Available at www.ComputerScienceWeb.com
POWERED BY SCIENCE DIRECT®

Pattern Recognition Letters 24 (2003) 895–902

Pattern Recognition Letters

www.elsevier.com/locate/patrec

Fuzzy discretization of feature space for a rough set classifier

Amitava Roy [a], Sankar K. Pal [b,*]

[a] Variable Energy Cyclotron Centre, 1/AF Bidhan Nagar, Calcutta 700064, India
[b] Machine Intelligence Unit, Indian Statistical Institute, 203, Barrackpore Trunk Road, Calcutta 700035, India

Abstract

A concept of fuzzy discretization of feature space for a rough set theoretic classifier is explained. Fuzzy discretization is characterised by membership value, group number and affinity corresponding to an attribute value, unlike crisp discretization which is characterised only by the group number. The merit of this approach over both crisp discretization in terms of classification accuracy, is demonstrated experimentally when overlapping data sets are used as input to a rough set classifier. The effectiveness of the proposed method has also been observed in a multi-layer perceptron in which case raw (non-discretized) data is considered as input, in addition to discretized ones.

Keywords: Classification; Discretization; Data mining; Rough sets; Fuzzy sets

1. Introduction

Data mining and knowledge discovery in database is an intelligent method of discovering unknown or unexplored relationship within a large database. It uses the principles of pattern recognition and machine learning to discover the knowledge, and various statistical and visualisation techniques to present knowledge in a comprehensible form.

The theory of rough sets (Pawlak, 1991) offers a theoretical basis for reasoning about data and is found to be an effective tool for the decision support system. This theory gives a set-theoretic definition of knowledge, based on equivalence relation and provides algorithms for reduction of number of attributes, rule generation and classification related to any information system. Rough set theoretic classifiers perform better with the discrete valued (symbolic) attributes or features. These can be applied to continuous valued attributes using a process called (crisp) discretization (Lenarcik and Piasta, 1992; Nguyen and Skowron, 1995; Nguyen and Nguyen, 1998). The crisp discretization is a method of generating a set of values or the 'cuts' of attributes within the dynamic ranges of the corresponding attributes. The intervals formed by the adjacent values of the cuts, become the discrete groups for the continuous valued attributes. The positions of cuts are very sensitive to the subsets of the information system, which are used to generate the cuts, as well as to the methodology adopted. The position sensitivity of cuts may make the classification accuracy adversely affected. In order to avoid this problem the

* Corresponding author. Tel.: +91-33-577-8085x3100; fax: +91-33-556-6680/6925 and +91-33-578-3357.
E-mail address: sankar@isical.ac.in (S.K. Pal).

0167-8655/03/$ - see front matter
PII: S0167-8655(02)00201-5

present article introduces the concept of fuzzy discretization, which uses the cuts obtained from crisp discretization and transforms a decision table of continuous valued attributes to a fuzzy discretized decision table. This incorporates the positional information of the samples within an interval. The superiority of the proposed scheme in terms of both producer accuracy (PA) and user accuracy (UA) is established on speech and hepatobiliary disorder data. For this purpose we have used both rough set theoretic classifier and multilayer perceptrons (MLP) which are capable of handling symbolic input.

2. Rough sets

Let us consider a finite, non-empty set U of objects called the universe. Any subset $X \subseteq U$ of the universe is called a concept and the family of concepts is called the abstract knowledge. The concept of class emerges if we deal with a certain universe U, in which the families of subsets $C = \{X_1, X_2, \ldots, X_n\}$ are disjoint partitions such that $\cup X_i = U$. The partitions X_i are the equivalence class derived through a set of equivalence relations R. The concept of rough set was introduced to approximate a set $X \subseteq U$ which is not a partition directly but can be approximated by the set of equivalence relations R which generate the family of equivalence classes.

An information system (IS) is a pair, $\mathrm{IS} = (U, A \cup \{d\})$, where U is the universe, A is the set of conditional attributes and d is the set of decision attributes. If A and d are sets of n and m attributes respectively then every n-tuple is a pattern vector and the corresponding m-tuple is its identity. Thus an information system can be viewed as a decision table or a set of patterns with their identities.

