Rough Sets Used in the Measurement of Similarity of Mixed Mode Data

Sarah Coppock
Lawrence Mazlack
Applied Artificial Intelligence Laboratory
ECECS Department
University of Cincinnati
Cincinnati, Ohio 45220
{coppocs,mazlack}@uc.edu

Abstract

Similarity is important in knowledge discovery. Cluster analysis, classification, and granulation each involve some notion or definition of similarity. The measurement of similarity is selected based on the domain and distribution of the data. Even within a specific domain, some similarity metrics may be considered more useful than others. There is an amount of uncertainty in quantitatively measuring the similarity between records of mixed data. The uncertainty develops from the lack of scale that both nominal and ordinal data have. Rough set theory is one tool developed for handling uncertainty. Rough sets can be used in dissimilarity analysis of qualitative data. It would seem that rough sets could be applied in measuring similarity between records containing both quantitative and qualitative data for the purpose of clustering the records.

1. Introduction

Similarity metrics are used in many fields. When determining the similarity between records in a data set that contains different kinds of data, a certain amount of uncertainty is introduced. While metrics such as Euclidean and its generalized Minkowski metrics can be used when all of the data is quantitative (both discrete and continuous), it is not as easy to usefully combine scalar metrics representing qualitative data (both nominal and ordinal).

Data can be categorized into qualitative and quantitative data. Qualitative data can be further described as either ordinal or nominal. Ordinal data has order without scale, e.g., *small*, *medium*, *large*. Nominal data has no order and no scale, e.g., *Cincinnati*, *Tampa*, *Atlanta*. Data such as cities and colors can be argued as having some "order," e.g., latitudes or longitudes and frequencies. However, in the case of unsupervised learning, such knowledge is not explicitly known by the learning algorithms. For a more detailed discussion of data varieties, see [1] [2].

Similarity is important in knowledge discovery. Cluster analysis, classification, and granulation each involve some notion or definition of similarity. Measuring similarity between multidimensional, multi-modal data is difficult, but offers information. The information provided by clustering records based on similarity measurements includes an overall distribution of the data and discovery of possible outliers. The measurement of similarity may be appropriately selected based on the domain and distribution of the data. Even within a domain, there may be some similarity metrics considered more useful than others.

There is an amount of uncertainty in quantitatively measuring similarity (or dissimilarity) between records of mixed kinds of data. The uncertainty develops from the fact that both nominal and ordinal data lack a natural fixed scale. For example, should we say that the similarity between *red* and *orange* is more or less than (or equal to) the similarity between *blue* and *green*? Some metrics assign a Boolean value for whether the values match. The similarity would be considered equal.

Rough set theory is one of the tools developed for handling uncertainty. Pawlak [4] demonstrates how rough sets can be used in dissimilarity analysis of qualitative data. It would seem that rough sets could be applied in measuring similarity between records containing both quantitative and qualitative data.

2. Rough sets in dissimilarity analysis

Rough sets are built on the notion of indiscernibility. There is an equivalence relation imposed on the items based on attribute values. For example, in Table 1, r_1, r_3, and r_6 are considered indiscernible with respect to a_{10}, denoted $IND(a_{10})$. A similar idea is used in similarity metrics for nominal and ordinal data. A simple matching technique where a Boolean *0* or *1* is assigned based on whether two attribute values are the same for two records.

Pawlak [4] describes applying rough sets to measure dissimilarity between records of Boolean values. A brief description of Pawlak's method to measure dissimilarity using rough sets follows using his Middle East Situation example. This example is given in Table 1.

Any attributes that have the same values as another attribute for all records, i.e. equivalent attributes, to one another, are disregarded. For example, a_1, a_7 and a_8 in Table 2 have the same value for each record. Only one of

a_1, a_7 and a_8 would then be considered further in the process, but not all three. This is because once one of the attributes is taken into account; the other two do not offer any more information in computing the dissimilarity. In computing similarity, it would be seem desirable to take into account that multiple attributes are equal. That is, the more values two records have in common the greater the similarity is between the records.

