International Journal of Computational Intelligence Systems, Vol.1, No. 4 (December, 2008), 285-298

Clustering feature vectors with mixed numerical and categorical attributes

Roelof K Brouwer Senior Member IEEE
Visiting Professor
Department of Mechanical and Mechatronics Engineering
Stellenbosch University
South Africa

Received: 31-03-2008
Revised: 07-11-2008

Abstract

This paper describes a method for finding a fuzzy membership matrix in case of numerical and categorical features. The set of feature vectors with mixed features is mapped to a set of feature vectors with only real valued components with the condition that the new set of vectors has the same proximity matrix as the original feature vectors. This new set of vectors is then clustered using fuzzy c-means. Simulations show the method to be very effective in comparison with other methods.

Keywords : Fuzzy clustering, gradient descent, categorical, nominal clustering, fuzzy c-means

1. Introduction

The first stage of knowledge acquisition and reduction of complexity concerning a group of objects is to partition or divide the objects into groups based on their attributes or characteristics. “Organizing data into sensible groupings is one of the most fundamental modes of understanding and learning” [1]. “A fundamental operation in data mining is the partitioning of a set of objects represented by data into homogeneous groups or clusters” [2]. Identification of object types is one of the first steps of knowledge acquisition.

Clustering [1, 3, 4] is a popular approach to implementing the partitioning. It is unsupervised classification, aggregation and segmentation [5]. It is the problem of partitioning a set of objects into classes (called clusters) so that (i) the objects belonging to the same class are *similar* and (ii) the objects belonging to different classes are *dissimilar.* By partitioning objects into clusters, interesting groups may be found such as the groups of consumers having a particular property useful for market analysis [6]. Objects can be partitioned into clusters (or called groups) according to their proximity in terms of features. Clusters are then information granules, with each cluster equating to a granule, and hence can be used in computationally intelligent systems as units of learning and reasoning. The word “proximity” is used as a general term representing both similarity and dissimilarity.

Unsupervised clustering has been extensively studied in machine learning, databases, and statistics from various perspectives. Many applications of clustering have been discussed and many clustering techniques have been developed. There are two main clustering methods. Statistical clustering methods [3, 4, 7] partition objects according to some proximity measures, whereas conceptual clustering methods cluster objects according to the concepts the objects carry [8, 9]. For automatic clustering, both a method of determining proximity (similarity or dissimilarity) between feature vectors and a method for determining representatives (prototypes) of clusters are generally required.

Proximity of objects is generally based on their feature vectors. These feature values or attributes may be of ratio scale, interval scale, ordinal scale or categorical scale [10]. Grouping the first two together, feature values can also be divided into 3 types : numerical, ordinal and categorical. Numerical feature values are well understood. Examples of values of an ordinal feature are labels such as *Infant*, *Child* and *Adult* that can be ordered. An example of a categorical feature is colour with values such as *Red*, *Blue* and *Yellow* where no ordering is sensible unless based on position in the frequency spectrum.

The problem of clustering becomes very challenging when the data is ordinal or categorical, that is, when there is no inherent proximity measure between data values. Often ordinal features are treated as either numeric or categorical. Both of these approaches are incorrect. Numerical feature values are easily handled in clustering because values can be summed and compared. Summing and division allows the center (centroid or prototype) of a set of values to be computed. The dissimilarity between two numerical values can be easily computed by taking the difference. Proximity is a general term for similarity and dissimilarity but without loss of generality we will restrict ourselves to dissimilarity from now on. Categorical feature values do not have any relationship except for equality among them, and hence the dissimilarity between two values cannot be readily defined except in terms of equality. There is no ordering. The distance between two different values, e.g., *Male* and *Female*, can be defined as 1, and the distance between two identical values can be defined as 0. Ordinal feature values are similar to categorical values in that arithmetic operations do not make sense and a method for finding the center of a set of values is not obvious.

Clustering may be either crisp or fuzzy. In the former case, each object is placed in one and only one cluster. In the latter case, an object is assigned to all clusters to varying degrees. This degree may be close to zero and even zero for some feature vectors and clusters. A commonly used method of strictly crisp clustering is the k-means. Fuzzy versions of the k-means algorithm are due to Ruspini [11] and Bezdek [12], where each object is allowed to have membership values for all clusters rather than having a distinct membership in exactly one cluster. The membership matrix is a generalization of the membership matrix obtained in crisp clustering in which case it may be called a characteristic matrix. If, after fuzzy clustering, it is still desired to put an object into a single cluster the cluster/class/group assigned to an object can be chosen to be the cluster in which the object has maximum membership value. The process is then that of fuzzy clustering but the result is crisp clusters. Fuzzy clustering in this instance is used as an intermediate step to reduce the effect of noise. Working only on numeric data limits the use of these k-means-type algorithms in such areas as data mining where large categorical data sets are frequently encountered. Categorical and ordinal data is plentiful in real-world databases.

Representing clusters requires determination of cluster prototypes as clustering takes place. Using cluster prototypes has both an advantage and a disadvantage. The advantage lies in the fact that it is not required that dissimilarity between all object pairs be determined. The disadvantage is that a way of aggregating feature vectors is required. Thus two operations rather than one have to be defined. This may not represent a problem in case of feature vectors where all the features are of the ratio scale or interval scale variety. In that case, addition of numbers is a logical form of aggregation. In case of ordinal and categorical features however, we may have a problem and it may be useful to drop the aggregation requirement. Sometimes feature vectors are not even available. The data in some psychometric applications are collected as proximities only [1]. In that case relational clustering methods apply[13-17]. Converting a matrix of feature vectors into a proximity matrix requires that dissimilarity between individual feature values is measurable. The dissimilarity between feature vectors is then an aggregation of the initial proximities.

Following this introduction the paper commences with a brief review of the literature. Next is a brief description of fuzzy c-means. This is followed by a description of the proposed method. A discussion on clustering quality measures follows next. Results of simulations and experiments are provided next followed by a description of future work, conclusion and summary.

To allow variable names of more than one letter and thereby permit variable names to be mnemonic, all operations are denoted by explicit operators. Multiplication, for example, is defined explicitly using the operator ×. Implicit multiplication will not exist and *na* for instance is just a variable name. Names of arrays are in bold font. Rank-1 array variables are in lower case and names of arrays of rank greater than 1 are in capital. Division and multiplication between arrays are generally between the components of the arrays.

2. Previous work

Many algorithms have been developed for clustering categorical data [4, 8, 18-43]. Algorithms for clustering categorical data include hierarchical clustering methods using Gower's proximity coefficient [42] or other proximity measures [24], the PAM algorithm [4], the fuzzy-statistical algorithms [43], and the conceptual clustering methods [8].

