Fuzzy Sets and Systems 152 (2005) 553–564

Symbolic classification, clustering and fuzzy radial basis function network

Kalyani Mali[a], Sushmita Mitra[b,*]

[a]*Department of Computer Science, Kalyani University, Kalyani 741 235, India*
[b]*Machine Intelligence Unit, Indian Statistical Institute, 203 B. T. Road, Calcutta 700 108, India*

Abstract

Symbolic fuzzy classification is proposed using fuzzy radial basis function network, with fuzzy *c*-medoids clustering at the hidden layer. Symbolic objects include linguistic, nominal, boolean and interval-type of features, along with quantitative attributes. Classification and clustering in this domain involve the use of symbolic dissimilarity between the objects. Fuzzy memberships are used for appropriately handling uncertainty inherent in real-life decisions. The fuzzy radial basis function (FRBF) network here comprises an integration of the principles of radial basis function (RBF) network and fuzzy *c*-medoids clustering, for handling non-numeric data. The optimal number of hidden nodes is determined by using clustering validity indices, like normalized modified Hubert's statistic and Davies–Bouldin index, in the symbolic framework. The effectiveness of the symbolic fuzzy classification is demonstrated on real-life benchmark data sets. Comparison is provided with the performance of a decision tree.

Keywords: Radial basis function network; Fuzzy clustering; Symbolic object; Symbolic classification; Fuzzy classification; Validity index

1. Introduction

Symbolic or categorical clustering refers to the clustering of symbolic (or categorical) data. This is important from the point of view of data mining, where one has to mine for information from a set of

symbolic objects. Symbolic objects are defined as the logical conjunction of events linking values and variables. The following are two examples of events: $e_1 = [\text{color} = \{\text{white, blue}\}]$, $e_2 = [\text{height} = [1.5 - 2.0]]$. Here e_1 indicates that the variable color takes a value either white or blue, while e_2 indicates that the variable height takes a value between 1.5 and 2.0. For simplicity, we can drop the variable name and only take the value of that feature variable. Symbolic objects are defined by attributes that can be quantitative (numeric or intervals) as well as qualitative. The similarity and dissimilarity measures between symbolic objects are determined based on their position, span and content [4].

Soft computing is a consortium of methodologies that works synergistically and provides flexible information processing capability for handling real-life ambiguous situations [14]. Its aim is to exploit the tolerance for imprecision, uncertainty, approximate reasoning and partial truth in order to achieve tractability, robustness and low-cost solutions. Neuro-fuzzy computing [12] is the earliest and most widely reported hybridization in this framework. This integration provides intelligent systems, in terms of parallelism, fault tolerance, adaptivity and uncertainty management, in order to handle real life recognition/decision-making problems.

The radial basis function (RBF) [11] is a three-layered network, typically used for supervised classification. The hidden layer performs crisp clustering using Gaussian basis function at the nodes. The output layer performs a linear combination of the weighted activations from the hidden layer. The fuzzy RBF (FRBF) [10] is designed by integrating the principles of RBF network and the fuzzy c-means (FCM) algorithm [1]. It incorporates fuzzy set-theoretic concepts at the input, output and hidden layers. The model can handle both linguistic and numeric inputs, and provides soft decision in case of overlapping pattern classes at the output. The use of FCM in the hidden layer allows the network to provide a more accurate representation of real-life situations, where a pattern can have finite non-zero membership to two or more classes. The architecture of the network is suitably modified at the hidden layer to realize the fuzzy clustering algorithm.

Computation of means is often not feasible in the symbolic framework, where numeric data may not be involved. Instead of the mean, one may use another central tendency termed *medoid*. This has wide applications in handling of multimedia data, as required in data mining [9]. In the present article, the hidden layer of the network is adapted to model the clustering around medoids. The hard and fuzzy c-medoids clustering [7] is incorporated at the hidden layer of the RBF and FRBF, respectively. The novelty of the proposed method lies in its ability to handle optimal number of clusters, followed by an effective classification in the symbolic framework.

