

MNV for Clustering based on Non Symmetric Symbolic Proximity

Bapu B Kiranagi¹, D S Guru², S Manjunath² and B S Harish²

¹HCL Technologies, Bangalore, India

Email: bapu_b@yahoo.com

²Department of Studies in Computer Science, University of Mysore, Mysore, India

Email: {dsg@compsci.uni-mysore.ac.in, manju_uom@yahoo.co.in, bsharish@ymail.com}

Abstract— In this paper, we bring out the importance of non-symmetric proximity values among symbolic objects in simulating the reality during clustering. The concept of Mutual Neighborhood Value (MNV) has been exploited on non-symmetric proximity values. The results of the experiments conducted reveal that the approaches based on non-symmetric proximity measures best suite the reality than the symmetric proximity measures.

Index Terms— symbolic objects, non-symmetric proximity values, mutual neighborhood value, clustering.

I. 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 symbolic object representation has an explanatory power, where as in conventional data the objects are individualized which have less explanatory power. Normally, symbolic data are more unified by means of relationship which is self explanatory as we human perceive in our daily life. The relationships between symbolic data appear in the form of continuous ratio, discrete absolute, interval, modal, multivalued and also multivalued data with weights. Hence to be realistic in nature the data are to be represented using such unconventional representation i.e., symbolic representation. Example, representing symbolically the chemical properties of fat/oil which contains multi type attributes, city temperatures which are of type interval, or representing handwritten numeral and signature which varies with time within the same person may be effective than using conventional data. In particular these representations are clustered for further processing such as grouping of cities based on their city temperature, recognition of numerals, retrieval and verification of signatures etc. In such case we need to cluster these symbolic objects.

However using the conventional clustering algorithms such as hierarchical, partitional clustering algorithms consider only conventional data sets of discrete type whose proximity are symmetric and hence their application on symbolic data or realistic data may not always yield the desired output [1]. Thus in order to make clustering models more realistic the proximity measures have to be modified or new proximity measures that can heed such complex data sets have to be proposed. In this direction many proximity measures have been proposed [1] [2] [3] [4] in the literature. However all of the aforementioned works assumes that the proximity matrix is symmetric in nature which is not realistic. Hence in [5] [7] the authors have brought out the idea of non symmetricity in designing proximity measures. Similarly in this work also we bring out the

importance of non symmetric proximity values among symbolic objects in simulating the reality during clustering.

For instance, two persons A and B group together as close friends if they mutually feel that the other is closest friend. If A feels that B is not such a close friend to him, then even though B may feel that A is his closest friend, the bond of friendship between them is comparatively weak. If any person feels that the other is not his close friend, then the two do not group together as close friends. In other words, the authors in [7] claim that the strength of the bond of friendship between two persons is a function of mutual feelings rather than one-way feeling. Similarly two objects form a cluster if they are mutually near neighbors rather than simply near neighbors. The level at which two samples are merged is decided by the Mutual Neighborhood Value (MNV) between them. Thus, [7] have proposed a model which instead of using actual proximity (similarity/dissimilarity) among the objects to be clustered, uses the MNV computed for each pair of objects as the proximity values. In [8] a theoretical analysis of the threshold selection of the Mutual Neighborhood Clustering Algorithm (MNCA) under the hypothesis of random data was presented. The method presented in [8] yields a theoretical minimum value of the threshold value below which even unclustered data are broken into separate clusters. Similar to work in [7] a nonparametric pairwise distance based clustering was proposed in [9].

However, the motivation behind the concept of MNV to consider the individual proximities is not entirely realized as all the aforementioned works are based on symmetric proximity matrix. Therefore, in this paper we present a method of clustering symbolic objects based on MNV by the use of non-symmetric proximity measure. The concept of Mutual Neighborhood Value (MNV) has also been applied for both symmetric and non symmetric proximity matrices to reflect how the MNV concept will have better meaning in case of non symmetric proximity values. The results of the experiments conducted reveal that the proposed non symmetric proximity based approaches best suite the reality.

The organization of the paper is as follows, in section 2 the mutual neighborhood for symbolic object clustering is proposed and illustrated on fat oil data set in section 3. Section 4 gives the obtained results based on microcomputer, and microprocessor data along with the comparative study with the existing methods. The paper is concluded in section 5.

