International Journal of Engineering and Advanced Technology (IJEAT)
ISSN: 2249-8958 (Online), Volume-1 Issue-6, August 2012

K- Means Algorithm with Different Measurements in Clustering Approach

P.C.Chenna Reddy, R. Siva Sankara Reddy

Abstract: ***Clustering techniques have been used by many intelligent software agents in order tretrieve, filter and categorize document available on the World Wide Web .Clustering isalso useful in extracting salient features of related web documents to automatic ally formulate queries and search for other similar documents on the Web. In this paper, we introduce two new clustering algorithm withs K-Means Clustering in GeneLinker™ that can effectively cluster documents, even in the presence of a veryhighdimensional feature space.These clusteringtechniques, which are based on generalizations of graphpartitioning, do not require prespecifiedad hoc distance functions, and are capable of automatically discovering document similarities or associations. We conduct several experiment son real Web data using various feature selection and find out the no off clusters in the data documenting this paper also discuss about the real example. In this example we are find out the no.Of clusters.***

Keywords: clustering, categorization, World Wide Web documents, K-means alogrithmg, Genelinker ™

I. INTRODUTION:

K-Means clustering generates a specific number of disjoint, flat (non-hierarchical) clusters. It is well suited to generating globular clusters. The K-Means method is numerical, unsupervised, non-deterministic and iterative.

1.1 Definition:

This nonheirarchial method initially takes the number of components of the population equal to the final required number of clusters. In this step itself the final required number of clusters is chosen such that the points are mutually farthest apart. Next, it examines each component in the population and assigns it to one of the clusters depending on the minimum distance. The centroid's position is recalculated everytime a component is added to the cluster and this continues until all the components are grouped into the final required number of clusters.

Example:Let us apply the k-Means clustering algorithm to the same example as in the previous page and obtain four clusters :

Food item #	**Protein content, P**	**Fat content, F**
Food item #1	1.1	60
Food item #2	8.2	20
Food item #3	4.2	35
Food item #4	1.5	21
Food item #5	7.6	15
Food item #6	2.0	55
Food item #7	3.9	39

us plot these points so that we can have better understanding of the problem. Also, we can select the three points which are farthest apart.

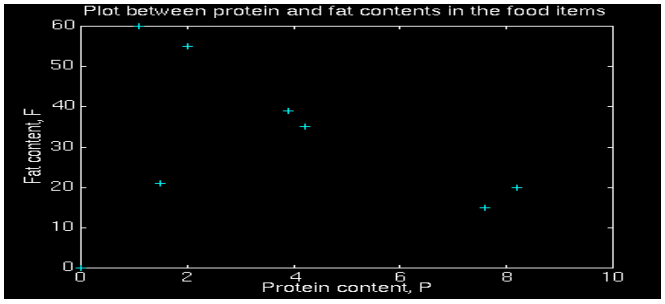

Fig:Plot between protein and fat contents in the food items

We see from the graph that the distance between the points 1 and 2, 1 and 3, 1 and 4, 1 and 5, 2 and 3, 2 and 4, 3 and 4 is maximu

Thus, the four clusters chosen are :

Cluster number	**Protein content, P**	**Fat content, F**
C1	1.1	60
C2	8.2	20
C3	4.2	35
C4	1.5	21

Also, we observe that point 1 is close to point 6. So, both can be taken as one cluster. The resulting cluster is called C16 cluster.

The value of **P** for C16 centroid is (1.1 + 2.0)/2 = 1.55 and **F** for C16 centroid is (60 + 55)/2 = 57.50.

Upon closer observation, the point 2 can be merged with the C5 cluster. The resulting cluster is called C25 cluster.

The values of **P** for C25 centroid is (8.2 + 7.6)/2 = 7.9 and **F** for C25 centroid is (20 + 15)/2 = 17.50

The point 3 is close to point 7. They can be merged into C37 cluster.

The values of **P** for C37 centroid is (4.2 + 3.9)/2 = 4.05 and **F** for C37 centroid is (35 + 39)/2 = 37.

The point 4 is not close to any point. So, it is assigned to cluster number 4 i.e., C4 with the value of **P** for C4 centroid as 1.5 and **F** for C4 centroid is 21.

Finally, four clusters with three centroids have been obtained.

Manuscript published on 30 August 2012.
***** Correspondence Author (s)
P.C.Chenna Reddy*, HOD,Asso.Professor,Department of CSE, JNTU College of engineering,Pulivendula (A.P.), India.
R. Siva Sankara Reddy, S/W Engineer,TCS,Hydarabad (A.P.), India.

Cluster number	Protein content, P	Fat content, F
C16	1.55	57.50
C25	7.9	17.5
C37	4.05	37
C4	1.5	21

In the above example it was quite easy to estimate the distance between the points. In cases in which it is more difficult to estimate the distance, one has to use *euclidean metric* to measure the distance between two points to assign a point to a cluster.

II. K-MEANS CLUSTERING ALGORITHM:

2.1Data Mining with the k-means Algorithm:

k-means clustering is a data mining/machine learning algorithm used to cluster observations into groups of related observations without any prior knowledge of those relationships. The k-means algorithm is one of the simplest clustering techniques and it is commonly used in medical imaging, biometrics and related fields.

2.2 K-Means Algorithm Properties:

- There are always K clusters.
- There is always at least one item in each cluster.
- The clusters are non-hierarchical and they do not overlap.
- Every member of a cluster is closer to its cluster than any other cluster because closeness does not always involve the 'center' of clusters.

