. 1999;286(5439):531–537.
[7]. L. Kaufman and P. J. Rousseeuw, Finding Groups in Data: an Introduction to Cluster Analysis, John Wiley & Sons, 1990.
[8]. MacQueen, J.B. (1967). Some Methods for Classification and Analysis of Multivariate Observations. In Proc. of 5th Berkley Symposium on Mathematical Statistics and Probability, Volume I: Statistics, pp. 281–297
[9]. Nielsen T.O, West R.B, Linn S.C, et al. Molecular characterisation of soft tissue tumours: a gene expression study. Lancet2002.
[10]. Yeung K.Y, Haynor D.R, Ruzzo W.L. Validating clustering for gene expression data. Bioinformatics. 2001.
[11]. ParveshKumar, Sirikrishnan wasan Comparative Analysis of k-mean Based Algorithms, IJCSNS International Journal of Computer Science and Network Security, VOL.10 No.4, April 2010 314.
[12]. Greg Hamerly "Making k-means evenfaster"2010academic.research.microsoft
[13]. Federico Ambrogi, Elena Raimondi, Daniele Soria, PatriziaBoracchi and Elia Biganzoli1 Cancer profiles by Affinity Propagation University of Nottingham, School of Computer Science, Jubilee Campus, Wollaton Road, Nottingham, NG8 1BB,
[14]. Jinze Liu, Jiong Yang and Wei Wang, "Biclustering in Gene Expression Data by Tendency". Paul Bunn, RafailOstrovsky "Secure Two-Party k-Means Clustering" 2007.
[15]. Margaret H.Dunham,Sridharz "Data Mining Introductory and Advanced Topics" Dorling Kindersley (India) Pvt. Ltd., 2006.
[16]. Lai, J.Z.C.[Jim Z.C.], Liaw, Y.C.[Yi-Ching], Improvement of the k-means clustering filtering algorithm, PR(41),No.12, December2008.
[17]. Hui Li, Sourav S. Bhowmick, Aixin Sun, Blog Cascade Affinity: Analysis and Prediction portal.acm.org/ft_gateway.
[18]. Federico Ambrogi, Elena Raimondi, Daniele Soria, PatriziaBoracchi and Elia Biganzoli1 Cancer profiles by Affinity Propagation University of Nottingham, School of Computer Science, Jubilee Campus, Wollaton Road, Nottingham, NG8 1BB.
[19]. Wu Jiang, Fei Dingy, Qiao-Liang Xiang An Affinity Propagation Based Method for Vector Quantization Codebook Design College of Computer and Information Science, North-eastern University.
[20]. Zhenjie Zhang, Bing Tian Dai, Anthony K.H. Tung On the Lower Bound of Local Optimums in K-Means Algorithm2003.