Some of the work will now be briefly described as noted by the authors in their abstracts. Ralambondrainy [41] presents an approach to using the k-means algorithm to cluster categorical data. His approach is to convert multiple category attributes into binary attributes (using 0 and 1 to represent either a category absent or present) and to treat the binary attributes as numeric in the k-means algorithm. A drawback is that the cluster means, given by real values between 0 and l, do not indicate the characteristics of the clusters. Huang in his paper [20] presents two algorithms that extend the k-means algorithm to categorical domains and domains with mixed numeric and categorical values. The k-modes algorithm uses a simple matching dissimilarity measure to deal with categorical objects, replaces the means of clusters with modes, and uses a frequency-based method to update modes in the clustering process to minimize the clustering cost function. With these extensions the k-modes algorithm enables the clustering of categorical data in a fashion similar to k-means.

The authors of [26] also describe extensions to the fuzzy k-means algorithm for clustering categorical data. By using a simple matching dissimilarity measure for categorical objects and modes instead of means for clusters, a new approach is developed, that permits the use of the k-means paradigm to efficiently cluster large categorical data sets. Guan etal. [27] define new distance based on the improved Levenshtein distance with the tolerance relation for incomplete categorical data, and a new dissimilarity strategy for incomplete numerical data. Li etal. [29] present a novel clustering algorithm for mixed data sets by modifying the common cost function; the trace of the within cluster dispersion matrix. A genetic algorithm (GA) is used to optimize the new cost function to obtain valid clustering result.

In [32] the clustering technique uses an FCM-type simple iterative algorithm that includes a quantification step. In the quantification step, the category scores are derived so that they suit FCM clustering considering cluster centers and memberships. Li etal. [34] in their paper study the entropy-based criterion in clustering categorical data. They show that the entropy-based criterion can be derived in the formal framework of probabilistic clustering models and establish the connection between the criterion and the approach based on dissimilarity coefficients. The authors in [35] introduce clustering based on compressed data that is an extension of the Birch algorithm. Its main characteristics refer to the fact that it can be especially suitable for very large databases and it can work both with categorical attributes and mixed features. The authors in [36] develop an ensemble based mixed attribute cluster model for mixed numeric and categorical databases based on the cluster ensemble method. The method has excellent scalability according to its originators.

Ramakrishna and Minho [37] propose an algorithm, although hierarchical in essence, that avoids the characteristic error propagation through reassignment and deletion of bad clusters. They also propose new indices for cluster validation in categorical datasets, an area that is almost unexplored.

Ahmad and Dey [39] propose a method to compute distance between two attribute values of the same attribute for unsupervised learning. This approach is based on the fact that similarity of two attribute values is dependent on their relationship with other attributes. They use the proposed distance measure with k-mode clustering algorithm to cluster various categorical data sets.

3. Fuzzy c-means

The traditional fuzzy clustering method is called the fuzzy c-means (FCM) [12, 44]. Fuzzy clustering is less sensitive to noise than crisp clustering and may be used even when the desired result is crisp clusters. This method of clustering requires that attributes be numeric. FCM in the main consists of repeatedly determining prototypes for the clusters to be found and calculating membership values for the feature vectors in the clusters. Every attribute value in a prototype is the weighted mean over all members of the data set to be clustered with weights equal to a power of the degree to which an object belongs to a cluster.

Formally, let the weights or membership values for clusters be designated by $\boldsymbol{M}_{p,k}$; the membership of feature vector p in cluster k. Also let $\boldsymbol{F}_{p,a}$ represent the parameters for the attribute values themselves within feature vectors. The a^{th} attribute of the prototype for cluster k is

$$\boldsymbol{P}_{k,a} = \frac{\sum_{p=1}^{np} \boldsymbol{M}_{p,k}^{m} \times \boldsymbol{F}_{p,a}}{\sum_{p=1}^{np} \boldsymbol{M}_{p,k}^{m}} \qquad (1)$$

The subscripts have the following meaning in (1) and throughout the paper.

$p = 1..np$	identifies feature vectors
$a = 1..na$	identifies attributes
$k = 1..nk$	identifies clusters
m	fuzziness parameter

The exponent m relates to the of degree of fuzziness desired. Crisp clustering is achieved with m equal to 1. A commonly used value for fuzzy clustering is 2.The cluster membership values are calculated in terms of dissimilarity between feature vectors and the prototypes as

$$\boldsymbol{M}_{i,j} = \frac{\dfrac{1}{\boldsymbol{D}_{i,j}^{\frac{2}{m-1}}}}{\sum_{l} \dfrac{1}{\boldsymbol{D}_{i,l}^{\frac{2}{m-1}}}} \qquad (2)$$

$\boldsymbol{D}_{p,k}$ is the distance or dissimilarity between feature vector *p* and prototype *k.* Aggregation of feature vectors and therefore of feature values is required in (1); thereby constraining the use of FCM. As argued before aggregation may not always be possible. The required arithmetic operations may not be defined or the feature vectors may not be defined and only dissimilarities between feature vectors are known. In this case another approach has to be pursued. This is the essence of this paper and is explained next.

4. Proposed Method

This section introduces a method for clustering of categorical feature vectors that is based on replacing the original set of feature vectors that contain both categorical and numeric features with another set of vectors in $(\Re^{+})^{q}$ where *q* is some fraction of the number of components in the original feature vectors. This new set of vectors, that has only numeric components, is then clustered using FCM. Applying FCM at this juncture is now possible because the new set of vectors, the rows of ***F`***, have only numeric components. FCM is desirable since it is very effective if the input is of the right form ie. all components are of interval scale.

4.1. *Distance between two mixed feature vectors*

Unless a dissimilarity matrix, $\boldsymbol{D}^{(\boldsymbol{F})}$, for the original set of feature vectors to be clustered is available the distance between two mixed feature vectors is found as follows.

Distance for mixed feature vectors

Consider two feature vectors with both numerical and categorical features $(n_{1,1}..n_{1,nna},\ c_{1,1},..c_{1,nca})$ and $(n_{2,1},..n_{2,nna},\ c_{2,1},..c_{2,nca})$. Then the distance is

$$\sqrt{\sum_{i=1}^{nna}(\boldsymbol{n}_{1,i}-\boldsymbol{n}_{2,i})^{2}+\sum_{i=1}^{nca}(\boldsymbol{c}_{1,i}\neq\boldsymbol{c}_{2,i})}=\sqrt{dn^{2}+dc^{2}} \qquad (3)$$

$\boldsymbol{n}_{1,i},\boldsymbol{n}_{2,i}\quad i=1..nna$ and $\boldsymbol{c}_{1,i},\boldsymbol{c}_{2,i}\quad i=1..nca$ are the numerical and categorical attributes respectively for the two feature vectors.

The numerical attributes should be pre-processed such that all are in [0,1]. This is so that values of numerical attribute differences will be of the same order of magnitude as values of the categorical attribute differences. A definition of aggregation of categorical feature values is not required here since prototypes do not need to be determined.

