A New Symbolic Dissimilarity Measure for Multivalued Data Type and Novel Dissimilarity Approximation Techniques

Bapu B Kiranagi
Imaging COE,
HCL Technologies, Bangalore

D S Guru
DoS in Computer Science
University of Mysore
Manasagangotri, Mysore, India

ABSTRACT
In this paper a new statistical measure for estimating the degree of dissimilarity between two symbolic objects whose features are multivalued type is proposed. In addition two new simple representation techniques viz., interval type and magnitude type for the computed dissimilarity between the symbolic objects are introduced. The dissimilarity matrices obtained are not necessarily symmetric. Hence, clustering algorithms to work on such unconventional approximated matrices, by introducing the concept of mutual average dissimilarity value and magnitude average dissimilarity respectively for interval type and magnitude type approximation representations are also proposed.

Categories and Subject Descriptors
Artificial Intelligence, Pattern Recognition and Database systems.

General Terms
Algorithms, Measurement, Experimentation and Theory.

Keywords
Symbolic Data Analysis, Proximity Approximation, Clustering Algorithms

1. INTRODUCTION
A symbolic object is defined by its intent which contains a way of finding its extent. For instance, the description of the inhabitant of a region and the way of allocating an individual to this region is called intent, the set of individual which satisfies this intent is called extent. The syntax of symbolic objects must have an explanatory power. Symbolic data are extensions of classical data types. In conventional datasets, the objects are individualized whereas in symbolic datasets they are more unified by means of relationship. The relationships between symbolic objects may appear in the form continuous ratio, discrete absolute, interval, modal, multivalued and also multivalued data with weights. However, in the literature many proximity measures have been proposed [1] [2][3][4][5][6][7]. These proximity measures except for [5][6], compute the proximity between symbolic objects in crisp symmetric form. However in [5][6], the proximity is approximated in multivalued type which is not necessarily symmetric. In this paper the measure proposed in [6] is modified to suit the multivalued data type. The proposed measure is used to compute the proximity of multivalued symbolic data types and the measure proposed in [6] for interval symbolic data. Furtherance, new proximity approximation techniques viz., interval data type and crisp data type is proposed in this paper. In addition, unlike conventional proximity matrices, the dissimilarity matrices obtained by the newly proposed approximations are not necessarily symmetric. Methods of clustering symbolic data based on the obtained unconventional dissimilarity matrices are also explored in this paper by introducing the concepts of Mutual average dissimilarity value and magnitude average dissimilarity value which are similar to the concept of MDV proposed in [6] for multivalued approximation technique. These new concepts preserve the degree of mutual farness possessed by two symbolic objects, there by retaining the naturalistic proximity characteristics of the objects. Experiments on standard benchmark dataset have been conducted in order to study the efficacy of the proposed methodology. The proposed methodology, in fact possess two important principles of symbolic data analysis, namely, the coherence and explicability principle [2]. The coherence principle states that the input and output should be expressed by the same kind of symbolic objects and the explicability principle states that the results must be easily interpretable by the user even if they are less efficient [2][3][4].

The paper is organized as follows. Section 2 proposes the dissimilarity measure for estimating the degree of dissimilarity between multivalued features of symbolic objects. In addition a brief of dissimilarity measure proposed in [6], for estimating the degree of dissimilarity between interval type symbolic objects is given. Clustering the symbolic objects whose proximities are approximated in interval form and crisp magnitude form are explained in Section 3. The section 4 introduces the concept of MADV and MgAdv for clustering the symbolic objects whose proximities are respectively approximated in interval and crisp form. The results of the experiments conducted are presented in the section 5. Finally conclusion is given in section 6.

2. SYMBOLIC DISSIMILARITY FOR MULTIVALUED DATA TYPE
In this section we propose the dissimilarity measure to compute the degree of dissimilarity between symbolic objects with mulitvalued features and in conjunction, the symbolic dissimilarity measure proposed in [6] to compute the degree of dissimilarity between symbolic objects suitable for interval features is also briefed.