Attributes that have the same value for all records are disregarded. For example, a_5 in Table 2 would be disregarded since there is only one value, *1*, for all records. The attribute does not offer any information in discerning between any of the records. Attributes that are the negation of another, such as a_4 with a_1, a_7 and a_8 in Table 2, are also disregarded. Only one of a_1, a_4, a_7 or a_8 in Table 2 would be considered.

Table 1. Middle East Situation example

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	a_8	a_9	a_{10}
r_1	0	1	1	1	1	1	0	0	1	1
r_2	1	1	1	0	1	1	1	1	1	0
r_3	1	1	0	0	1	1	1	1	1	1
r_4	1	1	0	0	1	1	1	1	1	0
r_5	1	1	0	0	1	1	1	1	0	0
r_6	1	1	1	0	1	1	1	1	1	1

Table 2. Small example

	a_1	a_2	a_3	a_4	a_5
r_1	0	0	1	0	1
r_2	1	1	0	1	1

Table 1 is modified for measuring dissimilarity. One of the possible resulting modified sets is given in Table 3.

Table 3. Possible modified set from Table 1

	a_1	a_3	a_9	a_{10}
r_1	0	1	1	1
r_2	1	1	1	0
r_3	1	0	1	1
r_4	1	0	1	0
r_5	1	0	0	0
r_6	1	1	1	1

A graph is then constructed from the modified table. There is a node for each record and a labeled edge between the nodes if removing an attribute would put the records in the same equivalence class. For example, an edge is between r_1 and r_6 with the label a_1 since these records would be in the same equivalence class, $IND(a_1)$. Figure 5 shows the graph for Table 3. The dissimilarity between two records is computed by determining the length of the shortest path between the nodes in the graph corresponding to the records. For example, the dissimilarity between r_1 and r_2 would be 2.

Table 4. Core values for Table 3

	a_1	a_3	a_9	a_{10}
r_1	0	-	-	-
r_2	-	1	-	0
r_3	-	0	-	1
r_4	-	0	1	0
r_5	-	-	0	-
r_6	1	1	-	1

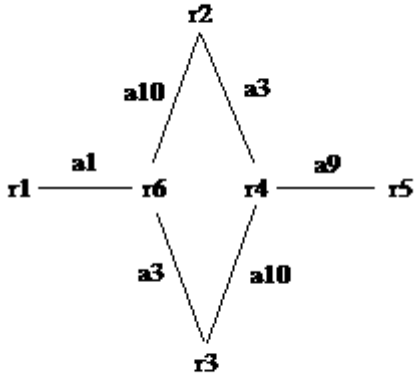

Figure1. Graph for Table 3

Dissimilarity is the complement to similarity. Because of the relationship between dissimilarity and similarity, we could modify the above approach to quantify similarity between records.

A consideration to make in modifying this approach is the generalization to multi-valued attributes. For example, if one attribute has more than two or three values such as the make of a car {*Ford*, *GM*, *Toyota*, *Nissan*, *BMW*}.

3. Converting quantitative attributes into qualitative

A common approach to measure similarity between records containing mixed data is to add the measurements of qualitative similarity and of quantitative similarity. Without knowledge of the domain and specifically the data set description, finding an appropriate weighting to give reasonable results would be computationally expensive.

Methods to cluster quantitative data have been developed. One possibility for the discovery of similar records in multi-modal data would be to convert the quantitative attributes into one qualitative attribute according to the "natural" clusters in the quantitative attributes. The modified rough set dissimilarity analysis approach can then be applied.

Table 5 gives an example mixed data set with one nominal (a_1), one ordinal (a_2: {*A*, *B*, *C*, *D*, *F*} in order) and one discrete quantitative (a_3) attribute. Table 6 is the modified data set with the quantitative attribute, a_3, in Table 5 clustered into c_1 and c_2. The attribute a'_3 is the label in which the value for the record belongs.