If ***F*** is the matrix whose rows are the feature vectors with mixed numerical and categorical features then $\boldsymbol{D}^{(\boldsymbol{F})}$ a dissimilarity matrix based on feature vectors, is defined by (4).

$$\boldsymbol{D}^{(\boldsymbol{F})}_{i,j}=\boldsymbol{D}(\boldsymbol{F}_{i,\cdot},\boldsymbol{F}_{j,\cdot})=\sqrt{\sum_{s=1}^{nk}\left(\boldsymbol{F}_{i,s}-\boldsymbol{F}_{j,s}\right)^{2}}\quad i,j=1..np(4)$$

$\boldsymbol{F}_{i,s}$ is the value for the s^{th} attribute of the i^{th} feature.

4.2. *Mapping the original set of Feature Vectors, F, to another set of Feature vectors F' using Gradient Descent*

A method for determining ***F'*** from ***F*** or from $\boldsymbol{D}^{(\boldsymbol{F})}$, if it is available, is to use gradient descent on the sum of the squares of the errors between the two dissimilarity matrices, $\boldsymbol{D}^{(\boldsymbol{F})}$ and $\boldsymbol{D}^{(\boldsymbol{F}')}$ as in

$$e=\frac{1}{2}\times\sum_{i,j}\boldsymbol{E}_{i,j}^{2} \qquad (5)$$

With

$$\boldsymbol{E}=\boldsymbol{D}^{(\boldsymbol{F})}-\boldsymbol{D}^{(\boldsymbol{F}')} \qquad (6)$$

since dissimilarity has to be preserved. $\boldsymbol{D}^{(\boldsymbol{F}')}$ and $\boldsymbol{D}^{(\boldsymbol{F})}$ are the distance matrices for ***F'*** and ***F*** respectively. Since the distance function is symmetric and distances between identical feature vectors are zero we really only need to be concerned with the part of the matrices above the main diagonal. The error, *e*, will be 0 if and only if $\boldsymbol{D}^{(\boldsymbol{F}')}=\boldsymbol{D}^{(\boldsymbol{F})}$.

Now

$$(\nabla_{\boldsymbol{E}}e)_{i,j}=\frac{\partial e}{\partial\boldsymbol{E}_{i,j}}=\boldsymbol{E}_{i,j} \qquad (7)$$

To ensure non-negative values of the components of ***F'*** we can define it in terms of an auxiliary matrix variable ***X*** as ***F'=X*** ^{2}. The values of ***X*** are updated according to (8).

$$\boldsymbol{X}\leftarrow\boldsymbol{X}-\mu_{0}\times\nabla_{\boldsymbol{X}}e \qquad (8)$$

The expression for the gradient is given by (9)

Published by Atlantis Press

$$(\nabla_{X} e)_{i,j} = -2\times\left(\sum_{q}\left(\left(\frac{E}{D^{(F')}}\right)_{i,q}+\left(\frac{E}{D^{(F')}}\right)_{i,q}^{T}\right)\times\left(F_{i,j}-F_{q,j}\right)\right)\times X_{i,j} \quad (9)$$

Derivation of $\nabla_{X} e$

According to the chain rule

$$(\nabla_{X} e)_{i,j} = \frac{\partial e}{\partial X_{i,j}} = \sum_{r,s}\frac{\partial e}{\partial F'_{r,s}}\times\frac{\partial F'_{r,s}}{\partial X_{i,j}} = \sum_{r,s}(\nabla_{F'} e)_{r,s}\times(\nabla_{X} F')_{r,s,i,j} \quad (10)$$

and

$$(\nabla_{F'} e)_{r,s} = \frac{\partial e}{\partial F'_{r,s}} = \sum_{p,q}\frac{\partial e}{\partial D^{(F')}_{p,q}}\times\frac{\partial D^{(F')}_{p,q}}{\partial F'_{r,s}} = \sum_{p,q}(\nabla_{D^{(F')}} e)_{p,q}\times(\nabla_{F'} D^{(F')})_{p,q,r,s} \quad (11)$$

and

$$(\nabla_{D^{(F')}} e)_{p,q} = \frac{\partial e}{\partial D^{(F')}_{p,q}} = \sum_{t,u}\frac{\partial e}{\partial E_{t,u}}\times\frac{\partial E_{t,u}}{\partial D^{(F')}_{p,q}} = \sum_{t,u}(\nabla_{E} e)_{t,u}\times(\nabla_{D^{(F')}} E)_{t,u,p,q} \quad (12)$$

Now

$$(\nabla_{E} e)_{t,u} = E_{t,u} \quad (13)$$

It can be shown that

$$(\nabla_{D^{(F')}} E)_{t,u,p,q} = -\delta_{t,p}\times\delta_{u,q} \quad (14)$$

$\delta_{i,j}$ is the Kronecker delta function that is equal to 1 if *i* is equal to *j* and 0 otherwise. By substitution of (13) and (14) into (12) we get

$$(\nabla_{D^{(F')}} e)_{p,q} = -\sum_{t,u} E_{t,u}\times\delta_{t,p}\times\delta_{u,q} = -E_{p,q} \quad (15)$$

Now it can be shown that

$$(\nabla_{F'} D^{(F')})_{r,s,p,q} = \frac{(F'_{p,s}-F'_{q,s})\times(\delta_{p,r}-\delta_{q,r})}{D^{(F')}_{p,q}} \quad (16)$$

By substitution of (15) and (16) into (11) we get

$$(\nabla_{F'} e)_{r,s} = -\sum_{p,q}\frac{E_{p,q}}{D^{(F')}_{p,q}}\times(F'_{p,s}-F'_{q,s})\times(\delta_{p,r}-\delta_{q,r})$$

$$= -\left(\sum_{q}\frac{E_{r,q}}{D^{(F')}_{r,q}}\times\left(F'_{r,s}-F'_{q,s}\right)+\sum_{p}\frac{E_{p,r}}{D^{(F')}_{p,r}}\times\left(F'_{r,s}-F'_{p,s}\right)\right)$$

$$= -\sum_{q}\left(\frac{E_{r,q}}{D^{(F')}_{r,q}}+\left(\frac{E_{r,q}}{D^{(F')}_{r,q}}\right)^{T}\right)\times\left(F'_{r,s}-F'_{q,s}\right) \quad (17)$$

It can be shown that

$$(\nabla_{X} F')_{r,s,i,j} = 2\times X_{r,s}\times\delta_{r,i}\times\delta_{s,j} \quad (18)$$

By substitution of (17) and (18) into (10) we get

$$(\nabla_{X} e)_{i,j} = -2\times\sum_{r,s}\left(\sum_{q}\left(\frac{E_{r,q}}{D^{(F')}_{r,q}}+\left(\frac{E_{r,q}}{D^{(F')}_{r,q}}\right)^{T}\right)\times\left(F_{r,s}-F_{q,s}\right)\right)\times X_{r,s}\times\delta_{r,i}\times\delta_{s,j}$$

$$= -2\times\left(\sum_{q}\left(\left(\frac{E}{D^{(F')}}\right)_{i,q}+\left(\frac{E}{D^{(F')}}\right)_{i,q}^{T}\right)\times\left(F_{i,j}-F_{q,j}\right)\right)\times X_{i,j} \quad (19)$$

Q.E.D.