II. PROPOSED METHODOLOGY

In this section we introduce a new mutual neighborhood value based clustering method. In which the proposed method works on the non symmetric positional matrix which will be

later converted into symmetric mutual neighborhood value proximity matrix to cluster the data.

Let there be 'm' number of objects say $O_1, O_2, O_3, \dots, O_m$ defined by 'n' dimensional features of composite types of interval and multivalued data which need to be clustered into k clusters. Among all pair of m objects we compute a mutual neighborhood value matrix obtained by the positional values of mutual nearest obtained using similarity values as follows.

To compute similarity between symbolic objects which can be of type interval, multivalued we use the similarity measures proposed in [5]. The similarity measures proposed in the work [5] computes the degree of similarity among symbolic objects which are described by both interval and multivalued features. In case of features of interval type, the degree of similarity between objects is estimated based on degrees of overlapping of the interval types features and in case of multivalued the degree of similarity is estimated based on the commonality between the multivalued type features. As the relative overlapping (in case of interval type features) and relative commonality (in case of multivalued type features) are not necessarily equal and hence the degree of similarity between two symbolic objects may not necessarily be symmetric.

Using the [5] similarity measure we compute the similarity between objects with respect to each feature. The obtained similarity value among features is approximated multivalued data type and this approximation is further used for clustering. However, in this paper in order to suite the MNV type clustering we propose a new approximation technique.

A. Approximation through crisp data type.

The degree of similarity between objects O_i and O_j is as shown below

Let O_i and O_j be two symbolic objects described by n interval valued features as follows

$$O_i = \{F_{i1}, F_{i2}, \dots, F_{in}\}$$

$$i.e. F_i = \{(f_{i1}^-, f_{i1}^+), (f_{i2}^-, f_{i2}^+), \dots, (f_{in}^-, f_{in}^+)\}$$

$$O_j = \{F_{j1}, F_{j2}, \dots, F_{jn}\}$$

$$i.e. F_j = \{(f_{j1}^-, f_{j1}^+), (f_{j2}^-, f_{j2}^+), \dots, (f_{jn}^-, f_{jn}^+)\}$$

The degree of similarity between objects say O_i and O_j is estimated based on degrees of overlapping and their separability, if any in each feature of the patterns. The degree of similarity of each feature value of O_i with respect to the corresponding feature value of O_j is estimated as follows:

The k^{th} features $F_{ik} = [f_{ik}^-, f_{ik}^+]$ and $F_{jk} = [f_{jk}^-, f_{jk}^+]$ may or may not overlap. The degree of similarity of O_i to O_j , with respect to the k^{th} feature (irrespective of overlapping or no overlapping) is characterized by

$$s_{i \rightarrow j}^k = \frac{\text{Overlapping portion between } F_{ik} \text{ and } F_{jk}}{\text{Length of } F_{jk}}$$

$$i.e., s_{i \rightarrow j}^k = \left(\frac{|F_{ik} \cap F_{jk}|}{|F_{jk}|} \right)$$

here, $F_{ik} \cap F_{jk}$ represents the overlapping portion of the interval features F_{ik} and F_{jk} , and $| \cdot |$ represents the length of an interval.

Thus the similarity of the object O_i to O_j with respect to all n features turned out to be multivalued and is given by

$$S_{i \rightarrow j} = [s_{i \rightarrow j}^1, s_{i \rightarrow j}^2, s_{i \rightarrow j}^3, \dots, s_{i \rightarrow j}^n] \tag{2}$$

Similarly, the similarity of the object O_j to the object O_i with respect to all n features is $S_{j \rightarrow i} = [s_{j \rightarrow i}^1, s_{j \rightarrow i}^2, s_{j \rightarrow i}^3, \dots, s_{j \rightarrow i}^n]$. It can be perceived from the above that the similarity matrix apart from being multivalued, is not necessarily symmetric.

However, this type of approximation is not suitable for MNV clustering. Hence we propose a new approximation technique. In this type of approximation the total degree of similarity from the objects 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 degree of similarity between the objects and is given by (1)

$$S_{i \rightarrow j} = \sqrt{\sum_{k=0}^n (s_{i \rightarrow j}^k)^2} \tag{1}$$

In a similar way, the total degree of similarity of the object O_j so the object O_i with respect to all n features is given by

$$S_{j \rightarrow i} = \sqrt{\sum_{k=0}^n (s_{j \rightarrow i}^k)^2} \tag{2}$$

It is obvious that the similarity matrix obtained through this approximation technique is also non symmetric. However, it is of crisp type.

