A Data Mining Approach for Fuzzy Classification Rule Generation

Dianhui Wang  Tharam S. Dillon  Elizabeth J. Chang*

Department of Computing, The Hong Kong Polytechnic University
Hunghom, Kowloon, Hong Kong, P. R. China
* Dept of CS & SE, Newcastle University, Australia

Abstract

This paper aims at developing a data mining approach for fuzzy classification rule generation. A regularization theory based theoretical framework for refining fuzzy classification rules is proposed. Our fuzzy rule induction methodology has four phases, namely: (1) ellipsoidal crisp rule generation with membership function assignment, (2) generic hyperbox fuzzy rule (GHR) derivation,(3) refinement of the GHR using a regularization model, and (4) simplification of the GHR by selecting an informative subset of premises out of the initial set.

1 Introduction

The concept of fuzzy sets or fuzzy logic has received considerable attention since Zadeh proposed it [1]. In the past few decades, as one of the important tools in the study of intelligent systems, fuzzy logic have been extensively used in many domains including decision-making theory, pattern recognition, image processing, software engineering, nonlinear system modeling (approximation) and control engineering et al [4, 9, 11]. When domain experts are available they can help specify some fuzzy rules and further refinement on this fuzzy rule set can be made by learning approaches. Unfortunately, domain experts do not always exist in most situations. Also if the dimension of input variables involved in a system is larger than 3 it may be difficult to specify the rules because the complexity will exceed a human being’s mental power. This drives researchers to explore the problem of fuzzy rule extraction from numerical data using algorithms [7, 8, 10, 12].

Roughly speaking, there are two kinds of fuzzy rules, i.e., Mamdani-type fuzzy rule (MR) and Sugeno-type fuzzy rule (SR). They can be represented by the following expressions, respectively [5]:

MR: IF $x_1$ is $A_1$ AND $x_2$ is $A_2$ AND $x_n$ is $A_n$, THEN $y = a_0 + a_1 x_1 + \cdots + a_n x_n$.

SR: IF $x_1$ is $A_1$ AND $x_2$ is $A_2$ \ldots AND $x_n$ is $A_n$, THEN $y$ is class.

where $x_j (j = 1, 2, \ldots, n)$ denotes the input, $y$ denotes the output of the rule, $A_j (j = 1, 2, \ldots, n)$ and $B$ are the fuzzy membership functions associated with linguistic terms, and $a_j$ are real numbers.

There are different advantages associated with each of these two types of knowledge representation. If one cares about readability and comprehensibility of the rule, then MR can be adopted; otherwise, the SR can be used if one focuses on the precision issue without considering its linguistic interpretability aspect. Recently, a new fuzzy model is popular in both approximation and classification domains [4, 11, 12]. The fuzzy rule in this case can be described by

Rule: IF $x_1$ is $A_1$ AND $x_2$ is $A_2$ \ldots AND $x_n$ is $A_n$, THEN $y$ is class,

where class denotes a class related to the fuzzy rule which can be either a real number or a class label. Indeed, this model is a tradeoff between MR and SR, i.e., uses a crisp form as the consequent part in the MR model and ignores the individual terms in the SR model. Moreover, the membership functions (MFs) in the fuzzy rule are derived by projecting an ellipsoid in n-dimensional space onto each input axis [11].

For modeling problems, supervised learning can be applied to adjust the parameters involved in the fuzzy system in order to obtain a better accurate approximation [5, 11, 12]. However, if one wants to address some performance aspects in classification problems, no cost function is currently available which has embeded performance information in it. It is known that the representation of fuzzy rules with projection MFs is not totally acceptable although the fuzzy rules work in some sense. Figure 6 in [11] showed an example giving the same fuzzy rules with same membership functions but from different ellipsoidal regions. This is not acceptable in
the pattern recognition domain because the rule may cover patterns from different classes at the same time, and hence is not really useful for classification.