Feedback for controlling the number of iterations is provided by the root mean squared error defined by (20).

$$RMSE = \sqrt{\frac{\sum_{i,j=1}^{np}(D^{(F')}_{i,j}-D^{(F)}_{i,j})^{2}}{np\times np}} \quad (20)$$

The algorithm may be summarized as follows. For controlling the number of iterations the root mean square error defined in (20) is used.

```
                         Algorithm
Initialize
    $D^{(F)}$ using (3) and (4)
    $X$ using random numbers
    $F' = X^2$
    $D^{(F')}$ using (3) and (4)
    $E$ using (6)
    RMSE using (20)
while (RMSE > predetermined amount) and (number of iterations < than predetermined number)
        determine
            $\nabla_X e$ using (9)
            X using (8)
            $F' = X^2$
            $D^{(F')}$ using (3) and (4)
            $E$ using (6)
            RMSE using (20)
end while
```

The number of components required for ***F'*** is not large and 5 components is sufficient to capture the dissimilarities between feature vectors as demonstrated by simulations

Another method for determining vectors given inter vector distances is a class of methods called multidimensional scaling [45].

5. Clustering Quality Measures Utilized for Comparing Methods

Determining the quality of clustering can be done by comparing the result of clustering, the cluster partition, to a known correct partition, the class partition, if it is available. A key to comparing two crisp partitions, $\boldsymbol{P}^{(1)}$ and $\boldsymbol{P}^{(2)}$, is the contingency table, $\boldsymbol{C}$. Entry $\boldsymbol{C}_{i,j}$ is the number of objects in subset i of $\boldsymbol{P}^{(1)}$ and subset j of $\boldsymbol{P}^{(2)}$. Interestingly $\boldsymbol{P}^{(1)}$ and $\boldsymbol{P}^{(2)}$ can also be determined from $\boldsymbol{C}$ to within a labelling of the elements.

Several well known measures for comparing two partitions are obtained from the contingency table including the Rand index (*RI*) [46], Adjusted Rand index(*ARI*) [47-49] and Jaccard index (*JI*) [34]. If one of the partitions is an assumed correct class partition and the other partition is the result of clustering, then these measures become clustering quality measures. They are defined in terms of the following parameters as shown. The quality indices are measured in terms of bonded pairs of elements. For two elements to be bonded in a partition means that they are in the same subset of the partition.

The total number of unordered pairs of elements, including both bonded and unbonded, is

$$np = \binom{n}{2} \tag{21}$$

The total number of unordered pairs of elements bonded in both $\boldsymbol{P}^{(1)}$ and in $\boldsymbol{P}^{(2)}$ is

$$n^{(1,2)} = \sum_j \sum_i \binom{C_{i,j}}{2} \tag{22}$$

$\binom{\boldsymbol{C}_{i,j}}{2}$ is the number of unordered pairs in set i of $\boldsymbol{P}^{(1)}$ and in set j of $\boldsymbol{P}^{(2)}$. The total number of unordered pairs of elements bonded in $\boldsymbol{P}^{(1)}$ is

$$n^{(1)} = \sum_i \binom{\sum_j C_{i,j}}{2} \tag{23}$$

The total number of unordered pairs of elements bonded in $\boldsymbol{P}^{(2)}$ is

$$n^{(2)} = \sum_j \binom{\sum_i C_{i,j}}{2} \tag{24}$$

The following parameters that make up the indices can then be defined

$$a = n^{(1,2)} \tag{25}$$

a is the number of unordered pairs of objects whose components are bonded in $\boldsymbol{P}^{(1)}$ and also bonded in $\boldsymbol{P}^{(2)}$

$$b = n^{(1)} - n^{(1,2)} \tag{26}$$

b is the number of unordered pairs of objects whose elements are bonded only in $\boldsymbol{P}^{(1)}$

$$c = n^{(2)} - n^{(1,2)} \tag{27}$$

c is the number of unordered pairs of objects whose elements are bonded only in $\boldsymbol{P}^{(2)}$

$$d = np - (n^{(1)} + n^{(2)} - n^{(1,2)}) \tag{28}$$

d is the number of unordered pairs of objects whose elements are not bonded in $\boldsymbol{P}^{(1)}$ and not bonded in $\boldsymbol{P}^{(2)}$.

The Rand index(*RI*)[46] is then defined as

$$RI = \frac{a+d}{a+b+c+d} \tag{29}$$

The adjusted Rand index (*ARI*) [47-49] corrects the *RI* to give a constant expected value of 0 and may be calculated according to the formula (30)[50]

$$ARI = \frac{2\left(a\times d - b\times c\right)}{c^2+b^2+2\times a\times d+\left(a+d\right)\times\left(c+b\right)} \tag{30}$$

or more intuitively

$$ARI = \frac{n^{(1,2)} - \dfrac{n^{(1)}\times n^{(2)}}{np}}{\dfrac{n^{(1)}+n^{(2)}}{2} - \dfrac{n^{(1)}\times n^{(2)}}{np}} \tag{31}$$

The Jacquard index(*JI*) [47] defined as:

$$JI = \frac{a}{a+b+c} \tag{32}$$

Another commonly accepted clustering quality measure based on comparing a cluster partition to an assumed correct class partition is the cluster purity index. A cluster is defined as pure if all the elements in the cluster come from one class. If all the clusters are pure then each column in the contingency matrix will have exactly one non-zero entry. A measure of this is clustering purity (*CPI*) defined as in (33).

$$CPI = \frac{\sum_j \max_i \boldsymbol{C}_{i,j}}{n} \tag{33}$$

The variable n is the total number of elements. The maximum value for *CPI* is 1 when the maximum in each column is the only non-zero entry. This measure assumes that one of the partitions, the class partition, is correct . If no such assumption is made and both partitions play the same role as they do in the *RI*, the *ARI* and *JI* we can define an index , called purity index, that is the average of class and cluster purity as in

$$PI = \frac{\sum_j \max_i \boldsymbol{C}_{i,j} + \sum_i \max_j \boldsymbol{C}_{i,j}}{2\times n} \tag{34}$$

This will be 1 if there is exactly one non-zero entry in each row and each column. The *PI* penalizes for the number of sets in the partitions not being equal as do the *RI*, *ARI*, and *JI*.

Huang etal. in [26] and later Kim et al. in [51] a use a measure that is equivalent to Cluster Purity and called Huang's accuracy measure by Kim etal.

$$H = \frac{\sum_j a_j}{n} \tag{35}$$

a_j is the number of elements in cluster j that belong to the correct class. The correct class for a cluster is defined to be the class with the maximum number of elements in the cluster.

An example of a contingency matrix for 71 objects is given below in Table 1. In this case *CPI* is equal to 34/56 or 0.607 if we assume that $\boldsymbol{P}^{(1)}$ is the class partition and that $\boldsymbol{P}^{(2)}$ is the cluster partition.