B. Mutual neighborhood value based clustering.

From the obtained non symmetric similarity matrix we compute the proximity position matrix as follows. For each object O_i , we obtain the proximity position for all other objects by sorting them according to decreasing order of the similarity that the object O_i possesses with the others. Let O_i be the s^{th} nearest neighbor of O_j (O_i is at s^{th} position for O_j), and let O_j be the t^{th} nearest neighbor of O_i (O_j is at t^{th} position for O_i). The mutual neighborhood value (MNV) between O_i and O_j is defined to be the sum of their nearest neighborhood position values.

$$i.e., MNV(O_i, O_j) = s + t, \text{ where } s, t \in \{1, 2, \dots, m\} \tag{3}$$

Once the MNV between each pair of objects is computed, we have a matrix called MNV matrix of size $m \times m$ in which the diagonal elements are not defined (or we can assume that diagonal elements are zero indicating that each object is its own zeroeth neighbor). It shall be noticed that the MNV is symmetric and thus, any conventional clustering technique could be employed on this MNV matrix as if it is a conventional crisp matrix.

III. ILLUSTRATION

To illustrate the proposed method we use fat oil data set used in [7]. The fat oil data set has been used by several researchers as a typical example. It is composed of eight patterns described by four features of interval valued and one feature of multivalued qualitative type. To illustrate the working principle of the proposed MNV based clustering approach we have used fat oil data set.

Using the similarity measure proposed in [5] and applying the crisp approximation technique proposed in section A, the crisp magnitude type similarity matrix on fat oil data is given in Table I.

I TABLE.

CRISP MAGNITUDE TYPE SIMILARITY MATRIX OF FAT OIL DATA

Object No	0	1	2	3	4	5	6	7
0	1.00	0.70	0.49	0.52	0.67	0.55	0.40	0.37
1	0.59	1.00	0.54	0.56	0.63	0.59	0.43	0.39
2	0.36	0.69	1.00	0.72	0.77	0.59	0.53	0.47
3	0.27	0.62	0.53	1.00	0.63	0.53	0.30	0.31
4	0.23	0.26	0.34	0.34	1.00	0.32	0.17	0.18
5	0.27	0.53	0.67	0.57	0.89	1.00	0.40	0.37
6	0.27	0.43	0.53	0.34	0.40	0.44	1.00	0.54
7	0.36	0.49	0.59	0.42	0.55	0.53	0.61	1.00

With respect to each object in the type similarity matrix (Table I) the position (neighborhood) values for other objects are computed. Based on this position information we have constructed a matrix called a position matrix in which (i, j)th element refers to the neighborhood value of the jth object with respect to ith object and the same is given in Table II. It shall be observed that this position matrix is non symmetric.

Now, the MNV matrix of the fat oil data based on the position matrix obtained (Table. II) is constructed in which (i, j)th element is the sum of the (i, j)th and (j, i)th elements of the position matrix. Table III presents the constructed MNV matrix for fat oil data. The conventional agglomerative single linkage cluster approach is then applied on the MNV matrix and the obtained clusters are shown in Table VI.

II TABLE.

POSITION MATRIX CONSTRUCTED USING NON SYMMETRIC SIMILARITY MATRIX ON FAT OIL DATA

Object No	0	1	2	3	4	5	6	7
0	1	2	6	5	3	4	7	8
1	5	1	6	3	2	4	7	8
2	8	5	1	3	2	4	6	7
3	8	2	4	1	3	5	7	6
4	6	5	3	4	1	2	7	8
5	8	5	3	4	2	1	6	7
6	8	5	3	7	6	4	1	2
7	8	6	3	7	4	5	2	1

III TABLE.

MNV MATRIX CONSTRUCTED USING THE POSITION MATRIX ON FAT OIL DATA

Object No	0	1	2	3	4	5	6	7
0	-	7	14	13	9	12	15	16
1	7	-	11	5	7	9	12	14
2	14	11	-	7	5	7	9	10
3	13	5	7	-	7	9	14	13
4	9	7	5	7	-	4	14	11
5	12	9	7	9	4	-	10	12
6	15	12	9	14	14	10	-	4
7	16	14	10	12	11	12	4	-

IV TABLE.

