Fuzzy-Symbolic Analysis for Classification of Symbolic Data*

M.S. Dinesh[1], K.C. Gowda[2], and P. Nagabhushan[3]

[1] Siemens Information Systems Ltd.,Bangalore-560 100, India
dineshms@rocketmail.com
[2] Jnana Sahyadri, Kuvempu University, Shimoga-577451, India
[3] Department of studies in computer science, University of Mysore, India

Abstract. A recent study on symbolic data analysis literature reveals that symbolic distance measures are playing a major role in solving the pattern recognition and analysis problems. After a careful study on the existing symbolic distance measures, we have identified that most of the existing symbolic distance measures either suffer from generalization or do not address object variability. To alleviate these problems we are proposing new generalized Similarity symbolic distance measure. The proposed distance measure is asymmetric, addresses object variability, and obeys partial order. To leverage the advantages of both fuzzy set theory and symbolic data analysis, conventional classification algorithm that works on the principles of fuzzy equivalence relation has been extended to handle Symbolic data. Efficacies of the proposed techniques are validated by conducting several experiments on the well-known assertion type of symbolic data sets with known classification results.

Keywords: Fuzzy-Symbolic data analysis, Fuzzy hierarchical analysis, Symbolic distance measures.

1 Introduction

From the literature it is evident that the two fields in Pattern Recognition (PR) namely, Fuzzy Data Analysis (FDA) and Symbolic Data Analysis (SDA) have been individually supplementary to the growth of PR, while seeming to have remained complementary to each other. Thus the essence of this paper is to leverage the advantages of both Symbolic Data Analysis and the Fuzzy Set Theory.

Distance measure plays a key role in Clustering or Classification of Data and gives an index of proximity, or alikeness, or affinity, or association between pairs of patterns. With the use of a proper distance measure, a proximity matrix can be computed from the pattern matrix where, proximity index is used to represent either dissimilarity or similarity between the patterns/objects/samples [2]. Most of the existing symbolic distance measures are metric in nature and therefore fail to grasp the asymmetric relation between the objects.

* This work was carried out at SJCE, Mysore, India.

S.K. Pal et al. (Eds.): PReMI 2005, LNCS 3776, pp. 338 – 343, 2005.

Asymmetric relation between the objects exists due to object variability in size and for many other inherent reasons. Detailed study on the object variability is available in the reference paper [7]. In this paper, we have proposed non-metric similarity distance measure, which successfully overcome the drawbacks of the existing Symbolic Distance Measures. Proposed distance measure has been experimented with the data sets of known classification results and these results are compared with the existing distance measures that are available in literature.

2 Proposed Symbolic Similarity measure

2.1 Feature Space

Let the symbolic object be described with respect to d features $X_1, X_2 \ldots X_d$ and U_k denote the domain of the feature X_k. The domain U_k is assumed to be a bounded closed interval and is of the form $U_k[a_k, b_k]$ where a_k and b_k are minimum and maximum possible values for X_k, when X_k is continuous quantitative, discrete quantitative and ordinal qualitative. On the other hand, U_k is a finite set of all possible values, when X_k is a nominal qualitative. Then the feature space is the Cartesian product of $U_1, U_2 \ldots U_d$ that is,

$$U^{(d)} = U_1 \text{ x } U_2 \text{ x } \ldots \text{ x } U^d. \quad (1)$$

2.2 Similarity Measure

The similarity measure S between two Symbolic objects $A = A_1 \text{ x } A_2 \text{ x } \ldots \text{ x } A_d$ and $B = B_1 \text{ x } B_2 \text{ x } \ldots \text{ x } B_d$ in U_d is written as:

$$S(A, B) = \frac{1}{d} \sum_{k=1}^{d} \frac{W_k}{U_k} S(A_k, B_k) \quad (2) \qquad \sum_{k=1}^{d} W_k = 1 \ , \ W_k > 0, k = 1, \ldots d,$$

thus $0 < S(A, B) < 1$. Weighting constant (W_k) controls the relative importance of the features and U_k helps in the normalization of the output proximity values. For the k^{th} feature, $S(A_k, B_k)$ is defined using two components such as $S_p(A_k, B_k)$ due to position p. and $S_s(A_k, B_k)$ due to span S.