Table 1 Example contingency matrix

	$\boldsymbol{P}^{(2)}$			
$\boldsymbol{P}^{(1)}$	3	2	1	8
	7	3	3	1
	2	1	9	2
	1	10	2	1
	5	3	5	2

In case of the example *RI*=0.700, *ARI*=0.200, *JI*=0.242, *CPI*=0.607 and *PI*=. 0.514

There are more clusters than classes in the example and CPI does not penalize the situation where the number of clusters is unequal to the number of classes.

In [52] the authors recommended the adjusted Rand index as the index of choice after many different indices were evaluated for measuring agreement between two partitions in cluster analysis with different numbers of clusters [53].

6. Simulations

Simulations are performed to determine the effectiveness of the proposed method that is applied to both artificial and real data sets. The proposed method, **Method 0**, is compared to other methods implemented by the author here that will be referred to as **Methods 1** and **2**. **Method 1** consists of replacing each categorical attribute with a number of binary attributes equal to the number of categories as proposed by Ralambondrainy [31]. **Method 2**, the naïve method, consists of replacing categories by numbers that are then treated as being numerical. Results show that this simple method can do just as well or better than some more complex methods and is therefore included in the comparison.

Results for the proposed method are also compared to results obtained by other researchers using their own proposed methods. The methods proposed by these researchers are not implemented and only their values for certain indices that they have provided are used. This means that values for some indices for some methods are not available since contingency tables are not provided so that the values can be calculated.

The fuzziness index, *m,* used in FCM, is kept at 2 throughout. The real data sets are from the UCI Machine Learning Repository [55].

6.1. *Methods 0,1, and 2 applied to an Artificial Data set*

The first simulation involves applying the method of this paper to an artificial data set. This data set is generated by first producing 5 vectors with 20 randomly produced components of integers between 0 and 19. Each of these five vectors is then copied 20 times producing a matrix of dimension 100 by 20 and 5 classes with 20 elements in each. To this matrix of feature vectors is added a binary matrix of the same dimension. The probability of a 1 in this matrix is 80%. The resulting matrix is the data set that is used with the integers treated as categories. Adding a 1 is as drastic as adding any other number except for 0 since not differences but only equality is considered for measuring distance between categorical values in **Method 0**.

For the learning of the distance-equivalent feature vectors, the learning rate was 0.001. The total number of iterations, to obtain ***F'***, was 1000. The number of components in the original feature vectors was 20. The number of components in the distance-equivalent feature vectors is set to 5. The reduction in the root mean square error over the iterations is shown in **Figure 1**. The root mean square error is over the differences between the components of the two dissimilarity matrices as defined by (20).

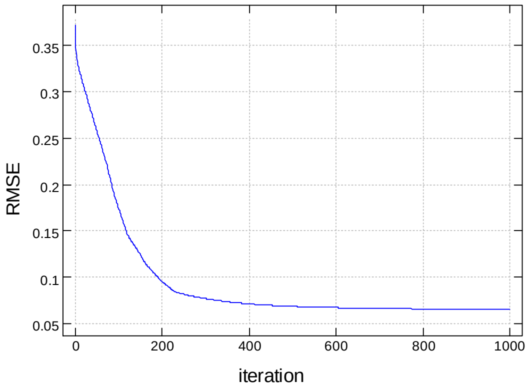

Figure 1 Training schedule to obtain distant equivalent vectors for artificial data set

Below in **Table 2** is the contingency matrix for method 0 applied to the artificial data set.

Table 2 Contingency matrix for artificial data and method 0

cluster class	0	1	2	3	4
0	0	0	0	**20**	0
1	**20**	0	0	0	0
2	0	**20**	0	0	0
3	0	0	0	0	**20**
4	0	0	**20**	0	0

The contingency matrix for method 1 and the same input is in **Table 3**.

Table 3 Contingency matrix for artificial data and method 1

Cluster class	0	1	2	3	4
0	0	0	0	0	**20**
1	0	0	0	5	**15**
2	6	**14**	0	0	0
3	**8**	6	0	0	6
4	6	0	0	0	**14**

There are 3 maximum in column 4 corresponding to class 4. The contingency matrix for method 2 is in **Table 4**. The contingency matrix is the basis for all the clustering quality measures here.

Table 4 Contingency matrix for artificial data and method 2

cluster class	0	1	2	3	4
0	**15**	0	0	0	5
1	5	0	0	0	**15**
2	0	**20**	0	0	0
3	0	0	0	**20**	0
4	0	0	**20**	0	0

A comparison of the 3 methods applied by the author is given in **Table 5** where CPI means cluster purity, *RI* is the rand index, *ARI* is the adjusted rand index and *JI* is the Jaccard index.

Table 5 Comparison of methods applied by author to the artificial data set

Quality Measures Method	*CPI*	*RI*	*ARI*	*JI*
0	1	1	1	1
1	0.47	0.66	0.20	0.26
2	0.90	0.94	0.80	0.73

This shows that **Method 0** is far superior to the other two methods. What is interesting is that the naiive method, 2, is superior to method 1 proposed by Ralambondrainy in this case.

6.2. *Small Soybean data set*

The second simulation involves applying the method of this paper to a small subset of the original soybean database. The soybean disease data set [12] is used because all attributes of the data can be treated as categorical. The data base is from the UCI Machine Learning Depository[54]. The data set has 47 records, each being described by 35 attributes. The number of missing attribute values is zero. Each record is labeled with one of the four diseases: Diapehhe Stem Canker, Charcoal Rot, Rhizoctonia Root Rot, and Phytophthora Rot. Except for phytophthora Rot which has 17 records, all other diseases have 10 records each.

A plot of a set of feature vectors with two components and the same distance matrix as the distance matrix for the original set of feature vectors is shown in Figure 2. This 2-dimensional plot is also obtained by multidimensional scaling. The plot shows an inherent clustering.

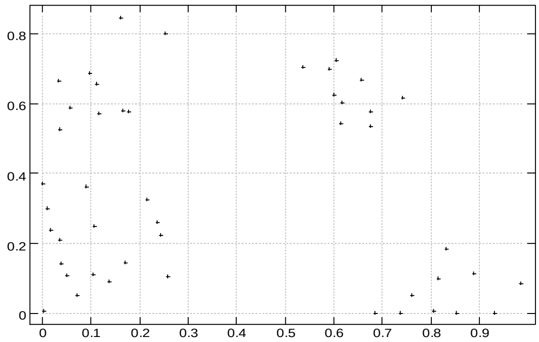

Figure 2 Plot of 2-D distance- equivalent feature vectors for small soybean data set

To obtain an equivalent set of feature vectors with just 4 real components, rather than 5, using the method of this paper required the training schedule as demonstrated in Figure 3.

