International Journal of Advanced Science and Technology
Vol.64 (2014), pp.101-118
http://dx.doi.org/10.14257/ijast.2014.64.08

Ensemble Clustering based on Heterogeneous Dimensionality Reduction Methods and Context-dependent Similarity Measures

Augustine S. Nsang[1], Irene Diaz[2] and Anca Ralescu[3]

[1]CS Department, School of Inf. Tech. and Computing, American University of Nigeria
Yola By-Pass, PMB 2250, Yola, Nigeria
[2]Department of Computer Science, University of Oviedo, Spain
[3]EECS Department, University of Cincinnati, Cincinnati, OH 45221-0030, USA

augustine.nsang@aun.edu.ng, sirene@uniovi.es, anca.ralescu@uc.edu

Abstract

This paper discusses one method of clustering a high dimensional dataset using dimensionality reduction and context dependency measures (CDM). First, the dataset is partitioned into a predefined number of clusters using CDM. Then, context dependency measures are combined with several dimensionality reduction techniques and for each choice the data set is clustered again. The results are combined by the cluster ensemble approach. Finally, the Rand index is used to compute the extent to which the clustering of the original dataset (by CDM alone) is preserved by the cluster ensemble approach.

Keywords: *dimensionality reduction, context dependency measures, rand index, cluster ensemble approach*

1. Introduction

Given a collection of n data points (vectors) in high dimensional space, it is often helpful to be able to project it into a lower dimensional space without suffering great distortion [1]. In other words, it is helpful to embed a set of m points in n-dimensional space into r-dimensional space, where r << n. This operation is commonly known as *dimensionality reduction*.

There are many known methods of dimensionality reduction. In some of these methods, each attribute in the reduced set is a linear combination of the attributes in the original data set. Such methods include Random Projection (RP), Singular Value Decomposition (SVD), Principal Component Analysis (PCA), Kernel Principal Component Analysis (KPCA), Discrete Cosine Transform (DCT) and Latent Semantic Analysis (LSA) [5]. Therefore, in effect, to compute a data point in the new dimensions one needs to actually know all of the original dimensions.

Other dimensionality reduction methods, however, reduce a dataset to a subset of the original attribute set [6]. These include those studied in [16], namely, the Direct Approach (DA), the Combined Approach (CA), the Variance Approach (VAR), the Top-Down Approach (TD), the Bottom-Up Approach (BU), the Weighted Attribute Frequency Approach (WAF) and the Best Clustering Performance Approach (BCP) which will be illustrated further in this paper. Other filtering approaches for high dimensional data can be found in [9] [11], [12], [2].

An application of dimensionality reduction is the clustering of high dimensional data. A common view of clustering is as the assignment of a data set into subsets (called clusters) so

ISSN: 2005-4238 IJAST

that data in the same cluster are similar in some sense [4], [8], [17] and data assigned to different clusters are as dissimilar as possible. Clustering is a method of unsupervised learning, and much used for data analysis in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. This paper discusses a method of partitioning an n-dimensional data set D containing m data points, into k clusters, k << n, based on integrating context dependency measures and dimensionality reduction techniques, such that the clusters in the original dimensions are preserved as much as possible by the clusters in the reduced dimensions. From this point on, the paper is organized as follows. Section 2 describes the concepts of *context-based divergence* and *context-based similarity*. Sections 3 and 4 illustrate the use of these concepts in connection with *context-based cluster structure*, the computation of the similarity and divergence between a data sample and a cluster and between two clusters in a context-based environment. Section 5 reviews the dimensionality reduction approaches used in this study and in Section 6 these are used in conjunction with the context dependency measures in a cluster ensemble approach. Section 7 presents illustrative experimental results on a small artificial data set. Section 8 provides brief concluding remarks.

2. Context-Based Proximity Measurements [19]

The distance between two data vectors, as measured by various distance formulae is static. Once computed, it does not change as long as the vectors do not change. Moreover, there could be cases where the vectors change but the distance between them remains constant. This is possible when the two vectors are translated by a fixed amount in the same direction in multidimensional space. But the perception of distance is not static. For example, consider the problem of measuring the proximity between two geographical locations, say two cities, for example Cincinnati, Ohio and Indianapolis, Indiana, based only on their geographical information, and not taking into consideration any other locality. Since the locations of these cities are different, each of these two cities perceives the other as most different. Now consider adding a third location, for example, Los Angeles, California. Now, the perceived proximity between Cincinnati and Indianapolis is modified - has become larger, since Los Angeles is much farther than any of these two cities, and by comparison Cincinnati and Indianapolis seem to be almost the same. Similarly, if now another city, say, Lagos Nigeria is added, proximity between the three American cities is further increased. This modification of proximity perception is evident when people discuss about distances. It is not uncommon for someone from Lagos to feel, for example that if one traveled to Cincinnati, one has also almost "arrived" in Los Angeles. Yet, once in Cincinnati this feeling is not present anymore, since in the context of distances in the US, Cincinnati and Los Angeles are pretty far. This example illustrates the effect of the context in perception of proximity. If proximity perception is to be measured in the interval [0, 1], then in the first case described above, the proximity between Cincinnati and Indianapolis is almost 0. As other, more distant cities are added, it increases almost to 1 (almost as if Cincinnati and Indianapolis were co-located). The context-based divergence is an approach (proposed in [19] and further explored in [20, 21]), which captures and explores this notion of perceptual proximity and it will be reviewed in this paper.

Terminology

Unless otherwise specified, the following terminology will be used in this paper:

- $X = X_{i,e}, i = 1,\dots,m; e = 1\dots,n$ denotes a data matrix each with m data vectors, each with n attributes or dimensions. $X_{i,e}$ can take any real value consistent with its semantics. $X_{i,*}$ denotes the i th data sample (row), and $X_{*,e}$ denotes the column corresponding to the e th variable.

- C_m^l denotes a cluster identified by label 1, consisting of m data samples (cluster members). That is, $|C_m^l| = m$.
- C^l denotes a cluster of unspecified size identified as l, and $i \in C^l$ refers to the i th member in C^l.
- C_m stands for an unlabeled cluster of size m, such that $i \in C_m$ refers to the i th element in it.

Definition 1. [19] The vertical context (v-context) for a sample data vector is the collection of the remaining data vectors. Generalizing, the v-context for a collection of data vectors is the intersection of the v-contexts of the individual vectors. Mathematically:

$$v - context_X(X_i, *) = \{x \in X : x \neq X_{i,*}\}$$
$$v - context_X(N \subset X) = \cap_{x \in N} v - context_X(x)$$

In the example used earlier, while measuring the proximity between two cities, the vertical context consists of all the other cities in the data set. Similarly, the horizontal context of a sample attribute is the collection of the remaining attributes.

Definition 2. [19] The context dependent divergence between data vectors $X_{i,*}$ and $X_{j,*}$ with respect to attribute e is defined by:

$$D_e(i, j) = \frac{[X_{i,e} - X_{j,e}]^2}{\delta_X^2(i,e)}$$

where $\delta_X^2(i,e)$, the variance of $X_{i,e}$ about the i^{th} point along the e^{th} dimension, is defined by:

$$\delta_X^2(i,e) = \sum_{k=1}^{m} (X_{i,e} - X_{k,e})^2$$

Generalizing over all the attributes, the *overall context dependent divergence* between data vectors $X_{i,*}$ and $X_{j,*}$ is defined as the arithmetic mean of their divergences over all the attributes. That is:

$$D(i,j) = \frac{1}{n}\sum_{e=1}^{n} D_e(i,j)$$

Finally, the similarity (or proximity) between two data vectors i and j is given by:

$$S(i,j) = 1 - D(i,j)$$

To illustrate how the above definitions capture the effect of context in evaluating the proximity between two data points, consider the data set X_δ given by 3×2 matrix in the equation below:

$$X_\delta = \begin{bmatrix} 1 & 4 \\ 2+\delta & 3+\delta \\ 50 & 100 \end{bmatrix}$$

In this example, $X_{\delta\,(2,*)}$, the second row, is the $v-$context for the rows one and three. The context divergence between the 1st and 3rd row of this matrix, as a function of δ is computed to be

$$D_\delta(1,3) = \frac{1}{2}\left(\frac{49^2}{(1+\delta)^2+49^2} + \frac{96^2}{(1-\delta)^2+96^2}\right)$$

Obviously, $\lim\limits_{\delta\to\infty} D_\delta(1,3) = 0$ and hence, $\lim\limits_{\delta\to\infty} S_\delta(1,3) = 1 - \lim\limits_{\delta\to\infty} D_\delta(1,3) = 1$.

