Gorti Satyanarayana Murty, Dr. V. Vijaya Kumar, Tangudu Naresh / International Journal of Engineering Research and Applications (IJERA) ISSN: 2248-9622 www.ijera.com
Vol. 2, Issue 2,Mar-Apr 2012, pp.041-045

Learning Number of Clusters in Unlabeled Dataset using Rotation Estimation

Gorti Satyanarayana Murty*, Dr. V. Vijaya Kumar, Tangudu Naresh*****

*(Assoc.Professor, AITAM, Tekkali, A.P, India-532201)
**(Dean-computer Sciences & Head-SRRF, GIET,Rajahmundry,A.P,India)
***(Asst.professor ,AITAM, Tekkali,A.P,India-532201)

Abstract:

Most of the Clustering algorithms in Cluster Analysis, partition a dataset into a fixed number of clusters supplied by the user manually i.e. learning by Observation [1]. The estimation of number of clusters for partitioning the dataset is difficult in the case of large datasets, which leads to inefficient data distribution or majority outliers. Hence, in this paper we propose a novel method using rotation estimation also called Cross-Validation [2,3,4] which identifies a suitable number of clusters in a given unlabeled dataset without using prior knowledge about the number of clusters. In this paper, the k-means and Expectation Maximization(EM) Clustering Techniques are optimized and enhanced for typical applications in Data Mining.

Key *Words*: Cluster analysis, unlabeled dataset, rotation estimation,cross-validation,unsupervised learning.

I. Introduction

Clustering may be defined as the optimal partitioning of a given set of *n* data points into *c* subgroups, such that data points belonging to the same group are as similar to each other as possible, whereas data points from two different groups share the maximum difference [5,6].

Most of the unsupervised clustering algorithms like K-means or EM algorithms, however, assume a prior knowledge of the number of clusters, while in practical situations, the appropriate number of clusters may be unknown or impossible to determine even approximately. Finding an optimal number of clusters in a large dataset is usually a challenging task and for which here we propose optimal solution using rotation estimation also called as Cross- validation is a technique for assessing how the results of a statistical analysis will generalize to an independent data set. It is mainly used in settings where the goal is prediction, and one wants to estimate how accurately a predictive model will perform in practice. One round of cross-validation involves partitioning a sample of data into complementary subsets, performing the analysis on one subset (called the *training set*), and validating the analysis on the other subset (called the *validation set* or *testing set*).To reduce variability, multiple rounds(ten rounds) of cross-validation are performed using different partitions, and the validation results are averaged over the rounds.

The remainder of the paper is organized as follows. Section-2 briefly reviews commonly used clustering techniques like k- means and EM algorithms. In Section-3, we describe our cross-validation technique for data clustering. Section-4 provides the experimental results from the proposed model and compared with other clustering techniques using benchmark datasets. Finally, we conclude this paper with a discussion of the model and its implications in Section-5**.**

II. Background

In this section, we briefly review the well known unsupervised clustering techniques reported in the literature. A more comprehensive review can be found in [6]. The K-means algorithm is one of the oldest unsupervised algorithm [7]. The idea is to group data into k-clusters (known priori) using k-centroids (one for each cluster). The performance of clusters thus obtained depends on the initial centroid values. The aim of this algorithm is to minimize the Euclidean distance between the data points and the corresponding cluster centroid, which is achieved by minimizing the objective function:

$$\sum_{j=1}^{k}\sum_{i=1}^{n}\left\|x_i^{(j)}-c_j\right\| \quad -----(1)$$

where $x_i^{(j)}$ is the data point and c_j is the j^{th} cluster center.

An inherent assumption built into the k-means algorithm is that the data points are independent. Consequently, there is degradation in the algorithm effectiveness (accuracy)when the data points are highly dependent on each other. The k-means algorithm is not able to find the optimal configuration compared with the global objective function minimum.