Let O_i and O_j be two symbolic objects in n-dimensional space described by n = u + v number of symbolic features out of which u are of type interval and v are of type multivalued.

In case of features of type interval, the degree of dissimilarity of each feature value of O_i with respect to the corresponding feature value of O_j is estimated as follows:

i.e., $O_i = I_{i1}, I_{i2}, \ldots, I_{iu}, M_{i1}, M_{i2}, \ldots, M_{iv}$,

where, $I_{ik} = [f_{ik}^{-}, f_{ik}^{+}] \forall k = 1,2,...,u$ are of type interval and $M_{i1} = (m_{il}^{1}, m_{il}^{2}, ..., m_{il}^{x_{il}}) \forall l = 1,2,...,v$ are of type multivalued and

$O_j = I_{j1}, I_{j2}, \ldots, I_{ju}, M_{j1}, M_{j2}, \ldots, M_{jv}$

where, $I_{jk} = [f_{jk}^{-}, f_{jk}^{+}] \forall k = 1,2,...,u$ are of type interval and

$M_{j1} = (m_{jl}^{1}, m_{jl}^{2}, ..., m_{jl}^{x_{jl}}) \forall l = 1,2,...,v$ are of type multivalued.

Here, x_{pl} denotes number of multiple values describing the l^{th} multivalued feature of p^{th} object in general.

The degree of dissimilarity between two objects with respect to the interval valued features is estimated by using the dissimilarity measure proposed in [6]. The degree of dissimilarity object O_i to the object O_j with respect to the k^{th} interval features is characterized by

$$d_{i \to j}^{k} = \frac{|I_{ik}| + (\text{Max}(f_{ik}^{-}, f_{jk}^{-}) - \text{Min}(f_{ik}^{+}, f_{jk}^{+}))}{|I_{jk}|} \quad \ldots(1)$$

Similarly, the degree of dissimilarity object O_j to the object O_i with respect to the k^{th} interval features is given by

$$d_{j \to i}^{k} = \frac{|I_{jk}| + (\text{Max}(f_{ik}^{-}, f_{jk}^{-}) - \text{Min}(f_{ik}^{+}, f_{jk}^{+}))}{|I_{ik}|} \quad \ldots(2)$$

It shall be observed that the separability between the objects with respect to k^{th} features is generalized as $(Max(f_{ik}^{-}, f_{jk}^{-}) - Min(f_{ik}^{+}, f_{jk}^{+}))$ in the numerator of eqn (1). This term will be the compute the common part between I_{ik} and I_{jk} in case of overlapping and it will be the separability between I_{ik} and I_{jk} in case of non overlapping feature intervals.

2.1 Computation of degree of dissimilarity with respect to multivalued valued features

The degree of dissimilarity between two objects with respect to the multivalued features is estimated by computing the farness interms of the non common portion between the multivalued features of the object O_i and O_j. The multivalued features may or may not have elements in common. In case of commonality, the dissimilarity of the object O_i to the object O_j with respect to the l^{th} multivalued feature is given by the ratio of the non common portion of M_{il} with M_{jl} to $|M_{jl}|$. Hence the degree of dissimilarity of the object O_i to the object O_j with respect to the l^{th} multivalued features which are as shown in Fig 1 (a, b, c) is given by

$$d_{i \to j}^{l} = \frac{\text{Non common portion of } M_{il} \text{ with respect to } M_{jl}}{\text{Cardinality of } M_{jl}}$$

$$\Rightarrow d_{i \to j}^{l} = \left(\frac{|M_{il}| - |M_{il} \cap M_{jl}|}{|M_{jl}|} \right) \quad \ldots(3)$$

here, $M_{il} \cap M_{jl}$ represents the intersection of the sets M_{il} and M_{jl}, and $|.|$ represents the cardinality of the interval.

Further, in case of Fig 2, where M_{il} and M_{jl} have no element in common, the degree of dissimilarity is given by $\frac{|M_{il}|}{|M_{jl}|}$, which can also be expressed by $\frac{|M_{il}| - |M_{il} \cap M_{jl}|}{|M_{jl}|}$, because $|M_{i1} \cap M_{j1}| = \Phi$.