The evolution of the values of X , $D_\delta(1,3)$ and $S(1,3)$ when $\delta = 0,100,200,\ldots,1000$ is shown in the table below.

δ	0	100	…	1000
X_δ	$\begin{bmatrix} 1 & 4 \\ \mathbf{2} & \mathbf{3} \\ 50 & 100 \end{bmatrix}$	$\begin{bmatrix} 1 & 4 \\ \mathbf{102} & \mathbf{103} \\ 50 & 100 \end{bmatrix}$	…	$\begin{bmatrix} 1 & 4 \\ \mathbf{1002} & \mathbf{1003} \\ 50 & 100 \end{bmatrix}$
$D_\delta(1,3)$	0.9997	0.3376	…	0.0058
$S_\delta(1,3)$	0.0003	0.6624	…	0.9942

Figure 1 illustrates the perceptual proximity for the data in this example.

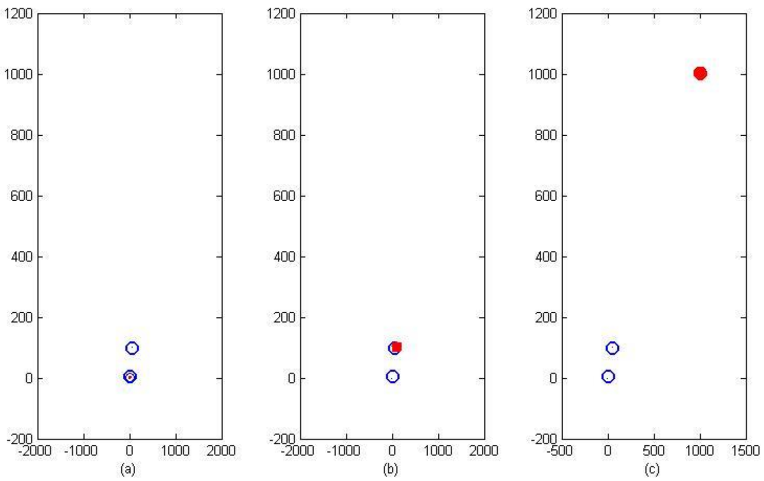


Figure 1. Illustration of the change in perception of proximity for the data used in the above example.

The values of δ are 0, in (a), 100 in (b) and 1000 in (c). In each figure the blue circles show the rows one and three of X_{δ}, while the red dots (very small in (a)) correspond to the changing positions of the second row as δ changes.

3. Context-dependent Cluster Structure

Taking into account context the notion of *context dependent clustering* can be defined (see [19]). Although a common approach used in pattern recognition, clustering remains a somewhat ill-defined problem: the objective is to find *good* groupings/clusters of data point with little or no information on the ground truth. A discussion of the issues underlying cluster validity can be found in [17]. The set of data vectors that satisfy certain criteria of proximity are grouped into clusters to reduce information load and facilitate information processing and interpretation.

Traditional approaches to clustering consider a cluster as a lump of data that is characterized by measures such as *within cluster distance* and *between cluster distance* [7][3] which measure respectively, the cluster tightness and separation. But these approaches are based on the assumption that all cluster members are equally significant from the point of view of measuring cluster properties and making decisions about cluster expansion [5][8]. As pointed out in [19], in real life, there are many situations in which assumptions of “equal significance” of all cluster members do not hold. If the clusters are represented as "clouds" in a multi-dimensional space, then the members near the center of a cluster may be considered more representative of the cluster properties. From the point of view of deriving cluster properties, the cluster members belonging to different regions of the clusters must be viewed differently. There should be a way to differentiate between the cluster members based on their significance to the cluster they belong to. One way to achieve this is by attaching weights to the cluster members.

In the spirit of using context to determine the significance of each member and not rely on static schemes of weighing cluster members, this section discusses a method of cluster representation which assigns weights to cluster members based on their context dependent divergence with other members in the cluster [19]. The idea is to define the significance of a given cluster member in a way which reflects its context dependent similarity with other cluster members.

Definition 3. (Perfect Cluster [19]) When all members of a cluster have identical attribute values, the cluster is called a *perfect cluster*. More precisely, a perfect cluster is a set of m vectors $v_i, i = 1, \ldots, m$ of dimension n (attributes of a data item), such that:

$$v_{i,k} = a_k, i = 1, \ldots, m; k = 1, \ldots, n$$

Thus, in fact, the perfect cluster is trivial - a collection of identical data points.

Definition 4. (Weight of a cluster member [19]) The weight, ρ_k, of the member $k \in C_m$, is defined as the *normalized similarity* of k with all the remaining members. More precisely:

$$\rho_k = \begin{cases} 1 & \text{if } m = 1 \\ \dfrac{\sum\limits_{l=1, l \neq k}^{m} S(l,k)}{\sum\limits_{j=1}^{m} \sum\limits_{l=1, l \neq j}^{m} S(l,j)} & \text{otherwise} \end{cases}$$

It is easy to see that

$$\sum_{k=1}^{m} \rho_k = 1$$

For example, when the cluster contains only two elements, i.e. becomes the above formula yields respectively:

$$\rho_1 = \frac{S(2,1)}{S(2,1) + S(1,2)}; \quad \rho_2 = \frac{S(1,2)}{S(2,1) + S(1,2)}$$

Therefore, the weights ρ_k define a convex combination. Moreover, if $S(i, j) = S(j,i)$, for all pairs of clusters elements $i, j \in C_m$, when $\rho_k = \dfrac{1}{m}$. In particular this is true for the perfect cluster.

4. Similarity Within and Between Clusters 0

To measure the tightness of and separation between clusters, a similarity measure must be defined. A context dependent similarity measure between a data item and a cluster, on one hand, and two clusters, on the other hand, can be defined based on the context dependent divergence and similarity measures defined above.

4.1. Similarity between a data item and a cluster

The similarity between a data item v and a cluster C_m, $S(v, C_m)$, is defined as convex aggregation of the similarity between v and the cluster members i, determined by the weights ρ_i.

That is,

$$S(\nu, C_m) = \sum_{i=1}^{m} \rho_i S(\nu, i)$$

Similarly, to account for the asymmetry, $S(C_m, \nu)$ can be defined as

$$S(C_m, \nu) = \sum_{i=1}^{m} \rho_i S(i, \nu)$$

4.2. Similarity between clusters

Similarly to data item to cluster similarity, the cluster to cluster similarity is defined as a convex combination of similarities. Let C and C' denote two clusters.
Then $S(C, C')$ is defined as

$$S(C, C') = \sum_{i \in C} \rho_i S(i, C')$$

which can be further written as

$$S(C, C') = \sum_{i \in C} \rho_i \sum_{j \in C'} \rho_j S(i, j)$$

From the asymmetry of the item-to-item similarity, it follows that

$$S(C, C') \neq S(C', C)$$

Remark 1. In the above, for ease of notation, the same symbol, S, was used to denote three different similarity measures, item-to-item similarity, item-to-cluster similarity, and cluster-to-cluster similarity although these measures of similarities are defined between different structures.

