A QUADRATIC MARGIN-BASED MODEL FOR WEIGHTING FUZZY CLASSIFICATION RULES INSPIRED BY SUPPORT VECTOR MACHINES

M. TAHERI, H. AZAD, K. ZIARATI AND R. SANAYE

Abstract. Recently, tuning the weights of the rules in Fuzzy Rule-Base Classification Systems is researched in order to improve the accuracy of classification. In this paper, a margin-based optimization model, inspired by Support Vector Machine classifiers, is proposed to compute these fuzzy rule weights. This approach not only considers both accuracy and generalization criteria in a single objective function, but also is independent of any order in presenting data patterns or fuzzy rules. It has a global optimum solution and needs only one regularization parameter C to be adjusted. In addition, a rule reduction method is proposed to eliminating low weighted rules and having a compact rule-base. This method is compared with some greedy, reinforcement and local search rule weighting methods on 13 standard datasets. The experimental results show that, the proposed method significantly outperforms the other ones especially from the viewpoint of generalization.

1. Introduction

Classification, one of the most important fields of pattern recognition, is categorizing some patterns of data into disjoint labeled classes. Although this field is full of successful researches in proposing classifiers with good performance [9], [26], but most of them do not provide interpretable systems and clear information about properties of data cannot be extracted. Fuzzy systems not only consider uncertainty both in input data and system output, but also prepare interpretable fuzzy rules which can be used or modified by experts [29, 21]. These rules are constructed based on known fuzzy sets in the field of research which are associated with linguistic variables in describing attributes of data e.g. small, medium or large.
Importing fuzzy systems in the field of classification as Fuzzy Rule-Based Classification Systems (FRBCS), is not a modern approach [19, 4] but training a pregenerated rule-base from the perspective of the rule weights is recently under research [36, 35, 16]. In the normal form, each fuzzy rule plays a role similar to others but by a more concentration, it is obvious that some rules are more important than others in the classification process [16, 30]. Indeed, FRBCSes are weak classifiers in comparison with the ones which are not limited to be interpretable. Therefore, some investigations have been done to extract fuzzy rules from more

Received: April 2012; Revised: August 2012; Accepted: November 2012
Key words and phrases: Classification algorithms, Fuzzy systems, Margin maximization, Rule weighting, Support vector machines.

powerful classifiers e.g. Neural Networks [10, 20], Decision Trees and SVM [34]. Tuning a predefined rule-base is highly researched [12, 6] but specially tuning the weights of the fuzzy rules, without any change in other parameters of the rule-base, in order to improve the classification performance can gather both interpretability and performance in a classification system [16, 24].
Ishibuchi [24] proposed 4 heuristic functions for assigning the rule weights based on the confidence of associated rules. These methods are fast but with low precision. Nozaki [31] proposed a reward and punishment approach of tuning these rule weights. This process is not recommended by these authors due to its time consumption and unreliablity with noisy datasets. It is dependent on the order of patterns presented and, consequently, some noisy patterns presented in the final steps of learning, may be misleading. Also GA, as a stable search strategy in complex search spaces, is widely investigated upon the field of rule selection [23, 15, 25]. Then, Chen proposed a GA based method to improve the performance of the fuzzy systems by weighting the rules [1, 2, 3]. One of the most drawbacks of this approach is its time complexity.
These authors in [36] proposed a novel method of rule-weighting based on a greedy local search strategy with high accuracy in classification process. It is called, in this paper, Iterative Greedy Accuracy-based Rule-Weighting (IGARW). Its fast convergence to the local optima speeds up the procedure of weighting. Additionally, it removes irrelevant and redundant rules by setting their weights to zero as a side-effect. Despite the fact that this approach is not order-dependent from the viewpoint of patterns, it is highly sensitive to the order of weighting the rules and initial values of the rule weights.
Most of these approaches are aimed at improving the classification rate on training data neglecting their generalization for unseen data explicitly. This may lead to encountering over-fitting on the training data. Moreover, most of them are order-dependent from some perspectives or ramped to a local optima. But in many other fields of learning classifier parameters, there are, nowadays, researches in order to have a large-margin classification to improve the generalization of the classifier with global optimum solutions [14]. In the sections to come, a novel model of constraint programming has been put forward for rule-weighting. Due to have a convex optimization, the global optimum is guaranteed to be found fast without any order-dependency. Other than these, generalization perspectives have been explicitly utilized in the objective function. In addition, this method removes inaccurate rules by adjusting some of the rule weights to zero.
In the next section, FRBCS is briefly introduced and a novel linear programming model of rule weighting is proposed in the section 3 due to be familiar with some concepts. In section 4, SVM is briefly described and followed by a quadratic programming model of margin based rule-weighting inspired from SVM in section 5. Section 6 is dedicated to dual problems and some theoretical investigation in equality of the proposed model of rule weighting with SVM considering a special fuzzy kernel. In Section 7, the classification problem is extended for multi-class datasets