2.3The K-Means Algorithm Process:

- The dataset is partitioned into K clusters and the data points are randomly assigned to the clusters resulting in clusters that have roughly the same number of data points.
- For each data point:
  - Calculate the distance from the data point to each cluster.
  - If the data point is closest to its own cluster, leave it where it is. If the data point is not closest to its own cluster, move it into the closest cluster.
  - Repeat the above step until a complete pass through all thedata points results in no data point moving from one cluster to another. At this point the clusters are stable and the clustering process ends.
  - The choice of initial partition can greatly affect the final clusters that result, in terms of inter-cluster and intracluster distances and cohesion.

2.4 The k-means Algorithm:

The k-means algorithm is an evolutionary algorithm that gains its name from its method of operation. The algorithm clusters observations into k groups, where k is provided as an input parameter. It then assigns each observation to clusters based upon the observation's proximity to the mean of the cluster. The cluster's mean is then recomputed and the process begins again. Here's how the algorithm works:

1. The algorithm arbitrarily selects k points as the initial cluster centers ("means").
2. Each point in the dataset is assigned to the closed cluster, based upon the Euclidean distance between each point and each cluster center.
3. Each cluster center is recomputed as the average of the points in that cluster.
4. Steps 2 and 3 repeat until the clusters converge. Convergence may be defined differently depending upon the implementation, but it normally means that either no observations change clusters when steps 2 and 3 are repeated or that the changes do not make a material difference in the definition of the clusters.

III. K-MEANS CLUSTERING IN GENELINKER™:

The version of the K-Means algorithm used in GeneLinker™ differs from the conventional K-Means algorithm in that GeneLinker™ does not compute the centroid of the clusters to measure the distance from a data point to a cluster. Instead, the algorithm uses a specified linkage distance metric. The use of the Average Linkage distance metric most closely corresponds to conventional K-Means, but can produce different results in many cases.

3.1 Distance Measurements Between Data Points:

This parameter specifies how the distance between data points in the clustering input is measured. The options are:

- **Euclidean:** Use the standard Euclidean (as-the-crow-flies) distance.
- **Euclidean Squared:** Use the Euclidean squared distance in cases where you would use regular Euclidean distance in Jarvis-Patrick or K-Means clustering.
- **Manhattan:** Use the Manhattan (city-block) distance.
- **Pearson Correlation:** Use the Pearson Correlation coefficient to cluster together genes or samples with similar behavior; genes or samples with opposite behavior are assigned to different clusters.
- **Pearson Squared:** Use the squared Pearson Correlation coefficient to cluster together genes with similar or opposite behaviors (i.e. genes that are highly correlated and those that are highly anti-correlated are clustered together).
- **Chebychev:** Use Chebychev distance to cluster together genes that do not show dramatic expression differences in any samples; genes with a large expression difference in at least one sample are assigned to different clusters.
- **Spearman:** Use Spearman Correlation to cluster together genes whose expression profiles have similar shapes or show similar general trends (e.g. increasing expression with time), but whose expression levels may be very different.

3.2Distance Measurements Between Clusters:

This parameter specifies how the distance between clusters is measured. The options are:

- **Average Linkage**: The distance between two clusters is the average of the distances between all the points in those clusters.
- **Single Linkage**: The distance between two clusters is the distance between the nearest neighbors in those clusters.
- **Complete Linkage:** The distance between two clusters is the distance between the furthest points in those clusters.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)

3.3. Advantages to Using this Technique:

- With a large number of variables, K-Means may be computationally faster than hierarchical clustering (if K is small).
- K-Means may produce tighter clusters than hierarchical clustering, especially if the clusters are globular.

3.4. Disadvantages to Using this Technique:

- Difficulty in comparing quality of the clusters produced (e.g. for different initial partitions or values of K affect outcome).
- Fixed number of clusters can make it difficult to predict what K should be.
- Does not work well with non-globular clusters.
- Different initial partitions can result in different final clusters. It is helpful to rerun the program using the same as well as different K values, to compare the results achieved

IV. CHOOSING THE NUMBER OF CLUSTERS:

One of the main disadvantages to k-means is the fact that you must specify the number of clusters as an input to the algorithm. As designed, the algorithm is not capable of determining the appropriate number of clusters and depends upon the user to identify this in advance. For example, if you had a group of people that were easily clustered based upon gender, calling the k-means algorithm with k=3 would force the people into three clusters, when k=2 would provide a more natural fit. Similarly, if a group of individuals were easily clustered based upon home state and you called the k-means algorithm with k=20, the results might be too generalized to beeffective.

V. CONCLUSION:

In this paper mainly discuss about the clustering mechanism,how to apply these technique into non graphical method.In this paper mianly focus on the K-means algorithm with clustering mechanism.In this paper focus on the ao many clusters and different measurements.In this paper we can also expalin the one example,these example expalins the no of clusters and measurements.

REFERENCES:

[1]. Agrawal, H. Mannila, R. Srikant, H.Toivonen,and A.I.Verkamo Fast discovery of association rules.In U.M. Fayyad, G.Piatetsk.Shapiro, P.Smith, and R. Uthurusamy,editors,Advances in Knowledge Discovery and Data Mining. pages 307-328.1996.
[2]. Mark Achreman and ETet al. Learningprobalistic user profiles.AI Magazine.18 (2):47-56,1997.
[3]. D.L. Boley. Principal Direction Divisive Partitioning. Technical Report TR, Department of Computer Science. University of Minnesota.Minneapolis.1998
[4]. Andrei Z.Broder, Steven C, Glassman,and Mark S. Manasse..Syntactic clustering of the Web. In Proc of the International World WideWeb Conference. April.1983
[5]. J.Moore,E.han,D.Boley,M.Gini,R.Gross,K.Hastings,V.Kumar.Web apge categorization and feature selection usiong association rule and prinicipai component clustering. In & thwork shop on Information Technologies and Systems.Dec 1997.

Published By:
Blue Eyes Intelligence Engineering and Sciences Publication (BEIESP)