The context dependent measures (CDMs) of similarity are used in an agglomerative clustering procedure in the original high dimensional data set, and also in conjunction with several methods of dimensionality reduction. To this end, the CDMs are symmetrized, as follows:

$$S_{sym}(x, y) = \frac{S(x, y) + S(y, x)}{2}$$

where $S(x, y)$ denotes the item-to-item similarity.

5. Dimensionality Reduction Approaches

The basic description of the dimensionality reduction problem used in this paper is as follows: Given a data set represented as an $m \times n$ matrix D, find a representation of D as an $m \times r$ matrix D_r , $r < n$ (usually much smaller than n) subject to some constraints (or optimization criterion). Some of the dimensionality reduction techniques studied in [14 - 16] and used subsequently in this paper are briefly described below.

5.1. The Variance Approach (*VAR*)

The *Variance* approach starts with an empty set, I of dimensions and adds dimensions of D to this set in decreasing order of their variances [15]. That means that a set I_r with r dimensions contains the dimensions of top r variances. This approach assumes that the discriminatory power of a dimension depends directly on the variance of the data along that dimension; dimensions of low variance are left out as they would fail to discriminate between the data. Indeed, in an extreme case where all the values along a dimension are equal, and therefore, the variance is 0, it is impossible to distinguish between data points based on that dimension alone. Thus, letting $I_r = \{i_1, \ldots, i_r\} \subset \{1, \ldots, n\}$ be the collection of dimensions corresponding to the top *r* variances, the reduced database, D_r is obtained by extracting the data corresponding to the selected dimensions. That is, $D_r = D(:, I_r)$. Given that *D* is an *m x n* data set, D_r is an *m x r* data set whose i th column is the column of the original database of i th largest variance.

5.2. The Combined Approach (*CA*)

The *Combined Approach* selects the combination of *r* attributes which *best preserve the inter-point distances*, and reduces the dataset to a dataset containing only those *r* attributes. This approach consists of the following steps:

1. Determine the extent to which each attribute preserves the inter-point distances. In other words, for each dimension $i \in \{1, \ldots, n\}$ compute

$$m_i = \min_{u,v}\{\frac{\| f_i(u) - f_i(v) \|^2}{\| u - v \|^2}\}$$

$$M_i = \max_{u,v}\{\frac{\| f_i(u) - f_i(v) \|^2}{\| u - v \|^2}\}$$

where u and v are rows of D, $f_i(u)$ and $f_i(v)$ are the corresponding rows in the dataset reduced to the single attribute i.

2. Compute the average distance preservation, *mid_i*, for each attribute i:

$$mid_i = \frac{m_i + M_i}{2}$$

3. Reduce D to D_r as follows: Let $I_r \subset \{1, \ldots, n\}$ be defined as $I_r = \{i_1, \ldots, i_r\}$, where $i_1, \ldots, i_r$ is the collection of the r dimensions of *D* whose average mid_i value is maximum. In other words,

$$\mathrm{I_r} = \mathrm{argmax_C}\ (\frac{1}{|r|}\sum_{i \in C} mid_i\ )$$

Then $\mathrm{D_r}$ is defined as $\mathrm{D_r = D(:, I_r)}$.

5.3. The Direct Approach (*DA*)

Like for the *Combined Approach*, to reduce a dataset $D_{m \times n}$ to a dataset containing *r<n* columns, the *Direct Approach* selects the combination of *r* attributes which best preserve the

inter-point distances, and reduces the original dataset to a dataset containing only those *r* attributes. However this done in a different way according to the following steps:

1. Generate all possible combinations of *r* attributes from the original *n* attributes.

2. For each combination, C, generated at the previous step, compute m_C and M_C as follows:

$$m_C = \min_{u,v} \frac{\| f_C(u) - f_C(v) \|^2}{\| u - v \|^2}$$

$$M_C = \max_{u,v} \frac{\| f_C(u) - f_C(v) \|^2}{\| u - v \|^2}$$

where u and v are rows of D, and $f_C(u)$ and $f_C(v)$ are the corresponding rows in the dataset reduced to the attributes in C. Compute the average distance preservation for this combination of attributes as

$$mid_C = \frac{m_C + M_C}{2}$$

Select the subset of dimensions C_0 as

$$C_0 = \operatorname{argmin}_C | mid_C - 1 |$$

Thus D_r is given as $D_r = D(:, C_o)$.

The difference between the *Combined* and *Direct* approaches is that for the *Combined Approach*, to reduce a data set *D* containing *n* attributes to a data set D_r containing *r* attributes, the average distance preservation is first calculated for each attribute, and all combinations of *r* attributes (from the *n* attributes of D) are then generated. D_r then becomes the dataset containing the *r* attributes of *D* whose average mid_i value is maximum. With the *Direct Approach,* on the other hand, to reduce a data set *D* containing *n* attributes to a data set D_r containing *r* attributes, we first generate all combinations of *r* attributes from the *n* attributes of *D.* To find the average distance preservation for any combination of attributes, *C*, we reduce the original dataset directly to the dataset containing only the attributes in *C*, and then compute the average distance preservation for this combination using the formulas above. D_r then becomes the dataset containing the *r* attributes whose average distance preservation is maximum.

5.4. Top-down Approach (*TD*)

The *Top-Down* dimensionality reduction approach considers subsets of attributes reduced by one attribute at a time. To reduce a data set *D* containing *n* attributes to a data set D_r containing *r* attributes, the Top-Down Approach starts with the original *n* attributes of D, and reduces it successively to the subset of $n - i$, $i = 1, 2, \ldots$ attributes, such that a certain criterion is satisfied, until $n - i = r$. For example, the criterion could be to preserve as well as possible k-means clustering [14-16].

5.5. The Bottom-up Approach (*BU*)

Contrary to the *Top-Down* dimensionality reduction approach, the *Bottom-Up* approach considers subsets of attributes increased by one attribute at a time. Starting with the subset S_1 containing the single attribute, i_1, of D, which best preserves k -means clustering, it increases it to the subset S_2 of two attributes (i_1, i_2) which best preserve k -means clustering. The process continues until the subset S_r of r attributes $(i_1, \ldots, i_r)$ which best preserve k -means clustering has been obtained.

Remark 2. It should be noted that each of these approaches implements a heuristic and that the final subset of attributes obtained is not guaranteed to be optimal with respect to the criterion selected. However, in experiments, these approaches were found to be more effective.

5.6. The Weighted Attribute Frequency Approach (*WAF*)

In each of the approaches considered above, attributes/dimensions were considered implicitly equally important. On the other hand, from applying each of these approaches, and inspecting the corresponding selected attribute subsets it can often be seen (depending on the actual data set D) that some attributes appear more often than others. The weighted attribute frequency approach (WAF) aims at taking this into consideration. To reduce an $m \times n$ data set D to an $m \times r$ data set D_r, with $r < n$ WAF works as follows:

1. Generate *DA(r), CA(r), Var(r), TD(r),* and *BU(r)* each of which is the collection of attributes obtained when the given dimensionality reduction method is applied to reduce it to *r* attributes.
2. Assign weights to an attribute i according to its selection by each of these methods. From experimental studies, the weights were assigned as

$$w(i) = \begin{cases} 3 \text{ if } i \in \{NTD(r), NBU(r)\} \\ 1 \text{ if } i \in \{DA(r), CA(r), Var(r)\} \end{cases}$$

Thus *WAF(r)* becomes the collection of the *r* attributes of *D* attributes of largest weight.

5.7. The Best Clustering Performance Approach (*BCP*)

A consequence of distance preservation is that when used in algorithms where distance is computed, the original or reduced data should yield the same results in all algorithms based on inter-point distance. Such algorithms include classification and clustering. Given k, and the set DR = {CA(r), DA(r), Var(r), TD(r), BU(r)} whose elements are defined as in Section 5.6 above, $BCP(k, r)$ is the element of *DR* which best preserves k -means clustering with a subset of r attributes. For example, if the *Var* approach best preserves k-means clustering when *D* is reduced to five attributes, for $k = 3$, then $BCP(3,5) = Var(5)$. On the other hand, if the Top-Down Approach best preserves k-means clustering when *D* is reduced to three attributes, for *k* = 3, then $BCP(3,3) = TD(3)$.