Let $B \subseteq A$ and $X \subseteq U$ be in an information system. The set X is approximated using information contained in B by constructing B-lower and B-upper approximations sets:

$$\underline{B}X = \{x | [x]_B \subseteq X\} \quad \text{(B-lower approximation)}$$

and

$$\overline{B}X = \{x | [x]_B \cap X \neq \emptyset\} \quad \text{(B-upper approximation)}$$

The elements in $\underline{B}X$ can be classified as members of X by the knowledge in B, however the elements in $\overline{B}X$ can be classified as possible members of X by the knowledge in B. The set $BN_B(X) = \overline{B}X - \underline{B}X$ is called the B-boundary region of X and it consists of those objects that cannot be classified with certainty as members of X with the knowledge in B. The set X is called rough (or roughly definable) with respect to the knowledge in B if the boundary region is non-empty.

Rough set theoretic classifiers use the concept of rough set usually in reducing the number of attributes in a decision table (computation of "reducts" (Pawlak, 1991)) and in handling inconsistent decision tables. It accepts discretized (symbolic) input.

3. (Crisp) discretization

When the value set of any attribute in a decision table is continuous valued or real numbers, then it is likely that there will be very few objects that will have the same value of the corresponding attribute. In such a situation the number of equivalence classes based on that attribute will be large and there will be very few elements in each of such equivalence class. This leads to the generation of a large numbers of antecedents in the classification rule, thereby making rough set theoretic classifiers inefficient. Discretization is a process of grouping the values of the attributes in intervals in such a way that the knowledge content or the discernibility is not lost.

Let the information system be $\mathrm{IS} = (U, A \cup \{d\})$, where $V_a = [v_a, w_a)$ is an interval of reals which is to be partitioned by a set P_a of V_a for any $a \in A$. The process of discretization finds a partition of V_a defined by a sequence of the so-called cuts $v_1 < v_2 \cdots < v_k$ from V_a satisfying some natural condition such as preserving the discernibility of the information system.

The process of discretization can be local (univariate) or global (multi-variate). In local discretization each attribute is discretized independently satisfying the constraints locally for that attribute. The constraints however may not be satisfied on the final discretized decision table.

Therefore the process may result in loss of discernibility. In the case of global discretization, all the attributes are considered together and discretization process does not result in loss of discernibility. In global discretization the decision table however may not result in the formation of grouping and there may be a one-to-one correspondence with the non-discretized decision table. The problem of finding the optimal set of cuts in global discretization is NP-hard. Most of the discretization algorithms are heuristic and the obtained partitions are sub-optimal (Komorowski et al., 1999).

4. Fuzzy discretization

Any realistic multi-dimensional data of a physical process is primarily continuous valued and is overlapped in feature space relative to some classes. In other words, in rough set terminology we get an inconsistent decision table of continuous valued attributes. As mentioned in the previous section, since the crisp discretization algorithms do not usually find the optimal set of cuts, the positions of cuts tend to be sensitive to the subsets considered of any information system. We conjecture, this sensitiveness would be more pronounced in the case of an inconsistent decision table. This means any attribute value that is close to any cut can fall either way if some different set of data is considered, or some different heuristic is applied in the process of discretization. Therefore it may not be appropriate to consider crisp discretization always.

In this article we propose a method of fuzzy discretization that works on the available set of cuts found from crisp discretization. We consider that the positions of cuts are fuzzy, or in other words, the degree of belonging of the value of any attribute to any interval, defined by consecutive cuts, is not crisp and is defined by a membership value. Let c_i and c_{i+1} be two consecutive cuts. Let the attribute value v be designated not only by the group number, g, but also by a membership value, $m \in [0,1]$ depending on the position of v in the interval $[c_i, c_{i+1}]$. We have also added a third component called the "affinity", a, which discriminates the data having the same membership value in the given interval but closer to a particular cut. Thus a point is converted, after fuzzy discretization, to a triplet $\{m, g, a\}$, instead of a singlet as in the case of crisp discretization.

