# Multi-Modal Data Fusion

Sarah Coppock

A Dissertation Proposal
Submitted to the Department of Electrical and Computer Engineering and Computer Science in partial fulfillment of the requirements for the degree of Doctor of Philosophy

September, 2002

**Error! Bookmark not defined.**

# Abstract

Clustering groups items together that are most similar to each other and sets those that are least similar into different clusters. Methods have been developed to cluster records in a data set that are of only qualitative or quantitative data. Data sets exist that contain a mix of qualitative (nominal and ordinal) and quantitative (discrete and continuous) data. Clustering records of mixed kinds of data is a difficult problem. A metric to measure the similarity between records of mixed data types is needed. Once a clustering is found, we do not know how to best evaluate the quality of the clustering when there is a mixture of data varieties. The proposed research is to find a useful way of clustering data sets containing both quantitative and qualitative data. This could be a developed metric general enough to be used in existing algorithms or the development of a new approach incorporating soft computing, specifically rough sets.

# 1. Introduction

Grouping together things described by different kinds of data is very important. Records have many diverse data types. We have a great need to differentially group records containing diverse data. In some sense, this can be considered as “data fusion.” Data fusion is a critical necessity for medicine, military sensor integration, and studies of the human population.

Clustering data records offers information. This information includes:

- Discovery of groups of data records that are similar to each other
- Discovery of one or more records that are *outliers*, records which are considered largely dissimilar to the other records. E.g., a fraudulent purchase on a customer’s credit card
- A description of similar records, e.g. what type of customer buys a particular product

Note the first two examples rely on the fact that clustering discovers the distribution of the data. The purpose of clustering is up to the user.

Consider a data set containing medical information regarding patients having a particular disease and who were given different treatments. There may exist a combination of different kinds of data: qualitative values such as blood type, and quantitative values such as age and weight. One could use classification on the data set to discover information such as which type of patient responds to a treatment. Clustering is classification’s unsupervised counterpart and this offers more potential information. Clustering does not group according to one attribute as classification does. This means that there is potential information offered by clustering that is not offered by classification.

The notion of clustering is closely related to classification and is also used in our own learning of concepts in the real world. Clustering is the grouping of objects into clusters such that the similarity among objects within the same cluster is maximized (*intra-cluster similarity*) and the similarity between objects in different clusters (*inter-cluster similarity*) is minimized (Everitt, 1993) (Jain, 1988).

Clustering in data mining and data analysis can discover the general distribution of the data. It allows discovery of similar objects described in the data set. Usually, a good characterization of the resulting clusters is also an objective. Another objective is scalability. An algorithm is considered *scalable* if its cost increases linearly with the number of records.

Though the idea of proximity and similarity in clustering rely on the context in which it is used, this experimental research does not seek to include contextual information other than the relative data distributions in the data set. This subjectivity is discussed in the following section.

Section 2 discusses the problem of grouping records together and the varieties of data. Section 3 discusses some of the previous research relevant to our work. Next, a brief overview of rough sets is given. Sections 5 and 6 discuss similarity metrics and problems with measuring similarity with mixed data. Section 7 discusses the open questions in grouping records of mixed data. This last section also discusses the question we are out to answer and our approach to answering this question.

# 2. Grouping Records Together

The problem of clustering can be defined as follows: we have a set of attributes, $A=\{A_1, A_2,\ldots,A_k\}$, where each attribute can take on a finite or infinite number of possible values, $Dom(A_i)=\{a_{i1},a_{i2},\ldots,a_{ij}\}$. The domains can be of different data types. *Data type* in the context of this paper refers to the classification of data as either qualitative or quantitative. A data set is composed of records, where each record, $r_i$ is a tuple of values from each attribute's domain. The goal is to cluster the records into groups such that the *intra-cluster similarity*, similarity between records in the same cluster, is maximized and the *inter-cluster similarity*, similarity between clusters, is minimized (Everitt, 1993) (Jain, 1988). Clustering methods that utilize a similarity or distance function assume all domains are of the same kind of data.

### 2.1 Kinds of Data

Data can be classified by its type and scale, i.e. qualitative or quantitative (Everitt, 1993). Most current clustering algorithms deal with quantitative data. This includes continuous values, such as a person's *height*, and discrete values, such as the *number of cars sold*. Qualitative data on the other hand, is symbols or names with no natural scale between the values. This includes nominal data such as the *color of a car* and ordinal data such as the *doneness of a burger*: *rare*, *medium*, *well*.

A consequence of the lack of a fixed scale for qualitative data is the difficulty in quantitatively measuring the similarity between two qualitative values. It is common to use *simple matching* that assigns a 1 if two values match and 0 if two values don't match (or the converse for measuring dissimilarity). For the similarity of two quantitative values, the magnitude difference between the two values is usually used.

There are different approaches to clustering (Everitt, 1993) (Jain, 1988). Most current research that seeks to cluster records of qualitative data, explicitly or implicitly assume only qualitative data (Ganti, 1999) (Gibson, 2000) (Guha, 2000) (Wang, 1999) (Zhang, 2000). This is discussed further in section 3.

The proposed research seeks to group records of both quantitative and qualitative data such as in *Figure 2.1*.

| | **Color** | **Weight** | **Height** | **Edible** |
|---|---|---|---|---|
| **r1** | orange | .001 | 12” | no |
| **r2** | red | .002 | 12” | no |
| **r3** | orange | 10.2 | 12” | no |
| **r4** | green | 9.8 | 11” | yes |
| **r5** | orange | .98 | 4” | yes |

*Figure 2.1. Example data set with qualitative and quantitative data*

*Color* and *Edible* are qualitative and *Weight* and *Height* are quantitative. One possible clustering could be:

{r1,r2},{r3},{r4,r5} if the physical attributes (*Height* and *Weight*) are at least equally important with the *Edible* attribute;

or:

{r1,r2,r3},{r4,r5} if the *Edible* attribute dominates, or is most important, and *Height* and *Weight* are irrelevant. The proposed research does not seek to limit any results to only one specific clustering if two or more are found to be reasonable.

There are many possible partitions of the data set. When grouping together multi-modal records in an unsupervised manner, we do not know for what purpose the partition will be used. That is, if in the above example, we do not know whether edibility is more important than the physical features or vice versa.