The similarity component due to "position" arises only when the feature is quantitative interval or quantitative absolute/ratio type. It indicates the relative positions of two feature values. The similarity component due to "span" indicates the relative sizes and overlaps of the feature values. Computation of span component is required for both quantitative and qualitative types of features.

Let, a_m = Median value of interval A_k, b_m = Median value of interval B_k, Ø = Cartesian join operator,

Ø (x) = Cardinal(x), if x is categorical,
Ø (x) = Length(x), if x is quantitative.
Where x = A_k or B_k or A_k Ø B_k

Similarity due to position is defined as:

$$S_p(A_k, B_k) = \frac{1}{1 + / A_m - B_m /} \quad (3) \qquad S_p(A_k, B_k) = S_p(B_k, A_k) \quad (4)$$

For the special cases given in Eqn 4 a_m and b_m the position component will change as follows:

Case 1: When the values of $a_m = b_m$, a_m and b_m in the position component become $a_m = a_L$ and $b_m = b_L$

Similarity due to span is given by:

$$S_s(B_k, A_k) = \frac{\phi(B_k)}{\phi(A_k \oplus B_k)} \quad (5) \qquad S_s(A_k, B_k) = \frac{\phi(A_k)}{\phi(A_k \oplus B_k)} \quad (6)$$

Net Similarity between Ak and Bk is:

$$S(B_k, A_k) = \frac{S_p(B_k, A_k) + S_s(B_k, A_k)}{2.0} \quad (7) \qquad S(A_k, B_k) = \frac{S_p(A_k, B_k) + S_s(A_k, B_k)}{2.0} \quad (8)$$

Concepts of Similarity, Dissimilarity, and Cartesian join operations illustrated through figures are given below:

Let the object be described in terms of d features X_k, k=1, 2, . . . d. and E_k is the feature value taken by the feature X_k. Then we represent a pattern by a Cartesian product set $E = E_1 \times E_2 \times - - - \times E_d$.

Let $A = A_1 \times A_2 \times \ldots \times A_d$ and $B = B_1 \times B_2 \times \ldots \times B_d$ be a pair of events of $U^{(d)}$.

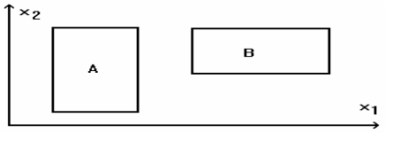

Fig. 1. Events in the Euclidean plane

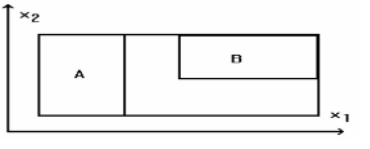

Fig. 2. Cartesian Join operator (A⊕B)

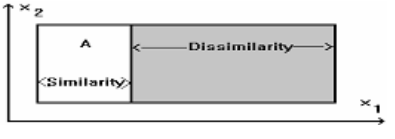

Fig. 3. Similarity/Dissimilarity from A to B

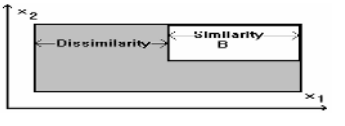

Fig. 4. Similarity/Dissimilarity from B to A

2.3 Specific Features

The proposed distance measures possess the following properties:

- Conveniently takes care of mixed features like, quantitative interval, quantitative absolute/ratio, and qualitative types.
- Satisfy the rules of partial order (reflexive, asymmetric and transitive).

- As a special case the equality $S_{AB} = S_{BA}$ occurs only when the object A has the same description as B, meaning that the two objects A and B having the attributes of same size are exactly identical in all respects.
- Distance measures produce normalized output in the range [0 -1], which helps to employ the fuzzy concepts for further analysis.
- Weightage factors in the distance measure helps in assigning the relative importance to the features.
- Distance measures can be decomposed into symmetric and skew-symmetric parts, this feature will be useful to study the object variability due to unequal spread of feature values.