4.1. Algorithm

In order to explain the process and the algorithm of the fuzzy discretization, we have shown in Fig. 1, as an example, the distribution of a set of n (in this example, $n = 5$) data points $\{P_1, P_2, \ldots, P_n\}$ (as marked by solid dots) of a continuous valued attribute. The members of the set $C = \{c_0, c_1, \ldots, c_{n-1}\}$ are the cuts on the attribute. We have considered a trapezoidal membership function between the cuts as shown in the figure. In this function, membership value varies between 'base' (in this example, base = 0.2) and 1. The fractions of the slanted and flat portions of the trapezoidal function with respect to the width of the base are marked as s and f respectively. In the open intervals $\langle -\infty, c_0 \rangle$ and $\langle c_5, \infty \rangle$ the membership value is considered to be constant (=1) instead of trapezoidal. The affinity of the points having membership value of 1 is considered to be the same as the group number. For other points, the affinity value is the same as the adjacent group number close to it. The five values P_1 to P_5 have therefore been transformed as follows:

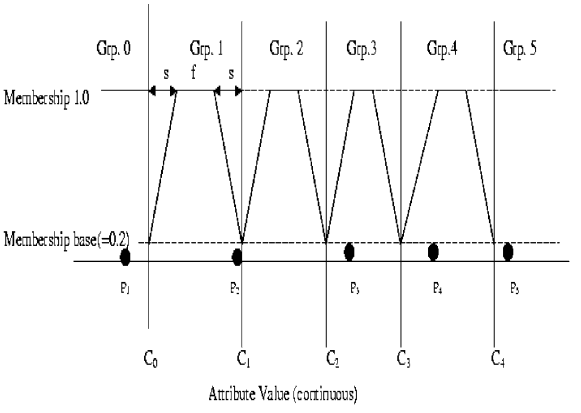

Fig. 1. An example.

$P_1 \to \{1,0,0\},\ P_2 \to \{0.2,1,2\},$
$P_3 \to \{0.8,3,2\},\ P_4 \to \{1,4,4\},\ P_5 \to \{1,5,5\}$

The algorithm for converting an attribute value v to a triplet $\{m,g,a\}$ is described below:

For each attribute value v do the following:

```
if v <= c_0 {
        m = 1.0;
        g = 0;
        a = g;
}
else if v > c_{n-1} {
        m = 1.0;
        g = n;
        a = g;
}
else {
   done = false;
   i = 0;
   do {
      t = s * (c_{i+1} - c_i);
      left = c_i + t;
      right = c_{i+1} - t;
      if v < c_{i+1} {
         if v < left {
            m = base + (1 - base) * (v - c_i)/t;
            g = i + 1;
            a = g - 1;
         }
         else if v < right {
                m = 1;
                g = i + 1;
                a = g;
         }
         else {
            m = base +
            (1 - base) * (c_i+1 - v)/t;
            g = i + 1;
            a = g + 1;
         }
         done = true;
      }
      i = i + 1
   } while(i < n - 1 and done == false);
}
```

4.2. Effect of fuzzification

To illustrate how the proposed fuzzification scheme can alter the feature space, we consider two overlapping classes in a single attribute (feature). In Fig. 2 we have shown the distribution of such attribute values. The members of the set $C = \{21.5, 33.5, 49.5, 67.4, 100.5, 126.5, 224\}$ representing the cuts obtained by the global discretization process, are shown by vertical lines. The overlap nature of the values for the two classes is evident from the distribution. Fig. 3 shows the distribution of values of the same patterns after crisp discretization. In Figs. 4 and 5 we have shown the projections of the distribution of the same patterns in m–g plane and m–a plane after fuzzification. Figs. 4 and 5, together demonstrates the enhanced discernibility of patterns, as compared to the original (Fig. 2) and crisp discretized version (Fig. 3).

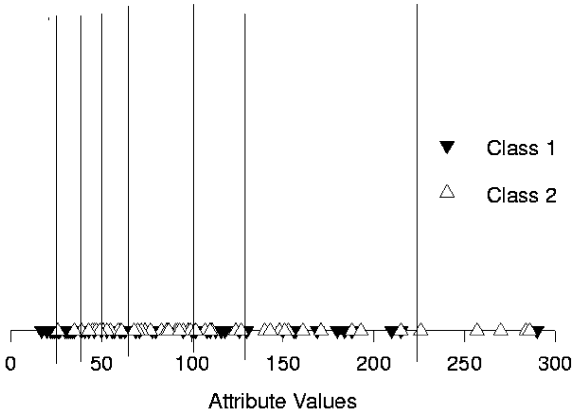

Fig. 2. Distribution of attribute values of two overlapping classes.

