A Data Driven Rule-Base Inference Approach for Classification Systems

Shuwei Chen, Jun Liu, Hui Wang, and Juan Carlos Augusto
School of Computing and Mathematics
University of Ulster at Jordanstown
Newtownabbey, Northern Ireland, UK
chensw915@gmail.com

Abstract—**This paper proposes a generic data driven inference methodology for rule-based classification systems. The generic rule base is in a belief rule base structure, where the consequent of a rule takes the belief distribution form. Other knowledge representation parameters such as the weights of both input attributes and rules are also considered in this framework. In an established rule base, the matching degree of an input between the antecedents of a rule is firstly computed to get the activation weight for the rule. Then a weighted aggregation of the consequents of activated rules is used for the inference process. Two numerical examples are provided to illustrate the proposed method.**

Keywords—Rule-based systems; belief distribution; data driven; classification; aggregation

I. Introduction

Rule is one of the most common forms for representing various kinds of knowledge. Rule-based systems (or knowledge-based systems), usually constructed from human knowledge in forms of if-then rules, are often applied to classification problems, such as safety analysis, biology, and medicine [1].

For a simple system, the rule base is usually obtained from human experts. However, this is not applicable when the system is complex, where the expert experience is incomplete. Many methods have been proposed in literature for generating or learning rule base from data, including heuristic approach [2], neural network technique [3], genetic algorithm approach [1, 4], support vector machine technique [5], particle swarm optimization method [6], and rough set based method [7]. Although there are already a lot of approaches proposed for learning rule base, it is still difficult to generate a fair rule base from real data due to the complexity and uncertainty of real situations. Furthermore, the rules extracted from the data are always representing the dominant features of the data by ignoring the minor ones, which will cause loss of information to some degree. So, it will be promising if all the data can be used directly and properly for predicting the class of the input without extracting or learning rules from them.

Most existing methods on rule-based classification systems do not use attribute (feature) weights and rule weights. Sometimes, this can be done for fuzzy rule-based classification systems by adjusting the membership functions of antecedent attributes [8]. This paper will provide a generic data driven rule-base inference methodology for classification system, which will take into account both the attribute weights and rule weights. In this framework, the rule base is expressed as a belief rule base [9], with the consequent of a rule taken the form of belief distribution. This method can not only be used for ordinary rule-based systems, but also be used directly on the data, which will be illustrated in the numerical study part.

The paper is organized as follows. Section II proposes the data driven rule-base inference approach, which consists of four parts: the generic rule-base structure, matching degree computation of input, activation weights for a rule, and aggregation of weighted consequents. Numerical study is given in Section III to illustrate the methodology. Section IV comes to the concluding remarks.

II. Data Driven Rule-Base Inference Approach

A. Rule-Base Structure

The rule base used in this paper takes the similar structure with the belief rule base proposed in [9], which is designed on the basis of belief structure. In a belief rule base, input for each antecedent is transformed into a distribution on the "referential values" [9] of this antecedent. This distribution describes the degree of each antecedent being activated. The activation weight of a rule can be generated by aggregating the degrees to which all antecedents in the rule are activated. The consequent of each rule is in a belief distribution form which is shown in (1). The weights of both input attributes and rules are also considered in this structure.

Suppose that there are *N* classes in a dataset. The *k*th rule in a general belief rule base for classification in forms of a conjunctive rule can be expressed as

$$R_k\text{: if } A_1^k \wedge A_2^k \wedge \cdots \wedge A_{T_k}^k \text{, then, } (D_1^k, D_2^k, \cdots, D_N^k)$$

$$\text{with a rule weight } \theta_k \text{ and attribute weights } \delta_1^k,\ \delta_2^k,\ \ldots,\ \delta_{T_k}^k,\ k \in \{1, 2, \cdots, n\}, \quad (1)$$

where A_i^k (i=1, …, T_k) is the referential value of the *i*th antecedent attribute in the *k*th rule, T_k is the number of antecedent attributes used in the *k*th rule, D_i^k (i=1, …, N) is the

978-1-4244-9698-3/11/$26.00 ©2011 IEEE

belief degree of the consequent belongs to the *i*th class, and $0 \le \theta_k \le 1$, $0 \le \delta_i^k \le 1$.

B. Matching Degree of Input to a Rule

Before an inference process can start, the relationship between an input (fact) and the antecedents in a rule needs to be determined. The matching degrees to which an input is consistent with the antecedents in a rule are processed to generate an activation weight for the rule, which is used to measure the degree to which the packet antecedent of the rule is activated by the input.

The determination of matching degree can be done through many different ways. In a belief rule base, the matching degree is obtained through referential values of the attributes. The basic idea is to examine all the referential values of each attribute in order to determine a matching degree to which an input belongs to a referential value. This is equivalent to transforming an input into a distribution on referential values using belief degrees [9, 10]. For example, one may use such linguistic terms as "highly good," "good," "fair," "poor," and "very poor." These linguistic terms are the referential values for an antecedent attribute "comfort." In a general rule base, the set of referential values may be numerical or linguistic.