The hard and fuzzy c-medoids algorithms are used to generate the optimal number of clusters, as determined by the validity indices. Different indices [2,6], like Normalized modified Hubert's statistic and Davies–Bouldin index, are used in the symbolic framework [8], with the dissimilarity measure expressed in terms of symbolic distance. They help determine the optimal number of hidden nodes for the FRBF and RBF networks. These models are used for classifying symbolic data. Results are provided on real-life benchmark data sets.

The major contribution of this article lies in incorporating clustering and validity indices in the symbolic domain (as described in Ref. [8]) into the radial basis function framework. Fuzzy set-theoretic concepts are used for fuzzy clustering as well as classification, thereby enabling more efficient handling of real-life overlapping data. Both the hard and fuzzy c-medoid algorithms are used in the RBF and FRBF, respectively, to generate the optimal number of hidden nodes. Finally, the output layer of the FRBF performs fuzzy classification. The FRBF is observed to perform consistently better over all datasets.

The article is organized as follows. Section 2 provides a basic understanding of the fuzzy radial basis function (FRBF) network. The notion of symbolic partitions and different validity indices is described in Section 3. The design of the symbolic FRBF, with medoids-based clustering at the hidden layer, is discussed in Section 4. The results of the proposed system are presented in Section 5, along with a comparison with the performance of a decision tree, viz., ID3 [13]. Section 6 concludes the article.

2. Fuzzy radial basis function network

The FRBF network [10] incorporates an amalgamation of FCM clustering at the hidden layer of the RBF network. The input space is partitioned using overlapping linguistic sets, thereby utilizing more local information that aids in better classification. Each input feature X_j is expressed in terms of membership values to each of the three linguistic property sets *low*, *medium* and *high*. Therefore, an n-dimensional pattern $\vec{X}_i = [X_{i1}, X_{i2}, \ldots, X_{in}]$ is represented as a $3n$-dimensional vector

$$[x_1, \ldots, x_{3n}] = \left[\mu_{low(X_{i1})}(\vec{X}_i), \ldots, \mu_{high(X_{in})}(\vec{X}_i)\right]. \tag{1}$$

Here, the linguistic properties *low, medium* and *high* are modeled using $1 - S$, π and S functions [12], respectively. The output is provided in terms of class membership values to the l classes, such that $0 \leqslant \mu_k(\vec{x}_i) \leqslant 1$ for $k = 1, \ldots, l$. This is proportional to the weighted distance of the training pattern from the kth class mean.

The input-hidden layer weights are initialized by cluster centers using fuzzy c-means, instead of the more conventional hard c-means. The intermediate (hidden) layer is suitably modified to incorporate FCM clustering [1] during learning, such that each output node receives the weighted membership value (as opposed to a Gaussian function-based measure of proximity) of the enhanced input vector within each cluster. The resultant FRBF architecture is depicted in Fig. 1 [10].

In the fuzzy c-means algorithm, the membership value of any pattern vector $\vec{x}_j$ to a class k is represented as

$$u_{kj} = \frac{1}{\sum_{i=1}^{c} \left(\frac{d_{kj}}{d_{ij}}\right)^{(2/m-1)}}, \tag{2}$$

where d_{kj} is the distance of the pattern vector $\vec{x}_j$ from the center $\vec{v}_k$ of the kth cluster. Here

$$\vec{v}_k = \frac{\sum_{j=1}^{N} (u_{kj})^m \vec{x}_j}{\sum_{j=1}^{N} (u_{kj})^m}, \tag{3}$$

with fuzzifier $1 < m < \infty$, such that $u_{kj} \in [0, 1]$, for c clusters $1 \leqslant k \leqslant c$, for N pattern points $1 \leqslant j \leqslant N$, with $\sum_{k=1}^{c} u_{kj} = 1$, for $1 \leqslant j \leqslant N$, and $\sum_{j=1}^{N} u_{kj} > 0$, for $1 \leqslant k \leqslant c$.