### 2.1 Approaches to Clustering

There is more than one approach to clustering. Two of the more commonly found approaches are hierarchical clustering and clustering using partitioning. While hierarchical based clustering is more dependent on similarity between two single records (or a record and a cluster representation), the partitioning approach relies on a similarity function over all the records, i.e. a function of the intra- and inter-cluster similarities.

#### 2.1.1 Hierarchical Clustering

There are two types of hierarchical approaches to clustering: *agglomerative* and *divisive* (Jain, 1988) (Han, 2000). Agglomerative begins with all objects in their own cluster and combines clusters for which the similarity is the greatest. This is done repeatedly until all objects are in the same cluster. Divisive begins with all objects in the same cluster and works in the reverse direction, until all records are in their own cluster.

Because these approaches are based on a similarity metric, an appropriate similarity metric must be defined for measuring the similarity between records (or cluster representation) of any combination of data. Such metrics already exist for records that contain only qualitative or quantitative data, e.g. Minkowski metric for quantitative records (Han, 2001) or Simple Matching Coefficient for qualitative records (Sneath, 1973). In deciding to merge clusters, a representation for clusters may be used. How best to represent clusters with multiple records which contain both qualitative and quantitative data is also an open problem.

### 2.1.2 Clustering by Partitioning

Another approach to clustering is to use an initial, possibly arbitrary, partition of the records and to refine this initial partition (Han, 2000) (Jain, 1988). The refinement of the clustering is achieved by redistributing records to other clusters according to some similarity criterion. In this approach, it is common to use a representation for each cluster, e.g. a mean, in order to decide how to redistribute the records. Alternatively, the algorithm can use an overall evaluation function of the goodness of the clustering to search for an optimal clustering.

Usually, this approach requires the number of clusters to be known a priori—which can become a difficulty if the decided number is not appropriate to the data distribution. This problem exists independent of the kinds of data, but the difficulty in evaluating the goodness of the clustering and finding a suitable cluster representation is not completely solved for records composed of different kinds of data.

Clustering qualitative and quantitative data is a difficult problem. When all attributes are of the same kind then the inter- and intra-cluster similarity can be defined according to a similarity measure between records. Similarity metrics are defined for records of one data type. When attributes are of different data types, we do not have a sufficiently defined similarity metric to use in measuring the similarity between two records. This makes it difficult to generalize clustering algorithms to cluster records of mixed data.

## 3. Background

Although problems still exist with clustering data sets of quantitative data, algorithms have been developed to cluster records of quantitative data meaningfully and efficiently (Everitt, 1993) (Jain, 1988). Research has now tried to incorporate clustering of categorical values in data sets (Gowda, 1991) (Gupta, 1999) (Hirano, 2001) (Huang, 1999) (Li, 1998). Other algorithms to cluster only qualitative data have been developed by (Ganti, 1999) (Gibson, 2000) (Guha, 2000) (Wang, 1999) (Zhang, 2000). As of yet, clustering data sets of mixed data remains a difficult problem due to the difference in data characteristics such as scale and order.

The approaches that cluster only qualitative data use co-occurances as a measure of similarity (Gibson, 2000) (Han, 1997) (Wang, 1999) (Zhang, 2001). For example, if for two attributes $A_1$ and $A_2$, the values $a_1$ and $a_2$ occur frequently together, then these values will be considered “more similar” than say $a_1$ and $b_2$, which do not occur frequently together[1]. For example, in the approach by Gibson (2000) weight is propagated among the values by means of a dynamical system. Ganti (1999) mentions the need for a post-processing step in the approach by (Gibson, 2000), but it is not immediately clear what needs to be done.

Another approach to cluster qualitative records by Wang (1999) uses set theory to develop a function to evaluate the quality of the clustering. This approach actually was developed to cluster transactions. That is, the algorithm clusters records with varying dimensions such as $r_1$={apples, bananas} and $r_2$={oranges, pears, grapes}. In the context of this paper, items in transactions can be equated with attributes in records. In the evaluation of the clustering, values that are *large* within each cluster are found. A value is *large* if it is present within a particular cluster above a user-supplied support threshold. A value is considered *small* if it is not labeled *large*. Two sets for each cluster are determined, one containing *large* values and one containing *small* values. Then a cost function is defined using these sets. The cost function takes a possible clustering and results in a quantitative measure of the quality of the clustering. The portion of the cost function that represents the intra-cluster similarity is a count of all values in the clusters that are considered *small*. Therefore, when considering a two-item cluster (or the similarity between two records) where the minimum to be considered *large* is 2, the similarity is simply a form of simple matching.

One approach to clustering records of a variety of data is to extend the distance-based k-means algorithm to handle qualitative data in addition to the quantitative data (Gupta, 1999) (Huang, 1999) (Ralambondrainy, 1995). The distance metric is redefined as a sum of two measures, one for the qualitative attributes and one for the quantitative attributes. The hard part of combining metrics like this is that an appropriate weighting of the measures needs to be derived for the overall measure to be useful. This is explained further in section 5.2. A weighting is suggested with these algorithms, but it is uncertain whether the suggested weighting is appropriate. Finding an appropriate and computationally feasible weighting is a problem.

| | **$a_1$** | **$a_2$** | **$a_3$** |
|---|---|---|---|
| **$r_1$** | 1 | a | 2 |
| **$r_2$** | 1 | b | 2 |
| **$r_3$** | 4 | a | 2 |
| **$r_4$** | 2 | a | 1 |
| **$r_5$** | 1 | b | 4 |
| **$r_6$** | 3 | a | 3 |

*Figure 3.1. An example data set with two discrete quantitative ($A_1$ and $A_3$) and one qualitative attribute ($A_2$).*

Li and Biswas (1998) developed an hierarchical algorithm based on Goodall's similarity metric (Goodall, 1966). The classification tree when applying SBAC (Li, 1998) to the example in *Figure 3.1* is shown in *Figure 3.2*. For each node, *D* represents the dissimilarity at which the child nodes merge. For example, records 2 and 5 merge with dissimilarity of .10. The best way of selecting threshold for determining the level of the clustering is not clear. It is possible that the selection is closely related to the *cluster indicator value* proposed by (Gowda, 1991).