6. Clustering in a High Dimensional Space based on Clustering in Reduced Dimensions

The context dependent procedure (CDM) defined here is an agglomerative clustering algorithm which starts with clusters corresponding to each data item. Clusters are merged based on their context dependent similarity until the desired number of clusters is obtained. Algorithm 1 shows the exact description of the CDM based clustering procedure.

Algorithm 1: Cluster formation using the context dependent similarity measure

```
Input: $X : m \times n$ matrix
       $k$ : the desired number of clusters ($k < m$)
Output: a partition of $X$ into $k$ clusters
Procedure: Agglomerative grouping using symmetric CDMs
    $l = m$
    # Initialize $l$ (number of clusters) to the $m$ data vectors.
    # That is, $c_i = X(i,:)$ for $i = 1, \dots, m$
    Repeat
        Step 1. Compute similarity between every two clusters using the symmetric CDMs
        Step 2. Find the most similar pair of clusters $c_i$ and $c_j$ in $X$
        Step 3. Merge $c_i$ and $c_j$ and decrement $l$ by 1
    Until $l \le k$
    Return all nonempty clusters
```

Let *DR = {CA, DA, Var, TD, BU, WAF, BCP, PCA}.* For each dimensionality reduction technique $\mathrm{C} \in DR$, clustering is obtained based on CDMs agglomerative clustering Algorithm 1. Then the ensemble clustering technique [4] is used to aggregate the multiple clustering results. The approach used in the current ensemble approach is inspired from and similar to that in [4], with the difference that in [4] all the clusters were represented by the same probabilistic model, determined by a parameter θ as a mixture of Gaussians.

For a data point x_i, $P(l \mid i, \mathrm{C})$, denotes the probability that x_i belongs to cluster l, $l = 1, \dots, k$ under the dimensionality reduction C. This probability is defined as

$$P(l \mid i, \mathrm{C}) = \begin{cases} 1 \text{ if data point } i \text{ belongs to cluster } l \text{ under C} \\ \quad 0 \qquad \text{otherwise} \end{cases}$$

Next, assuming independence, the probability that the data points x_i and x_j belong to cluster l under C is

$$P(l \mid i, j, \mathrm{C}) = P(l \mid i, \mathrm{C}) \times P(l \mid j, \mathrm{C}),$$

the probability that data points x_i and x_j belong to the same cluster under C is

$$P(i, j \mid \mathrm{C}) = \sum_{l=1}^{k} P(l \mid i, j, \mathrm{C}),$$

and finally, the probability that they belong to the same cluster is the average over clustering under all of the dimensionality reduction methods in *DR* is

$$P(i,j) = \frac{1}{|DR|} \sum_{C \in DR} P(i,j \mid \mathrm{C})$$

Note that in fact, from the definition of $P(l \mid i, \mathrm{C})$ it follows that $P(i,j)$ has the following properties:

1. $0 \le P(i,j) \le 1$,
2. $P(i,j) = 0$ if the data points x_i and x_j never belong to the same cluster when CDM clustering is run with any of the dimensionality reduction methods in DR
3. $P(i,j) = 1$ if the data points x_i and x_j belong to the same cluster when CDM clustering is run with every dimensionality reduction method in DR.

As in [4], the matrix $P(i,j)$, $i, j = 1, \ldots, m$ is used to define a conservative cluster similarity measure

$$Sim(c_i, c_j) = \min_{x_{i_1} \in c_i, x_{j_1} \in c_j} P(i_1, j_1)$$

where c_i and c_j are clusters, which is used in an agglomerative clustering procedure identical to that Algorithm 1 in which Step 1 computes the similarity matrix P, and Step 2 evaluates $Sim(c_i, c_j)$.

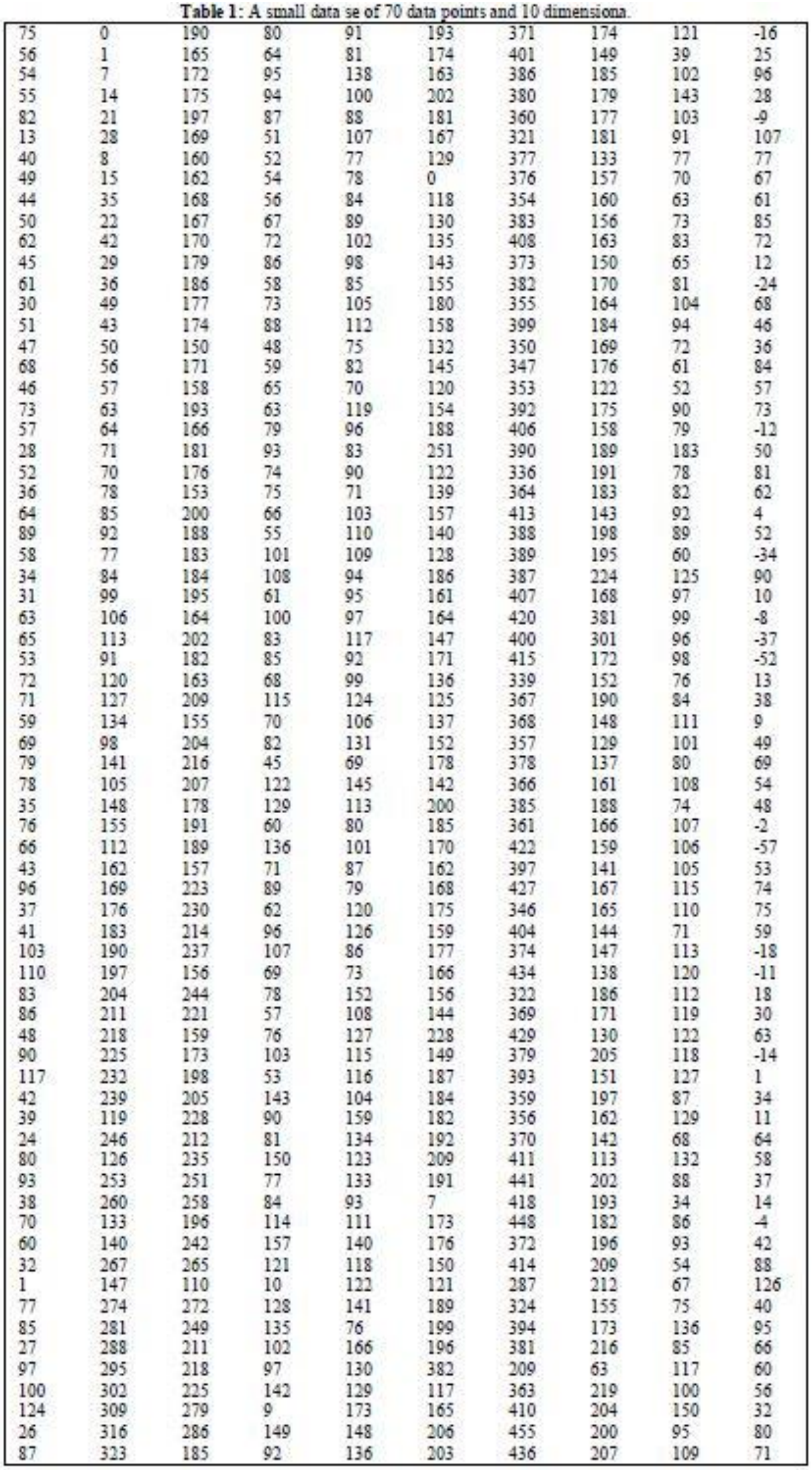
Outliers

Outliers are defined here as those points for which maximum similarity to any other data point is very small. These data points are initially discarded in the process of cluster formation. Approximately 10% of data points are removed this way in the current study. Then these points are assigned to the most similar of the final clusters obtained.