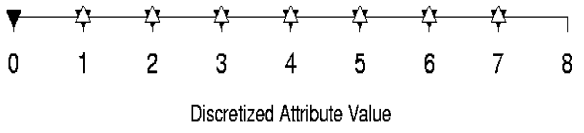

Fig. 3. Distribution of attribute values after crisp discretization.

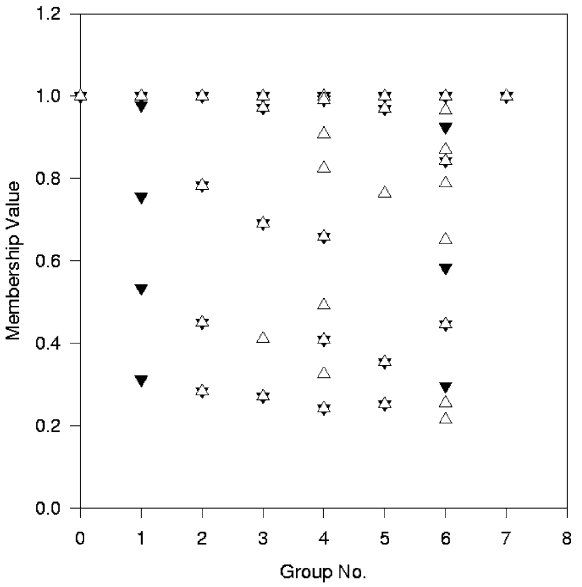

Fig. 4. Distribution of attribute values in m–g plane after fuzzification.

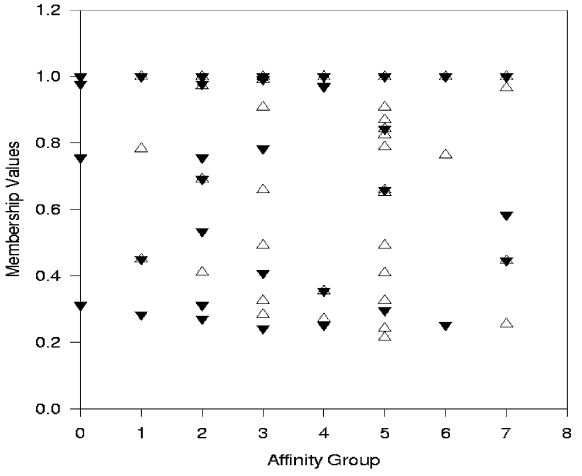

Fig. 5. Distribution of attribute values in m–a plane after fuzzification.

5. Experimental results

To demonstrate the effectiveness of fuzzy discretization, we have considered here the problem of pattern recognition using a rough set classifier (RSC). We have also used one of the widely used pattern recognition tools, MLP, to check the consistency of the effectiveness. Two overlapping data sets, namely, vowel (Pal and Dutta Majumder, 1977) and hepatobiliary disorder (Hyashi, 1994; Mitra, 1994) are used. Both the data sets have continuous valued attributes.

Vowel data consists of 871 patterns. There are six overlapping classes $(\partial, a, e, i, o, u)$ and three input features (formant frequencies F_1, F_2 and F_3). All entries are integers. Hepatobiliary disorders data consists of 536 patterns. There are four hepatobiliary disorders (classes) and nine features (symptoms). Out of these we have considered four best features (Pal et al., 1999) for our experiment. These are glutamic oxalacetic transaminate (GOT, Karmen unit), glutamic pyruvic transaminase (GPT, Karmen unit), lactate dehydrase (LDH, iu/l), and mean corpuscular volume of red blood cell (MCV, fl). The four classes represent hepatobiliary disorders namely, alcoholic liver damage (ALD), primary hepatoma (PH), liver cirrhosis (LC) and cholelithiasis (C). The overlapping nature of the data sets is evident from their projections on a subset of the feature space as seen in Figs. 6 and 7.