[1] “frequent” in this context means the value’s frequency in the data set

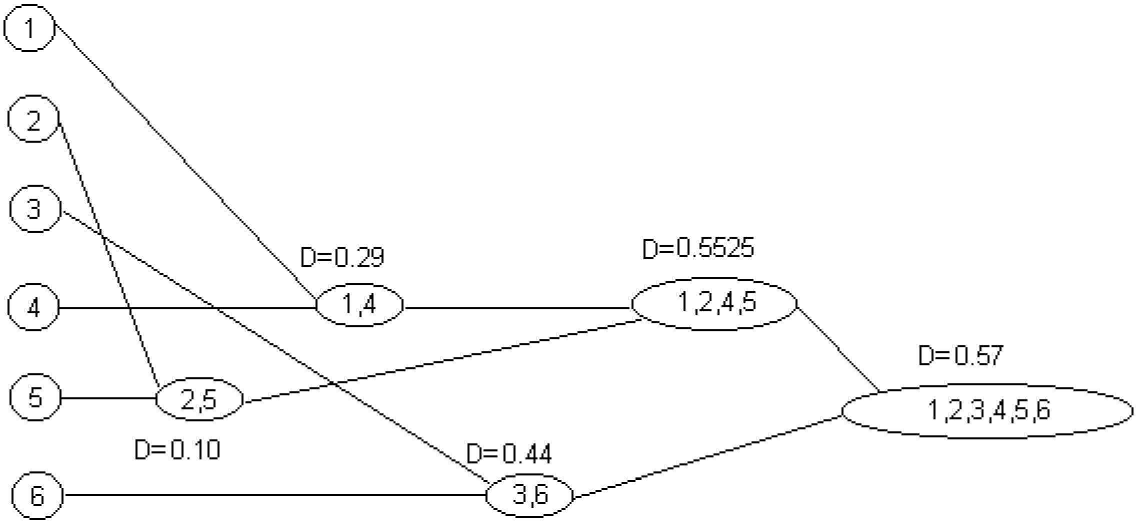


*Figure 3.2. Classification tree of SMAC applied to Figure 3.1.*

Gowda (1991) defined algorithm which utilizes a metric for multiple types of data. *Figure 3.3* shows the resulting tree with the *cluster indicator value* for each stage. The proposed clustering is the stage at which this value is highest (in this case, the first level).

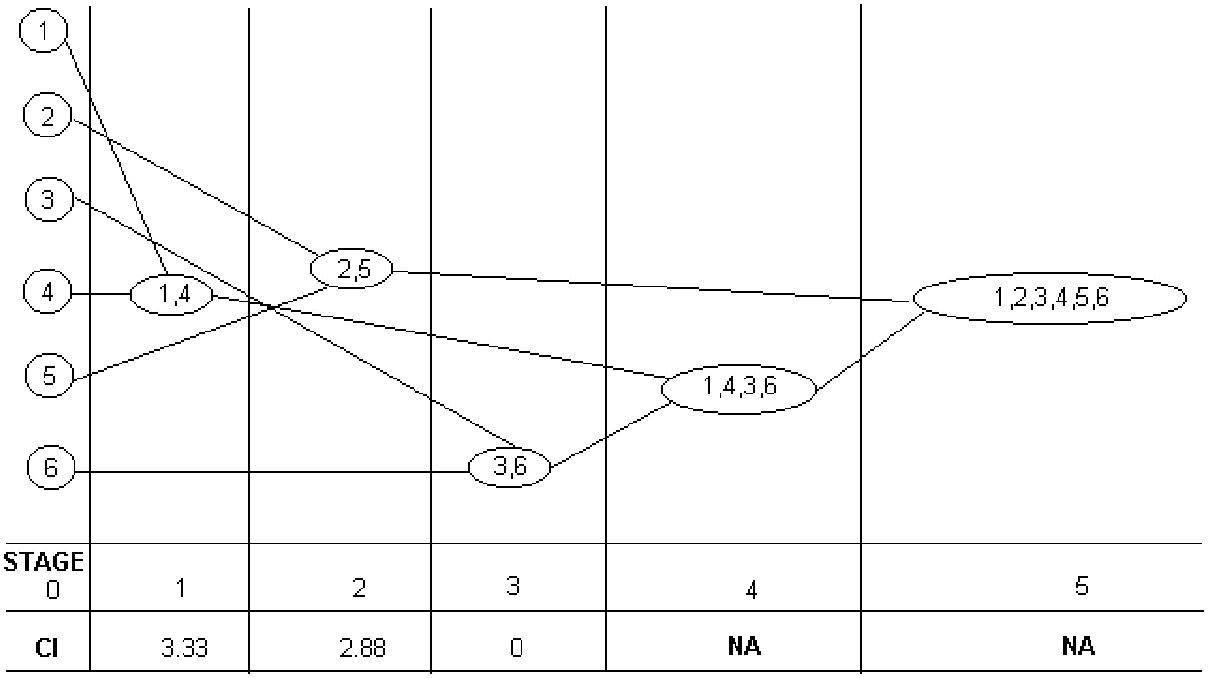
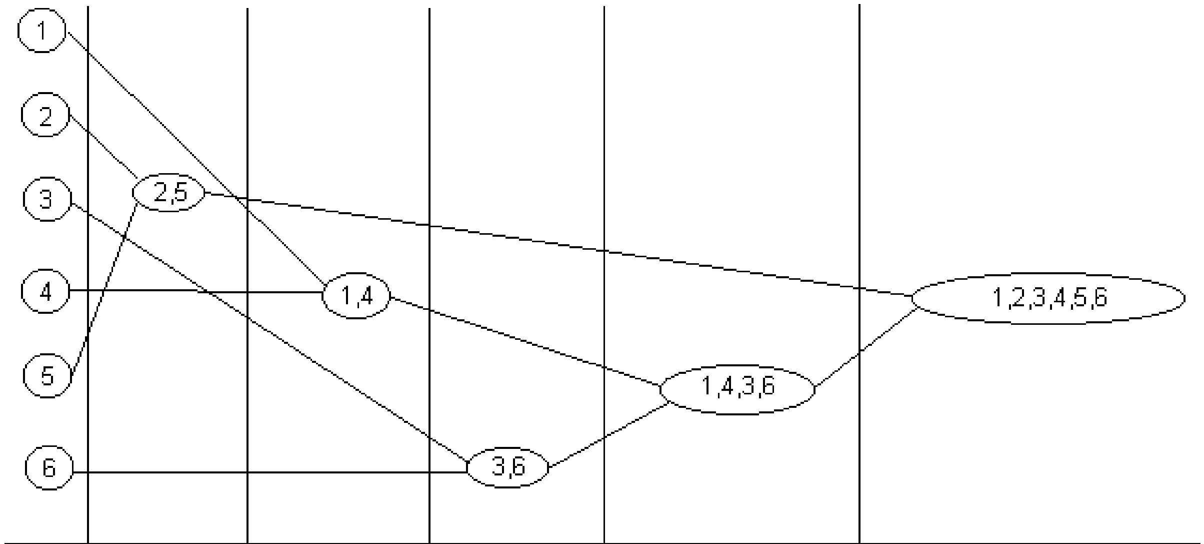