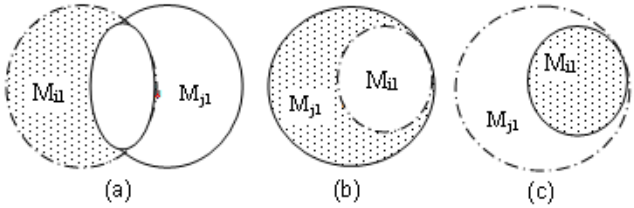

Fig 1 Shaded region showing the component of the dissimilarity between the objects

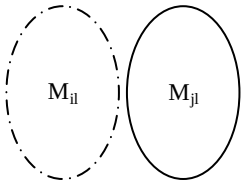

Fig 2 Possibility of non overlapping portion

Thus in general, let F_{ik} and F_{jk} be the k^{th} (k = 1, …, n) feature of the objects O_i and O_j, when F_{ik} and F_{jk} are of type multivalued, the dissimilarity is given by

$$d_{i \to j}^{k} = \left(\frac{|F_{ik}| - |F_{ik} \cap F_{jk}|}{|F_{jk}|} \right) \quad \ldots(4)$$

Similarly, on the other way round, the degree of dissimilarity of the object O_j to the object O_i, for multivalued type features, the degree of dissimilarity of the object O_j to the object O_i is given by

$$d^{k}_{j\to i} = \left(\frac{|F_{jk}| - |F_{ik} \cap F_{jk}|}{|F_{ik}|} \right) \quad \ldots(5)$$

It is evident from eqns (4) and (5) that $d^{k}_{i\to j}$ and $d^{k}_{j\to i}$ for features of type multivalued are not necessarily equal.

Once the degree of dissimilarity between two objects O_i and O_j with respect to each feature is estimated, we recommend to approximate the degree of dissimilarity from O_i to O_j and also from O_j to O_i as follows.

3. APPROXIMATION FOR CLUSTERING SYMBOLIC OBJECTS

3.1 Approximation through Interval Data Type

In this type of approximation the minimum and maximum of all the computed n dissimilarity values (due to all n features) respectively represent the minimum and maximum degree of dissimilarity from the object O_i to the object O_j.

Hence, the total degree of dissimilarity from the object O_i to the object O_j with respect to all n features is of interval type and is given by

$$D_{i\to j} = [D^{-}_{i\to j}, D^{+}_{i\to j}] \quad \ldots(6)$$

where $D^{-}_{i\to j} = \min\{d^{k}_{i\to j}, \forall k = 1,2,...,n\}$ and

$D^{+}_{i\to j} = \max\{d^{k}_{i\to j}, \forall k = 1,2,...,n\}$, represent the minimum and maximum degree of dissimilarity of the object O_i to the object O_j due to all n features.

Similarly, the degree of dissimilarity of the object O_j to the object O_i with respect to all n features is given by

$$D_{j\to i} = [D^{-}_{j\to i}, D^{+}_{j\to i}] \quad \ldots(7)$$

where $D^{-}_{j\to i} = \min\{d^{k}_{j\to i}, \forall k = 1,2,...,n\}$ and

$D^{+}_{j\to i} = \max\{d^{k}_{j\to i}, \forall k = 1,2,...,n\}$, represent the minimum and maximum degree of dissimilarity of the object O_j to the object O_i due to all n features.

It can be observed that the dissimilarity matrix through this approximation is of type interval and is non symmetric.

3.2 Approximation through Crisp Data Type

In this type of approximation the total degree of dissimilarity from the object O_i to the object O_j with respect to all n features is approximated to be a single crisp value computed by taking the magnitude of the vector representing the dissimilarity of the corresponding features between the object O_i and the object O_j and is given by

i.e. $$D_{i\to j} = \sqrt{\sum_{k=0}^{n} (d^{k}_{i\to j})^2} \quad \ldots(8)$$

On the other way round, the total degree of dissimilarity of the object O_j to the object O_i with respect to all n features is given by

$$D_{j\to i} = \sqrt{\sum_{k=0}^{n} (d^{k}_{j\to i})^2} \quad \ldots(9)$$

It is obvious that the dissimilarity matrix through this approximation technique is non symmetric.