3 Fuzzy Agglomerative Symbolic Classification

Based on the similarity relation proposed by Lofti Zadeh, Shinuchi Tamura et. al.,[11]; Dunn[6], Abraham Kandel et. al.[1], and Bezdek et. al.[3] we have employed fuzzy hierarchical techniques to analyze symbolic data sets.

Algorithm

1. Let { $y_1, y_2, \ldots, y_n$}be a set of symbolic objects in 'd' dimensions and the initial number of clusters/classes be 'n'.
2. Compute the symbolic similarities between all pairs of symbolic objects in the data set (similarity measure between objects is computed as described in section 2.2).
3. Asymmetric similarity measure is decomposed into symmetric and skew symmetric parts.
4. Check the Similarity values computed in step 2 for transitiveness. If the similarity relation is not transitive, perform the transitive closure on the similarity values to make the similarity measure as fuzzy equivalence relation($R_{T)}$).
5. Since R_T is symmetric, consider either lower or upper triangle elements as distinct α - cut values.
6. Arrange all α - cut values in descending order.
7. Apply each α - cut, one by one on the data set and obtain the partitions.
8. From the partitions construct a dendrogram.
9. Merge all the symbolic objects in a class to form a Composite Symbolic Object (CSO). CSO represents the class description.

4 Experiments and Relation to Other Works

Experiments are conducted on the well-known Symbolic Data Sets whose classification results are known [8, 9, 10]. Assertion type of symbolic data sets of Fat-oil, Microprocessors and Microcomputer are used for the experiments. Fuzzy hierarchical classification scheme described has been extended used to obtain different clusters/classes of the Symbolic Data Sets. The results are compared and contrasted with the existing symbolic clustering techniques.

Experiment No. 1: The data set used for this experiment consists of information about fats and oils[10]. Fat-oil data set consists of four quantitative features of interval type and one nominal qualitative feature. Dendrogram shown in Fig.5. has been obtained after applying the proposed algorithm on Fat-Oil data. By cutting the dendrogram at an appropriate level we can obtain different classes/clusters. The samples grouped for two class and three class are: Two Class: {0,1,2,3,4,5,}, {6,7}, Three class:{0, 1}, {2, 3, 4 ,5}, {6, 7}. The results obtained by Ichino, Ichino & Yaguchi[10] on Fatoil data for two classes are identical with the results obtained by the proposed method. Results obtained by Gowda & Diday[8] on the same data set for three classes are identical with the results obtained by the proposed method.

Experiment No. 2: Experiment on the microprocessor data resulted in the dendrogram as shown in Fig. 6. Samples obtained for two and three classes are: Two class: {0, 1, 2, 3, 4, 5, 6, 7}, {8}, Three Class: {0, 1, 4, 7}, {2, 3, 5, 6}, {8} . The classification results obtained using Ichino's [17] method resulted in three clusters as {0, 1, 4, 5}, {2, 3, 6,}, and {7, 8}. The classification result obtained using the Gowda & Diday dissimilarity measure [8] resulted in two clusters as {0, 1, 2, 3, 4, 5, 6}, {7, 8}. The classification results obtained using the Gowda & Ravi [9] method resulted in three clusters as {0, 1, 4, 5, 7}, {2, 3, 6,}, and {8}. The results of the proposed method vary marginally when compared with the results of Ichino, Gowda and Diday, and Gowda and Ravi.

Experiment No. 3: Application of the proposed algorithm on microcomputer data set produced the dendrogram as shown in Fig. 7. Samples grouped for three classes are {0, 1, 3, 4, 5, 7, 9, 10, 11}, {2, 8}, {6}. The classification results obtained using the Gowda and Ravi [9] divisive algorithm and Ichino's method resulted in two clusters as {0, 1, 2, 4, 5, 7, 8, 9, 10, 11}, and {6} and the classification result obtained using Gowda Diday dissimilarity measure [8] resulted in four clusters as {0, 1, 9, 10}, {6}, {2, 8}, {3, 4, 5, 7, 11}. The results obtained by the proposed method vary marginally when compared with the results of Ichino, Gowda and Diday and Gowda and Ravi.