A general input form corresponding to all antecedent attributes is given as

$$(A_1^*, \varepsilon_1) \wedge (A_2^*, \varepsilon_2) \wedge \cdots \wedge (A_T^*, \varepsilon_T), \tag{2}$$

where ε_i expresses the belief degree assigned to the input value A_i^* of *i*th attribute, and T is the total number of different antecedent attributes involved in all the rules in a rule base.

By using the distribution assessment approach [9, 10], a referential value of an attribute may in general be regarded as an evaluation grade, and the input (A_i^*, ε_i) for the *i*th attribute can be transformed to a distribution on the referential values of the attribute using belief degrees as

$$R(A_i^*, \varepsilon_i) = \{(A_{ij}, \alpha_{ij})\}, \tag{3}$$

where A_{ij} is the *j*th referential value of the *i*th attribute, α_{ij} the degree to which the input A_i^* belongs to the referential value A_{ij} with $\alpha_{ij} \ge 0$ and $\sum_{j=1}^{J_i} \alpha_{ij} \le 1$. α_{ij} could be generated using various ways, depending on the nature of an antecedent attribute. For instance, in evaluation of qualitative antecedent attributes, subjective judgments could be used. In assessment of the quality of a product (if its referential set is {poor, indifferent, average, good, excellent}), for example, examiners may give judgment as

> "40% sure that its quality is at the average level and 60% sure that it is good."

Hence, the set of referential values of "quality" is *R*(quality)={(poor, 0), (indifferent, 0), (average, 0.4), (good, 0.6), (excellent , 0)}.

In fuzzy rule-based systems, the matching degree to which an input is consistent with the antecedents in a rule can also be computed via the similarity measure between the input fuzzy set and the corresponding antecedent fuzzy set in a rule. For example, a simple similarity measure can be used between fuzzy sets *A* and *B* is:

$$S(A,B) = \frac{|A \cap B|}{|A \cup B|}, \tag{4}$$

where $|\cdot|$ denotes the cardinality of a set, and the $\cap$ and $\cup$ operators represent the intersection and union, respectively.

The matching degree between numerical input *A* and the corresponding antecedent A_i^k in the *k*-th rule can be obtained by a normalized difference between the corresponding numbers, one may use the simple one as:

$$S(A, A_i^k) = 1 - \frac{|A - A_i^k|}{\max\{A_1^k, \cdots, A_n^k\}}, \tag{5}$$

C. Activation Weight for a Rule

Consider an input given in a format shown in (3) corresponding to the *k*th rule defined as in (1)

$$(A_1^k, \alpha_1^k) \wedge (A_2^k, \alpha_2^k) \wedge \cdots \wedge (A_{T_k}^k, , \alpha_{T_k}^k), \tag{6}$$

where $A_i^k \in \{A_{ij}; j = 1, \cdots, J_i\}$ and $\alpha_i^k \in \{\alpha_{ij}; j = 1, \cdots, J_i\}$.

The total degree α_k to which the input matches the packet antecedent A^k in the *k*th rule can be calculated using the following formula:

$$\alpha_k = \varphi((\delta_1^k, \alpha_1^k), \cdots, (\delta_{T_k}^k, \alpha_{T_k}^k)). \tag{7}$$

Here, φ is an aggregation function that reflects the relationship among the T_k antecedents in the *k*th rule.

Suppose the "$\wedge$" connective is used for all antecedents in a rule, such as "if $A \wedge B \wedge C$." In such cases, one may use the max-min one or the following simple weighted multiplicative aggregation function to calculate α_k :

$$\alpha_k = \prod_{i=1}^{T_k} (\alpha_i^k)^{\delta_i^k} . \tag{8a}$$

If the "$\vee$" connective is used for all antecedents in a rule, such as "if $A \vee B \vee C$," then one may use the following recursively defined weighted product–sum aggregation function proposed in [10] to calculate α_k :

$$\alpha_{k(1)} = h_1^k = \delta_1^k \times \alpha_1^k$$

$$\alpha_{k(i+1)} = \alpha_{k(i)} + (1 - \alpha_{k(i)}) h_{i+1}^k \quad \text{for } i = 1, \cdots, T_k - 1$$

$$\alpha_k = \alpha_{k(T_k)}, \tag{8b}$$

where $h_j^k = \delta_j^k \times \alpha_j^k$, *j*=1, 2, …,$T_k$.