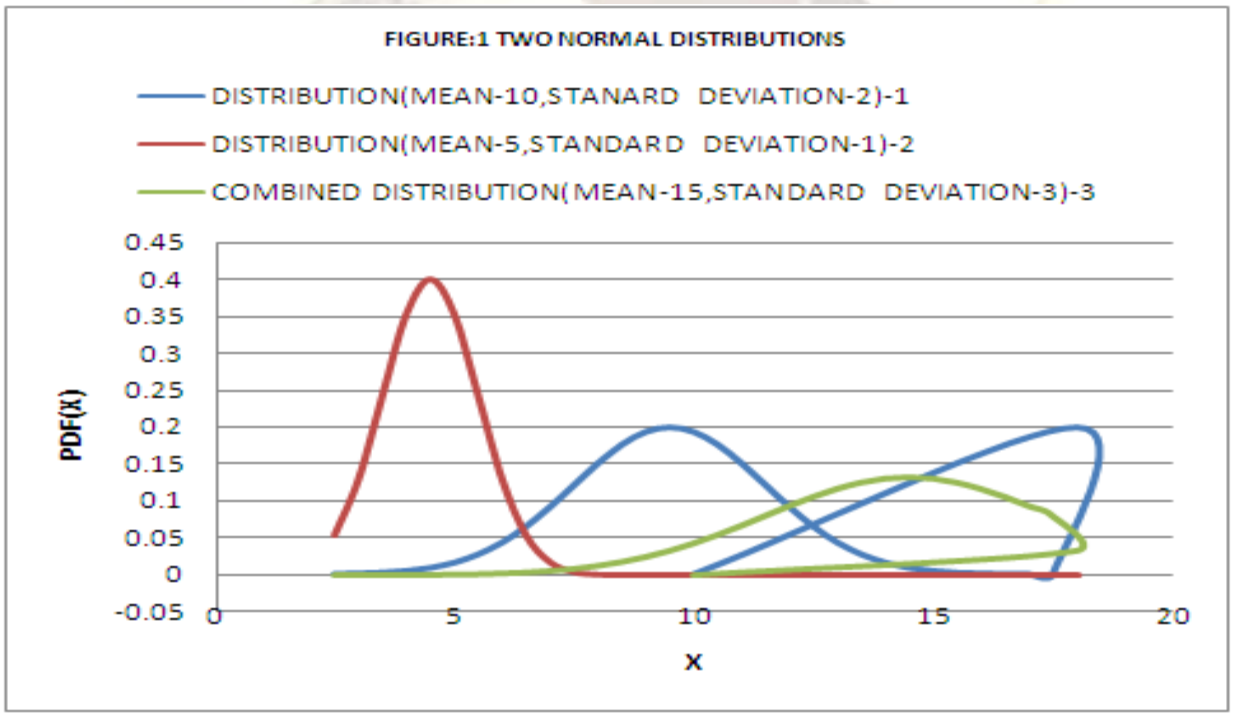
A. Extensions and Generalizations. The EM (expectation maximization) algorithm extends the basic approach of k-means to clustering in two important ways:

1. Instead of assigning cases or observations to clusters to maximize the differences in means for continuous variables, the EM clustering algorithm computes probabilities of cluster memberships based on one or more probability distributions. The goal of the clustering algorithm then is to maximize the overall probability or likelihood of the data, given the (final) clusters.

2. Unlike the classic implementation of *k*-means clustering, the general *EM* algorithm can be applied to both continuous and categorical variables (note that the classic *k*-means algorithm can also be modified to accommodate categorical variables).