While using RSC and MLP we used both the original and crisp discretized data sets as input for

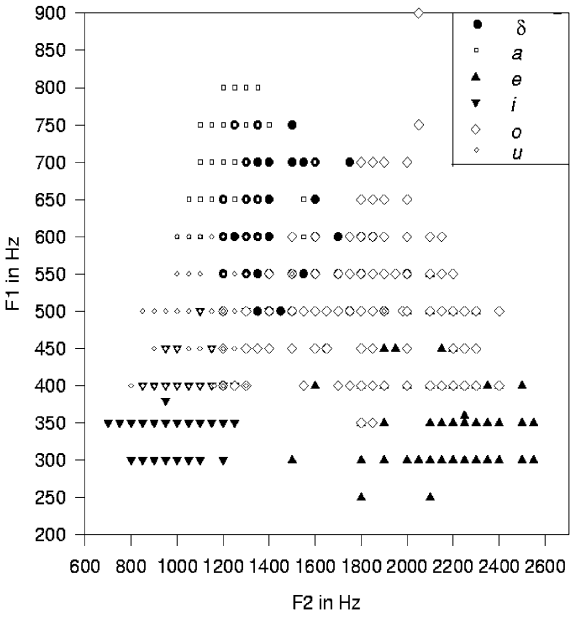

Fig. 6. Projection of vowel data in $F1$–$F2$ plane.

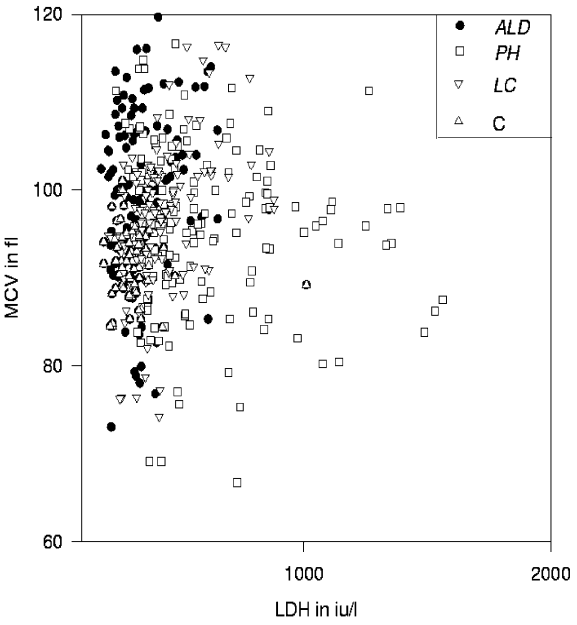

Fig. 7. Projection hepatobiliary disorder data in MCV–LDH plane.

comparing the results with that of fuzzy discretized data. In the case of MLP, we have also added the comparison with the raw (non-discretized) data as MLP has no restriction on the type of inputs. The performance is compared in terms of PA and UA, described in Appendix A. Results shown here are the average values computed over five different runs.

Table 1
RSC: overall PA in percentage

Discretization method	Number of patterns 330	Number of patterns 440	Number of patterns 500
Crisp	71.57	71.56	74.37
Fuzzy	73.31	74.91	78.06

5.1. Results with vowel data

Tables 1 and 2 show the performance for vowel data using RSC. Here a rough set explorer ROSE2Lite (Predki et al., 1998; Predki and Wilk, 1999) was used for both crisp discretization and classification. It splits a data set into n different folds, uses $n-1$ folds as the training set and the rest as the test set, and it repeats the same n times rotating across the folds. We have kept the number of folds as 2. It means the training and test sets are both 50%. (ROSE2Lite has a limitation of handling maximum of 500 samples and 20 attributes.)

In Table 1 we have shown the variation of overall PA with the number of patterns in the data set as 330, 440 and 500. Table 2 demonstrates the class-wise performance in terms of both PA and UA for 500 patterns corresponding to Table 1. From Table 1 it is seen that fuzzy discretization performs better than the crisp discretization in all the cases. This is also true for class-wise performance in terms of PA and UA, except for class ∂ (Table 2).