Data mining is a process of transferring and analyzing available sets of specific data and extracting the information and knowledge in the form of relationships, patterns or clusters for decision-making, classification, prediction and control [2, 4]. Construction of if-then fuzzy rules can be viewed as a data mining task which is concerned with the exploitation of information inherent in databases. This is often achieved by clustering data points that are close to one another according to some metric or criteria [3]. Fuzzy rule generation can be categorized into two broad classes, i.e., direct and indirect methods. In the direct method, the cluster centers for a linguistic concept in high dimensional space for a fuzzy rule can be obtained by clustering techniques, and then a MF associated with this concept is assigned using some parameterized functions. A further refinement for the cluster center and parameters involved in the MF can be carried out to meet some criteria [6, 7, 9, 11, 12]. The indirect method encodes domain knowledge expressed using linguistic concepts in various neural-net models and then updates the structures and weights of the neural-nets during the refinement purpose. The second method can automatically provide one with explanatory functional fuzzy rules and has a fast inference process due to the connectionist models obtained. For more details, readers may refer to a recent survey paper [5].

This paper aims at developing a data mining approach for fuzzy classification rule (FCR) generation. We here refer to fuzzy rule generation in terms of both rule extraction and rule optimization. A generic representation for FCRs is proposed using coordinate transformation. A method for constructing the MFs is also given. There are four phases in our proposed methodology, namely, (a) ellipsoidal crisp rule generation with MFs assignment using self-organizing map networks (SOM); (b) generic hyperbox fuzzy rules (GHR) derivation from the ellipsoidal fuzzy rules; c) refinement of the GHR using regularization theory and genetic algorithms (GA); and (d) simplification of the GHR by selecting an informative subset of premises out of the initial set.

Our main contributions include:

1. Generic FCR representation
2. Regularization theory for FCR refinement
3. Data mining methodology for FCR generation

The remainder of this paper is arranged as follows. Section 2 gives a unified framework for FCR representation. Section 3 presents the details of our regularization theory for FCR refinement. Section 4 describes a data mining approach for FCR generation. We conclude this paper in the last section.

2 Fuzzy Rule Representation

For simplicity, we consider a two-class classification problem and only focus on the connective fuzzy rule generation issue. The method proposed here can be extended to complex situations directly. To proceed our discussions, we first introduce the notation used in this paper.

Let \( \mathbb{R}^{n \times m} \) denote the space for the set of \( n \times m \) real matrices space. If \( M \in \mathbb{R}^{n \times m} \), then \( M^T \) denotes the transpose of \( M \); \( M^{-1} \) is the inverse of \( M \) if it is nonsingular and \( Tr(M) \) the trace of \( M \) if \( m = n \), respectively. Let \( S \) be a set with finite elements, we use \( |S| \) to denote the cardinality of \( S \), i.e., the number of elements contained in the set \( S \). For a given positive definite matrix \( M \in \mathbb{R}^{n \times n} \), \( X,Y \in \mathbb{R}^{n} \) the Mahalanobis Distance Function (MDF) is defined by \( d_M(X,Y) = (X-Y)^TM^{-1}(X-Y) \), and the Euclidean Distance Function, denoted by \( d(X,Y) \), is given by setting \( M \) to be the identity matrix in the MDF. The 2-norm of a vector \( X \in \mathbb{R}^{n} \) is denoted by \( ||X|| \) and calculated as \( ||X|| = \sqrt{d(X,0)} \).

Fuzzy rules are usually modeled by a Cartesian product, namely, \( A = A_1 \times A_2 \times \cdots \times A_r \) of the fuzzy sets for the input space. This Cartesian product defines a fuzzy partition of the input space. From a mathematical point of view, the search or mining of relevant fuzzy classification rules can be formulated as a search of the regions in the feature space where some meaningful or typical patterns with a high density of samples appear. This is related to clustering techniques. Note that the description of a fuzzy rule contains two aspects, i.e., a point associated with a linguistic concept and a membership function characterizing the state of pattern distribution around this region. Note that the distribution of patterns from two classes in high dimensional space can be complicated so that it is very possible to have overlap projection ranges on all axes. Although the patterns from different classes do not share the same cluster center, a high misclassification rate will appear in such a situation. This is because the classification criterion depends upon the firing strength of the fuzzy rules. This observation is meaningful and significant to fuzzy classifier design.

To overcome this problem, we use a coordinate transformation technique which results in a new type of fuzzy rule which is described as follows:

\[
\text{WR: IF } z_1 \text{ is } A_1 \text{ AND } z_2 \text{ is } A_2 \ldots \text{ AND } z_n \text{ is } A_n \text{ THEN } y \text{ is class}_r.
\]
where \( z_j = \sum_{i=1}^{n} \alpha_{ij} x_i \alpha_{ji} \) are real numbers, \( A_j (j = 1, 2, \cdots, n) \) are MFs associated with the rule \( WR \) and dependent on the new variable \( z_j \).