*Figure 3.3. One clustering tree from Gowda (1991).*



| STAGE 0 | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| CI | 2.84 | 3.34 | 0 | NA | NA |

*Figure 3.4. Second method of merging clusters with algorithm by (Gowda, 1991)*

In our example, there are three possible ways to merge in the first stage. *Figures 3.3 and 3.4* give the clustering tree of two. In *Figure 3.3*, $r_1$ and $r_4$ are merged first and in *Figure 3.4*, $r_2$ and $r_5$ are merged first. As can be seen from the figures, in both cases, the number of clusters is given at the same level, level 1.

Hirano (2001) developed an algorithm utilizing rough set theory to cluster mixed data. For each record, $r_i$, two equivalence relations are derived according to the similarity between $r_i$ and all other records. The similarity function defined by Hirano is a combination of two existing metrics, the Hamming and Mahalanobis distances. Hamming distance is defined as the number of attributes for which two records differ. Mahalanobis distance is defined as: $M(x,y) = (x-y)\ C^{-1}\ (x-y)$ where $C^{-1}$ is the inverse of the covariance matrix. The threshold by which the initial equivalence relations are defined is supplied by the user. A separate threshold is needed in a modification step. This lies between 0 and 1. The closer to 1 the second threshold, the more modification is made. The idea is to maximize the following function: $V(R')=1/N \sum_{k=1}^{N} (\frac{\#\underline{Ck}}{\#\overline{Ck}} \times \#Ck)$ , where R' is the modified equivalence relations, N is the number of clusters given by R', $\underline{Ck}$ is the lower approximation, $\overline{Ck}$ is the upper approximation for the $k^{th}$ cluster, and *#X* is the cardinality of a set *X*. This will be maximized when the boundary is minimized over all equivalence relations.

When we apply Hirano's algorithm to our small example in *Figure 3.1* using the similarity threshold between -2.0 and 2.0 and the threshold for modification between 0.1 and 0.9, there are two partitions that have the maximum goodness, V function value. The two are: $\{\{r_1,r_2,r_4\}\{r_3,r_5,r_6\}\}$ and $\{\{r_1,r_2\},\{r_3,r_4,r_5,r_6\}\}$. One interesting thing to note from this example is the number of clusters suggested by each algorithm. Hirano's algorithm suggests two clusters while Gowda's and Li's algorithms suggest 5 clusters.

Applying existing algorithms to our example demonstrates the difficulty in interpreting the resulting clustering. In the Hirano's and Li's approaches, how do we set the thresholds? With any of the previous approaches, how do we know this is the best or even reasonable clustering?

# 4. Similarity Metrics

To be able to use developed clustering algorithms, it would be useful to have a similarity metric that is useful on any mix of data types. Most current similarity metrics use some combination of the similarity between individual attribute values to derive the overall similarity between records (Everitt, 1965).

| | | | |
|---|---|---|---|
| $\mathbf{r_1}$ | $a_1$ | $b_1$ | $c_1$ |
| $\mathbf{r_2}$ | $a_2$ | $b_2$ | $c_2$ |

*Figure 4.1*

For example, the similarity between the two records in *Figure 4.1* would be defined as $sim(a_1,a_2) \oplus sim(b_1,b_2) \oplus sim(c_1,c_2)$ where ⊕ indicates some combination operator. Typically, the combination operator is the addition operator (Gower, 1971) (Gowda, 1991) (Sneath, 1973).

Although the measure itself can be either quantitative or qualitative, most current metrics derive a quantitative measure. It is assumed that finding a quantitative measure of similarity would be best for the purpose of clustering records. Similarity between values is according to the type of data the values are. For example, in *Figure 6*, sim($a_1$,$a_2$) is defined according to the kinds of data of $a_1$ and $a_2$.

# 5. Similarity Between Records Containing Mixed Data Types

Data sets with a mixture of types of data are common. Applying an existing algorithm that assumes only one type of data would not be meaningful. In order to cluster these records using an existing clustering algorithm, a meaningful way of measuring the similarity between records must be developed. Therefore, measuring the similarity or distance between two records requires that the metric must be able to handle a mixture of types of data. There are at least two possible ways of developing a useful metric:

1. a metric developed for one type of data, e.g. Euclidean distance for quantitative data, can be extended by some form of mapping to include both types, or
2. a metric that usefully combines two or more metrics, some qualitative and some quantitative, can be developed.

Both of these approaches to solving the problem of clustering mixed data have difficulties. It is possible that when developing a similarity metric for mixed data types, the utility of the metric is compromised.

### 5.1 Extending Quantitative Metrics

The most straightforward way to extend metrics developed for quantitative data is to use some form of mapping from the qualitative data to quantitative values. The difficulty in doing this is discovery of a useful mapping. This is due to the lack of scale between qualitative values. With nominal data, the lack of order along with lack of scale causes difficulty.