The objective is to do fuzzy partitioning of the data in the hidden layer of the FRBF network. In order to perform the local computation of Eq. (2), a modified architecture is used. Eq. (2) is rewritten as

$$u_{kj} = \frac{h_k^{(j)}}{\sum_{i=1}^{c} h_i^{(j)}}, \tag{4}$$

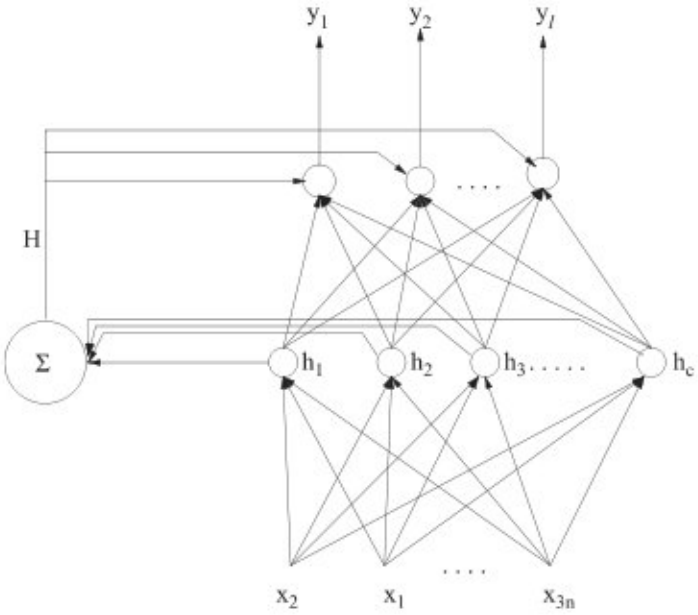

Fig. 1. Fuzzy radial basis function (FRBF) network.

where

$$h_i^{(j)} = \left(\frac{1}{d_{ij}}\right)^{(2/m-1)}. \tag{5}$$

The activation of each node in the output layer is given as

$$y_i^{(p)} = \sum_{j=1}^{c} W_{ij} u_{jp}, \tag{6}$$

where $y_i^{(p)}$ is the response of the ith output node when $\vec{x}_p$ is present at the input of the network, u_{jp} is the output of the jth hidden node and W_{ij} is the corresponding connection weight (typically initialized to a random value lying in the range [0, 0.5]). From Eqs. (4) and (5), $y_i^{(p)}$ can be written as

$$y_i^{(p)} = \frac{1}{H_a^{(p)}} \sum_{j=1}^{c} W_{ij} h_j^{(p)}, \tag{7}$$

where

$$H_a^{(p)} = \sum_{i=1}^{c} h_i^{(p)}. \tag{8}$$

Eqs. (5) and (7) reveal that the h_is can be computed locally in the hidden nodes and the activation of the output nodes can be computed from the hidden node activations with an additional normalization by the total output in the hidden layer (H_a). An auxiliary hidden node is used in the FRBF (as shown in Fig. 1)

to compute the total activation in the hidden layer and feed it to the output layer. The weights of the links from all hidden nodes to the auxiliary hidden node are set to unity. Note that the membership value u_{kj} of Eq. (2) is implicitly included in the network architecture in terms of the hidden node activations.

We minimize the error $E = \frac{1}{2}\sum_p \sum_i (*y_i^{(p)} - y_i^{(p)})$ during training. The rule for updating the weights, using the least mean square algorithm, is given as

$$\Delta W_{ij}^{(p)} = \frac{\eta}{H_a^{(p)}} \left(*y_i^{(p)} - y_i^{(p)}\right) h_j^{(p)}, \tag{9}$$

where η is the learning rate and $*y_i^{(p)}$ is the target output in terms of class membership of the training pattern. Here $\Delta W_{ij}^{(p)}$ is the change in W_{ij} during training when $\vec{x}_p$ is presented at the input in the $3n$-dimensional form of Eq. (1).

3. Symbolic partitions

In this section we describe dissimilarity measure between objects, along with some validity indices in the symbolic framework. The dissimilarity between two symbolic objects A and B is defined as [4,5]

$$D(A, B) = \sum_{i=1}^{n} D(A_i, B_i), \tag{10}$$

where

$$D(A_i, B_i) = D_p(A_i, B_i) + D_s(A_i, B_i) + D_c(A_i, B_i)$$

with D_p, D_s and D_c (normalized to [0,1]) indicating the components due to position, span and content, respectively.