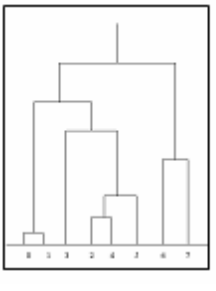

Fig. 5. Fat-Oil Data

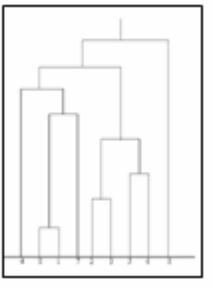

Fig. 6. Microprocessor Data

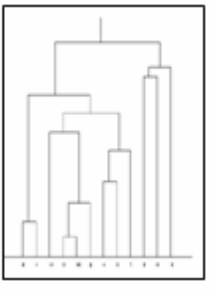

Fig. 7. Macrocomputer Data

5 Summary

New similarity measure for Symbolic objects is presented. Conventional algorithm, which works on the principles of Fuzzy equivalence relation, has been extended to

handle Symbolic data. To validate the proposed distance measures, they are applied on the well-known assertion type of Symbolic Data Sets with known classification results and these results are compared with the existing techniques. Proposed techniques were also applied and validated on the large data sets like multi spectral satellite images and Magnetic resonance images. Results on large data set are not discussed in this paper due to space constraints. However, authors are planning to discuss these results during the presentation.

References

1. Abraham Kandel and Lawrence Yelowitz,"*Fuzzy Chains,"* IEEE Tran. Sys., Man, Cybern., Vol. SMC-4, No. 5, pp 472-475, Sept. 1974.
2. Anil K. Jain and Richard C. Dubes,"Algorithms for Clustering Data," Prentice Hall, Englewood Cliffs, New Jersey, 1988.
3. J. C. Bezdek, J. Douglus Haris, "*Fuzzy Partitions and Relations: An Axiomatic Basis For Clustering,"* Fuzzy Sets and Systems, Vol. 1, pp. 111-127, 1978.
4. E. Diday, C. Hayashi, M. Jambu and N. Ohsumi, Eds, “Recent Developments in Clustering and Data Analysis, Academic Press, New York, 1987.
5. E. Diday, "*Knowledge representation and symbolic data analysis*," Proc. 2nd int. workshop on Data, Expert Knowledge, and Decision, Hamburg, Sep. 1989a.
6. J.C. Dunn, "Some Recent Investigations of a New Fuzzy Partitioning Algorithms and its Application to Pattern Classification Problems," Journal of Cybernetics, Vol. 4, pp. 1-15, 1974.
7. Francesco Palumbo and Maria Benedetto, "*A Generalization Measure for Symbolic Objects,"* Proc. of KESDA’98 Conference, Luxembourg, April 1998.
8. K. C. Gowda and E. Diday, "Symbolic Clustering Using a New Similarity Measure," IEEE Trans. Sys. Man Cybern. 22, 368-378, 1992.
9. K. C. Gowda, and T.V.Ravi, "*Agglomerative clustering of symbolic objects using the concepts of both similarity and dissimilarity*," Pattern Recognition Lett. 16, 647-652, 1995.
10. M. Ichino and H. Yaguchi, Generalized Minkowski Metrics for Mixed Features, Trans. IECE Jpn, J72-A, 398-405(in Japanese), 1989.
11. Shinichi Tamura, Seihaku Kiguchi, and Kokichi Tanaka, "*Pattern Classification Based on Fuzzy Relations,"* IEEE Trans. Sys., Man, Cybern., Vol. SMC-1, No. 1, pp. 61-66, Jan. 1971.
12. L.B. Turksen, "*Interval Valued Fuzzy Sets Based on Normal Forms,"* Fuzzy Sets and Systems, Vol. 20, pp 191-210, 1986.