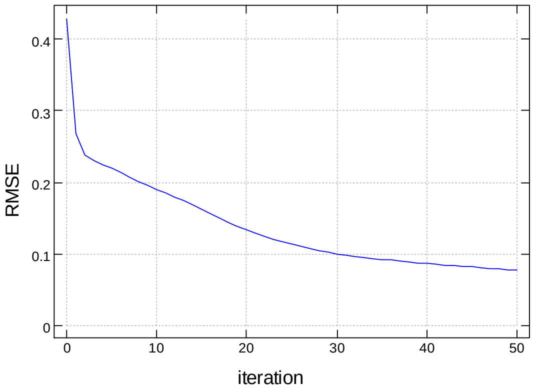

Figure 3 Training schedule for determining the set of distance equivalent feature vectors for the small soybean data set

The result after FCM was applied is the contingency matrix in **Table 6**. This shows a clustering result identical to the classes. Less than 4 components was not found to be sufficient. The learning rate was obtained by trial and error.

Table 6 Contingency matrix using proposed method (Method 0) on small soybean data set

clusters Classes	0	1	2	3
0	0	**10**	0	0
1	0	0	**10**	0
2	**10**	0	0	0
3	0	0	0	**17**

The various clustering indices for the 3 methods are as in Table 7. Three other methods for which only cluster purity measures are provided are found in [26]. The fuzzy k-modes has been developed by Huang and Ng in the previously mentioned paper. In the table NA means not available in reference [26].

Table 7 Comparison of clustering quality indices for the various methods using small soybean data set

Quality Measures Method	CPI	RI	ARI	JI
0	1	1	1	1
1	0.81	0.85	0.65	0.61
2	0.79	0.83	0.55	0.49
3 Fuzzy k-modes[26]	0.79	NA	NA	NA
4 (Conceptual k-means) [26]	0.70	NA	NA	NA
5 (Hard k-modes) [26]	0.78	NA	NA	NA

Again Method 0 is superior in terms of all the cluster quality indices. Methods 1 and 2 are similar in terms of the indices.

6.3. *Large soybean data set*

The same methods as above were also applied to the large soybean data set also found in the machine learning data base. The number of records in this case is 307. The attributes are the same as before. There are 19 classes in this case. The folklore seems to be that the last four classes are unjustified by the data since they have so few corresponding instances. There are 705 missing attribute values , that in the authors methods, rightly or wrongly, are treated as a distinct category. The plot below is a plot of 2-dimensional vectors that have the intra distances as the original data vectors. It is obvious that clusters are not very definite in this case.

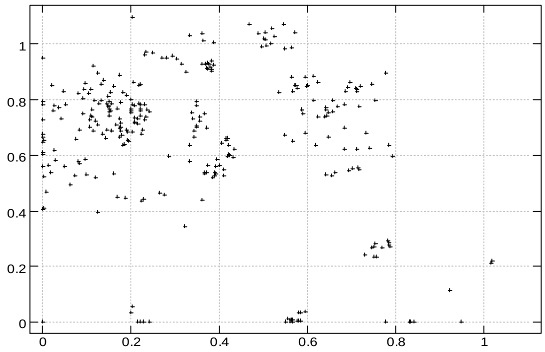

Figure 4 Plot of equivalent vectors for feature vectors in large soybean data set

The total number of iterations required to obtain equivalent feature vectors with 4 components is 5000. The learning rate was 0.0001. The learning schedule required to obtain the distant equivalent feature vectors is shown in **Figure 5**. Using distant-equivalent feature vectors with fewer than 4 components was not found to be successful.

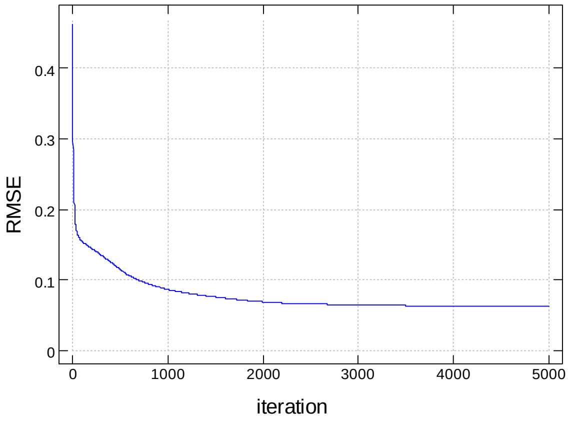

Figure 5 Schedule for determining distant equivalent feature vectors

A comparison of the 3 methods applied by the author is given in **Table** 8.

Table 8 Comparison of the three methods applied by author to large soybean data set

Quality Index Method	*CPI*	*RI*	*ARI*	*JI*
0	0.62	0.92	0.35	0.24
1	0.27	0.56	0.14	0.15
2	0.49	0.79	0.24	0.20

Again **Method 0** is superior by all counts to the other 2 methods. The naiive method, method 2, is superior to method 1 proposed by Ralambondrainy in this case.

6.4. *Zoo Data Base*

Another data set to which the method of this paper is applied is also from the Machine Learning Database. It is referred to as the Zoo database. The number of instances is 101. The number of attributes is 16. Even though one of the attributes, the number of legs, is numeric, differences in this value are not meaningful as far as class is concerned and this attribute is therefore treated as categorical. The number of classes of animals is 7.

To determine the distance equivalent vectors with 4 components 5000 iterations were used. The learning rate used was 0.001. The number of clusters allowed was 7. The contingency matrix is in **Table 9.**

Table 9 Contingency matrix for zoo database and method 0

cluster	0	1	2	3	4	5	6

class							
0	0	22	19	0	0	0	0
1	0	0	0	0	0	0	13
2	0	0	0	20	0	0	0
3	0	0	0	0	2	8	0
4	0	0	0	0	8	0	0
5	2	0	0	0	0	1	0
6	4	0	0	0	0	0	1

A comparison of results for 5 methods including 2 provided by another author is shown in **Table 10**.

Table 10 Comparison of methods applied by author and other researchers to zoo data set

Quality index method	*CPI*	*RI*	*ARI*	*JI*
0	0.94	0.90	0.69	0.60
1	0.87	0.69	0.04	0.12
2	0.86	0.68	0.02	0.12
3 (Entropy based) [34]	0.90	NA	NA	NA
4 (k-means) [34]	0.84	NA	NA	NA
5 k-mode [51]	0.60	NA	NA	NA
6 Fuzzy k-modes[51]	0.64	NA	NA	NA
7Kim's fuzzy centroids [51]	0.75	NA	NA	NA

In the table, NA, means not available. The results in 3 and 4 are from paper [34] and in rows 5-7 are from [51] . Again Method 0 proves to be better in terms of partitioning accuracy. Methods 1 and 2 are similar in terms of the indices. Reference [34] that describes an entropy-based method, did not have values for RI, ARI, and JI nor did the other methods.