POSITION MATRIX CONSTRUCTED USING SYMMETRIC SIMILARITY MATRIX ON FAT OIL DATA

Object No	0	1	2	3	4	5	6	7
0	1	2	4	6	3	5	8	7
1	2	1	3	4	6	5	8	7
2	8	5	1	4	3	2	7	6
3	6	3	2	1	5	4	8	7
4	5	6	3	4	1	2	8	7
5	8	4	3	5	2	1	7	6
6	7	4	3	6	8	5	1	2
7	8	5	3	7	6	4	2	1

V TABLE.

MNV MATRIX CONSTRUCTED USING THE POSITION MATRIX ON FAT OIL DATA

Object No	0	1	2	3	4	5	6	7
0	-	4	12	12	8	13	15	15
1	4	-	8	7	12	9	12	12
2	12	8	-	6	6	5	10	9
3	12	7	6	-	9	9	14	14
4	8	12	6	9	-	4	16	13
5	13	9	5	9	4	-	12	10
6	15	12	10	14	16	12	-	4
7	15	12	9	14	13	10	4	-

Out of interest in bringing out the fact that the non symmetricity of a proximity measure has significant contribution in arriving, sometimes, altogether different clusters, we have converted the non symmetric similarity matrix into symmetric similarity matrix using the concept of mutual similarity value proposed in [5]. Using this converted symmetric similarity matrix we constructed the position matrix and also the corresponding MNV matrix which are respectively shown in Table IV and Table V. Here also the conventional agglomerative single linkage cluster approach is applied on the MNV matrix that is shown in Table V and obtained cluster are shown in Table VI. From the Table VI, the proposed non-symmetric based clusters are not similar to symmetric based clusters. Thus showing the difference.

VI TABLE.

SUMMARIZES THE CLUSTER OBTAINED BY THE USE OF TYPE-2 SIMILARITY MATRIX

Levels	Fat oil data	
	Non symmetric	Symmetric
A stage just before two cluster level	{0} {1,3} {2,4,5} {6,7}	{0,1,2,4,5} {3} {6,7}
Two cluster level	{0,1,2,3,4,5} {6,7}	{0,1,2,3,4,5} {6,7}

IV. EXPERIMENTAL ANALYSIS

Similarly we have conducted the same experiments on the other two standard data sets viz., microcomputer and microprocessor data sets. Table VII presents the clusters obtained using proposed technique.

For the purpose of establishing the superiority of the proposed clustering models over other existing methods, we consider the work of [2] [3] [4] [5]. These methods are specifically chosen for this comparative study as they provide results on the same standard data sets.

Table VIII summarizes the cluster obtained by the proposed methods and the other existing methods on fat oil, microcomputer and microprocessor data set. The entries not-available (NA) in Table VIII denotes that the corresponding result has not been shown in the respective research work. We have not computed the same during experimentation as those methodologies are either parametric or require prior knowledge of the number of samples in each pattern which is indeed a real drawback of those approaches. However, the entry not available corresponding to some of the proposed unsupervised algorithms denotes that the clusters formation process did not stop at the level specified and hence cluster are not obtained. Further, the clustering methods [2] [3] [4] [5] work on symmetric proximity matrix. However, it can be clearly seen that the proposed MNV on non symmetric matrix differs and as explained earlier (section 1) the clusters obtained are closer to reality.

V CONCLUSION

In this paper we explored the concept of mutual neighborhood value (MNV) by using the non symmetric proximity measure. In addition, we propose a approximation technique though crisp upholds the basic property of symbolic data analysis (i.e., non symmetricity). Further, the attention is focused towards bringing out the significance of non symmetricity of the proposed proximity measure in obtaining the different types of realistic clusters. The proposed approach is expected to find their application in social life examples, such as clustering people based on friendship, which are indeed the applications of the exploratory/symbolic data analysis.

REFERENCES

[1]. H. H. Bock and E. Diday, "Analysis of symbolic data," Springer Verlag, 2000.
 [2]. K. C. Gowda and E. Diday, "Symbolic clustering using a new dissimilarity measure," Pattern Recognition, Vol. 24(6), pp 567 – 578, 1991.
 [3]. K. C. Gowda and T. V. Ravi, "Agglomerative clustering of symbolic objects using the concepts of both dissimilarity and dissimilarity," Pattern Recognition Letters, Vol. 16, pp 647 – 652, 1995(a).