As the dissimilarity between corresponding features of the objects is non symmetric, it is obvious that the presented approximation techniques are non symmetric. Therefore, the dissimilarity matrices obtained through this alternative approximation is also non symmetric. In view of this, here we propose clustering techniques to work on such unconventional dissimilarity matrices.

4. CLUSTERING METHODOLOGY

It can be observed that dissimilarity matrix obtained through the proposed approximations is not necessarily symmetric and will have the elements which are crisp or interval. But, most of the clustering algorithms so far proposed in literature insist that the dissimilarity matrix must have crisp values in addition to being symmetric. In reality, we can expect two different proximities between two symbolic objects and those proximities need not be equal and the same has been reflected by the dissimilarity measure and the approximation techniques proposed in the previous section. Thus, the development of clustering algorithms to work on such unconventional dissimilarity matrices has received a considerable attention in today's research. In view of this, in this section, we extend the existing agglomerative clustering techniques [8] by introducing the concepts of Mutual Average Dissimilarity Value (MADV) and Magnitude Average Dissimilarity Value (MgADV) which are similar to Mutual Dissimilarity value (MDV) based clustering technique proposed in [6] respectively for the Dissimilarity matrices obtained from crisp magnitude and interval.

4.1 Mutual Average Dissimilarity Value for Clustering

In this section, we present a clustering technique which works on the proposed interval type dissimilarity matrix. The interval type approximation is obtained by taking the minimum and maximum of the computed n dissimilarity values due to all n features between the object O_i to the object O_j.

The proposed method is based on a newly introduced concept called mutual average dissimilarity value (MADV). The MADV between two objects is defined to be the average of the midpoints of the intervals representing the degree of dissimilarity possessed by the objects with each other. That is, the MADV between the object O_i and the object O_j is given by

$$MADV_{ij} = \frac{[D^{-}_{i\to j} + D^{+}_{i\to j}] + [D^{-}_{j\to i} + D^{+}_{j\to i}]}{4} \quad \ldots(10)$$

The proposed clustering methodology is a modified version of an existing agglomerative clustering technique. The conventional agglomerative clustering technique looks at a dissimilarity value which is a crisp (a single value) inorder to merge the corresponding two objects into one cluster, where as our methodology computes MADV between two objects using the interval valued type data representing the degrees of dissimilarities between the objects inorder to merge them into one cluster. Since it is an agglomerative clustering, initially m clusters, each consisting an individual object are created, where m is the

total number of objects. Two objects belonging to different clusters possessing the maximum MADV are chosen and subsequently the corresponding clusters are merged together into a single cluster. If there are many more number of such pairs of clusters, then they are merged together at the same stage. This process of merging is continued till the desired number of classes are obtained or a class consisting of all the objects is obtained. Thus, the proposed methodology for clustering symbolic objects based on MADV is as trivial as follows.

Algorithm 1: MADV based agglomerative clustering

Input: The dissimilarity matrix $D_{ij} = [D^{-}_{i\to j}, D^{+}_{j\to i}] \forall i, j = 1,2,...,m$ of size m × m where m is the total number of objects.

Output : C = Collection of clusters of objects.

Method :

Let C_1, C_2,…,C_m, be m number of clusters each containing an individual object.

Repeat

Merge two clusters C_p and C_q if there exist two objects O_i and O_j respectively in C_p and C_q possessing minimum MADV, computed using eqn (10)

Until

(Desired number of clusters are obtained

OR

A Single cluster containing all the objects is obtained.)

Algorithm ends.