Tables 3 and 4 correspond to the results using MLP for vowel data. Since MLP can work, unlike the rough classifier, on raw data as the input, we have included the results using raw data also in Tables 3 and 4. In Table 3 we have shown the variation of overall PA with training set for vowel

Table 3
MLP: overall PA in percentage

Discretization method	Trn. set 20%	Trn. set 30%	Trn. set 40%	Trn. set 50%
Raw data	81.73	82.62	82.84	83.46
Crisp	83.19	82.52	82.96	83.37
Fuzzy	82.17	84.83	85.72	86.48

Test set: 52%.

Table 2
RSC: class-wise PA vs. UA

Discretization method	Class ∂		Class a		Class e		Class I		Class o		Class u	
	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA
Crisp	50.4	69.4	73.1	87.3	81.2	86.7	81.5	76.4	72.0	78.3	75.9	74.2
Fuzzy	46.2	63.2	86.1	88.8	86.3	90.1	83.3	85.2	75.8	80.0	76.1	78.6

Number of patterns: 500.

Table 4
MLP: class-wise PA vs. UA

Discretization method	Class ∂		Class *a*		Class *e*		Class *I*		Class *o*		Class *u*	
	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA	PA	UA
Raw data	46.6	72.9	86.4	89.6	83.4	90.2	83.9	90.8	89.3	77.3	87.2	83.7
Crisp	50.8	69.6	81.7	84.6	84.1	89.3	87.0	87.9	91.5	76.9	81.9	87.7
Fuzzy	64.1	74.6	87.8	88.2	86.9	90.1	86.7	91.0	88.5	82.0	88.4	87.8

Training set: 40%.

data using MLP with one hidden layer containing ten nodes. This configuration of the network was found earlier (Pal and Mitra, 1992) to produce the best performance among many others. The same test set consisting of 52% samples is used throughout the experiment. In all the cases (except for 20% training set) fuzzy discretization is found to be the best.

Table 4 shows the class-wise performance for both PA and UA corresponding to the 40% training set of Table 3. The fuzzy discretization is seen to be superior in most cases. The class ∂ having maximum overlapping nature (Fig. 6) shows significant improvement in PA without sacrificing UA.

5.2. Results with hepatobiliary disorder data

Tables 5 and 6 show the results corresponding to RSC, whereas Tables 7 and 8 refer to those using the MLP. The overall recognition scores (PA) in the case of fuzzy discretization are found to be the best for both classifiers (Tables 5 and 7) using different training sets. As in the case of vowel data, the class-wise performance (using PA and UA) is also seen to be superior in most of the cases in fuzzy discretization (Tables 6 and 8).

6. Conclusions

A concept of fuzzy discretization is introduced. It has three components, namely, membership value, group number and affinity value, unlike crisp discretization which is characterised only by the group number. This provides a better tool for handling inconsistent decision tables arising from overlapping pattern classes. It is evident experimentally using a RSC (which accepts symbolic input) that fuzzy discretization has an edge over

Table 5
RSC: overall PA in percentage

Discretization method	Number of patterns 300	Number of patterns 400	Number of patterns 500
Crisp	49.00	49.94	49.74
Fuzzy	49.88	49.05	51.01

Table 7
MLP: overall PA in percentage

Discretization method	Trn. set 20%	Trn. set 30%	Trn. set 40%	Trn. set 50%
Raw data	47.15	50.95	52.88	53.33
Crisp	56.46	60.91	63.53	62.25
Fuzzy	58.78	63.60	68.05	69.88

Test set: 60%.

Table 6
RSC: class-wise PA vs. UA

Discretization method	Class ALD		Class PH		Class LC		Class C	
	PA	UA	PA	UA	PA	UA	PA	UA
Crisp	27.64	47.04	58.09	64.03	26.80	44.17	82.19	66.49
Fuzzy	28.98	48.39	61.63	65.12	24.79	44.51	82.07	67.39

Number of patterns: 500.