7. Illustrative Experimental Results

The approach described in the preceding sections is applied to a 70×10 data set shown in Table 1. The data set was clustered into 20 clusters in the original dimensions using Algorithm 1 and the context dependent measure of similarity. The clusters obtained are shown in Table 2. The data set was then reduced to seven dimensions, successively, using each of the eight dimensionality reduction techniques from the set *DR* (in Section 6 above), and clustered using the context dependent measures of similarity for each reduction. Tables 3 – 10 show the clusters obtained for each choice of dimensionality reduction technique. The similarity matrix P is then computed and used to aggregate the clustering results in Tables 3 – 10. The final clusters obtained by the Cluster Ensemble Approach are shown in Table 11.

Table 1: A small data se of 70 data points and 10 dimensiona.



75	0	190	80	91	193	371	174	121	-16
56	1	165	64	81	174	401	149	39	25
54	7	172	95	138	163	386	185	102	96
55	14	175	94	100	202	380	179	143	28
82	21	197	87	88	181	360	177	103	-9
13	28	169	51	107	167	321	181	91	107
40	8	160	52	77	129	377	133	77	77
49	15	162	54	78	0	376	157	70	67
44	35	168	56	84	118	354	160	63	61
50	22	167	67	89	130	383	156	73	85
62	42	170	72	102	135	408	163	83	72
45	29	179	86	98	143	373	150	65	12
61	36	186	58	85	155	382	170	81	-24
30	49	177	73	105	180	355	164	104	68
51	43	174	88	112	158	399	184	94	46
47	50	150	48	75	132	350	169	72	36
68	56	171	59	82	145	347	176	61	84
46	57	158	65	70	120	353	122	52	57
73	63	193	63	119	154	392	175	90	73
57	64	166	79	96	188	406	158	79	-12
28	71	181	93	83	251	390	189	183	50
52	70	176	74	90	122	336	191	78	81
36	78	153	75	71	139	364	183	82	62
64	85	200	66	103	157	413	143	92	4
89	92	188	55	110	140	388	198	89	52
58	77	183	101	109	128	389	195	60	-34
34	84	184	108	94	186	387	224	125	90
31	99	195	61	95	161	407	168	97	10
63	106	164	100	97	164	420	381	99	-8
65	113	202	83	117	147	400	301	96	-37
53	91	182	85	92	171	415	172	98	-52
72	120	163	68	99	136	339	152	76	13
71	127	209	115	124	125	367	190	84	38
59	134	155	70	106	137	368	148	111	9
69	98	204	82	131	152	357	129	101	49
79	141	216	45	69	178	378	137	80	69
78	105	207	122	145	142	366	161	108	54
35	148	178	129	113	200	385	188	74	48
76	155	191	60	80	185	361	166	107	-2
66	112	189	136	101	170	422	159	106	-57
43	162	157	71	87	162	397	141	105	53
96	169	223	89	79	168	427	167	115	74
37	176	230	62	120	175	346	165	110	75
41	183	214	96	126	159	404	144	71	59
103	190	237	107	86	177	374	147	113	-18
110	197	156	69	73	166	434	138	120	-11
83	204	244	78	152	156	322	186	112	18
86	211	221	57	108	144	369	171	119	30
48	218	159	76	127	228	429	130	122	63
90	225	173	103	115	149	379	205	118	-14
117	232	198	53	116	187	393	151	127	1
42	239	205	143	104	184	359	197	87	34
39	119	228	90	159	182	356	162	129	11
24	246	212	81	134	192	370	142	68	64
80	126	235	150	123	209	411	113	132	58
93	253	251	77	133	191	441	202	88	37
38	260	258	84	93	7	418	193	34	14
70	133	196	114	111	173	448	182	86	-4
60	140	242	157	140	176	372	196	93	42
32	267	265	121	118	150	414	209	54	88
1	147	110	10	122	121	287	212	67	126
77	274	272	128	141	189	324	155	75	40
85	281	249	135	76	199	394	173	136	95
27	288	211	102	166	196	381	216	85	66
97	295	218	97	130	382	209	63	117	60
100	302	225	142	129	117	363	219	100	56
124	309	279	9	173	165	410	204	150	32
26	316	286	149	148	206	455	200	95	80
87	323	185	92	136	203	436	207	109	71

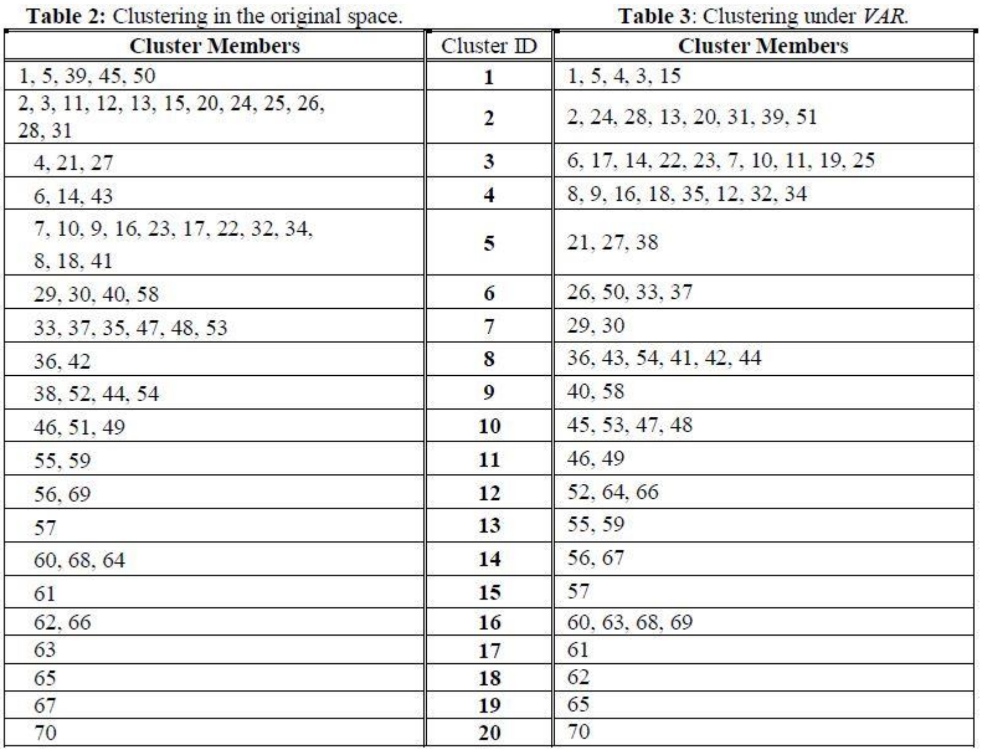


Table 2: Clustering in the original space. **Table 3**: Clustering under *VAR*.