Let $\{X_1, \ldots, X_{c_k}\}$ be a set of symbolic objects lying in a cluster U_k. Then the *average scatter* within the cluster U_k is expressed as

$$S_a(U_k) = \frac{\sum_{i,i'} ||X_i - X_{i'}||}{|c_k|(|c_k| - 1)}, \tag{11}$$

where $X_i, X_{i'} \in U_k$, $i \neq i'$, $|c_k|$ is the number of samples in cluster U_k and $||\cdot||$ indicates the symbolic dissimilarity of Eq. (10). The *between cluster scatter* is defined as

$$d_a(U_k, U_l) = \frac{\sum_{i,j} ||X_i - X_j||}{|c_k||c_l|}, \tag{12}$$

where $X_i \in U_k$, $X_j \in U_l$, such that $k \neq l$. We have used S_a and d_a in our computations, in terms of the symbolic dissimilarity D of Eq. (10).

To select the best among different partitioning, each of these can be evaluated using some validity index. The procedure is repeated for $c = 2, \ldots, \sqrt{N}$ number of clusters, where N is the size of the data set. Some validation methods, like normalized modified Hubert's statistic and Davies–Bouldin index, are used in the symbolic framework [8].

Let X_i be the ith object and $L(i) = k$ if $X_i \in U_k$. The modified Hubert's Γ statistic for a particular cluster structure is expressed in terms of symbolic dissimilarity as

$$\Gamma = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} D(X_i, X_j) d_a(U_{L(i)}, U_{L(j)}). \tag{13}$$

If X_i and X_j lie in two different clusters, d_a is computed using Eq. (12). However, when they belong to the same cluster, $d_a = 0$. From this, we get normalized modified Hubert's statistic $\hat{\Gamma}$ as

$$\hat{\Gamma} = \frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \frac{(D(X_i, X_j) - \overline{D})(d_a(U_{L(i)}, U_{L(j)}) - \overline{d_a})}{s_D s_{d_a}}, \tag{14}$$

where

$$\overline{D} = \frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} D(X_i, X_j),$$

$$\overline{d_a} = \frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_a(U_{L(i)}, U_{L(j)}),$$

$$s_D^2 = \frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} D^2(X_i, X_j) - \overline{D}^2,$$

and

$$s_{d_a}^2 = \frac{1}{M} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} d_a^2(U_{L(i)}, U_{L(j)}) - \overline{d_a}^2,$$

where $M = [N(N-1)]/2$ is the total number of terms under the double summation. Note that $M = N^2$ if the matrix under summation is not symmetric. The optimal partitioning occurs at $c = c_0$ for which $\Delta(\Delta\hat{\Gamma})$ is minimum. This corresponds to a sharp change in slope of the piecewise linear graph for the normalized modified Hubert's statistic.

The Davies–Bouldin index is a function of the ratio of the sum of within-cluster scatter to between-cluster separation. The best clustering, for $c = c_0$, minimizes

$$\frac{1}{c} \sum_{k=1}^{c} \max_{l \neq k} \left\{ \frac{S_a(U_k) + S_a(U_l)}{d_a(U_k, U_l)} \right\}, \tag{15}$$

for $1 \leqslant k, l \leqslant c$. Here, the within-cluster scatter is minimized and the between-cluster separation is maximized. The index is expressed in the symbolic framework.

4. Symbolic classification using medoids-based clustering

Since computation of means is often not feasible in the symbolic framework, we incorporate the c-medoids and fuzzy c-medoids algorithms at the hidden layer of the RBF and FRBF (Fig. 1), respectively. The use of fuzzy memberships in FRBF helps in better modeling of uncertain real-life data. Before moving on to the modeling of the FRBF in the symbolic domain, we provide a brief description of the hard and fuzzy c-medoids algorithms.