6.5. *Australian Credit Data Base*

This dataset contains results of credit card applications. There are 690 instances with 6 numerical and 8 categorical attributes. There are 2 classes. This data base was generated by Quinlan [56, 57] and may be retrieved from [54]. Kim etal. [51] refer to the Australian credit database in their test but do not use the entire data set but only 202 applicants and 9 attributes as opposed to 690 instances and 14 attributes. The authors do not describe how this subset is obtained. The results are in **Table 11**.

Table 11 Cluster quality measures for the Australian Credit Data Base

Quality index method	*CPI*	*RI*	*ARI*	*JI*
0	0.81	0.70	0.39	0.54
1	0.56	0.51	0.03	0.50
2	0.56	0.51	0.03	0.50
3 k-mode [51]	0.66	NA	NA	NA
4 Fuzzy k-modes[51]	0.74	NA	NA	NA
5Kim's fuzzy centroids [51]	0.80	NA	NA	NA

Again method 0 fares better than the others.

7. Summary and Conclusion

A method for clustering feature vectors with mixed numeric and categorical attributes has been described. In the proposed method, a set of feature vectors, of very small dimension (3-5 components), equivalent to the original set of feature vectors in terms of inter feature vector distances but having only numeric components is generated and then FCM is applied to the new set of feature vectors. The method that is used for finding the new set of feature vectors is gradient descent. The method has been shown to be very effective in terms of the resulting cluster partition.

There is potential for improvement however. The idea of mapping to a FCM clusterable set of feature vectors appears to be sound but computationally intensive and it is therefore appropriate to look for other means of doing so. The error function may have a large number of local optima and therefore other optimization methods should be considered. The main effort is in determining the feature vectors from a distance matrix which is in the domain of multidimensional scaling.

Another drawback of the proposed method is that when the number of records is very large the dissimilarity matrix will be extremely large and data reduction methods will have to be considered. One data reduction method consists of clustering and then replacing the original data set by the centroids of the clusters. However this takes us back to the original problem of finding centroids when the feature vectors are mixed categorical and numeric.

Another area for future work is the development of methods in the case where there is a combination of ordinal and categorical features. Feature vectors generally consist of all 3 types of features including ordinal and categorical. For ordinal neither distance nor aggregation is defined. For categorical features distance is defined but aggregation is not defined.

Future work will therefore consist of applying a more efficient multidimensional scaling method and of combining the method for dealing with ordinal features proposed by the author in other papers [58-61] with the method for dealing with categorical features proposed here.

Acknowledgements

The support of the Canadian Government in terms of NSERC grant 227338-04 and the support of the Department of Mechanical and Mechatronic Engineering of the University of Stellenbosch is gratefully acknowledged. The work of the reviewers is also much appreciated.

8. References

[1] A. K. Jain and R. C. Dubes, *Algorithms for Clustering Data*. Upper Saddle River, NJ: Prentice Hall, 1988.

[2] W. Klosgen and J. M. Zytkow, "Knowledge discovery in databases terminology," in *Advances in Knowledge Discovery and Data Mining*, U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds.: The MIT Press, 1996, pp. 573-92.

[3] M. R. Anderberg, *Cluster Analysis for Applications*: Academic Press, 1973.

[4] L. Kaufman and P. J. Rousseeuw, *Finding Groups in Data: An Introduction to Cluster Analysis*. New York: John Wiley and Sons, 1990.

[5] R. M. Cormack, "A review of classification," *J. Roy. Statist. Soc*, vol. Series A, pp. 321-67, 1971.

[6] G. J. Williams and Z. Huang, "A case study in knowledge acquisition for insurance risk assessment using a KDD methodology," in *Pacific Rim Knowledge Acquisition Workshop*, Sydney, Australia, 1996, pp. 117-29.

[7] B. Everitt, *Cluster Analysis. Heinemann Educational Books Ltd.*, 1974.

[8] R. S. Michalski and R. E. Stepp, "Automated construction of classifications: Conceptual clustering versus numerical taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. PAMI-5, pp. 396-410, July 1983.

[9] D. H. Fisher, "Knowledge acquisition via incremental conceptual clustering," in *Machine Learning*. vol. 2(2), 1987, pp. 139-72.

[10] S. S. Stevens, "On the Theory of Scales of Measurement," *Science,* vol. 103, pp. 677 - 680, 1946.

[11] E. H. Ruspini, "A new approach to clustering ." *Information and Control,* vol. 15, pp. 22-32, 1969.

[12] J. Bezdek, *Pattern Recognition with Fuzzy Objective Function Algorithms*: Plenum, 1981.

[13] H. Vinod, "Integer programming and theory of grouping," *Journal of the american statistical association,* vol. 64, pp. 506-17, 1969.

[14] E. Ruspini, "Numerical methods for fuzzy clustering," *Information Sciences,* vol. 2, pp. 319-50, 1970.

[15] M. P. Windham, "Numerical classification of proximity data with assignment measures," *Journal of Classification,* vol. 2, pp. 157-72, 1985.

[16] M. Roubens, "Pattern classification problems and fuzzy sets," *Fuzzy Sets and Systems,* vol. 1, pp. 239-53, 1978.

[17] J. C. Bezdek, R. J. Hathaway, and M. P. Windham, "Numerical comparison of the RFCM and AP Algorithms for clustering relational data," *Pattern Recognition,* vol. 24, pp. 783-91, 1991.

[18] D. Barbara, J. Couto, and Y. Li "COOLCAT:An entropy-based algorithm for categorical clustering," in *Information and Knowledge Management*, 2002, pp. 582-9.

[19] D. Gibson, J. Kleinberg, and P. Raghavan, "Clustering categorical data: An approach based on dynamical systems," in *24rd VLDB Conference* 1998, pp. 311-22.

[20] Z. Huang, "Extensions to the k-Means Algorithm for Clustering Large Data Sets with Categorical Values," *Data Mining and Knowledge Discovery,* vol. 2, pp. 283-304, 1998.

[21] V. Ganti, J. Gehrke, and R. Ramakrishnan, "CACTUS - clustering categorical data using summaries," in *Fifth ACM SIGKDD*, 1999, pp. 73-83.

[22] S. Guha, R. Rastogi, and K. Shim, "ROCK A robust clustering algorithm for categorical attributes," *Information Systems,* vol. 25, pp. 345-66, 2000.

[23] M. Gyllenberg, Koski, T, & Verlaan, M., "Classification of binary vectors by stochastic complexity," *Journal of Multivariate Analysis,* pp. 47-72, 1947.

[24] K. Gowda and E. Diday, "Symbolic clustering using a new dissimilarity measure," *Pattern Recognition*, vol. 24, pp. 567-578, 1991.

[25] Y. El-Sonbaty and M. A. Ismail, "Fuzzy clustering for symbolic data," *Fuzzy Systems, IEEE Transactions on,* vol. 6, pp. 195-204, 1998.

[26] Z. Huang and M. K. Ng, "A fuzzy k-modes algorithm for clustering categorical data," *Fuzzy Systems, IEEE Transactions on,* vol. 7, pp. 446-52, 1999.