Cluster Members	Cluster ID	Cluster Members
1, 5, 39, 45, 50	**1**	1, 5, 4, 3, 15
2, 3, 11, 12, 13, 15, 20, 24, 25, 26, 28, 31	**2**	2, 24, 28, 13, 20, 31, 39, 51
4, 21, 27	**3**	6, 17, 14, 22, 23, 7, 10, 11, 19, 25
6, 14, 43	**4**	8, 9, 16, 18, 35, 12, 32, 34
7, 10, 9, 16, 23, 17, 22, 32, 34, 8, 18, 41	**5**	21, 27, 38
29, 30, 40, 58	**6**	26, 50, 33, 37
33, 37, 35, 47, 48, 53	**7**	29, 30
36, 42	**8**	36, 43, 54, 41, 42, 44
38, 52, 44, 54	**9**	40, 58
46, 51, 49	**10**	45, 53, 47, 48
55, 59	**11**	46, 49
56, 69	**12**	52, 64, 66
57	**13**	55, 59
60, 68, 64	**14**	56, 67
61	**15**	57
62, 66	**16**	60, 63, 68, 69
63	**17**	61
65	**18**	62
67	**19**	65
70	**20**	70

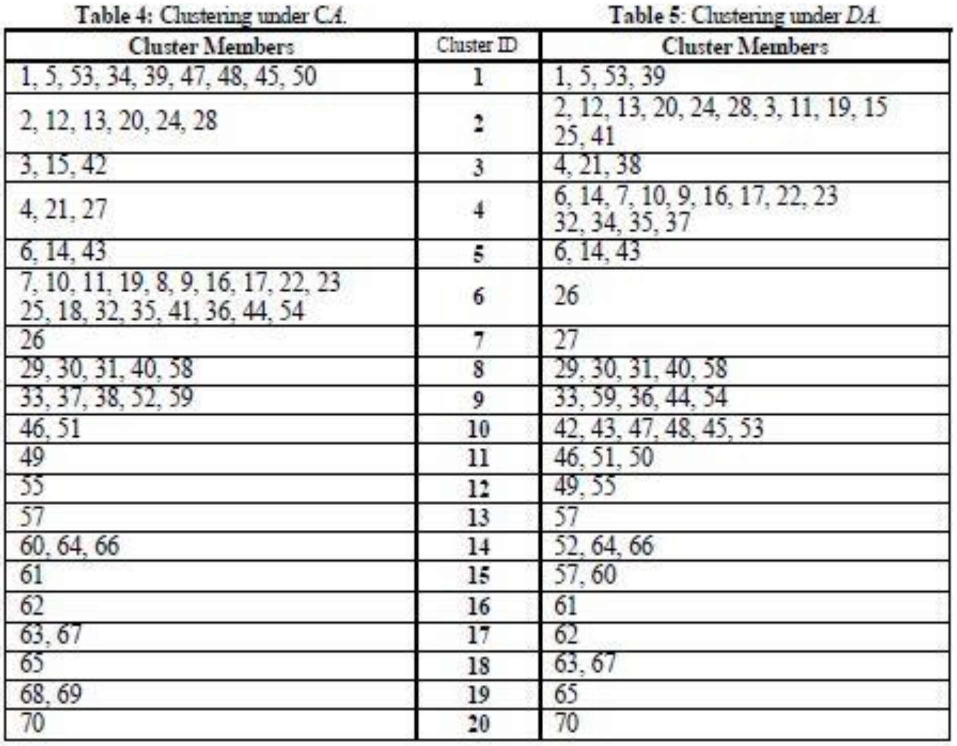


Table 4: Clustering under *CA*. **Table 5**: Clustering under *DA*.

Cluster Members	Cluster ID	Cluster Members
1, 5, 53, 34, 39, 47, 48, 45, 50	**1**	1, 5, 53, 39
2, 12, 13, 20, 24, 28	**2**	2, 12, 13, 20, 24, 28, 3, 11, 19, 15 25, 41
3, 15, 42	**3**	4, 21, 38
4, 21, 27	**4**	6, 14, 7, 10, 9, 16, 17, 22, 23 32, 34, 35, 37
6, 14, 43	**5**	6, 14, 43
7, 10, 11, 19, 8, 9, 16, 17, 22, 23 25, 18, 32, 35, 41, 36, 44, 54	**6**	26
26	**7**	27
29, 30, 31, 40, 58	**8**	29, 30, 31, 40, 58
33, 37, 38, 52, 59	**9**	33, 59, 36, 44, 54
46, 51	**10**	42, 43, 47, 48, 45, 53
49	**11**	46, 51, 50
55	**12**	49, 55
57	**13**	57
60, 64, 66	**14**	52, 64, 66
61	**15**	57, 60
62	**16**	61
63, 67	**17**	62
65	**18**	63, 67
68, 69	**19**	65
70	**20**	70

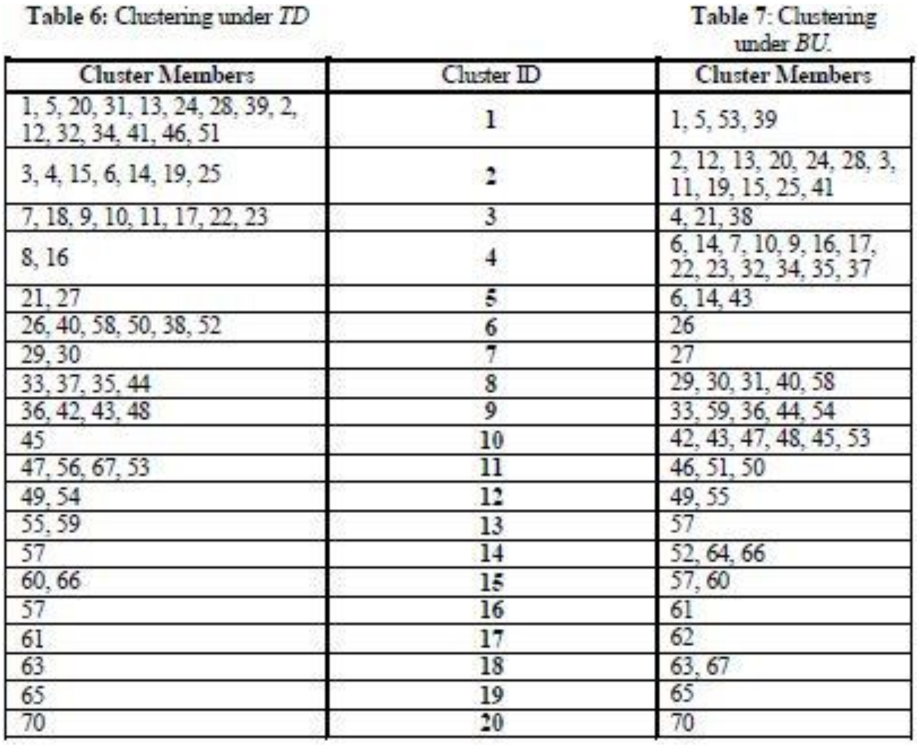


Table 6: Clustering under *TD*

Table 7: Clustering under *BU*.

Cluster Members	Cluster ID	Cluster Members
1, 5, 20, 31, 13, 24, 28, 39, 2, 12, 32, 34, 41, 46, 51	1	1, 5, 53, 39
3, 4, 15, 6, 14, 19, 25	2	2, 12, 13, 20, 24, 28, 3, 11, 19, 15, 25, 41
7, 18, 9, 10, 11, 17, 22, 23	3	4, 21, 38
8, 16	4	6, 14, 7, 10, 9, 16, 17, 22, 23, 32, 34, 35, 37
21, 27	5	6, 14, 43
26, 40, 58, 50, 38, 52	6	26
29, 30	7	27
33, 37, 35, 44	8	29, 30, 31, 40, 58
36, 42, 43, 48	9	33, 59, 36, 44, 54
45	10	42, 43, 47, 48, 45, 53
47, 56, 67, 53	11	46, 51, 50
49, 54	12	49, 55
55, 59	13	57
57	14	52, 64, 66
60, 66	15	57, 60
57	16	61
61	17	62
63	18	63, 67
65	19	65
70	20	70