Unfortunately, it does not make sense to map these values into a form appropriate for the typical distance measures. Rather, it is difficult to discover a meaningful mapping even if we have contextual information regarding the data.

| | fruit | color | bag size |
|---|---|---|---|
| $r_1$ | apple | red | 5 |
| $r_2$ | orange | orange | 3 |
| $r_3$ | apple | green | 5 |

*Figure 5.1*

Let one arbitrary mapping ($\pi_1$) be:

*fruit*={orange: 1, apple: 2}, *color*={red: 1, orange: 2, green: 3}

and another arbitrary mapping ($\pi_2$) be:

*fruit*={orange: 2, apple: 1}, *color*={red: 0, orange: 1, green: 6}

for the nominal values in *Figure 5.1*.

Euclidean distance is defined as:

$d(X,Y)= (” ([x_i-y_i])^2 )^{1/2}$ , where $X$ and $Y$ are the records being compared, $x_i$ and $y_i$ are the $i^{th}$ attribute values of $X$ and $Y$, and $d(X,Y)$ is the distance between records $X$ and $Y$.

Using this distance metric, *Figure 5.2* gives the respective distance between the records in *Figure 5.1*:

| distance using $\pi_1$ | | | | | distance using $\pi_2$ | | | |
|---|---|---|---|---|---|---|---|---|
| | $r_1$ | $r_2$ | $r_3$ | | | $r_1$ | $r_2$ | $r_3$ |
| $r_1$ | 0 | 2.45 | 2 | | $r_1$ | 0 | 2.24 | 2.45 |
| $r_2$ | 2.45 | 0 | 2.45 | | $r_2$ | 2.24 | 0 | 2.83 |
| $r_3$ | 2 | 2.45 | 0 | | $r_3$ | 2.45 | 2.83 | 0 |

*Figure 5.2*

In this case, the quantitative values are arbitrarily assigned to the qualitative values. Note from above,

$d(r_1,r_2) > d(r_1,r_3)$ with $_1$

but

$d(r_1,r_2) < d(r_1,r_3)$ with $_2$.

The difficulty is that the metric is defined for quantitative values, but we are imposing an artificial ordering and a scale. Therefore, we can select an arbitrary mapping, but we can't be sure about the usefulness of the resulting measure. If we put records in order of similarity from an arbitrary single record, i.e. $r_1<r_2<\ldots<r_n$ where $r_i<r_j<r_k$ indicates that $r_j$ is more similar to $r_i$ than $r_k$, then we can change the similarity order according to the mapping used. This could then affect the resulting clustering.

### 5.2 Extending Qualitative Metrics

Extending qualitative metrics to deal with quantitative data is also difficult. Most metrics useful for records of qualitative data use a form of matching of values to decide the measure (Sneath, 1973). Usually, this measure is some proportion of the total number of attribute values. For example, the simple matching coefficient (Sneath, 1973) is defined as the number of attribute-value pairs that the two records have in common divided by the total number of attributes in the records.

Extending a metric such as this to quantitative data results in a loss of information; this would not be desirable. For example, if using simple matching, as defined in section 3.1, between the quantitative values, 3.4, 3.5, and 4.2, the similarity measure will be the same between each pair. Information such as the fact that 4.2 is more dissimilar to 3.4 than it is to 3.5 is lost.

Even if matching were to be used on quantitative ranges or intervals such as [3.0,3.5], it is possible that information would still be lost. For example, if two different values lay in the same interval, then the information that they are dissimilar will be lost when measuring similarity. In addition, discovering the optimal intervals in order to make this useful would be difficult and computationally expensive.

### 5.3 Combining Existing Metrics

Since extending existing metrics for one data type has difficulties, the next possible solution to determine similarity between records of mixed data type is to look at combining existing metrics. For example, extending the k-means algorithm to handle both qualitative and quantitative data has been attempted (Huang, 1999) (Ralambondrainy, 1995) (Gupta, 1999). This particular approach is further discussed in section 5.3.1.

Since extending one metric to handle multiple data types is difficult, the other alternative is to combine two types of metrics, one measure for the qualitative attributes and one for the quantitative attributes. But, combining in a simple manner causes the utility of the measure to be compromised.

#### 5.3.1 Similarity Metric Utility

Although the following discussion is about distance measures, the same argument can be made for any pair of similarity measures. This is because distance is a *mathematically metric* measure of dissimilarity and dissimilarity is the complement of similarity. By *mathematically metric*, it is meant that the measure meets the following criteria:

- measure(x,y) ≥ 0
- measure(x,x) = 0
- measure(x,y) = measure(y,x)
- measure(x,y) + measure(y,z) ≤ measure(x,z)

This definition will be used in later sections.

Huang (1997) (1999) extends the distance-based method k-means algorithm to handle categorical data. By using an integer value, 1 or 0, to indicate non-matching and matching respectively, a categorical attribute is incorporated into the distance metric. In (Huang, 1997), similarity is computed as the sum of square differences for the numerical attributes simply added to a weighted summation of matches for the categorical attributes. In other words, the similarity is a sum of two metrics, one for quantitative and one for qualitative. This is like adding apples to oranges. This is because quantitative values contribute their magnitude whereas the qualitative attributes have no magnitude to contribute (simply an integer value between 0 and the number of qualitative attributes). The magnitudes of the quantitative attributes therefore contribute to the measure differently, a point previously made by Goodall (1966).

|  | **$a_1$** | **$a_2$** | **$a_3$** |
|---|---|---|---|
| **$r_1$** | 1 | a | 2 |
| **$r_2$** | 1 | b | 2 |
| **$r_3$** | 4 | a | 2 |
| **$r_4$** | 2 | a | 1 |
| **$r_5$** | 1 | b | 4 |
| **$r_6$** | 3 | a | 3 |

*Figure 3.1. An example data set with two discrete quantitative ($A_1$ and $A_3$) and one qualitative attribute ($A_2$).*

The respective distance using Huang's approach with weight of 1 is displayed in *Figure 5.4* for the example in *Figure 5.3*:

| | $r_1$ | $r_2$ | $r_3$ | $r_4$ | $r_5$ | $r_6$ |
|---|---|---|---|---|---|---|
| **$r_1$** | 0 | 1 | 9 | 2 | 5 | 5 |
| **$r_2$** | 1 | 0 | 10 | 3 | 4 | 6 |
| **$r_3$** | 9 | 10 | 0 | 5 | 14 | 2 |
| **$r_4$** | 2 | 3 | 5 | 0 | 11 | 5 |
| **$r_5$** | 5 | 4 | 14 | 11 | 0 | 6 |
| **$r_6$** | 5 | 6 | 2 | 5 | 6 | 0 |