4.2 Magnitude Average Dissimilarity Value for Clustering

The clustering technique proposed in this section works on crisp dissimilarity matrix. The dissimilarity value approximated by the crisp magnitude value is non symmetric. In order to work on this type of proximity matrix we introduce the concept of magnitude average dissimilarity value (MgADV). The MgADV between the objects O_i and O_j is defined to be the average of the magnitude of the corresponding features dissimilarity possessed by the object O_i with the object O_j and the magnitude of the corresponding feature dissimilarity possessed by the object O_j with the object O_i.

$$MgADV_{ij} = \tfrac{1}{2}(D_{i\to j} + D_{j\to i})$$

$$= \tfrac{1}{2}\left(\sqrt{\sum_{k=0}^{n} (d^{k}_{i\to j})^2} + \sqrt{\sum_{k=0}^{n} (d^{k}_{j\to i})^2} \right) \quad \ldots(11)$$

Hence our methodology computes MgADV between two objects using the non symmetric dissimilarity values representing the degrees of dissimilarities between the objects inorder to merge them into one cluster. Since it is an agglomerative clustering, initially m clusters, each consisting an individual object are created, where m is the total number of objects. Two objects belonging to different clusters possessing the maximum MgADV are chosen and subsequently the corresponding clusters are merged together into a single cluster. If there are many more number of such pairs of clusters, then they are merged together at the same stage. This process of merging is continued till the desired number of classes are obtained or a class consisting of all the objects is obtained. It shall be noticed that, similar to MDV [6] the MADV and MgADV concept are not a step in the computation of the proximity matrix, instead, a first step in the proposed modified conventional agglomerative clustering approach. The concepts are introduced specifically to adapt the conventional clustering algorithms to work on unconventional type proximity matrices.

- $MADV_{ij} = \dfrac{\alpha(D^{-}_{i\to j} + D^{+}_{i\to j}) + \beta(D^{-}_{j\to i} + D^{+}_{j\to i})}{4}$
- $MgADV_{ij} = \tfrac{1}{2}(\alpha D_{i\to j} + \beta D_{j\to i})$

where α and β are scalars.

It is obvious that the MADV and MgADV matrices are crisp and symmetric. But, unlike other algorithms where the proximity value is directly derived to be a crisp assuming that the measure is symmetric, in the proposed methods similar to MDV [6], MADV and MgADV are derived by the use of corresponding non-symmetric values $D_{i\to j}$ and $D_{j\to i}$ with different weightages. If $D_{i\to j}$ and $D_{j\to i}$ are one and the same (as in conventional techniques), the weight factors do not convey any meaning. Indeed, deciding suitable weight factors for obtaining a better cluster of symbolic objects is a challenging task. In fact it has been perceived that this unconventional measure finds its importance in qualitative data analysis and also in pixels aggregation based on a seed point growing algorithm useful for image segmentation [9].

5. EXPERIMENTAL RESULTS

For the purpose of validating the proposed symbolic dissimilarity measure and the proposed clustering methodology, we have conducted a series of experiments on various data sets like fat oil data, microcomputer and microprocessor dataset. These are the standard datasets which are considered in many research works on clustering symbolic objects.

Fat oil, microcomputer and microprocessor data sets have been used by several researchers as a typical example for a data set involving both interval-valued features and multivalued qualitative features. The proposed method of estimating the degree of dissimilarity is employed. The clustering methodology is employed on the multivalued dissimilarity matrix. Table 1 presents the clustering results obtained on fat oil dataset through the MADV, MgADV and MDV. From Table 1, it can be observed that both MADV based clustering and MgADV based clustering methods result with similar clusters at the stage where two clusters are formed. But, at the stage where three clusters are formed, the resulting clusters through MgADV are different when compared to the clusters obtained through MADV and MDV methods. Similarly, in Table 2 and Table 3 the clusters obtained on microcomputer and microprocessor are provided. It can be seen that the results are very encouraging and authenticate the new analogy on symbolic data representation and analysis.