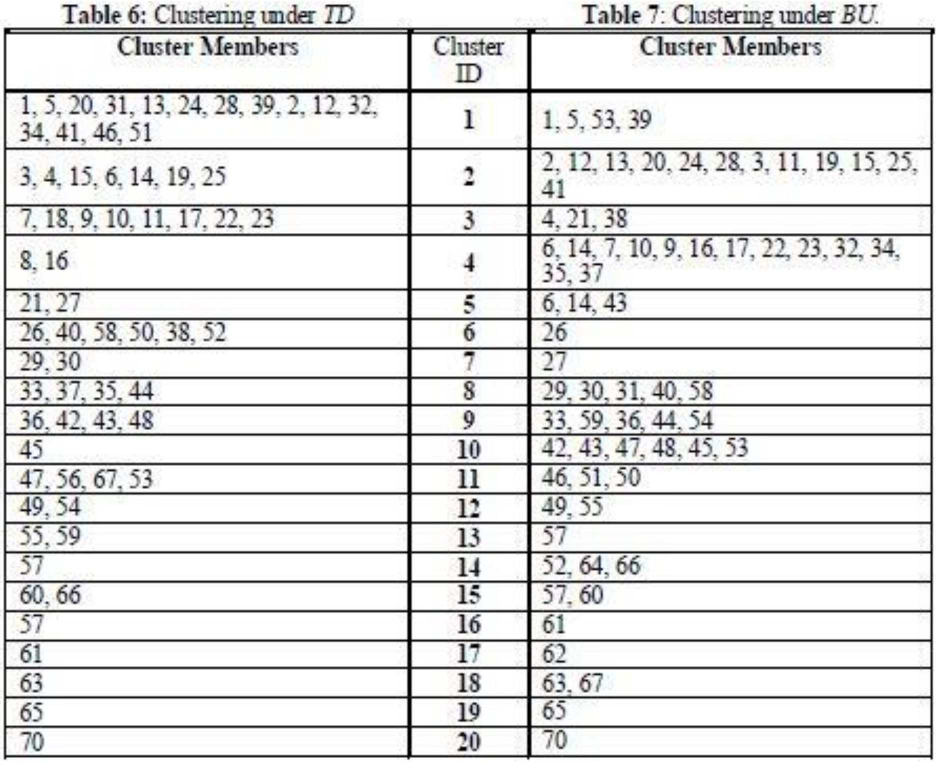


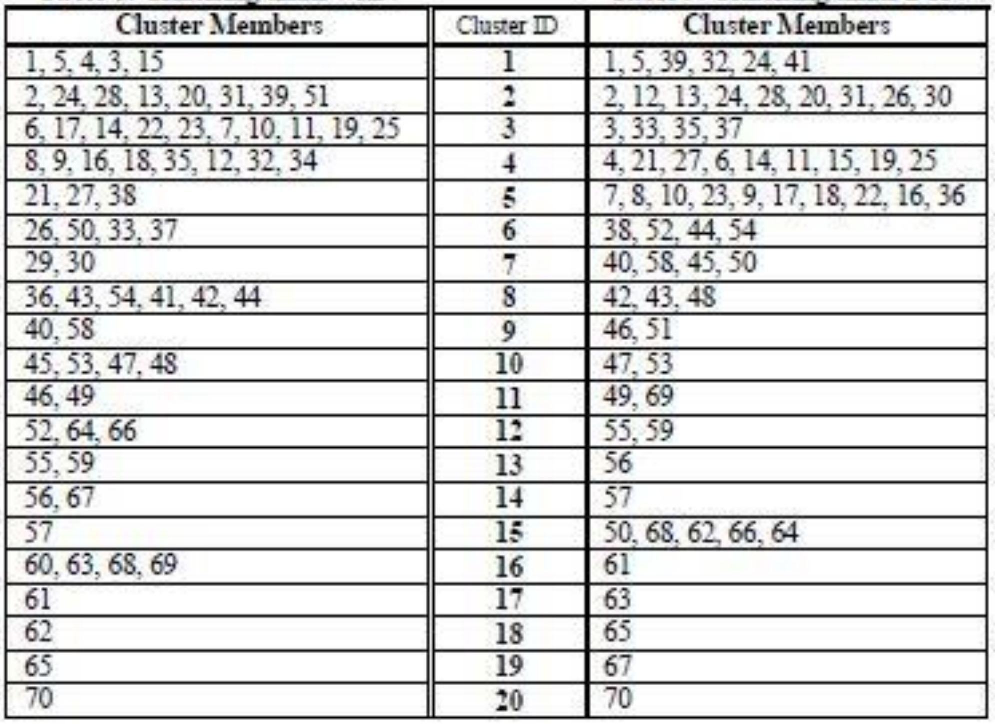
Table 6: Clustering under *TD*

Table 7: Clustering under *BU*.

Cluster Members	Cluster ID	Cluster Members
1, 5, 20, 31, 13, 24, 28, 39, 2, 12, 32, 34, 41, 46, 51	1	1, 5, 53, 39
3, 4, 15, 6, 14, 19, 25	2	2, 12, 13, 20, 24, 28, 3, 11, 19, 15, 25, 41
7, 18, 9, 10, 11, 17, 22, 23	3	4, 21, 38
8, 16	4	6, 14, 7, 10, 9, 16, 17, 22, 23, 32, 34, 35, 37
21, 27	5	6, 14, 43
26, 40, 58, 50, 38, 52	6	26
29, 30	7	27
33, 37, 35, 44	8	29, 30, 31, 40, 58
36, 42, 43, 48	9	33, 59, 36, 44, 54
45	10	42, 43, 47, 48, 45, 53
47, 56, 67, 53	11	46, 51, 50
49, 54	12	49, 55
55, 59	13	57
57	14	52, 64, 66
60, 66	15	57, 60
57	16	61
61	17	62
63	18	63, 67
65	19	65
70	20	70

Table 8: Clustering under *WAF*.

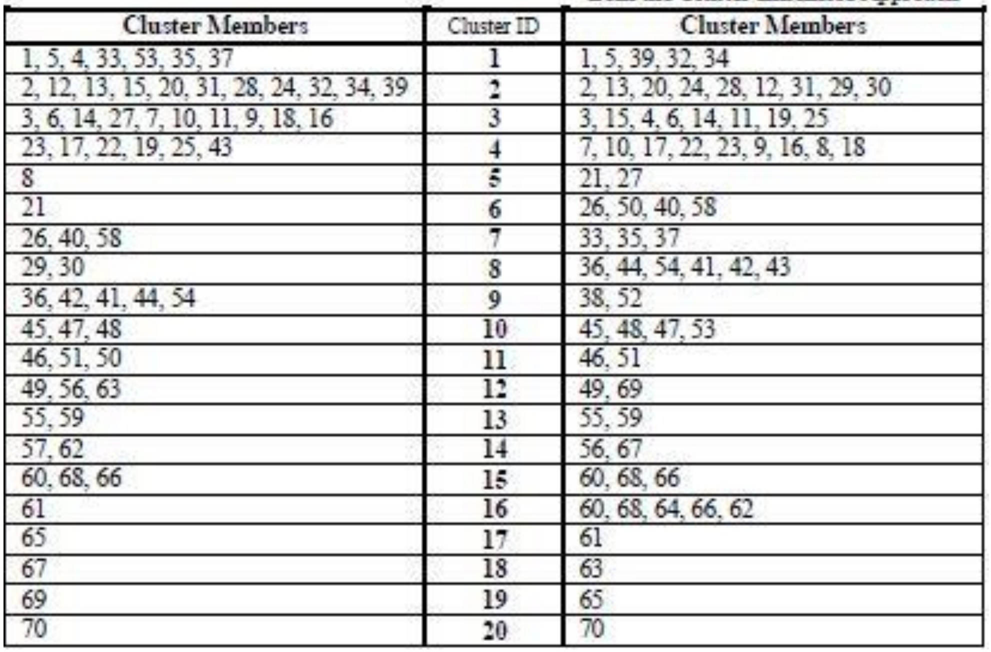
Table 9: Clustering under *BCP*.