*Figure 5.4 Huang's distance matrix for example in Figure 5.3*

$d(r_1,r_2)=1$ and $d(r_1,r_3)=9$ where $d(r_i,r_j)$ is the distance (dissimilarity) for objects $r_i$ and $r_j$. It is important to note that the suggested weight was approximately 1.27 and would make little, if any, difference in the following discussion. The question raised is: should the magnitude associated with a numerical attribute give considerably more (or less) weight to the (dis)similarity measure? Note that $r_1$ has two values in common with both $r_2$ and $r_3$. In this case, it makes more sense to find a quantitative metric consistent across all of the attributes being considered in the measure. This follows from the idea of standardizing the data before the computation of similarity (Everitt, 1993).

Li (1998) developed a clustering method using the Goodall similarity metric (Goodall, 1966). This metric measures the amount of weight that a categorical value contributes to the overall similarity measure. For example, if two records have the same value for an attribute k, then the similarity isn't necessarily 1 while most previous metrics allow only 0 or 1, non-match or match respectively. The given value for a match is therefore a real number between 0 and 1. The metric allocates a value proportional to the frequency as compared with other values. This weight allows for a more consistent contribution between the two types of attributes. The following table gives distance measures using Li's metric (Goodall's metric) for the records.

| | r1 | r2 | r3 | r4 | r5 | r6 |
|---|---|---|---|---|---|---|
| **r1** | 0.16 | 0.31 | 0.82 | 0.29 | 0.67 | 0.66 |
| **r2** | 0.31 | 0.03 | 0.71 | 0.31 | 0.10 | 0.93 |
| **r3** | 0.82 | 0.82 | 0.41 | 0.58 | 0.99 | 0.44 |
| **r4** | 0.29 | 0.31 | 0.58 | 0.82 | 0.92 | 0.52 |
| **r5** | 0.67 | 0.10 | 0.99 | 0.92 | 0.10 | 0.71 |
| **r6** | 0.66 | 0.93 | 0.44 | 0.52 | 0.71 | 0.82 |

*Figure 5.5 Goodall's distance matrix for example in Figure 5.3*

*Figure 5.5* gives the Goodall distance measure for the records in *Figure 5.3*. For the same three items we have $d(r_1,r_2) \approx 0.31$ and $d(r_1,r_3) \approx 0.82$ ($sim(r_1,r_2) \approx 0.69$ and $sim(r_1,r_3) \approx 0.18$). This metric takes the distribution of values into account along with the magnitude of the quantitative values. The chi-squared ($\chi 2$) statistic is used for computing the measure between records. Because this statistic is used, there is the assumption of independence among the attributes, which cannot be guaranteed. It also may not always be the best idea to have the probability distribution influence the measure of similarity. In other words, if two records have the same qualitative value for an attribute and the value happens to be less common, should it contribute more to the measure than if it were common? A possible solution would be to discover a meaningful and feasible weighting on the attributes if one exists (Modha, 2002).

The above example gives two different distance measures. Because of the different possible values for the measures, it does not make sense to compare the two directly. Huang's measure has no upper bound (it lies on the range of positive reals) where Goodall's measure is on the real line interval 0 to 1. In fact, Goodall's metric is not *mathematically metric* where as Huang's is. For example, with Goodall's metric the dissimilarity of $r_1$ with itself is 0.16 and not 0. This means that we cannot say whether the measures are comparable with regard to the similarity of the records. It in fact begs the question of how to decide the usefulness of the measures.

One approach that has been proposed by Hirano (2001) uses a metric similar to Huang's (1999), but differs only in the metric for measuring the distance between the quantitative attributes. Hirano (2001) uses the Mahalanobis metric. This metric for quantitative data solves the difficulty of the covariance of the data. The clustering uses this metric with rough sets to find an optimal clustering. The use of rough sets seems to be a useful approach to clustering records of mixed data types, but it is unclear whether this method works for mixed data.

In developing a suitable metric, the question of whether the metric should be mathematically metric arises. It would be natural to say that it should; yet, the answer to this question may be context-dependent. For example, the Goodall measure (Goodall, 1966) was developed with biological taxonomy in mind where the commonality of a characteristic influence how similar two objects with the characteristic are. This may or may not be desirable in some applications. In addition, the importance of one attribute may outweigh another.

Gower (1971) developed a metric similar to Goodall's in the field of taxonomy. His metric is defined as:

$\mathrm{sim}(X,Y) = \Sigma\ (\mathrm{wt}_{xy(i)}*\mathrm{sim}_{xy(i)})\ /\ \Sigma\ \mathrm{wt}_{xy(k)}$ where $\mathrm{wt}_{xy(j)}$ is the weight apportioned to the $j^{th}$ attribute. This weight is usually 0 or 1, and allows for the handling of missing attribute values of one or both records being compared. $\mathrm{sim}_{xy(k)}$ is the similarity measure between the $k^{th}$ attribute values for records $X$ and $Y$. For qualitative data, $\mathrm{sim}_{xy(k)}$ is either 1 or 0, whether the values match or not, respectively. For quantitative values, $\mathrm{sim}_{xy(k)}$ is the proportion of difference in magnitude and the range for the $k^{th}$ attribute. This metric may be more useful. The portion contributed by the quantitative attributes to the measure relies on the range of values, where as the portion contributed by the qualitative attributes is fixed for any qualitative attribute. It is unclear whether the difference between the two measures combined compromises the utility of the metric. It would seem that this is essentially the same difficulty: combining measures that lie in different spaces. As of yet, this metric has not been used in clustering data sets of mixed type.