Table 5. Example mixed data set

	a_1	a_2	a_3
r_1	Coke	B	4
r_2	Coke	C	2
r_3	Pepsi	B	1
r_4	Pepsi	A	1
r_5	Bud	F	2
r_6	Heineken	B	3

Table 6. Modified data set from Table 5

	a_1	a_2	a'_3
r_1	Coke	B	c2
r_2	Coke	C	c1
r_3	Pepsi	B	c1
r_4	Pepsi	A	c1
r_5	Bud	F	c1
r_6	Heineken	B	c2

A modified approach as in Section 2 may now be applied to determine pair-wise record similarity. Figure 2 shows the graph associated with Table 6. From the graph, it can be seen that some modification to handle multi-valued attributes needs to be made. The graph is not connected. Table 7 which provides the similarities also demonstrates this need. The similarities are computed as: $(D_{max}-D_{ij})/D_{max}$, where D_{max} is the maximum dissimilarity over all pairs and D_{ij} is the dissimilarity between r_i and r_j.

After following the method in a straight-forward manner at this point, it is unclear whether the difficulty in normalizing between attributes is handled. Regardless of the method used to cluster or granulize the data, it is difficult to evaluate whether the results are reasonable. That is, if we give a small data set such as the one in our example to a number of students, how would they group the records? It seems that this situation is more suited to having a fuzzy measure associated with a single particular grouping or a clustering of records. The kind of measure and how to define the measure function is unclear at this point. In this case, heuristics such as those used in [3] [6] can be used to restrict the search space to those that would be more likely to have a higher measure of certainty.

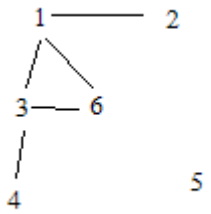

Figure 2. Graph associated with Table 5

Table 7. Pairwise similarities for Table 5

	r_1	r_2	r_3	r_4	r_5	r_6
r_1	1	2/3	2/3	1/3	1	2/3
r_2	2/3	1	1/3	0	1	1/3
r_3	2/3	1/3	1	2/3	1	2/3
r_4	1/3	0	2/3	1	1	1/3
r_5	1	1	1	1	1	1
r_6	2/3	1/3	2/3	1/3	1	1

4. Fusing quantitative and qualitative information

Metrics and methods have been developed to cluster data records that have only quantitative or only qualitative data. It is possible that information can be extracted by the "fusion" of the methods' results or the different measures.

Metrics are defined in or can be normalized to the interval [0,1]. Quantitative measures lie on the whole continuous interval while qualitative measures lie on a discrete linear subset of the interval.

4.1 Fusing quantitative and qualitative partitions

Metrics and methods have been developed to cluster records containing only one type of data. The results of these methods and metrics have different meanings. The characteristics that contribute to the similarity measures are different. It is possible that rough sets can be used in the fusion of the results of existing methods for the two sets of dimensions.

Let C_q (X) denote a clustering method for the quantitative dimensions of the data set X and $C_n(X)$ denote the clustering method for qualitative dimensions of the data set X. C_q clusters X based only on the quantitative attributes and C_n clusters X based only on the qualitative attributes. Let $C_q(X)=\{q_i,q_2,\ldots q_k\}$ and $C_n(X)=\{n_1,n_2,\ldots n_m\}$ where the sets of q_i and n_i are the clusters that result according to the quantitative and qualitative attributes respectively. Table 9 shows one possibility for the results of C_q and C_n applied to the simple example data set in Table 5. Note that the q_i and n_i

are arbitrary and are not a result of any specific metric or method.