Cluster Members	Cluster ID	Cluster Members
1, 5, 4, 3, 15	1	1, 5, 39, 32, 24, 41
2, 24, 28, 13, 20, 31, 39, 51	2	2, 12, 13, 24, 28, 20, 31, 26, 30
6, 17, 14, 22, 23, 7, 10, 11, 19, 25	3	3, 33, 35, 37
8, 9, 16, 18, 35, 12, 32, 34	4	4, 21, 27, 6, 14, 11, 15, 19, 25
21, 27, 38	5	7, 8, 10, 23, 9, 17, 18, 22, 16, 36
26, 50, 33, 37	6	38, 52, 44, 54
29, 30	7	40, 58, 45, 50
36, 43, 54, 41, 42, 44	8	42, 43, 48
40, 58	9	46, 51
45, 53, 47, 48	10	47, 53
46, 49	11	49, 69
52, 64, 66	12	55, 59
55, 59	13	56
56, 67	14	57
57	15	50, 68, 62, 66, 64
60, 63, 68, 69	16	61
61	17	63
62	18	65
65	19	67
70	20	70

Table 10: Clustering under *PCA*.

Table 11: Final Clustering obtained from the Cluster Ensemble Approach



Cluster Members	Cluster ID	Cluster Members
1, 5, 4, 33, 53, 35, 37	1	1, 5, 39, 32, 34
2, 12, 13, 15, 20, 31, 28, 24, 32, 34, 39	2	2, 13, 20, 24, 28, 12, 31, 29, 30
3, 6, 14, 27, 7, 10, 11, 9, 18, 16	3	3, 15, 4, 6, 14, 11, 19, 25
23, 17, 22, 19, 25, 43	4	7, 10, 17, 22, 23, 9, 16, 8, 18
8	5	21, 27
21	6	26, 50, 40, 58
26, 40, 58	7	33, 35, 37
29, 30	8	36, 44, 54, 41, 42, 43
36, 42, 41, 44, 54	9	38, 52
45, 47, 48	10	45, 48, 47, 53
46, 51, 50	11	46, 51
49, 56, 63	12	49, 69
55, 59	13	55, 59
57, 62	14	56, 67
60, 68, 66	15	60, 68, 66
61	16	60, 68, 64, 66, 62
65	17	61
67	18	63
69	19	65
70	20	70

The degrees to which the clustering of the original dataset (by *CDM*) is preserved by each individual dimensionality reduction approach and by the ensemble clustering approach are evaluated using the Rand index [Rand]. The results obtained are shown in Table 12.

Table 12. Agreement between clustering in the original space, each of the DR procedures considered, and ensemble clustering result.

Rand Index	**0.9284**	0.9275	0.9275	0.9275	0.9043	0.9023	0.9006	0.8928	0.8923
DR procedure	Ensemble Clustering	DA	BU	BCP	CA	Var	WAF	TD	PCA

As it can be seen from this table, the clusters obtained from the ensemble clustering approach have the highest agreement with the clusters obtained in the original dimensions. The clusters obtained from the PCA reduction have the lowest agreement with the clusters obtained in the original dimensions. This behavior of the PCA dimensionality reduction technique with respect to cluster preservation has also been observed in [4].

8. Conclusion

This study presents the initial results of the integration of context dependent similarity measures, dimensionality reduction and ensemble clustering approach. The context of two data points, defined with respect to the remaining data points, impacts the evaluation of their similarity. Several dimensionality reduction approaches are considered, each capturing different aspects of the data set. These include seven approaches which reduce the original set of dimensions to a proper subset, as well as the well-known PCA approach in which each dimension in the reduced set is a linear combination of the original dimensions. Experimental results on an artificial data set show that the ensemble clustering agrees the most with the clusters in the original data set. In addition, the clustering in a reduced dimensional space where the dimensions are a proper subset of the original dimensions agrees more with the clustering in the original dimensions set, than that in the set obtained by PCA reduction. These preliminary results warrant further investigation along the lines put forward in this study.

Acknowledgments

Anca Ralescu's contribution to this work was partially supported by University of Oviedo, University of Osaka.

References

[1] D. Achlioptas, "Random Matrices in Data Analysis", In Lecture Notes in Computer Science, vol. 3302, In Proceedings of the 8th European Conference on Principles and Practice of Knowledge Discovery in Databases, **(2004)**, pp. 1-7.

[2] E. F. Combarro, E. Montañés, I. Díaz, J. Ranilla and R. Mones, "Introducing a Family of Linear Measures for Feature Selection in Text Categorization", IEEE Transactions on Knowledge and Data Engineering, vol. 17, no. 9, **(2005)**, pp. 1223-1232.

[3] R. Duda and P. Hart, "Pattern Classification and Scene Analysis", John Wiley, New York, **(1973)**.

[4] X. Z. Fern and C. E. Brodley, "Random Projection for High Dimensional Data Clustering: A Cluster Ensemble Approach", In Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, **(2003)**, pp. 185-192.

[5] A. D. Gordon, "Classification: Monographs on Applied Probability and Statistics", Chapman and Hall, **(1981)**.

[6] J. A. Hartigan, "Clustering Algorithms", Wiley. MR0405726. ISBN 0-471-35645-X, **(1975)**.

[7] A. Jain and R. Dubes, "Algorithms for Clustering Data", Prentice Hall, New Jersey, **(1988)**.

[8] D. MacKay, "Information Theory, Inference and Learning Algorithms", Cambridge University Press, **(2003)**.

[9] B. Manley, "Multivariate Statistical Models: A Primer", Chapman and Hall, **(1986)**.

[10] E. Montañés, E. F. Combarro, I. Díaz and J. Ranilla, "Towards Automatic and Optimal Filtering Levels for Feature Selection in Text Categorization", Lecture Notes in Computer Science, vol. 3646, **(2005)**, pp. 239-248.

[11] E. Montañés, J. R. Quevedo and I. Díaz, "A Wrapper Approach with Support Vector Machines for Text Categorization", Lecture Notes in Computer Science, vol. 2686, (**2003**), pp. 230-237.

[12] E. Montañés, I. Díaz, J. Ranilla, E. F. Combarro and J. Fernández, "Scoring and Selecting Terms for Text Categorisation", IEEE Intelligent Systems, vol. 20, no. 3, **(2005)**, pp. 40-47.

[13] A. Nsang, "Novel Approaches to Dimensionality Reduction: An Empirical Study", Lambert Academic Publishing, Saarbrücken, Germany, **(2011)**.

[14] A. Nsang and A. Ralescu, "Approaches to Dimensionality Reduction to a Subset of the Original Dimensions", In Proceedings of the Twenty-First Midwest Artificial Intelligence and Cognitive Science Conference (MAICS 2010), **(2010)**, pp. 70-77.

[15] A. Nsang and A. Ralescu, "A Review Of Dimensionality Reduction Methods and their Applications", In Proceedings of the Twentieth Midwest Artificial Intelligence and Cognitive Science Conference, **(2009)**, pp. 118-123.

[16] A. Nsang and A. Ralescu, "More Dimensionality Reduction to a Subset of the Original Attribute Set", In Proceedings of the Twenty-First Midwest Artificial Intelligence and Cognitive Science Conference (MAICS 2010), **(2010)**, pp. 109-116.

[17] M. Rawashdeh and A. Ralescu, "A Pairwise Distance View of Cluster Validity", Advances on Computational Intelligence, Springer, **(2012)**, pp. 561-570.

[18] R. S'winiarski, K. Cios and W. Pedrycz, "Data Mining Methods for Knowledge Discovery", Kluwer Academic, ISBN 0-7923-8252-8, **(1998)**.

[19] W. D. Tembe, "Pattern Extraction Using Context Dependent Measure of Divergence and its Validation", MSc Thesis, University of Cincinnati, **(2001)**.

[20] W. Tembe and A. Ralescu, "Fuzzy algorithm for contextual character recognition", In Proceedings of the 2004 IEEE International Conference on Fuzzy Systems, vol. 3, **(2004)**, pp. 1733-1738.

[21] W. D. Tembe, "Proximity Metrics for Contextual Pattern Recognition", PhD Thesis, University of Cincinnati, **(2004)**.