The rest of this paper handles the problem of determination of parameters \( \alpha_{ij} \) and MFs \( A_j \) in the above fuzzy rule \( WR \).

3 Regularization Theory

The backbone of our rule induction approach is a two-phase process: a rule initialization stage followed by a rule optimization stage. A regularization model to tradeoff misclassification rate, recognition rate and generalization ability is proposed for refining the initial rules. In the first phase, we try to construct a crisp classification rule (CR) defined by an ellipsoidal region of the form below:

\[
\text{CR: IF } d_{A_j}(C_p, X) \leq 1 \text{ THEN } X \in \text{Class}_p.
\]

where \( p = 1, 2 \). The basic idea behind the regularization theory is to tradeoff two or more possibly conflicting criteria for performance in the design of a system. In data mining for pattern classification, we mainly consider three important criteria that the system is trying to meet, i.e., misclassification rate \( (\text{MisR}) \), an increased coverage rate \( (\text{CovR}) \) and an improved generalization power \( (\text{GenP}) \). The MisR is given by \( (\text{Mis/Tot}) \), where \( \text{Mis} \) and \( \text{Tot} \) stand for the number of patterns classified incorrectly by the rule set and the total number of patterns in the data set, respectively. The RR, which indicates the quality of a classification system, is a measure of the power to classify examples correctly for both training data and test data, and it is given by \( (\text{Con/Tot}) \), where \( \text{Con} \) stands for the number of patterns classified correctly by the rules. The GenP of a set of rules is a measure of the ability to classify unseen examples correctly. A good classifier will have lower MR and higher RR for both training set and test set and a higher GenP simultaneously.

As the ellipsoidal region in \( WR \) tends to be large enough to enclose many training patterns for a certain category of data, the misclassification rate could be increased for this class. It could lead to a lower recognition for other classes. On the other hand, if the size of region is reduced and only a small number of training data for a class is enclosed, a good generalization ability of the rule can not be expected because of over-fitting, also the recognition rate of this class could be reduced due to poor coverage. Moreover, the degree of complexity will increase with the number of rules. Therefore, a tradeoff between the size of the ellipsoidal region to achieve good generalization ability, a high coverage rate and a low misclassification rate should be addressed. Note that there is no direct way to exactly evaluate the GenP during the design phase. However, this measurement can be indirectly expressed by using the size of rule region. Recall that the trace is a reasonable measure to characterize the size of the ellipsoidal region, Thus, GenR and CovR for \( \text{Class}_p \) can be approximately evaluated by using \( |S_p(A_p, C_p)| + \text{Tr}(A_p) \). The MisR is naturally calculated by \( |S_p(A_p, C_p)|/|S_p(A_p, C_p)|, p \neq q \), where \( S_p(A_p, C_p) = \{ X : d_{A_p}(C_p, X) \leq 1, X \in \text{Class}_r \}, p, r = 1, 2 \). Using this notation, we define a regularization function (RF) as follows:

\[
F_p = \frac{|S_p(A_p, C_p)|}{|S_p(A_p, C_p)| + |S_p(A_p, C_p)| + \text{Tr}(A_p)}
\]

where \( \lambda > 0 \) is a regularizing factor. Figure 1 gives a graphical illustration of our regularization theory. It should be pointed out that the RF is not unique.

4 Fuzzy Rule Generation

This section provides an outline of our FCR induction techniques. The Kohonen net is a good tool for clustering, so we employ it in this work.

A. Normalization

The Kohonen approach works best with a normalized set of input features [3]. If the original features in the real world space are written as the vector \( X \), then we define a normalized set of features \( V \) as follows:

\[
V = \frac{1}{\sqrt{1 + ||X||^2}} \left[ \frac{X}{||X||} \right] = \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_{n+1} \end{bmatrix}
\]

Fig.1 Graphical illustration of regularization theory

A sub-optimal CR can be obtained by minimizing (1) for a given regularizing factor. The optimization problem above is ill-defined because the cost function is not differentiable or even continuous with respect to the involved parameters. Therefore, we solve it by using a GA algorithm [13].