There is one q_i and one n_i for every record. Let $s_i = q_i \bullet n_i$ for a given r_i. The set s_i contains all of the records considered similar to the record r_i according to some quantitative and/or some qualitative metric or method. There may be some order of the elements in q_i and n_i according to the similarity to r_i. That is, given $q_i=\{q_{i(1)}, q_{i(2)}, \ldots q_{i(k)}\}$ where $q_{i(j)}$ is the j^{th} record in the set q_i, it may be the case that $s(r_i,q_{i(j)}) \bullet s(r_i,q_{i(k)}) \bullet s(r_i, q_{i(m)})$ for any j, k, and m. The same may be true for the set n_i.

Table 8. Possible C_q and C_n for Table 6

C_q	r_1	r_2	r_3	r_4	r_5	r_6
q_1	1	1	0	0	0	0
q_2	0	0	1	1	0	0
q_3	0	0	0	0	1	1

C_n	r_1	r_2	r_3	r_4	r_5	r_6
n_1	1	0	0	0	0	1
n_2	0	1	1	1	1	0

Table 9. S_i for Table 8

	r_1	r_2	r_3	r_4	r_5	r_6
s_1	1	1	0	0	0	1
s_2	1	1	1	1	1	0
s_3	0	1	1	1	1	0
s_4	0	1	1	1	1	0
s_5	0	1	1	1	1	1
s_6	1	0	0	0	1	1

Table 9 gives the s_i for the example from Table 8. We can infer from the s_i, that the similarity between r_3 and r_4 is greater than the similarity between r_2 and r_3. Records r_3 and r_4 have the same membership for a greater number of s_i (all of the s_i) than the pair r_2 and r_3 which differ in s_1. We can also infer that r_3 and r_4 belong together in the overall clustering of the data set. For each s_i, r_3 and r_4 have the same membership.

Thus far, we have not addressed a weighting of attributes. For example, if there are 2 qualitative dimensions and 10 quantitative dimensions, it seems reasonable that the q_i would have more weight in determining the overall clusters. The overall clustering would be more like the resulting quantitative clusters.

The fact that there may exist an order to each of the sets leads to the idea that rough sets may be used in the development of a fuzzy measure. The measure may be either a specific group identified as being “similar” or an overall clustering of mixed data.

4.2 Fusing qualitative and quantitative measures

The sets C_q and C_n provide less information than having pair-wise similarity measurements. Suppose we are given the following: $C_n=\{\{x_1,x_4\}\{x_2,x_3\}\{x_5,x_6\}\}$ and $C_q=\{\{x_1,x_2,x_3,x_5\}\{x_4, x_6\}\}$. Both $\{x_1,x_4\}$ and $\{x_5,x_6\}$ are in different clusters in C_q. Suppose that the qualitative similarity between $\{x_1,x_4\}$ is maximal while the qualitative similarities between $\{x_2,x_3\}$ and $\{x_5,x_6\}$ are less than maximal. Suppose also that the quantitative similarity between $\{x_1,x_4\}$ is minimal while the quantitative similarity between $\{x_5,x_6\}$ is greater than minimal. We are not able to compare these similarities to determine if either pair should be kept together. It may be more useful to consider the pair-wise qualitative and quantitative similarities.

One can consider "rough sets" from the perspective of each record. In other words, there are those records which definitely belong in the same cluster as the record (lower approximation), those that definitely do not belong in the same cluster, and those that it is uncertain whether they belong in the same cluster (boundary). Each of these can be determined by given similarity values. For example, we can say that for any two records if the similarity measurement is less than some threshold then they are not in each others cluster approximation. One can define a similar threshold for those records that definitely belong in the same cluster. What these thresholds should be are subjective both to a particular domain and the metric that it used.

Suppose we have the following similarity matrices for qualitative and quantitative dimensions respectively in Table 10 and Table 11. The qualitative measure is computed as: number of matching attribute values / number of qualitative attributes. The quantitative measure is computed as:

$$1 - \sum_{\text{k quantitative}} \frac{|x_{ik} - x_{jk}|}{R_k}$$

where x_{mk} is the k^{th} attribute value for record m and R_k is the range of attribute k.