Table 8
MLP: class-wise PA vs. UA

Discretization method	Class ALD		Class PH		Class LC		Class C	
	PA	UA	PA	UA	PA	UA	PA	UA
Raw data	3.83	68.89	79.23	57.67	11.92	15.01	87.53	57.14
Crisp	43.32	66.33	74.85	61.64	44.99	49.71	84.35	79.16
Fuzzy	51.51	62.82	73.7	67.98	57.22	59.02	85.89	83.57

Traing set: 40%, test set: 60%.

crisp discretization when overlapping vowel data and hepatobiliary disorder data are used as input. Fuzzy discretization is also found to perform consistently better over crisp discretization as well as non-discretization (raw data) when MLP is used as a classifier. Although we have trapezoidal membership function, one may consider any π or triangular function depending on the problem.

Acknowledgements

The first author thanks Mr. S.K. De for his interest in this work. He also thanks Dr. Sushmita Mitra and Mr. Pabitra Mitra for their cooperation and help.

Appendix A. Producer accuracy and user accuracy

Let l be the number of classes and N be an $l \times l$ matrix whose (i,j)th element n_{ij} indicates the number of patters actually belonging to class i but classified as class j.

Producer accuracy: The PA or some times referred to as accuracy or classification score is defined as $n_{ii}/(n_i \times 100)$, where n_i is the number of patterns in class i, and n_{ii}, the number of these points which are correctly classified as class i.

User accuracy: The UA is defined as $n_{ii}/(n'_i \times 100)$, where n'_i is the number of patterns classified as class i. This gives a measure of the confidence that a classifier attributes to a region as belonging to a class. In other words, it denotes the level of purity associated with a region.

References

Hyashi, Y., 1994. Neural expert system using fuzzy teaching input and its application to medical diagnosis'. Inf. Sci. Appl. 1, 47–58.

Komorowski, J., Pawlak, Z., Polkowski, L., Skoworn, A., 1999. Rough sets: A tutorial. In: Pal, S.K., Skoworn, A. (Eds.), Rough Fuzzy Hybridization: A New Trend in Decision-Making. Springer, Berlin, pp. 3–98.

Lenarcik, A., Piasta, Z., 1992. Discretization of condition attributes space. In: Slowinski, R. (Ed.), Intelligent Decision Support—Handbook of Applications and Advances of the Rough Sets Theory. Kluwar Academic Publishrs, Dordrecht, pp. 373–389.

Mitra, S., 1994. Fuzzy MLP based expert system for medical diagnosis. Fuzzy Sets Syst. 65, 285–296.

Nguyen, N.H., Nguyen, N.S., 1998. Discretization methods in data mining. In: Skowron, A., Polkowski, L. (Eds.), Rough Sets in Knowledge Discovery, vol. 1. Physica Verlag, Heidelberg, pp. 451–452.

Nguyen, H.S., Skowron, A., 1995. Quantization of real valued attributes, rough set and Boolean reasoning approaches. In: Proc. Second Annual Conf. Information Sciences. Wrightsville Beach, NC, USA, pp. 34–37.

Pal, S.K., Dutta Majumder, D., 1977. Fuzzy sets and decision making approaches in vowel and speaker recognition. IEEE Trans. Syst., Man, Cybernet. 7, 625–629.

Pal, S.K., Mitra, S., 1992. Multi-layer perceptron, fuzzy sets and classification. IEEE Trans. Neural Networks 3, 683–697.

Pal, S.K., De, R.K., Basak, J., 1999. Unsupervised feature evaluation: A neuro-fuzzy approach. IEEE Trans. Neural Networks 12, 1429–1455.

Pawlak, Z., 1991. Rough Sets, Theoretical Aspects of Reasoning About Data. Kluwer, Dordrecht, The Netherlands.

Predki, B., Slowinski, R., Stefanowski, J., Susmaga, R., Wilk, Sz., 1998. ROSE—Software Implementation of the Rough Set Theory. In: Polkowski, L., Skowron, A. (Eds.), Rough Sets and Current Trends in Computing, Lecture Notes in Artificial Intelligence, 1424. Springer, Berlin, pp. 605–608.

Predki, B., Wilk, Sz., 1999. Rough Set Based Data Exploration Using ROSE System. In: Ras, Z.W., Skowron, A. (Eds.), Foundations of Intelligent Systems, Lecture Notes in Artificial Intelligence, 1609. Springer, Berlin, pp. 172–180.