## 6. Rough Sets

Rough set theory is a mathematical tool introduced by Pawlak (1982) and is used to deal with uncertainty and vagueness. What follows is a general introduction. For a more detailed, mathematical introduction, see (Pawlak, 1991). The basic idea behind rough sets is the notion of *indiscernibility* and *equivalence*. For a set X, an item is either in the set, not in the set or it's membership in the set is unknown. That is, if we are given the following set:

| | **Sky** | **Temp** | **Swim** |
|---|---|---|---|
| $\mathbf{r_1}$ | *rainy* | *warm* | *no* |
| $\mathbf{r_2}$ | *sunny* | *hot* | *yes* |
| $\mathbf{r_3}$ | *rainy* | *cold* | *no* |
| $\mathbf{r_4}$ | *sunny* | *hot* | *no* |
| $\mathbf{r_5}$ | *sunny* | *warm* | *yes* |

*Figure 6.1*

and we are interested in the set for swim=yes, {r2,r5}. This set is defined by:

- *lower approximation*: the records that are definitely in the set {r5}
- *boundary region*: those records that whose membership cannot be determined {r2,r4}and the
- *upper approximation:* those records definitely in the set and those in the boundary region {r2,r5,r4}

In this case, the set {r2,r5} is termed *rough*, the boundary region is not empty (Pawlak, 1991). Another way to view the concept of rough is visually, as in Figure 6.2. The concept, *A* is represented by the shaded irregular shape. The area in between the two rectangles, the *upper* and *lower* approximations, is the boundary region.

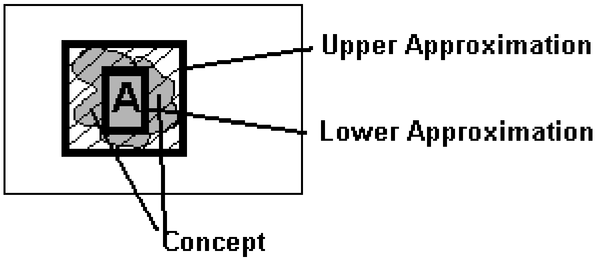


*Figure 6.2 Visual representation of a rough set.*

As discussed in section 5.3.1, there is an amount of uncertainty that arises when computing the similarity between records of mixed data. It is this uncertainty that leads us to apply soft computing techniques when clustering mixed data. Rough sets, in particular, lends itself, we believe, useful while less computationally expensive.

# 7. Proposed Research

In regard to knowledge discovery and to clustering in particular, it would be useful to have a fundamental similarity measure for data records. Unfortunately, few exist that account meaningfully for any combination of types of data. The more meaningful metrics known are restrictive to a particular area or science. How to combine difference in magnitude and simple matching so that it is general enough for mining is a question that is yet to be reasonably answered. In clustering records of mixed data types, questions exist that need to be answered.

## 7.1 Open Issues

When clustering data of mixed types, some questions arise which need to be answered. These questions include the following:

1. What is a useful representation for a cluster when there is a mix of data? When clustering records of mixed data types, how to best represent a cluster during the clustering process becomes a problem. Previous approaches use the mean of quantitative attributes with the mode for qualitative attributes. This may not be the most suitable method. For example, let *A* be a cluster representation, *B* be a record in the same cluster that *A* represents, and *C* be a record that is being compared with cluster representations in order to be clustered (or regrouped into a new cluster). Let j represent a qualitative attribute and $b_j$ be the next most frequent value for *A*'s cluster with $b_j$ not the same as $a_j$ (the mode for the $j^{th}$ attribute in the cluster). If $c_j$ is the same as $b_j$, the representation should give a higher similarity measure than if *C* had a $j^{th}$ value that was not present (or even frequent) in *A*'s cluster.
2. What metric would be most useful when records are of mixed data type? How can we be certain of the metric's utility? Some metrics are more suitable for some distributions than others. But, to analyze the data to the extent that one is relatively certain the measure selected is useful, defeats the purpose. That is, where is the line between utility or reasonableness and complexity?
3. When the clustering process is completed, how can we best evaluate the quality of the clustering? This includes being able to evaluate developed algorithms. This is related to question 2. Once a useful metric is available, it should be reasonable to evaluate the clustering produced. The problem is discovering what is useful and what is not.
4. Should the similarity (or dissimilarity) metric be *mathematically metric* or not? Is there a useful combination of two or more metrics—some *mathematically metric* and others not? Everitt (1997) discusses the relationship between distance and dissimilarity. In particular, he states that distance is a metric (*mathematically metric*) dissimilarity measure. It would seem that the answer to the question of whether similarity should be metric would be non-metric. Consider the similarity between three items, *i*, *j*, and *k*, should $\text{sim}(i,j) + \text{sim}(i,k) \geq \text{sim}(j,k)$? This would need to be examined further through experimentation, particularly the combining of metric and non-metric measures.
5. Can we ignore the context in the derivation of the similarity measure? For example, in one case, a qualitative attribute may have more impact on the measure than in another context. If we do ignore context, will the clustering quality be compromised? Consider the approach to extend the k-means. There is one specified weight for each of the qualitative attributes. When considering similarity between cars, should the color have the same contribution as the number of doors? If in fact context is ignored, as it usually is in any unsupervised approach, then the question raised is: when combining measures of quantitative and qualitative similarity, how should they be combined so that neither the quantitative nor the qualitative measure dominates the overall record measure?

The second and third questions are closely related. Once we can usefully determine the similarity between two records, the evaluation of clustering can be derived. The difficulty is in determining the similarity with mixed types of data.

### 7.2 Questions Proposed Research Seeks To Address

The objective is to be able to cluster records of mixed data types (quantitative and qualitative) efficiently and usefuly. An example of this type of data set is given in *Figure 5.3*. It is likely that not all questions stated will be dealt with in this research. For example, the question of context independence will not be a focus. The proposed research seeks to find a general clustering based on no prior information concerning the domain of the data.

A scalar metric may not be the answer because of the different properties of the data types. A weighting may not be computationally feasible because it is likely the weighting will depend on the proportion of data types in the records, the different distributions of each attribute and the different properties of the data types. This appears computationally costly because this will be dependent on each particular data set to be clustered. It may be that soft computing techniques can be used to cluster records of mixed data type.

Another possible approach to cluster records of mixed data types is to develop a hybrid algorithm—for example dividing the data set into sets containing qualitative and quantitative attributes and clustering the divided set. The resulting partitions can then be combined in some fashion for an overall clustering.