The partitioning around medoids (PAM) algorithm uses the most centrally located object in a cluster, the medoid, instead of the mean. The basic steps are outlined as follows:

- Choose the first initial medoid as the object that is most centrally located in the data set. Pick $c-1$ more objects successively as the subsequent medoids, such that each is most dissimilar to the medoids that have already been selected.
- Assign each remaining data object (pattern) to the cluster for the closest medoid.
- Replace each of the medoids by one of all the non-medoids (causing the greatest reduction in square error), as long as the quality of clustering improves.
- Iterate until the criterion function converges.

The fuzzy c-Medoids clustering performs a fuzzification of the c-medoids algorithm and is outlined as follows:

(i) Pick the initial medoids $\mathbf{v}_i$, $i = 1, \ldots, c$ (as described for PAM).
(ii) Repeat steps (iii)–(iv) until convergence.
(iii) Compute the fuzzy membership u_{ik}, for $i = 1, \ldots, c$ and $k = 1, \ldots, N$, using Eq. (2).
(iv) Compute new medoids

$$\mathbf{v}_i = \mathbf{X}_q, \tag{16}$$

where

$$q = \arg \min_{1 \leqslant j \leqslant N} \sum_{k=1}^{N} (u_{ik})^m ||\mathbf{X}_j - \mathbf{X}_k||^2. \tag{17}$$

Note that this boils down to the hard c-medoids with $u_{ik} = 1$, if $i = q$, and $u_{ik} = 0$ otherwise.

The FRBF is extended to work in the symbolic domain by incorporating the clustering validity indices of Section 3 [8] into the radial basis function framework. Fuzzy set-theoretic concepts are introduced in terms of fuzzy c-medoids clustering at the hidden layer. The use of fuzzy inputs and output class membership in the FRBF framework enables fuzzy classification, thereby leading to a more efficient handling of real-life overlapping data.

The hard c-medoids at the hidden layer of the RBF serve as a base for comparison. The inherent computational complexities of the hard/fuzzy c-medoids clustering, as compared to the hard/fuzzy c-means clustering, cannot be overcome. However, the utility of medoids in appropriately modeling non-numeric (symbolic) data, as compared to means, makes them a better candidate for extending the FRBF to work in the symbolic domain.

In this article, the fuzzy c-medoids clustering is used to determine the optimal number of hidden nodes of Fig. 1. The cluster medoids $\mathbf{v}_i$ are computed using Eq. (16). These are used as the cluster centers, in

place of Eq. (3). The mapping to the hidden nodes of the FRBF is made using Eqs. (4) and (5), with the distance d_{ij} of pattern $\mathbf{X}_j$ from cluster medoid $\mathbf{v}_i$ being computed in the symbolic framework of Eq. (10). Besides, the fuzzy membership u_{ik} of pattern $\mathbf{X}_k$ to cluster U_i [Eq. (2)] now involves this symbolic distance. The output node activations of the FRBF follow from Eqs. (6)–(8). The connection weight updates are controlled by Eq. (9).

5. Results

The symbolic classification was done on benchmark data, viz., *Zoo* and *Soybean* [3]. Table 1 illustrates the optimal clustering generated for the two symbolic data sets using the hard and fuzzy c-medoids algorithms, corresponding to the different clustering validity indices.

Tables 3–4 provide the classification performance of the RBF and FRBF on *Zoo* and *Soybean* data, for different training set sizes as well as different number of clusters (hidden nodes). Note that the number of clusters corresponds to the optimal values generated by the different validity indices in Table 1. The connection weights of the networks were initialized as random values lying in the range $-0.1, +0.1]$, the learning rate was set at $\eta = 0.5$ in Eq. (9) and the fuzzifier was chosen as $m = 1.2$ in Eq. (17) after several experiments. A random training set of size $x\%$, chosen classwise, is selected for training the network. The remaining $(100 - x)\%$ data serves as the test set. The results are provided in the tables for both the training and testing sets, individual classwise (1, 2, etc.) as well as total (*Net*) over all classes. The FRBF is found to perform consistently better than the RBF in all cases. This validates the necessity of using the FRBF in the symbolic framework. Table 5 provides a comparative performance of the decision tree algorithm ID3 [13] over the test set, for the *Zoo* and *Soybean* data.