and FRBCSes. A post-processing method is proposed in section 8, in order to reduce the number of the rules based on resulted rule weights. Experimental results are reported and explained in section 9 and is finally concluded.

2. **Fuzzy Rule-based Classifying Systems (FRBCS)**

A FRBCS is composed of three parts: Database, Rule-Base and Reasoning Method. The Database contains a series of fuzzy sets which are associated with specific linguistic variable e. g. small, medium, large. The Rule-Base is a collection of fuzzy if-then rules utilized for pattern classification. Each rule covers a region in the feature space, named its covering space, which is defined by antecedent part of the rule. In the field of pattern classification problems, various types of fuzzy rules have been proposed [19, 36, 16]. In this paper, the following type of fuzzy rules is used which is highly implemented in the researches from 1992 [22]:

$$Rule\,(r_k):\ If\ \left\langle x^1 = A_k^1 \right\rangle \& \left\langle x^2 = A_k^2 \right\rangle \&...\& \left\langle x^d = A_k^d \right\rangle \Rightarrow class\,(h_k)\,with\,(\omega_k) \quad (1)$$

where, r_k is the k^{th} rule in the rule-base, x_i is the i^{th} feature value of X as the pattern under classification i.e. $X = [x_1, x_2, ..., x_d]$. Also, d is the number of dimensions of the feature space concerned. A_k^i is the fuzzy set used for the i^{th} feature value in r_k. In addition, h_k is a class label as the consequence part of r_k. Each rule is associated with a certainty factor ω_k specifies how much this rule is trustable.
In order to specify how much a pattern X is compatible with a rule, a T-Norm function of its membership values in each antecedent of the rule is computed [36, 16] and addressed by $\mu_k(X)$ as the compatibility grade of pattern X with r_k. The Reasoning Method is the strategy of using both the rule-base and the database in the classification process. There are two major types of reasoning [24], i.e. Single Winner and Weighted Vote. In the Single Winner method, each pattern is classified by the rule which possesses the greatest value of weighted compatibility grade. Although the single winner method is a very simple reasoning strategy which is used in many classifiers (e.g. nearest neighbour), it is highly noise sensitive.
In the Weighted Vote method [24, 18, 5] all the rules vote in the classification process with their certainty factor. Finally, the class label with the largest value of accumulated votes is assigned to X as shown in (2). This accumulated vote of each class c for a pattern X is called here strength of the class c on X.

$$\begin{aligned} X\ is\ classified \quad & as\ c^* \in CL\ where \\ & c^* = \arg\max_{c \in CL} \left(\sum_{r_k \in R^c} \omega_k \mu_k(X) \right) \end{aligned} \quad (2)$$

in which $CL = \{c_1, c_2, ..., c_{|CL|}\}$) is the set of all class labels and $\mu_k(X)$ represents the compatibility grade of X with the rule r_k. Also R^c is the set of the rules with consequence class label c (i.e. $R^c = \{r_k | h_k = c\}$). It is shown in [18] that, voting can smooth the decision boundaries with higher generalization in classification. This is why, this paper concentrates on weighted vote reasoning.
Due to the role of the certainty factor in (2), it is coined in the following: "weight of the rule". Tuning these weights can improve the performance of the classification system. Indeed, rule-weighting is a special case of tuning fuzzy sets without any change in semantics of linguistic variables. Two samples of possible fuzzy sets are depicted in Figure 1, each one is associated with a linguistic variable i.e. Low and High. In this figure , $\eta\left(x^i, A\right)$ is the membership function of the feature value x^i in the fuzzy set A.