The activation weight ω_k of the packet antecedent A^k in the *k*th rule is generated by weighting and normalizing the matching degree α_k given by (8a) or (8b) as

$$\omega_k = \frac{\theta_k \alpha_k}{\sum_{i=1}^{n} \theta_i \alpha_i}. \tag{9}$$

where θ_k is the relative weight of the *k*th rule. Note that $0 \le \omega_k \le 1$ (*k*=1, …, *n*), and $\sum_{i=1}^{n} \omega_i = 1$.

D. Aggregation and Exploitation

With the activation weight ω_k for the *k*th rule in the rule base, then the output must be D_i to a certain degree. The degree is measured by both the degree to which the *k*th rule is important to the overall output and the degree to which the antecedents of the *k*th rule are activated by the actual input.

The final result (D_1, D_2, …, D_N) is computed through the weighted aggregation of the consequents of all activated rules, where the simple weighted addition can be used as in (10).

$$(D_1, D_2, \cdots, D_N) = \left(\sum_{k=1}^{n} \omega_k D_1^k, \sum_{k=1}^{n} \omega_k D_2^k, \cdots, \sum_{k=1}^{n} \omega_k D_N^k \right). \tag{10}$$

The output can also be interpreted again by a belief distribution format as in (1), like, “40% sure that it belongs to class 1 and 60% sure that it is in class 3.” One can also just select the one with maximum belief degree as the final decided class for the input.

III. Illustrative Examples

In this section, two examples about Iris data and wine data are given to illustrate the application of proposed method. These data sets are available from the UCI machine learning repository (http://archive.ics.uci.edu/ml/).

A. Iris Data Set

The best known Iris dataset consists of 3 classes of Iris flower: Setosa (class 1), Versicolour (class 2), and Virginica (class 3). Each class contains 50 samples and each sample is represented by 4 attributes: sepal length (x_1), sepal width (x_2), petal length (x_3), and petal width (x_4). The first class is linearly separable from the other two classes, while class 2 and class 3 are not separable from each other.

The 10-fold cross-validation is made on this dataset. The original dataset is randomly divided into 10 groups with each group containing the same proportions of the 3 types of class labels. In each validation, 9 from the 10 groups are selected for constructing the rule base, and the rest one is used for test. Some examples of the rule are listed in Table I.

TABLE I. Rule Base with Belief Structure

Number	θ	Antecedent				Consequent
		x_1	x_2	x_3	x_4	
1	1	4.9	3	1.4	0.2	(1, 0, 0)
2	1	5	3.2	1.2	0.2	(1, 0, 0)
3	1	4.5	2.3	1.3	0.3	(1, 0, 0)
4	1	5.5	2.4	3.8	1.1	(0, 1, 0)
5	1	6.8	2.8	4.8	1.4	(0, 1, 0)
6	0.6	5.6	3	4.5	1.5	(0, 1, 0)
7	1	7.2	3.2	6	1.8	(0, 0, 1)
8	0.8	6	2.2	5	1.5	(0, 0, 1)
9	1	6.7	3	5.2	2.3	(0, 0, 1)

In Table I, θ means the rule weight, and the consequents are the flower classes expressed by belief distribution. There are also attribute weights associated with each attribute as expressed in (1). Here, we assume that the attribute weights are (0.4, 0.1, 1, 0.8).

Now, suppose that we have a new flower with attribute values (5, 3.3, 1.4, 0.2), we will follow the steps in the proposed method to decide which class this flower belongs to.

Step 1. Input transformation

Taking into account that the attribute values of this example are numerical, we can use the normalized difference between the corresponding attribute values of the antecedents in the rules and the new flower as the matching degrees, which are shown in Table II.

TABLE II. Matching Degree

Number	Matching Degrees			
	α_1	α_2	α_3	α_4
1	0.99	0.93	1	1
2	1	0.98	0.97	1
3	0.94	0.77	0.99	0.96
4	0.94	0.80	0.65	0.64
5	0.77	0.89	0.51	0.52
6	0.92	0.93	0.55	0.48
7	0.71	0.98	0.33	0.36
8	0.87	0.75	0.48	0.48
9	0.78	0.93	0.45	0.16

Step 2. Activation weights for all rules

The total degrees α to which the input matches the packet antecedent in the rules can be calculated using formula (8a), for the relations between the antecedents are "and". The activation weights for the above 9 rues are α=(0.9878, 0.9688, 0.9049; 0.4342, 0.267, 0.2943; 0.1284, 0.2443, 0.0932).

Now, we can get the activation weights by using formula (9), and those for the above 9 rules are ω=(0.01596, 0.01565, 0.01462; 0.00701, 0.00259, 0.0475; 0.00207, 0.00316, 0.00151).

Step 3. Aggregation

This step is to aggregate the activation weights from step 2 and the corresponding consequents in belief distribution of the rule base to get the overall belief degree for each class by using formula (10) as:

$$(0.682, 0.227, 0.091) \tag{11}$$

Step 4. Exploitation (Ranking)

By a simple comparison, we can get the final decision: the new flower belongs to class 1. It can also be interpreted as:

"68.2% sure that the flower belongs to class 1, 22.7% sure that it is class 2, and 9.1% sure for class 3."

