Introduction falls into the general category of variance-based

INTRODUCTIONCLUSTERING problems arise nowadays in many differentapplications, such as data mining and knowledge discovery, data compression andvector quantization, and pattern recognition and pattern classification.

Thenotion of what constitutes a good cluster depends on the application and thereare many methods for finding clusters subject to various criteria, both adhocand systematic. These include approaches based on splitting and merging such asISODATA randomized approaches such as CLARA, CLARANS, methods based on neuralnets , and methods designed to scale to large databases, including DBSCAN, BIRCH, etc. Surrounded by clustering formulations that are dependon minimizing a proper objective function, possibly the most widely used andstudied is k-means clustering. Given a set of n data points in reald-dimensional space, Rd, and an integer k, the problem is to determine a set ofk points in Rd, called centers, so as to minimize the mean squared distancefrom each data point to its nearest centre. This measure is often called thesquared-error distortion and this type of clustering falls into the generalcategory of variance-based clustering. Clustering based on k-means is closely related to anumber of other clustering and location problems.

These include the Euclideank-medians in which the objective is to minimize the sum of distances to thenearest centre and the geometric k-centre problem in which the objective is tominimize the maximum distance from every point to its closest center. There areno efficient solutions known to any of these problems and some formulations areNP-hard. k-means is one of the simplestunsupervised learning algorithms that solve the well known

clustering problem. The procedure follows a simpleand easy way to classify a given data set througha certain number of clusters (assume k clusters) fixed apriori.

The main idea is to define k centers, one for eachcluster. These centers should be placed in acunning way becauseof different location causes different result. So, the better choice is to place them as much as possible far away from each other. The next step is to take each point belonging toa given data set and associate it to the nearest center. When nopoint is pending, the first step is completed and an early groupage is done.

At this point we need to re-calculate k new centroidsas barycenter of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to bedone between the same data set points and the nearestnew center. A loop has been generated. As a result of this loopwe may notice that the k centers change their location step bystep until no more changes are done or in other wordscenters do not move any more. Finally, this algorithm aims at minimizing an objective function know as squared error function givenby: '|| xi -vj||' is the Euclidean distance between xi and vj. ' ci' is the number of datapoints in ith cluster. ' c' is the number of cluster centers. Algorithmic stepsfor k-means clusteringLet X = {x1, x2, x3,.....

., xn}be the set of data points and V = {v1, v2,....., vc}be the set of centers.

1) Randomly select ‘ c’ clustercenters. 2) Calculate the distance betweeneach data point and cluster centers. 3) Assign the data point to thecluster center whose distance from the cluster center is minimum of all thecluster centers..

4) Recalculate the new cluster centerusing: 5) Recalculate the distance betweeneach data point and new obtained cluster centers. 6) If no data point was reassignedthen stop, otherwise repeat from step 3. For Example- Suppose that we have n sample feature vectors x1, x2,...

, xn all from the same class, and we know thatthey fall into k compact clusters, k < n. Let mi bethe mean of the vectors in cluster i. If the clusters are well separated, wecan use a minimum-distance classifier to separate them. That is, we can saythat x is in cluster i if || x - mi || is the minimum of all the k distances.

This suggests the following procedurefor finding the k means: Make initial guesses for the means m1, m2, ..., mk Until there are no changes in any mean Use the estimated means to classify the samples into clusters For i from 1 to k Replace mi with the mean of all of the samples for cluster i end_for end_untilHere isan example showing how the means m1 and m2 moveinto the centers of two clustersAdvantages1) Fast, robust and easier to understand. 2) Relatively efficient: O(tknd), where n is # objects, k is # clusters, d is # dimension ofeach object, and t is # iterations. Normally, k, t, d << n. 3) Gives best result when data set aredistinct or well separated from each other.

Note: For more detailed figure for k-means algorithm please referto k-means figure subpage. Disadvantages1) The learningalgorithm requires apriori specification of the number of cluster centers. 2) The use of Exclusive Assignment -If there are two highly overlapping data then k-means will notbe able to resolve that there are twoclusters. 3) Thelearning algorithm is not invariant to non-linear transformations i.

e. withdifferent representation of data we get different results (data represented in form of cartesian co-ordinates and polarco-ordinates will give different results). 4) Euclidean distance measures canunequally weight underlying factors. 5) The learningalgorithm provides the local optima of the squared error function. 6) Randomlychoosing of the cluster center cannot lead us to the fruitful result. 7) Applicable onlywhen mean is defined i.

e. fails for categorical data. 8) Unable tohandle noisy data and outliers. 9) Algorithm failsfor non-linear data set.