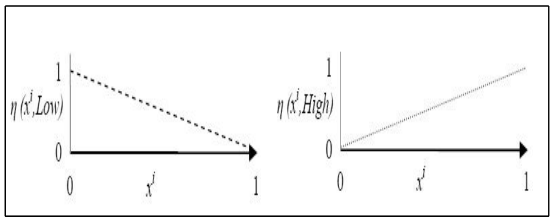

FIGURE 1. Membership Curve of Two Sample Fuzzy Sets
VS. Feature Value x^i: Low and High

3. **A Linear Model of Rule-weighting**

The major aim of rule-weighting is tuning the system in order to achieve a better classification rate on the training data. The set of training patterns is a collection of pairs $\langle X_p, y_p \rangle$, where X_p is the feature vector of p^{th} training pattern and y_p is its associated class label.
In this section, the problem is modeled in the format of a constraint linear programming. Suppose our training data are patterns form two classes +1 and -1 (for a binary dataset). Each pattern X_p is classified correctly by the application of weighted vote reasoning, if (3) is satisfied.

$$\left(\sum_{r_k \in R^{+1}} \omega_k \mu_k (X_p) - \sum_{r_k \in R^{-1}} \omega_k \mu_k (X_p)\right) y_p > 0 \Rightarrow \sum_{r_k \in R} h_k \omega_k \mu_k (X_p) > 0$$
$$h_k = +1 \; or \; -1, \; y_p = +1 \; or \; -1 \tag{3}$$

The dataset will be called separable if it is possible to tune the rule weights such that all training patterns are classified without any error. Otherwise, there is no feasible solution for constraints defined by (3). In many cases, the dataset is not separable, and constraints should be relaxed to have a feasible solution. Therefore, each constraint is permitted to be satisfied with an error, like shown in (4). In this case the goal is to minimize these errors.

$$\begin{aligned} &Minimize \sum_{X_p} \epsilon_p \\ &s.t. \qquad \forall X_p : \qquad \sum_{r_k \in R} h_k \omega_k \mu_k (X_p) \, y_p + \epsilon_p > 0 \& \omega_k, \epsilon_k \geq 0 \end{aligned} \tag{4}$$

whereof ϵ_p is a constraint relaxation parameter and considered directly as error in the objective function to be minimized. The problem (4) suffers from two drawbacks:

(1) It is trivial that $\omega_k = 0$ is the optimal solution of (4), not being desired here.
(2) It is desired to minimize just experimental risk on training data which may lead to decrease the generalization performance

To overcome these disadvantages, a margin-based model has been proposed in this paper. To make some comparisons between the proposed approach of rule-weighting and the well-known margin-based classifier SVM, a brief explanation of SVM model is presented in the next section.

4. **Support Vector Machine Classifier(SVM)**

SVM [32, 7] is a binary classifier that tends to classify patterns of two classes labeled by -1 and +1, utilizing a discriminant function. In the case of simple SVM, this function *f(X)* is linear respect to feature values of pattern X, forming a hyper-plane in the feature space according to (5).

$$f(X) = sgn\left(W^T X - b\right) \tag{5}$$

where, W is the perpendicular vector of discriminant hyper-plane biased with scalar value b. This hyper-plane divides the feature space into two parts to classify the patterns as +1 or -1. Also, $sgn(.)$ is the sign function that returns +1 and -1 for positive and negative values, respectively.

4.1. **Separable Datasets.** The aim of SVM is to find a hyper-plane with maximum symmetric margin on both sides, so that no training pattern is located in this margin as depicted in Figure 2 . Increasing the margin induces increase in generalization potential of the classifier. The signed Euclidian distance of a pattern X_p from the hyper-plane defined by W and b can be computed by (6) as $d^{W,b}(X_p)$.

$$d^{W,b}(X_p) = \left(W^T X - b\right) / \|W\| \tag{6}$$

of which $|W|$ is the 2-norm of W. Regardless of details described in [7], maximizing the margin is equal to minimizing $|W|$. Hence, the problem is formulated in (7).

$$\begin{array}{ll} Minimize & \frac{1}{2}W^2 \\ s.t. & \\ \forall X_p: & \left(W^T X_p - b\right) y_p \geq 1 \end{array} \tag{7}$$

4.2. **Non-separable Datasets.** Usually datasets are not separable into two disjoint groups of classes by a hyper-plane, as depicted in Figure 2. In these cases, some patterns are located in the margin or even misclassified. For the purpose of overcoming this challenge, some errors will be attributed to patterns if they enter to the margin. Consequently, (7) is converted into (8) as expressed in [7].