[4]. K. C. Gowda and T. V. Ravi, "Divisive clustering of symbolic objects using the concepts of both similarity and dissimilarity," Pattern Recognition, Vol 28(8), pp 1277 – 1282, 1995(b).
 [5]. D. S. Guru, B.B Kiranagi and P. Nagabhushan, "Multivalued type proximity measure and concept of mutual similarity value useful for clustering symbolic patterns," Journal of Pattern Recognition Letters, Vol 25, pp 1203 –1213, 2004.
 [6]. M. Ichino and H. Yaguchi, "Generalized minkowski metrics for mixed feature type data analysis," IEEE Transactions on system, man and cybernetics, Vol 24, No. 4, 698 – 708, 1994.
 [7]. K. C. Gowda and G. Krishna, "Agglomerative clustering using the concepts of mutual nearest neighborhood," Journal of Pattern Recognition, Vol. 10, No. 2, pp 105 – 112, 1978.
 [8]. S. Smith, "Threshold validity for mutual neighborhood clustering," IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 15, pp 89 – 92, 1993.
 [9]. S. Dubnov, R. El-Yaniv, Y. Gdalyahu, E. Schneidman, N. Tishby and G. Yona, "Clustering by Friends: A New Nonparametric Pairwise Distance Based Clustering Algorithm," Kluwer Academic Publishers, Netherlands, pp1–32,2000.

VII TABLE.

CLUSTERS OBTAINED THROUGH MNV BASED METHOD ON MICROCOMPUTER AND MICROPROCESSOR DATA USING TYPE-2 SIMILARITY MATRIX.

Levels	Microcomputer		Microprocessor	
	Non symmetric	Symmetric	Non symmetric	Symmetric
A stage just before two cluster level	{4,5,7,11} {0,1,2,3,8,9,10} {6}	{4,5,7,11} {0,1,2,3,8,9,10} {6}	{0,1,7,4}{2,3,6,5} {8}	{0,1,7,4}{2,3,6} {8,5}
Two cluster level	{4,5,7,11,0,1,2,3,8,9,10}{6}	{4,5,7,11,0,1,2,3,8,9,10}{6}	{0,1,7,4,2,3,6,5}{8}	{0,1,7,4,8} {2,3,6,5}

VIII TABLE

RESULT BASED COMPARISON OF VARIOUS CLUSTERING ALGORITHMS ON DIFFERENT DATA SETS.

Methodology	Fat oil data set		Microcomputer data set		Microprocessor data set	
	Description at 2 clusters level	Description at level just before 2 cluster level	Description at 2 clusters level	Description at level just before 2 cluster level	Description at 2 clusters level	Description at level just before 2 cluster level
Gowda and Diday [2]	NA	{0,1}{2,3,4,5} {6,7}	NA	{0,1,3,9,10}{6} {2,8}{4,5,11}{7}	NA	NA
Gowda and Ravi [3]	NA	{0,1}{2,3,4,5} {6,7}	NA	{0,1,3,5,7,8,9,10,11}{2}{6}{4}	NA	{0,1,4,5,7}{2,3,6}{7,8}
Gowda and Ravi [4]	{0,1,2,3,4,5} {6,7}	NA	{0,1,2,3,4,5,7,8,9,10,11}{6}	NA	{0,1,2,3,4,5,6,8}{4,7}	NA
Guru et al., [5]	{0,1,2,3,4,5} {6,7}	{0,1} {2,3,4,5}{6,7}	{0,1,2,3,4,5,7,8,9,10,11}{6}	{0,1,2,3,8,9,10} {4,5,7}{6}{11}	{0,1,4,7} {2,3,5,6,8}	{0,1,4,7}{2,3,5,6}{8}
Proposed (MNV)	{0,1,2,3,4,5} {6,7}	{0}{1,3} {2,4,5}{6,7}	{0,1,2,3,4,5,7,8,9,10,11}{6}	{4,5,7,11}{0,1,2,3,8,9,10}{6}	{0,1,2,3,4,5,6,7}{8}	{0,1,7,4}{2,3,6,5}{8}