The EM Algorithm [8,9]

The *EM* algorithm for clustering is described in detail in Witten and Frank (2001). The basic approach and logic of this clustering method is as follows. Suppose you measure a single continuous variable in a large sample of observations. Further, suppose that the sample consists of two clusters of observations with different means (and perhaps different standard deviations) within each sample, the distribution of values for the continuous variable follows the normal distribution. The resulting distribution of values (in the population) may look like this: Mixtures of distributions. The illustration in Fig-1 shows two normal distributions with different means and different standard deviations, and the sum of the two distributions. Only the mixture



(sum) of the two normal distributions (with different means and standard deviations) would be observed. The goal of *EM* clustering is to estimate the means and standard deviations for each cluster so as to maximize the likelihood of the observed data distribution. Put another way, the EM algorithm attempts to approximate the observed distributions of values based on mixtures of different distributions in different clusters.

With the implementation of the EM algorithm in some computer programs, you may be able to select (for continuous variables) different distributions such as the normal, log-normal, and

Poisson distributions. You can select different distributions for different variables and, thus, derive clusters for mixtures of different types of distributions.

B. Categorical variables.

The EM *algorithm* can also accommodate categorical variables. The method will at first randomly assign different probabilities (weights, to be precise)to each class or category, for each cluster. In successive iterations, these probabilities are refined (adjusted) to maximize the likelihood of the data given the specified number of clusters.

C. Classification probabilities instead of classifications.

The results of *EM* clustering are different from those computed by K- means clustering. The latter will assign observations to clusters to maximize the distances between clusters. The *EM* algorithm does not compute actual assignments of observations to clusters, but classification *probabilities*. In other words, each observation belongs to each cluster with a certain probability. Of course, as a final result you can usually review an actual assignment of observations to clusters, based on the (largest) classification probability.

III. V-fold Cross-Validation Technique for Data Clustering

The v-fold cross-validation algorithm is described in some detail in Classification Trees [10] and General Classification and regression Trees (GC&RT) [8]. The general idea of this method is to divide the overall sample into a number of v *folds*. The same type of analysis is then successively applied to the observations belonging to the *v-1* folds (training sample), and the results of the analyses are applied to sample *v* (the sample or fold that was not used to estimate the parameters, build the tree, determine the clusters, etc.; this is the testing sample) to compute some index of predictive validity. The results for the *v* replications are aggregated (averaged) to yield a single measure of the stability of the respective model, i.e., the validity of the model for predicting new observations. Cross validation has been tested extensively and found to generally work well when sufficient data is available and the value of 10 for v has been found to be adequate and accurate[11].

Cluster analysis is an unsupervised learning technique, and we cannot observe the (real) number of clusters in the data. However, it is reasonable to replace the usual notion(applicable to supervised learning)of "accuracy "with that of "distance." In general, we can apply the v-fold cross- validation method to a range of numbers of clusters in *k*-means or *EM* clustering, and observe the resulting average distance of the observations (in the cross-validation or testing samples) from their cluster centers (for *k*-means clustering); for *EM* clustering, an appropriate equivalent measure would be the average negative (log-)likelihood computed for the observations in the testing samples.

The following Figure-2 & Figure-3 clearly shows how the v-fold validation is implemented for determining the number of clusters.

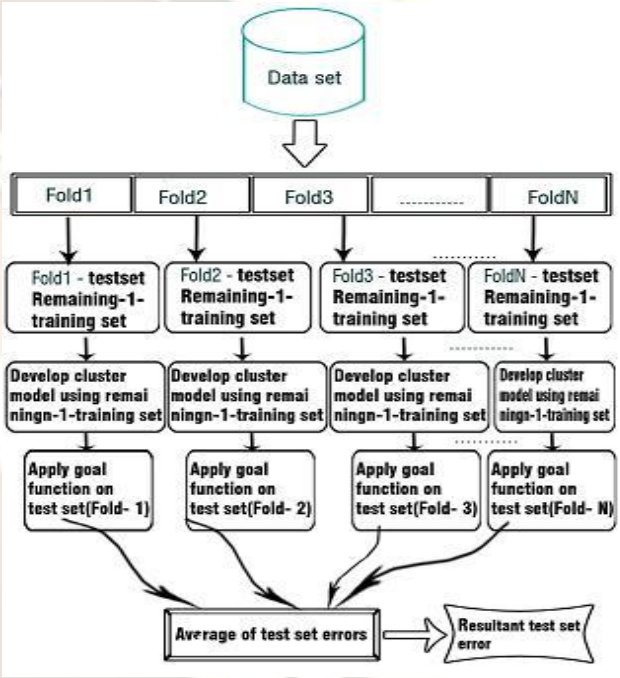


Fig-2 shows how to calculate test set error using goal function.