Step 5. Validation

The above 4 steps are implemented for each sample in the test group. Then the cross-validation process is repeated 10 times, with each of the 10 subsamples used exactly once as the validation data. The 10 results from the folds then can be averaged (or otherwise combined) to produce a single estimation. The average classification accuracy for the 10-fold cross-validation is 96.67%.

B. Wine Data Set

The wine data contains the chemical analysis of 178 wines grown in the same region in Italy but derived from three different cultivars, or three classes with 59 data for class 1, 71 for class 2, and 48 for class 3. The 13 continuous attributes are available for classification. We use $x1$, $x2$, $x3$, $x4$, $x5$, $x6$, $x7$, $x8$, $x9$, $x10$, $x11$, $x12$, $x13$ to represent these attributes: alcohol, malic acid, ash, alcalinity of ash, magnesium, total phenols, Paranoids, nonflavanoids phenols, proanthocyaninsm, color intensity, hue, OD280/OD315 of dilluted wines and proline respectively.

The 5-fold validation is made on this dataset along with the similar process to that for Iris dataset. Taking into account that the wine dataset is an unbalanced dataset, the dataset is randomly divided into 5 groups with each group containing different proportions of the 3 types of class labels. The rule weight for each rule is set to 1, and the attribute weights of 13 attributes are given as (0.8, 0.5, 0.5, 0.5, 0.4, 0.5, 0.8, 0.5, 0.5, 0.8, 0.7, 0.6, 1) according to the feature selection process in [1]. The average classification accuracy of the 5-fold cross-validation is 97.75%.

IV. Conclusions

A generic framework for data driven rule base inference approach for classification system has been proposed. The rule base used was in a belief structure with the consequent taking the form of belief distribution, and the weights of both input attributes and rules were also considered in this framework. Two numerical examples were given to illustrate the application of the methodology.

It should be noted that what presented in this paper is a general framework of data driven rule base inference methodology. Some related issues, such as feature selection and parameter optimization, are not discussed in detail. Future work will focus on these related issues, which will make the proposed methodology more applicable. Another interesting point is that the rule bases used in the two illustrative examples are directly interpreted from the data, *i.e.*, there is no learning or training process needed, while the accuracy rates of the two illustrative examples were similar to those in [1-8].

References

[1] J. A. Roubos, M. Setnes, and J. Abonyi, "Learning fuzzy classification rules from labeled data," Information Sciences, vol. 150, pp. 77–93, 2003.

[2] S. Abe, and M. S. Lan, "A method for fuzzy rules extraction directly from numerical data and its application to pattern classification," IEEE Trans. Fuzzy Syst., vol. 3, pp. 18–28, 1995.

[3] R. Setiono, B. Baesens, and C. Mues, "Recursive neural network rule extraction for data with mixed attributes," IEEE Trans. Neural Networks, vol. 19, pp. 299–307, 2008.

[4] A. Fernández, S. García, J. Luengo, E. BernadÓ-Mansilla, and F. Herrera "Genetics-based machine learning for rule induction: state of the art, taxonomy, and comparative study," IEEE Trans. Evolutionary Computation, vol. 14, pp. 913–941, 2010.

[5] Y. X. Chen, and J. Z. Wang, "Support vector learning for fuzzy rule-based classification systems," IEEE Trans. Fuzzy Syst., vol. 11, pp. 716–728, 2003.

[6] R.P. Prado, S. Garcia-Galan, J.E. Munoz Exposito, and A.J. Yuste, "Knowledge acquisition in fuzzy-rule-based systems with particle-swarm optimization," IEEE Trans. Fuzzy Syst., vol. 18, pp. 1083–1097, 2010.

[7] Y. N. Fan, T. L. Tseng, C. C. Chern, C. C. Huang, "Rule induction based on an incremental rough set," Expert Systems with Applications, vol. 36, pp. 11439–11450, 2009.

[8] H. Ishibuchi, and T. Yamamoto, "Rule weight specification in fuzzy rule-based classification systems," IEEE Trans. Fuzzy Syst., vol. 13, pp. 428–435, 2005.

[9] J. B. Yang, J. Liu, J. Wang, H. S. Sii, and H. W. Wang, "Belief rule-base inference methodology using the evidential reasoning approach—RIMER," IEEE Trans. Systems, Man Cybernet.—Part A, vol. 36, pp. 266–285, 2006.

[10] J. B. Yang, "Rule and utility based evidential reasoning approach for multi-attribute decision analysis under uncertainties," Eur. J. Oper. Res., vol. 131, no. 1, pp. 31–61, 2001.