This work will develop a novel approach to multi-modal clustering. This work may extend or combine existing methods or develop an entirely new approach. Although the question of a unifying metric will be explored, it may be the case that a methodology independent of a single metric may be developed. Existing methods will be tested prior to testing my method.

Testing will be performed on several data sets, for example, U. S. census data. Some other possible domains are historical, biological, political, and geological. In addition to real data sets, synthetic data will also be used as in Li (1998). Li (1998) created several synthetic data sets with a known classification and distribution for each class attribute pair. For example, there are x number of classes, y number of qualitative attributes and z number of quantitative attributes. For each class, a distribution is specified for each attribute. The data is then created according to this distribution. More data sets are then created through incorporating different levels of error into each of the attribute distributions during the creation process.

Testing will be for: evaluating whether there is variation in effectiveness of methods across domains, determining the type of multi-modal data which can be effectively handled by a particular method, and determining the validity of the clustering results.

As a goal of this work is to discover a metric or method to group records without semantic knowledge of the data, the methods developed will likely be heuristic as the problem seems to be NP-hard. Potentially, many competing clusters are possible. Whether any particular cluster is a good one needs to be evaluated. Because of the subjectivity of clustering, this work will test results using human evaluation. Human evaluation will be to determine whether any particular clusters can be reasonable. The results will be presented systematically to humans who will evaluate the appropriateness of the clusters. Although, it may be the case that the number of groups discovered is too large to humanly evaluate in a timely fashion. If this becomes the case, a comparison of results from 2 or 3 algorithms may be made. In comparing results, analytical insights into why one result is more or less reasonable than another will be included.

Scalability is a crucial issue in data mining. Therefore, a discussion of the scalability will also be included. It is known that as the amount of semantic knowledge is decreased, the complexity of a method increases. A complexity analysis of both the computation of the developed metric and the developed algorithm will be included.

This research will develop a method or metric(s) to cluster multi-modal data. This research will utilize experiments to investigate under which conditions various data fusion methods may be used. Intuitively, it would seem that different classes of problems might be better resolved by different techniques. However, only heuristic exploration through experimentation can tell.

## Bibliography

1. G. Biswas, J. Weinberg, and D. Fisher. "ITERATE: A Conceptual Clustering Algorithm for Data Mining", in *IEEE Records on Systems, Man and Cybernetics-Part C: Applications and Reviews*, Vol 28(2), pp.219-230. May, 1998.

2. B. Everitt. *Cluster Analysis*, 3rd ed. Hodder & Stoughton, London, 1993.

3. B. Everitt and S. Rabe-Hesketh, *The Analysis of Proximity Data*, Wiley, New York, 1997.

4. V. Ganti, J. Gehrke, and R. Ramakrishnan. "CACTUS: Clustering Categorical Data Using Summaries" in *Knowledge Discovery and Data Mining*, pp. 73-83, 1999.

5. D. Gibson, J. Kleinberg, and P. Raghavan. "Clustering Categorical Data: an Approach Based on Dynamical Systems" in *Proceedings of the 24th VLDB Conference*, Vol. 8(3/4), pp. 222-236, 2000.

6. D. Goodall. "A New Similarity Index Based On Probability" in Biometrics, Vol. 22(4), pp. 882-907, 1966.

7. K. Gowda and E. Diday. "Symbolic Clustering Using a New Dissimilarity Measure" in *Pattern Recognition*, Vol. 24(6), pp. 567-578, 1991.

8. J. Gower. "A General Coefficient of Similarity and Some of Its Properties" in Biometrics, Vol. 27(4), pp. 857-871, 1971.

9. S. Guha, R. Rastogi, and K. Shim, "ROCK: A Robust Clustering Algorithm for Categorical Attributes", in *Information Systems*, Vol. 25(5), pp. 345-366, 2000.

10. S. Gupta, K. Rao, and V. Bhatnagar, "K-Means Clustering Algorithm For Categorical Attributes", in *Data Warehousing and Knowledge Discovery*, pp. 203-208, 1999.

11. E. Han, G. Karypis, V. Kumar, and B. Mobasher. "Clustering Based on Association Rule Hypergraphs", in *Proceedings of SIGMOD '97 Workshop on Research Issues in Data Mining and Knowledge Discovery (DMKD'97)*, May 1997.

12. J. Han and M. Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers, San Fransisco, 2001.

13. S. Hirano, T. Okuzaki, Y. Hata, S. Tsumoto, and K. Tsumoto. "A Rough Set-Based Clustering Method with Modification of Equivalence Relations", in *PAKDD 2001*, D. Cheung et al eds., pp. 513-518, 2001.

14. Z. Huang and M. Ng. "A Fuzzy k-Modes Algorithm for Clustering Categorical Data", in *IEEE Records on Fuzzy Systems*, Vol 7(4), pp. 446-452, August 1999.

15. Z. Huang. "Clustering Large Data Sets With Mixed Numeric and Categorical Values", in *Proceedings of 1st Pacific-Asia Conference on Knowledge Discovery & Data Mining*, 1997.

16. Jain and R. Dubes. *Algorithms for Clustering Data*. Prentice Hall, New Jersey, 1988.

17. C. Li and G. Biswas. "Conceptual Clustering With Numeric-and-Nominal Mixed Data-A New Similarity Based System", in *IEEE Transcript on KCE* 1998.

18. Pawlak, Z. "Rough Sets", in *International Journal of Computer and Informational Sciences*, Vol. 11, no. 5, pp. 341-356, 1982.

19. Pawlak, Z. *Rough Sets: Theoretical Aspects of Reasoning about Data*. Kluwer Academic Publishers, Dordrecht, 1991.

20. H. Ralambondrainy. "A Conceptual Version of the K-Means Algorithm" in Pattern Recognition Letters, Vol. 16, pp. 1147-1157, 1995.

21. P. Sneath and R. Sokal. *Numerical Taxonomy*. Freeman and Company, San Fransisco, 1973.

22. K. Wang, C. Xu, and B. Liu. "Clustering Records Using Large Items", in CIKM 1999. pp. 483-490.

23. Y. Zhang, A. Wai-chee Fu, C. Cai, and P. Heng. "Clustering Categorical Data", in

*24. 16th International Conference on Data Engineering, 2000.*