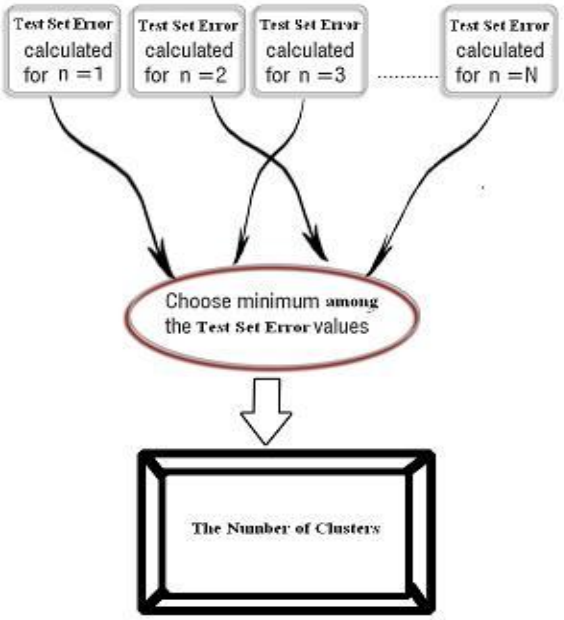


Fig-3 shows how to determine the number of clusters

IV. Experimental Results

Several artificially generated and real- life datasets were used to experimentally demonstrate that the Cross-Validation is able to find the proper cluster number for different types like continuous and categorical types of data sets.

The following Table-1 shows the results of identifying the number of clusters for k-means algorithm using the market analysis data set and Figure-3 & Figure-4 shows the distribution of data into clusters w.r.t AGE and ANNUALINC attributes

S.No	Name of the Relation	No of instances	No of Clusters	No of Attributes	Goal function (Min. Avg. SEED)
1	Marketing data set	6910	3	15	5.318099

Table-1: Results of Market Analysis DataSet for k-means

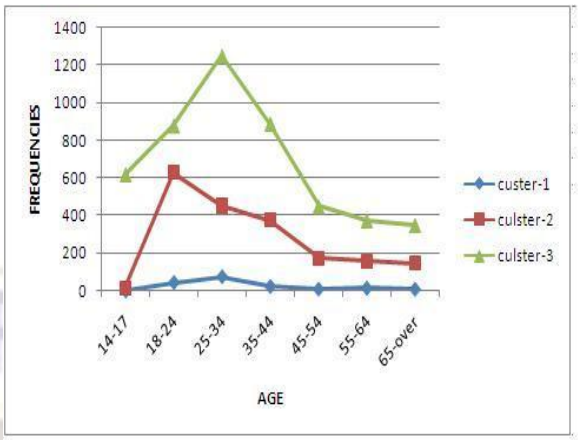


Fig-3: Graph of frequencies for AGE attribute

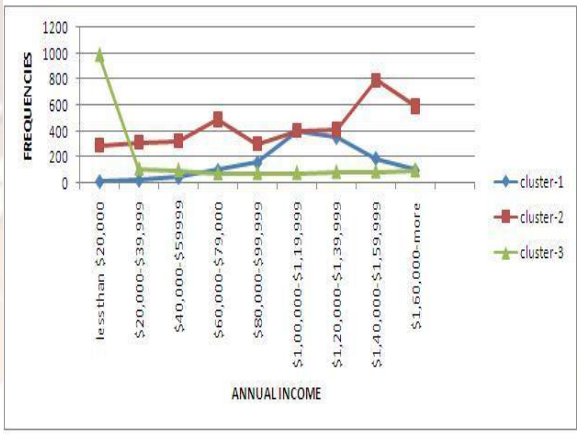


Fig-4:Graph of frequencies for ANNUAL INCOME

The following Table-2 shows the results of determining the number of clusters for EM algorithm using various datasets