5.1. Zoo data

The *Zoo* data [3] consist of 100 instances of animals with 17 features. The name of the animal constitutes the first attribute. There are 15 boolean features corresponding to the presence of hair, feathers, eggs, milk, backbone, fins, tail; and whether airborne, aquatic, predator, toothed, breathes, venomous, domestic, catsize. The character attribute corresponds to the number of legs lying in the set {0, 2, 4, 5, 6, 8}.

Table 2 provides a sample-detailed partitioning of the *Zoo* data, obtained using the Davies–Bouldin clustering validity index along with the fuzzy c-medoids algorithm. This generated three clusters, with 39, 29 and 32 objects each, corresponding to the last column of Table 1. This is used as the optimal number of hidden nodes for the FRBF, whose classification performance is demonstrated in Table 3.

From Tables 1 and 2, it is observed that in case of three clusters the partition is made based on criteria of "producing milk", "absence of feathers" and "not producing milk", and "presence of feathers" and "not producing milk". In case of four and five clusters the separation also involves "having tail".

Table 1
Optimal clusters with c-medoids algorithm

Data	Hard		Fuzzy	
	Huberts statistics	Davies–Bouldin	Huberts statistic	Davies–Bouldin
Zoo	5	4	5	3
Soybean	6	5	8	9

Table 2
Symbolic partitions on *Zoo* data using Davies–Bouldin index

Cluster No.	Animals
1 (39)	aardvark, bear, girl, boar, cheetah, leopard, lion, raccoon, wolf, lynx, mongoose, polecat, puma, mink, platypus, seal, sealion, antelope, buffalo, deer, elephant, giraffe, oryx, gorilla, wallaby, calf, goat, pony, reindeer, pussycat, cavy, hamster, fruitbat, vampire, squirrel, hare, vole, mole, opossum.
2 (29)	bass, catfish, piranha, chub, herring, carp, haddock, seahorse, sole, dogfish, pike, tuna, stingray, frog, toad, newt, tuatara, pitviper, clam, seawasp, crab, starfish, crayfish, lobster, slowworm, seasnake, dolphin, octopus, porpoise.
3 (32)	chicken, dove, parakeet, lark, pheasant, sparrow, wren, flamingo, ostrich, tortoise, crow, hawk, flea, vulture, kiwi, rhea, penguin, duck, swan, skua, termite, slug, worm, gnat, gull, skimmer, scorpion, ladybird, housefly, moth, honeybee, wasp.

Table 3
Recognition scores (%) with RBF and FRBF for *Zoo* data

Trn set (%)	No. of clust	RBF						FRBF					
		Training			Testing			Training			Testing		
		1	2	Net	1	2	Net	1	2	Net	1	2	Net
	3	91.7	100.	96.7	82.8	100.	93.0	91.7	100.	96.7	93.1	100.	97.2
30	4	91.7	100.	96.7	82.8	100.	93.0	91.7	100.	96.7	93.1	100.	97.2
	5	91.7	94.4	93.3	89.7	100.	95.8	91.7	94.4	93.3	89.7	100.	95.8
	3	93.8	100.	97.5	80.0	100.	91.8	93.8	100.	97.5	92.0	100.	96.7
40	4	93.8	100.	97.5	80.0	100.	91.8	93.8	100.	97.5	92.0	100.	96.7
	5	93.8	100.	97.5	84.0	100.	93.4	100.	91.7	95.0	92.0	100.	96.7
	3	85.0	100.	94.0	76.2	100.	90.2	95.0	100.	98.0	90.5	100.	96.1
50	4	95.0	100.	98.0	76.2	100.	90.2	100.	96.7	98.0	81.0	100.	92.2
	5	95.0	100.	98.0	81.0	100.	92.2	100.	96.7	98.0	81.0	100.	92.2

Table 3 provides the performance using the RBF and FRBF networks for different sizes of training sets. The classification is based on whether the animal is a mammal or not. The number of clusters in the hidden layer are determined by the hard and fuzzy c-medoids clustering algorithms, respectively. The optimal numbers of clusters from Table 1, obtained using the different validity indices, are used for evaluating the network performance. It is observed that the FRBF model, with embedded fuzzy c-medoids clustering, performs generally better over the test set.

