A Conceptual Framework For Extending Distance Measure Algorithm For Data Clustering.

A.M. Bagiwa[*1], S.I. Dishing[#2].

[#]*Mathematics Department, Ahmadu Bello University*
Zaria, Nigeria

***Abstract*— In this paper we look at data clustering as a problem that involves finding the relationship between data sets. The framework for this paper introduces an enhancement of the distance measure data clustering algorithm by adding some prior knowledge describing the domain of clusters. In this work we first take an overview of different data clustering algorithms. We then propose our approach for data clustering as an enhancement to the distance measure algorithm.**

Keywords*— Data, Cluster, Distance.*

I. INTRODUCTION

Data clustering is a common technique for data analysis, which is used in many fields of sciences that include machine learning, data mining, pattern recognition, image analysis and many other areas of research. Clustering can be considered as the most important unsupervised learning problem; so, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data.
A loose definition of clustering could be the process of organizing objects into groups whose members are similar. A cluster is therefore a collection of objects which are similar between them and are dissimilar to the objects belonging to other clusters. We can show this with a simple graphical example in the Figure below:

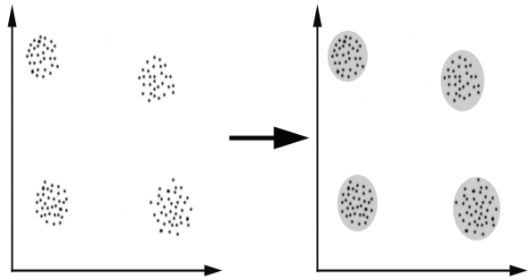


Figure 1: Sample Cluster

In this case we easily identify the 4 clusters into which the data can be divided; the similarity criterion is distance: two or more objects belong to the same cluster if they are close according to a given distance (in this case geometrical distance). This is called distance-based clustering.
The goal of clustering is to determine the intrinsic grouping in a set of unlabeled data.

II. OVERVIEW OF PRESENT DATA CLUSTERING ALGORITHMS.

In current literature, many approaches are given for clustering data. Many distance-based clustering algorithms [1] are proposed for transactional data. But traditional clustering techniques have the curse of dimensionality and the sparseness issue when dealing with very high-dimensional data such as market-basket data or Web sessions. Fuzzy c-means (FCM) is a method of clustering which allow one piece of data to belong to two or more clusters. This method developed by [2] and improved by [3] is frequently used in pattern recognition. It is based on minimization of the following objective function:

$$J_m = \sum_{i=1}^{N}\sum_{j=1}^{C} u_{ij}^{m}\left\|x_i - c_j\right\|^2 \quad , \quad 1 \le m < \infty$$

where m is any real number greater than 1, u_{ij} is the degree of membership of x_i in the cluster j, x_i is the ith of d-dimensional measured data, c_j is the d-dimension center of the cluster, and ||*|| is any norm expressing the similarity between any measured data and the center. As we know, data are bound to each cluster by means of a Membership Function, which represents the fuzzy behaviour of this algorithm.

K-means [4] is one of the simplest unsupervised learning algorithm that solve the well known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) . The K-Means algorithm has been adopted by replacing the cluster mean with the more robust notion of cluster medoid (that is, the object within the cluster with the minimal distance from the other points) or the attribute mode [5]. It can be viewed as a greedy algorithm for partitioning the n samples into k clusters so as to minimize the sum of the squared distances to the cluster centres. It does have some weaknesses: 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. Another way to deal with clustering problem is using a model-based approach, which consists of using certain models for clusters and attempting to optimize the fit between the data and the model.

In practice, each cluster can be mathematically represented by a parametric distribution, like a Gaussian (continuous) or a Poisson (discrete). The entire data set is therefore modelled by a *mixture* of these distributions. The algorithm works in this way: it chooses the component (the Gaussian) at random with probability $P(\omega_i)$; it samples a point $N(\mu_i, \sigma^2 I)$.

III. PROBLEMS

There are a number of problems with clustering. Among them:

1. Current clustering techniques do not address all the requirements adequately (and concurrently).
2. Dealing with large number of dimensions and large number of data items can be problematic because of time complexity.
3. The effectiveness of the method depends on the definition of distance (for distance-based clustering).
4. If an obvious distance measure doesn't exist we have to define it, which is not always easy, especially in multi-dimensional spaces;
5. The result of the clustering algorithm can be interpreted in different ways.

IV. DISTANCE MEASURE

An important component of a clustering algorithm is the distance measure between data points. If the components of the data instance vectors are all in the same physical units then it is possible that the simple Euclidean distance metric is sufficient to successfully group similar data instances. However, even in this case the Euclidean distance can sometimes be misleading. Figure 2 illustrates this with an example of the width and height measurements of an object. Despite both measurements being taken in the same physical units, an informed decision has to be made as to the relative scaling. As Figure 2 shows, different scaling can lead to different clustering.

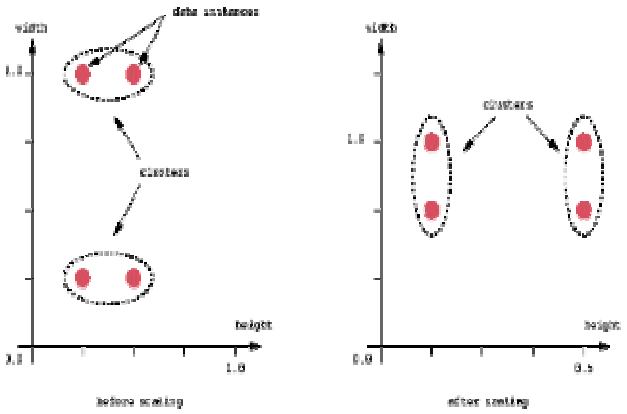
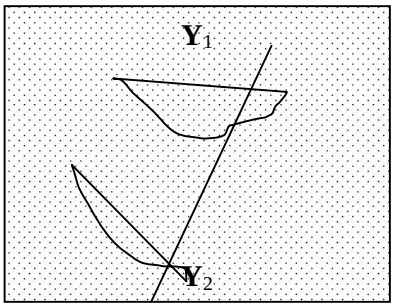


Figure 2: Distance Measure Cluster

Notice however that this is not only a graphic issue: the problem arises from the mathematical formula used to combine the distances between the single components of the data feature vectors into a unique distance measure that can be used for clustering purpose. Different formulas lead to different clusterings.
Again, domain knowledge must be used to guide the formulation of a suitable distance measure for each particular application.