S.No	Name of the Relatio	No of Instances	No of Clusters	No of Attributes	(Goal Function) Min. Avg. – LLH
1	contact-lenses	24	2	5	-3.82823
2	cpu	209	6	7	-39.69157
3	cpu with vendor	209	9	8	-40.81818
4	diabetes	768	9	9	-28.54483
5	glass	214	10	10	-2.99397
6	ionosphere	351	12	35	-10.97424
7	iris	150	5	4	-2.03504
8	labour	57	17	3	-17.21063
9	soyabean	683	36	14	-15.78766

Table-2:Results of identifying number of clusters for

EM

V. Discussion & Conclusions

This paper investigates a very good method for automatically estimating the number of clusters in unlabeled data sets. The goal function for example, cluster centers (for *k*- means clustering); for *EM* clustering, an appropriate equivalent measure would be the average negative (log-) likelihood computed for the observations in the testing samples are very important in estimating the suitable number of clusters.

A. Applications

Cross-validation can be used to compare the performances of different predictive modeling procedures. For example, suppose we are interested in optical character recognition, and we are considering using either support vector machines (SVM) or k nearest neighbors (KNN) to predict the true character from an image of a handwritten character. Using cross-validation, we could objectively compare these two methods in terms of their respective fractions of misclassified characters. If we simply compared the methods based on their in- sample error rates, the KNN method would likely appear to perform better, since it is more flexible and hence more prone to over fitting compared to the SVM method.

Cross-validation can also be used in variable selection. Suppose we are using the expression levels of 20 proteins to predict whether a cancer patient will respond to a drug. A practical goal would be to determine which subset of the 20 features should be used to produce the best predictive model. For most modeling procedures, if we compare feature subsets using the in-sample error rates, the best performance will occur when all 20 features are used. However under cross-validation, the model with the best fit will generally include only a subset of the features that are deemed truly informative.

References

[1] Data Mining:Concepts and Techniques,Second Edition,Jiawei Han,University of Illinois at Urbana-Champaign,Micheline Kamber, ISBN 13: 978-1-55860-901-3,ISBN 10: 1-55860-901-6,chapter-7 page No-384

[2] Geisser, Seymour (1993). Predictive Inference. New York: Chapman and Hall. ISBN 0412034719.

[3] Kohavi, Ron (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection". Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence **2** (12): 1137–1143.(Morgan Kaufmann, San Mateo)

[4] Devijver, P. A., and J. Kittler, Pattern Recognition: A Statistical Approach, Prentice-Hall, London, 1982

[5] Kaufman L. and Rousseeuw P. J. Finding Groups in Data: An Introduction to Cluster Analysis. Wiley,1990.

[6] Jain A. K., Murty, M. N. and Flunn P. J. DataClustering:AReview,ACM,ComputingSurveys,Vol.31(3),264323,1999.

[7] Hartigan J. A. and Wong M. A. Algorithm AS 136: a k-means clustering algorithm,Applied statistics,28,100108,1979.

[8]The on-line textbook: Information Theory, Inference, and Learning Algorithms, by David J.C. MacKay.

[9] EM algorithm and variants: an informal tutorial by Alexis Roche.

[10] Kohavi, Ron (1995). "A study of cross-validation and bootstrap for accuracy estimation and model selection" Proceedings of the Fourteenth International Joint

[11]Introduction to dataMining with case studies by g.k gupta second edition PHI publications(ISBN-978-81203-3053-5) chapter-3