A decision tree is another popular tool used for classification. We have compared the performance of the interactive dichotomizer ID3 [13], and the corresponding recognition scores are provided in Table 5. For training set sizes 30%, 40%, 50%, we find that the FRBF is consistently superior.

Table 4
Recognition scores (%) with RBF and FRBF for *Soybean* data

Trn (%)	Mod	Clus No.	Training					Testing				
			1	2	3	4	Net	1	2	3	4	Net
	R	5	90.0	100.	66.7	66.7	93.8	71.7	88.2	0.0	100.	81.7
	B	6	90.0	100.	66.7	66.7	93.8	69.6	84.0	0.0	100.	78.5
	F	8	90.0	100.	66.7	66.7	93.8	71.7	95.8	28.6	100.	87.6
30		9	90.0	100.	66.7	83.3	95.0	80.4	94.1	28.6	100.	88.7
	F	5	95.0	96.1	100.	66.7	93.8	82.1	99.2	28.6	85.7	91.4
	R	6	95.0	96.1	100.	83.3	95.0	84.8	94.1	57.1	92.9	90.9
	B	8	90.0	98.0	100.	83.3	95.0	86.9	98.3	28.6	92.9	92.5
	F	9	90.0	100.	100.	83.3	96.3	91.3	98.3	57.1	92.9	94.6
	R	5	92.3	95.6	100.	62.5	92.5	100.	99.1	16.6	75.0	91.3
	B	6	92.3	98.5	75.0	50.0	92.5	100.	92.2	16.7	75.0	90.0
	F	8	92.3	98.5	100.	87.5	96.2	100.	93.1	16.7	50.0	88.8
40		9	88.5	100.	75.0	87.5	95.3	100.	96.1	16.7	58.3	91.3
	F	5	96.2	95.6	100.	62.5	93.4	87.5	99.0	16.7	83.3	91.9
	R	6	96.2	97.1	100.	62.5	94.3	87.5	99.0	33.3	75.0	91.9
	B	8	96. 2	97.1	100.	87.5	96.2	97.5	95.1	33.3	75.0	91.9
	F	9	92.3	98.5	100.	87.5	96.2	97.5	95.1	33.3	91.7	93.1

5.2. Soybean data

There are 266 number of instances (samples) with 35 symbolic attributes. The four output classes (1,2,3,4) are categorized as the soybean diseases (*diaporthe stem canker*, *charcoal rot*, *rhizoctonia root rot*, *phytophthora rot*, *brown stem rot*); (*powdery mildew*, *downy mildew*, *brown spot*, *phyllosticta leaf spot*, *alternaria leaf spot*, *frog eye leaf spot*, *bacterial blight*, *bacterial pustule*); *purple seed stain*; and *anthracnose.*

Let us first analyze the optimal clusters generated in Table 1. Starting from two clusters, the partitioning separated classes 1,4 and classes 2,3. For five clusters, class 1 is split into "canker" and the various types of "rot"s; "anthracnose" is separated and some "pustule" samples comprise another group. Considering six clusters, some types of "spots" are put in a different cluster from class 2. The entire partitioning is rather fuzzy, implying improved performance for fuzzy classification and clustering.

Table 4 provides the performance of the RBF and FRBF networks on the *Soybean* data, for different sizes of training sets. The optimal numbers of clusters from Table 1, obtained using the different validity indices with the hard and fuzzy c-medoids algorithms, are used for evaluating the network performance. The FRBF model is found to have an overall better performance. Comparing with the performance of ID3 in Table 5, for training set sizes 30% and 40%, we find that FRBF is always superior.

6. Conclusions and discussion

Real-life data are essentially not restricted to the numeric domain. Hence, the need for symbolic processing to efficiently handle data like linguistic, nominal, boolean, interval, shape, color, etc. arises.