Here is an example showing how the means $\mathbf{y}_1$ and $\mathbf{y}_2$ move into the centers of two clusters.



V. DISTANCE MEASURE ENHANCEMENT

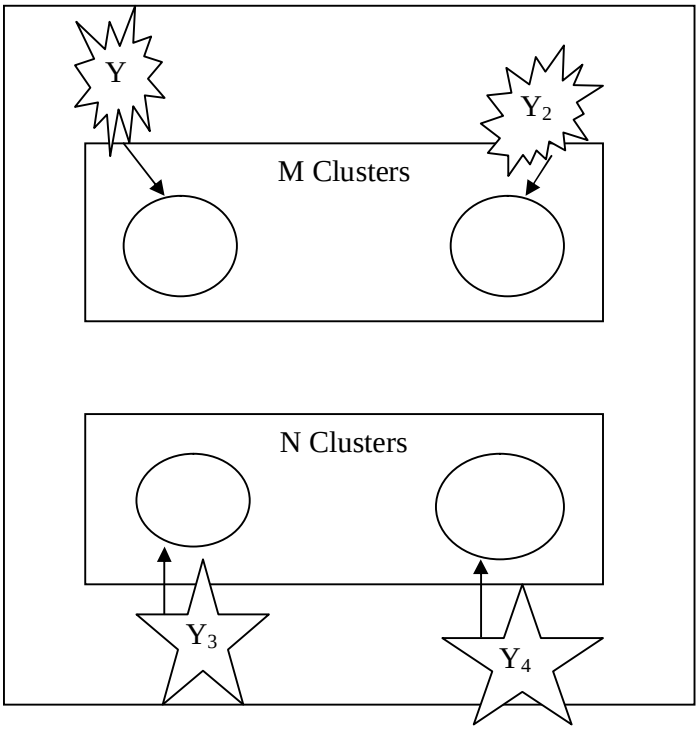
Our enhancement to the distance measure algorithm is the addition of some prior knowledge to those dataset that belong to a given cluster. The main idea here is to define a point of

focus for each data set, belonging to a cluster and assign some characteristics to that cluster. The next step is to take each point belonging to a given data set and associate it to the nearest data set. At this point a regroup is done. Suppose that we have n sample feature vectors $\mathbf{y}_1$, $\mathbf{y}_2$, $\mathbf{y}_3$ ..., $\mathbf{y}_n$ all from the same class, and we have the knowledge that they belong to **M** clusters, M < n. Let $\mathbf{m}_i$ be the mean of the vectors in cluster i. If the clusters are well separated, we can use a minimum-distance classifier to separate them. That is, we can say that **y** is in cluster i if $\| \mathbf{y} - \mathbf{m}_i \|$ is the minimum of all the k distances. This can easily be done when each sample feature vector has the prior information that describes the domain of its cluster. This suggest the following procedure for the enhancement:

Make initial guesses for the vector $\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_k$
Attach attributes to each vector describing the domain of its cluster
For i from 1 to k
Replace $\mathbf{m}_i$ with the mean of all of the samples for cluster i
end_for
end_until

VI. FRAMEWORK.

The framework for our implementation is given in the diagram below where each cluster is attached with a vector describing the domain of that dataset.



The enhancement provide a promising result but will require more time because of the elaborate computation it performs as a result of attaching each vector to the domain of its data set.

REFERENCES

[1]. R. Ng and J. Han, "CLARANS: A Method for Clustering Objects for Spatial Data Mining," IEEE Trans. Knowledge and Data Eng., vol. 14, no. 5, pp. 1003-1016, Sept./Oct. 2002.

[2] J. C. Dunn (1973): "A Fuzzy Relative of the ISODATA Process and Its Use in Detecting Compact Well-Separated Clusters", Journal of Cybernetics 3: 32-57

[3]. J. C. Bezdek (1981): "Pattern Recognition with Fuzzy Objective Function Algoritms", Plenum Press, New York

[4]. J. B. MacQueen (1967): "Some Methods for classification and Analysis of Multivariate Observations, Proceedings of 5-th Berkeley Symposium on Mathematical Statistics and Probability", Berkeley, University of California Press, 1:281-297

[5]. C. Gozzi, F. Giannotti, and G. Manco, "Clustering Transactional Data," Proc. Sixth European Conf. Principles and Practice of Knowledge Discovery in Databases (PKDD '02), pp. 175-187, 2002.

Mustapha A. Bagiwa is an M.Sc Student of Computer Science in Ahmadu Bello University, Zaria. Nigeria. He received his B.Sc Degree in computer science from Usmanu Danfodiyo University Sokoto, Nigeria in 2006. He is currently working as a Graduate Assistant in the Department of Mathematics, Ahmadu Bello University, Zaria. Nigeria

Salihu I. Dishing. He received his M.Sc Degree(RGU), in Aberdeen, United Kingdom. He received his B.sc Degree in computer science from Ahmadu Bello University, Zaria. Nigeria. He is currently working as an Assistant Lecturer in Ahmadu Bello University, Zaria. Nigeria.