Table 1. Clusters obtained at 2 and 3 cluster levels on fat oil data using the proposed dissimilarity measure

MADV based clustering method	At 2 cluster level	{0,1,2,3,4,5} {6,7}
	At 3 cluster level	{0} {1,2,3,4,5}{6,7}
MgADV based clustering method	At 2 cluster level	{0,1,2,3,4,5} {6,7}
	At 3 cluster level	{1} {0,2,3,4,5}{6,7}
MDV based clustering	At 2 cluster level	{0,1,2,3,4,5} {6,7}
	At 3 cluster level	{0} {1,2,3,4,5}{6,7}

Table 2. Clusters obtained at 2 and 3 cluster levels on microcomputer data using the proposed dissimilarity measure

MADV based clustering method	At 2 cluster level	{0,1,2,3,4,5,7,8,9,10,11} {6}
	At 3 cluster level	{0,1,2,8,9} {3,4,5,7,10,11}{6}
MgADV based clustering method	At 2 cluster level	{0,1,2,3,4,5,7,8,9,10,11} {6}
	At 3 cluster level	{0,1,2,8,9} {3,4,5,7,10,11}{6}
MDV based clustering	At 2 cluster level	{0,1,2,3,4,5,7,8,9,10,11} {6}
	At 3 cluster level	{0,1,2,8,9} {3,4,5,7,10,11}{6}

Table 3. Clusters obtained at 2 and 3 cluster levels on microprocessor data using the proposed dissimilarity measure

MADV based clustering method	At 2 cluster level	{0,1,4,5,7,8} {2,3,6}
	At 3 cluster level	{0,1,4,5,7} {2,3,6}{8}
MgADV based clustering method	At 2 cluster level	{0,1,4,5,7,8} {2,3,6}
	At 3 cluster level	{0,1,4,5,7} {2,3,6}{8}
MDV based clustering	At 2 cluster level	{0,1,4,5,7,8} {2,3,6}
	At 3 cluster level	{0,1,4,5,7} {2,3,6}{8}

6. CONCLUSION

In this paper, we have proposed a dissimilarity measure to compute the degree if dissimilarity between symbolic objects whose features are of type interval. In addition, we propose new ways of approximating the degree of dissimilarity between symbolic objects. It is found that the proximity computed using these methodology, unlike conventional is not necessarily symmetric. Thus, we have explored the concept of mutual average dissimilarity value (MADV) and magnitude average dissimilarity value (MgADV), which have driven us to propose agglomerative clustering techniques for clustering symbolic objects which are approximated by the proposed methods. The proposed clustering methods are in accordance with the MDV based clustering algorithm [6]. The experiments conducted on the standard benchmark datasets reveal that the proposed approaches are effective in classifying symbolic objects and give a new dimension to data analysis. Further studies reveal that the interval type of approximation helps in clustering data which have altogether different classes in a data set and in magnitude crisp type help in clustering data with leader oriented approach.

7. REFERENCES

[1] Bock, H.H., Diday, E.: Analysis of symbolic data. Springer Verlag (2000)

[2] Gowda, K.C., Diday, E.: Symbolic clustering using a new dissimilarity measure. Pattern Recognition 24(6) (1991) 567–578

[3] Gowda, K.C., Ravi, T.V.: Agglomerative clustering of symbolic objects using the concepts of both dissimilarity and dissimilarity. Pattern Recognition Letters 16 (1995(a)) 647–652

[4] Gowda, K.C., Ravi, T.V.: Divisive clustering of symbolic objects using the concepts of both similarity and

dissimilarity. Pattern Recognition 28(8) (1995(b)) 1277–1282

[5] Guru, D.S., Kiranagi, B.B., Nagabhushan, P.: Multivalued type proximity measure and concept of mutual similarity value useful for clustering symbolic patterns. Journal of Pattern Recognition Letters 25 (2004) 1203–1213

[6] Guru, D.S., Kiranagi, B.B.: Multivalued type dissimilarity measure and concept of mutual dissimilarity value for clustering symbolic patterns. Journal of Pattern Recognition 38(1) (2005) 151–156

[7] Ichino, M., Yaguchi, H.: Generalized minkowski metrices for mixed feature type data analysis. IEEE Transactions on system, man and cybernetics 24(4) (April 1994) 698 – 708

[8] Jain, A.K., Dubes, C.R.: Algorithms for Clustering Data. Prentice Hall, Engle Wood Cliffs (1998)

[9] Gonzalez, R.C., Woods, R.E.: Digital Image processing. Second edn. Prentice Hall,New Jersey (2001)