[27] T. Guan and B. Feng, "Fuzzy clustering of incomplete nominal and numerical data," in *Fifth World Congress on Intelligent Control and Automation*, 2004, pp. 2331-2334.

[28] M.-S. Yang, P.-Y. Hwang, and D.-H. Chen, "Fuzzy clustering algorithms for mixed feature variables," *Fuzzy Sets and Systems*, vol. 141, pp. 301-317, 2004.

[29] J. Li, X. Gao, and L. Jiao, "A GA-based clustering algorithm for large data sets with mixed numeric and categorical values," in *The International Society for Optical Engineering Multispectral Image Processing and Pattern Recognition*,. vol. 5286 Beijing, China, 2003, pp. 171-4.

[30] L. Vermeulen-Jourdan, C. Dhaenens, and E. G. Talbi, "Clustering Nominal and Numberical Data: A New Distance Concept for a Hybrid Genetic Algorithm," in *EvoCOP*, 2004, pp. 220-229.

[31] H. Ralambondrainy, "A clustering method for nominal data and mixture of numerical and nominal data," in *First Conference International Federation of Classification Sciences*, Aachen, 1987, pp. 368-76.

[32] K. Honda and H. Ichihashi, "Fuzzy c-means clustering of mixed databases including numerical and nominal variables," 2004, pp. 558-562 vol.1.

[33] O. M. San, V.-N. Huynh, and Y. Nakamori, "An alternative extension of the k-means algorithm for clustering categorical data," *International Journal of Applied Mathematics and Computing Science*, vol. 14, pp. 241-7, 2004.

[34] T. Li, S. Ma, and M. Ogihara, "Entropy-based criterion in categorical clustering " in *The Twenty-first International conference on Machine learning*, 2004, p. 68.

[35] E. Rendon and J. S. Sanchez, "Clustering based on compressed data for categorical and mixed attributes," Hong Kong, China, 2006, pp. 817-25.

[36] Y. Zhao, B. Li, X. Li, W. Liu, and S. Ren, "Cluster ensemble method for databases with mixed numeric and categorical values," *Journal of Tsinghua University (Science and Technology)*, vol. 46, pp. 1673-6, 2006.

[37] R. S. Ramakrishna and K. Minho, "Projected clustering for categorical datasets," *Pattern Recognition Letters*, vol. 27, pp. 1405-17, 2006.

[38] A. Nemalhabib and N. Shiri, "CLUC: a natural clustering algorithm for categorical datasets based on cohesion," Dijon, France, 2006, pp. 637-8.

[39] A. Ahmad and L. Dey, "A method to compute distance between two categorical values of same attribute in unsupervised learning for categorical data set," *Pattern Recognition Letters*, vol. 28, pp. 110-18, 2007.

[40] B. Andreopoulos, A. An, and X. Wang, "Bi-level clustering of mixed categorical and numerical biomedical data," *International Journal of Data Mining and Bioinformatics*, vol. 1, pp. 19-56, 2006.

[41] H. Ralambondrainy, "A conceptual version of the K-means algorithm," *Pattern Recognition Letters*, vol. 16, pp. 1147-57, 1995.

[42] J. C. Gower, "A general coefficient of similarity and some of its properties," *Biometrics*, vol. 27, pp. 857 - 871, 1971.

[43] M. A. Woodbury, and J.Clive, " Clinical pure types as a fuzzy partition," *Journal of Cybernetics*, vol. 4, pp. 111–121, 1974.

[44] F. Hoppner, F. Klawonn, R. Kruse, and T. Runkler, *Fuzzy Cluster Analysis*: John Wiley and Sons, 1999.

[45] Y. Takane, F. Young, and J. De Leeuw, "Nonmetric individual differences multidimensional scaling: an alternative least squares method with optimal scaling features," *Psychometrika*, vol. 42, pp. 7-67, 1976.

[46] W. M. Rand, "Objective Criteria for the Evaluation of Clustering Methods," *Journal of the American Statistical Association*, vol. 66, pp. 846-50, 1971.

[47] G. Saporta and G. Youness, "Comparing two partitions: Some Proposals and Experiments."" in *COMPSTAT, 15th Conference on Computational Statistics*, 2002, pp. 302-309.

[48] H. Li, K. Zhang, and T. Jiang, "Minimum Entropy Clustering and Applications to Gene Expression Analysis," in *IEEE Computational Systems Bioinformatics Conference*, 2004, pp. 142- 151.

[49] L. Hubert and P. Arabie, "Comparing partitions," *Journal of Classification*, vol. 2, pp. 193-8, 1985.

[50] D. Graves, "Clustering Quality Measures," Thompson Rivers University, Kamloops TR TRU-CIG-2006-07, July 2006.

[51] D.-W. Kim, K. H. Lee, and D. Lee, "Fuzzy clustering of categorical data using fuzzy centroids," *Pattern Recognition Letters*, vol. 25, pp. 1263-71, 2004.

[52] G. W. Milligan and M. C. Cooper, "A study of the comparability of external criteria for hierarchical cluster analysis.," *Multivariate Behavioral Research*, vol. 21, pp. 441-58, 1986.

[53] K. Y. Yeung and W. L. Ruzzo, "Principal component analysis for clustering gene

expression data," *Bioinformatics,* vol. 17, pp. 763-774, 2001.

[54] A. Asuncion and D. J. Newman, "Machine Learning Repository," University of California, Irvine, School of Information and Computer Sciences, 2007.

[55] "UCI Machine Learning Repository," http://mlearn.ics.uci.edu/MLRepository.html.

[56] R. Quinlan, "C4.5: Programs for Machine Learning " in *Macine Learning*: Morgan Kaufmann, 1992.

[57] R. Quinlan, "Simplifying decision trees," *Int J Man-Machine Studies* vol. 27, pp. 221-234, 1987.

[58] R. K. Brouwer, "A Method for Obtaining a Fuzzy Set Covering for Ordinal Attributes for Use in Fuzzy Clustering," Thompson Rivers University, Kamloops, Canada, Report BRO 2005 Department of Computing Science, Thompson Rivers University, 2005.

[59] L. Mahnhoon and R. K. Brouwer, "Fuzzy Clustering and Mapping of Ordinal Values to Numerical," in *FOCI 2007 IEEE Symposium on Foundations of Computational Intelligence*, 2007, pp. 538-43.

[60] R. K. Brouwer, "A Method for Fuzzy Clustering with Ordinal Attributes Replaced by Fuzzy Set Parameters," in *3rd International IEEE Conference on Intelligent Systems*, 2006, pp. 553-8.

[61] R. K. Brouwer, "Fuzzy Clustering of Feature Vectors with some Ordinal Valued Attributes Using Gradient Descent for Learning," *IJUFKS,* vol. 6, pp. 195-218, 2008.

[62] R. K. Brouwer, "Clustering without Use of Prototypes," Thompson Rivers University, Kamloops TR TRU-CS-CIG-2006-01, May 1 2006.