Authorized licensed use limited to: LA TROBE UNIVERSITY. Downloaded on December 15, 2009 at 19:34 from IEEE Xplore. Restrictions apply.
By transformation (2), the data set in the real world space can be mapped to a corresponding data set in V-space, which will be used to initialize rules. To recover the value of a vector $X$ in the real world from a value of a vector $V$ in the normalized feature space, we use the following inverse transformation:

$$X = \frac{v_n + 1}{1 - v_n} \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix}$$

(3)

B. CR Rule Generation

Let $Net = \{(i,j)\}$ denote the set of nodes in the Kohonen net, $W_{ij} : (i,j) \in Net$ the set of weights where $W_{ij}$ is the weight vector at each node $(i,j)$. A pattern $V_0$ in V-space is said to be fired by a node $(i_0,j_0) \in Net$, if $\|V_0 - W_{i_0,j_0}\| = \min(\|V_0 - W_{i,j}\|)$. The set of nodes fired by patterns from $Class_p$ can be transferred to the original feature space, and it is denoted by $Sw_p$. Let $I_w = Sw_1 \cap Sw_2$, $S_{wp} = Sw_p - I_w$, $p = 1, 2$. Then, an ellipsoid can be given by

$$A^0_p = \gamma(W_{ij} - C^0_p)(W_{ij} - C^0_p)^T, W_{ij} \in S_{wp},$$

$$C^0_p = \sum_{w_{ij} \in S_{wp}} W_{ij}$$

(4)

$$\sum_{w_{ij} \in S_{wp}} W_{ij}$$

(5)

where $\gamma > 0$ is a adjustable parameter.

By using the regularization function (1) and a real-coded GA algorithm, we can obtain a suboptimal CR rule, which is denoted by $(A^*_p, C^*_p)$.

C. Setting Membership Function

The crisp rule described in the last section provides a clear border for classification. However, it is usually sensitive to the data noise and this results in an increase in the rule numbers and a decrease of the power of the classifier. To ameliorate the performance of crisp classifiers so that it is more powerful in efficiency and robust with respect to the data noise, a fuzzy boundary for the crisp rule should be associated with a MF. The MF associated with a fuzzy subset can be approximated by interpolating some discrete grades of membership. Usually, this approximation curve is implemented by some specific types of functions with few adjustable parameters. In this paper, we give a method to construct the MF using a modified $\pi$-function.

Let $X \subseteq \mathbb{R}^n$ be the universe of discourse, and $0 \leq m_B(X) \leq 1$ the MF at point $X \in \mathbb{R}^n$ associated with a fuzzy subset $B$ of $X$. We call the set $\{X : \sigma < m_B(X) \leq 1\}$ a $\sigma$-cut set of the fuzzy subset $B$ and denoted by $B_\sigma$. Specifically, we name 0-cut set and 1-cut set as support set and core set of the fuzzy set $B$, respectively. Based on the refined crisp rule $(A^*_p, C^*_p)$, a MF is defined as follows:

$$\pi(X) = \begin{cases} 1 - (1 - \epsilon) d \frac{A^*_p}{A^*_p} (C^*_p, X), & \text{if } 0 \leq d \leq 1 \\ \frac{d^2}{(\delta - \delta)^2} [1 - \frac{\sqrt{d A^*_p(X,C^*_p)}}{\delta}]^2, & \text{if } 1 \leq d \leq \delta \\ 0, & \text{otherwise} \end{cases}$$

where $d = \sqrt{d A^*_p(X,C^*_p)}$, $0 < \epsilon \leq 1$ is a threshold given by designers, and $\delta$ is to be determined.

If the crisp rule contains no examples that are misclassified, then the core set is attainable by setting $\epsilon = 1$. However, we do not constrain the methodology to crisp rule generation with exactly zero-miscalssification rate, so the $\epsilon = 1$ is usually set to be a real number slightly less than 1. The corresponding $\epsilon$-cut set is denoted by $B_{\epsilon}$, which is exactly the ellipsoidal region defined by CR. To determine the support set for the rule, we need to select an appropriate parameter for $\delta$. To do this, a simple iterative algorithm is used such that

$$R(\delta) = \frac{|B_{\epsilon} \cap \Theta_p|}{|B_{\epsilon} \cap \Theta_q|} < \theta,$$

(6)

where $\theta > 0$ is a threshold given by designers, $\Theta_p$ denotes the training data set of $Class_p$, and $p \neq q$.