Table 12 and Table 13 give the approximations with the lower threshold 1/2 and the upper approximation threshold of 9/10. 0 denotes that the record is not in the approximation. 1 denotes that the record is in the lower approximation. Lastly, '--' denotes that the record is in the boundary. For example, in both Table 12 and Table 13, r_6 is in the boundary for r_1.

From Table 12 and Table 13 we can see that for the cluster including r_1, the most likely record in the same cluster would be r_6 since it is in both approximations. One could use a similar idea to section 4.1 and use the union of the upper approximations to determine likely clusters. For example, $\{r_1,r_2,r_3,r_6\}$ based on the sets for r_1 in both tables.

Table 10. Qualitative similarities for Table 5

	r_1	r_2	r_3	r_4	r_5	r_6
r_1	1	1/2	1/2	0	0	1/2
r_2	1/2	1	0	0	0	0
r_3	1/2	0	1	1/2	0	1/2
r_4	0	0	1/2	1	0	0
r_5	0	0	0	0	1	0
r_6	1/2	0	1/2	0	0	1

Table 11. Quantitative similarities for Table 5

	r_1	r_2	r_3	r_4	r_5	r_6
r_1	1	1/3	0	0	1/3	2/3
r_2	1/3	1	2/3	2/3	1	2/3
r_3	0	2/3	1	1	2/3	1/3
r_4	0	2/3	1	1	2/3	1/3
r_5	1/3	1	2/3	2/3	1	2/3
r_6	2/3	2/3	1/3	1/3	2/3	1

The difficulty in comparing different measures is still present because there still exists the problem of which approximation a resulting cluster should be more like. For example, since the thresholds and therefore the equivalence relations are based on two different measures, we cannot infer whether a likely result would be $\{r_1,r_2,r_3,r_6\},\{r_1,r_6\},\{r_1,r_2,r_3\},\{r_1,r_2,r_6\}$, or $\{r_1, r_3,r_6\}$. For this reason, it would seem that a fuzzy measure is needed for the unsupervised discovery of similar records in mixed data.

Table 12. Approximations for qualitative attributes

	r_1	r_2	r_3	r_4	r_5	r_6
r_1	1	--	--	0	0	--
r_2	--	1	0	0	0	0
r_3	--	0	1	--	0	--
r_4	0	0	--	1	0	0
r_5	0	0	0	0	1	0
r_6	--	0	--	0	0	1

Table 13. Approximations for quantitative attributes

	r_1	r_2	r_3	r_4	r_5	r_6
r_1	1	0	0	0	0	--
r_2	0	1	--	--	1	--
r_3	0	--	1	1	--	0
r_4	0	--	1	1	--	0
r_5	0	1	--	--	1	--
r_6	--	--	0	0	--	1

5. Summary

This paper discussed two approaches for determining similarity between records of mixed data. From both ideas, it can be seen that due to the uncertainty and vagueness of qualitative data and of trying to combine metrics leave rough set theory as an optional tool to be used. As concluded in the discussion, an additional or other approach is needed in the discovery of similar groups of records within data sets of mixed data.

References

[1] Everitt, B. Cluster Analysis, 3rd ed., Hodder & Stoughton, London, 1993.

[2] Han, J and Kamber, M. Data Mining: Concepts and Techniques, Morgan Kaufmann, San Francisco, 2001.

[3] He, A. Unsupervised Data Mining by Recursive Partitioning, Masters Thesis, University of Cincinnati, June, 2002, 102 pg.

[4] Pawlak, Z. Rough Sets: Theoretical Aspects of Reasoning About Data, Kluwer Academic Publishers, Dordrecht, 1991.

[5] Sneath, P and Sokal, R. Numerical Taxonomy, W. H. Freeman, San Francisco, 1973.

[6] Zhu, Y., Unsupervised Database Discovery Based on Artificial Intelligence Techniques, Masters Thesis, University of Cincinnati, June, 2002, 107 pg.