Table 5
Performance on test data with ID3

Data	Class	Training set (%)		
		30	40	50
Zoo	1	95.24	94.28	93.33
	2	96.43	96.15	95.24
	Net	95.77	95.08	94.12
Soybean	1	91.49	97.50	—
	2	89.07	92.16	—
	3	14.28	50.00	—
	4	85.71	83.33	—
	Net	86.63	91.25	—

Partitioning of such data demands the use of symbolic measures for determining the similarity and dissimilarity between objects. In this article we have demonstrated the effectiveness of symbolic fuzzy classification and clustering on benchmark data, viz., *Zoo* and *Soybean*. The FRBF, in the symbolic framework, performed consistently better over all data sets, even when compared to the decision tree based ID3.

Clustering has useful applications in data mining, pattern recognition, image segmentation, rule extraction and web mining. The importance of symbolic clustering in real world data is all the more evident, considering the availability of large volumes of mixed-media data that are distributed over the internet. We have used inter-cluster and intra-cluster scatter in the symbolic framework. Different clustering validity indices have been modified to incorporate the symbolic computations using dissimilarity measures. The generated clusters are observed to be *naturally* meaningful for the symbolic data used.

The novelty of the method lies in effectively clustering and classifying patterns in a symbolic framework. The hard and fuzzy c-medoids clustering algorithms have been used to determine the optimal number of hidden nodes of the RBF and FRBF networks, respectively. The incorporation of fuzzy membership of fuzzy c-medoids in the FRBF helps to handle better the uncertainties inherent in real-life data. In general the performance of the FRBF network has been found to be better for both symbolic data sets used.

Acknowledgements

The authors gratefully acknowledge the anonymous reviewers for their valuable suggestions toward improving this article.

References

[1] J.C. Bezdek, Pattern Recognition with Fuzzy Objective Function Algorithms, Plenum Press, New York, 1981.
[2] J.C. Bezdek, N.R. Pal, Some new indexes for cluster validity, IEEE Trans. Systems Man Cybernet. Part-B 28 (1998) 301–315.
[3] C.L. Blake, C.J. Merz, UCI repository of machine learning databases, Department of Information and Computer Sciences, University of California, Irvine, 1998. http://www.ics.uci.edu/ mlearn/MLRepository.html.

[4] K. Chidananda Gowda, E. Diday, Symbolic clustering using a new dissimilarity measure, Pattern Recognition 24 (6) (1991) 567–578.
[5] K. Chidananda Gowda, T.V. Ravi, Divisive clustering of symbolic objects using the concepts of both similarity and dissimilarity, Pattern Recognition 28 (8) (1995) 1277–1282.
[6] A.K. Jain, R.C. Dubes, Algorithms for Clustering Data, Prentice-Hall, Englewood Cliffs, NJ, 1988.
[7] R. Krishnapuram, A. Joshi, O. Nasraoui, L. Yi, Low complexity fuzzy relational clustering algorithms for web mining, IEEE Trans. Fuzzy Systems 9 (2001) 595–607.
[8] K. Mali, S. Mitra, Clustering and its validation in a symbolic framework, Pattern Recognition Lett. 24 (2003) 2367–2376.
[9] S. Mitra, T. Acharya, Data Mining: Multimedia, Soft Computing, and Bioinformatics, Wiley, New York, 2003.
[10] S. Mitra, J. Basak, FRBF: A fuzzy radial basis function network, Neural Comput. Appl. 10 (2001) 244–252.
[11] J. Moody, C.J. Darken, Fast learning in networks of locally-tuned processing units, Neural Comput. 1 (1989) 281–294.
[12] S.K. Pal, S. Mitra, Neuro-fuzzy Pattern Recognition: Methods in Soft Computing, Wiley, New York, 1999.
[13] Y.H. Pao, Adaptive Pattern Recognition and Neural Networks, Addison-Wesley, Reading, MA, 1989.
[14] L.A. Zadeh, Fuzzy logic, neural networks, and soft computing, Comm. ACM 37 (1994) 77–84.