In this way, the assignment of the MF can be completed by replacing the terminal value of $\epsilon$, denoted by $\sigma$, in the above $\pi$-function.

D. Initial WR Rule Derivation

Now, we have obtained two ellipsoids with the same center $C^*_p$ and different radii, that is

$$(X - C^*_p)^T (A^*_p)^{-1} (X - C^*_p) = 1$$

(7)

$$(X - C^*_p)^T (A^*_p)^{-1} (X - C^*_p) = \sigma^2$$

(8)

If the matrix $A^*_p$ is not positive definite, the ellipsoids will concentrate on a lower dimensional hyperplane. We here assume that $A^*_p$ is positive definite. Then, it can be transformed to be a diagonal matrix $\Lambda = diag(\lambda_1, \lambda_2, \ldots, \lambda_n)$ by using an orthogonal matrix $P$ whose columns are unit eigenvectors $e_1, e_2, \ldots, e_n$ of $A^*_p$, i.e., $P$ rotates the coordinate system to the eigenvectors to orient the ellipsoid. The Euclidean half lengths of axes equal $\sqrt{\lambda_1}, \sqrt{\lambda_2}, \ldots, \sqrt{\lambda_n}$ and $\sigma \sqrt{\lambda_1}, \sigma \sqrt{\lambda_2}, \ldots, \sigma \sqrt{\lambda_n}$ for (7) and (8), respectively. Let $z_j = e^T \cdot X, a_j = e^T \cdot C^*_p, j = 1, 2, \ldots, n$. Then the MF $A_j$ for variable $z_j, j = 1, 2, \ldots, n$ can be expressed by

$$m_A(z_j) = \begin{cases} 1, & \text{if } f_j^1 \leq z_j \leq g_j^1 \\ \frac{z_j - f_j^1}{g_j^1 - f_j^1}, & \text{if } f_j^2 \leq z_j \leq f_j^2 \\ \frac{z_j - g_j^2}{g_j^2 - f_j^2}, & \text{if } g_j^2 \leq z_j \leq g_j^2 \\ 0, & \text{otherwise} \end{cases}$$

(9)
where \( j = 1, 2, \cdots, n \) with
\[
\begin{align*}
f_1^j &= \alpha_j - \sqrt{x_j} + 1 - \epsilon_j, \\
g_1^j &= \alpha_j + \sqrt{x_j} - 1 + \epsilon_j.
\end{align*}
\]
This completes WR rule initialization.

**E. Rule Optimization**

The WR rule optimization involves both MF refinement and premise simplification here. It is well-known that performance of fuzzy classifiers is sensitive to MFs [10]. Hence, it is necessary to refine the MFs described by (9). A product combination operator is used for calculating the firing strength, and a presented pattern will be classified as one class that has the maximum value of a classification indicator (CI), here the CI is taken as the firing strength.

Let \( U_p \) denote a set of patterns (from training set) to be classified to be \( \text{Class}_p \) by the WR rules, \( T_p = U_p \cap \Theta_p, p = 1, 2 \). Next, a criterion is needed for adjusting the 4n parameters involved in the FMs. In this paper, the criterion is given by
\[
J_p = \frac{|U_p - T_p|}{|T_p|} + \frac{1}{|U_p|} \tag{10}
\]
The GA algorithm is utilized to minimize (10) in the parameter space.

The premise simplification is straightforward after completing the coordinate system rotation. Indeed, an attribute \( x_k \) can be removed out of \( z_j \) provided that the \( e_{jk} \) is small enough, where \( e_{jk} \) represents the k-th component of \( e_j \).

**5 Conclusions**

To construct fuzzy rules for multi-dimensional pattern recognition is significant in real world applications. Most of previous research on this issue concentrated on representation and determination of fuzzy subsets using partition, clustering and optimization techniques. A projection approach implies the use of hyperbox in characterizing the pattern distribution in feature space. This may result in either high misclassification rate or an increase of the number of rules.

Representation, generation and adaptation are three essential and important aspects in building up a fuzzy classification system. The regularization theory proposed in this work is an important contribution for this domain. We trust that the framework presented is valuable. Lastly, it should be pointed out that the use of Kohonen net for clustering purpose is not unique, and any other clustering approach could be used.

**References**