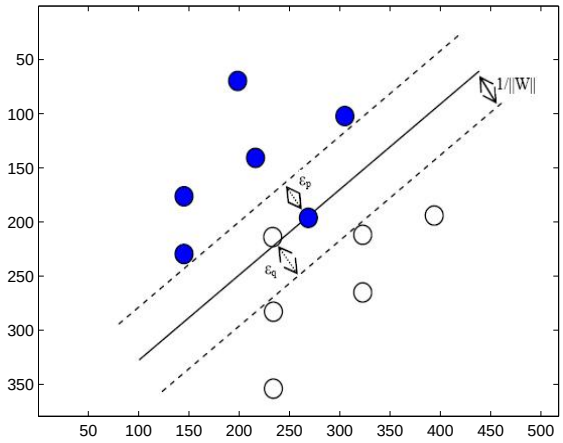

FIGURE 2. SVM Classifier with Maximum Margin for Inseparable Data Points. One of the Solid and One of the Transparent Patterns are (each) Located in the Margin with Errors ϵ_p and ϵ_q, Respectively

$$\begin{array}{ll} Minimize & \frac{1}{2}W^2 + C\sum_p \epsilon_p \\ s.t. & \\ \forall X_p: & \left(W^T X_p - b\right) y_p \geq 1 - \epsilon_p \\ & \epsilon_p \geq 0 \end{array} \tag{8}$$

where ϵ_p is the error associated with X_p and C is a regularization coefficient. Increasing C reduces the importance of generalization although finding a suitable C is itself a challenge.

4.3. **Non-linear SVM Classifiers.** Linear SVM classifiers are extended to nonlinear ones by transferring data patterns into high dimension with nonlinear functions. After this nonlinear transfer, new data patterns are discriminated by a hyper-plane as similar as previous subsection. By this conversion, data points in the higher dimension are likely to be more linear separable with less experimental error on training data; whereas the linear separator function in high dimensional space forms a nonlinear separation in the original space.
Assume $\phi(X)$ is a function transferring X to a high dimensional space. Also consider a special case of SVM which the discriminant hyper-plane is forced to cross the origin (it has no bias i.e. $b = 0$). Having unbiased SVM on $\phi(X)$, (8) is reformulated [7] as (9).

$$\begin{array}{ll} Minimize & \frac{1}{2}W^2 + C\sum_p \epsilon_p \\ s.t. & \\ \forall X_p: & \left(W^T\phi(X_p) - b\right) y_p \geq 1 - \epsilon_p \\ & \epsilon_p \geq 0 \end{array} \tag{9}$$

5. Proposed Quadratic Margin-based Rule-weighting

Increasing the size of the margin in order to improve the generalization, is inspired by SVM to be utilized in our rule-weighting process. In such an approach, the aim would be to tune rule weights so that training patterns could be classified correctly as much as possible. Moreover, it is desired that minimum distance of the patterns to the decision boundaries is maximized.
Decision boundary is a subspace which separates decision spaces of classes +1 and -1. Decision boundary in SVM is the discriminant function whereas, in FRBCSes, decision boundary H is a set of points which have equal voted compatibility grades (strength) with class -1 and +1 in (2), as demonstrated in (10). Here, the strength of the correct class minus strength of the opposite class on each pattern is considered as its distance from the decision boundary. Hence, discriminant function $\chi(X)$ and distance function $\delta^{\Psi}(X)$ are defined as shown in (10).

$$\begin{array}{l} \chi(X) = sgn\left(\sum_{r_k \in R} h_k \omega_k \mu_k(X)\right) \\ \delta^{\Psi}(X) = \frac{\chi(X)}{\|\Psi\|} \end{array} \tag{10}$$

according to which ψ is the vector of rule weights $\psi = \left[\omega_1, \omega_2, \ldots, \omega_{|R|}\right]$. Also, $\delta^{\psi}(X)$ is the distance of pattern X from decision boundary. Considering definitions of distances in (10) and (6), both of them are normalized by the norm of vectors ψ and W, respectively. Also, decision boundary ($\chi(X) = 0$) is a linear function of compatibility grades of points with each rule. This linearity motivates us to define a transfer function as following.
Assume each data pattern X is transferred by $M(X)$ to a new feature space such that it is represented by its signed compatibility grades with the whole rules in the rule-base. From now on, this new feature space is called *compatibility grade space*. $M(X)$ is defined in (11):

$$M(X)^T = \left[h_1\mu_1(X), \ldots, h_k\mu_k(X), \ldots, h_{|R|}\mu_{|R|}(X)\right] \tag{11}$$

Considering the transfer function (11), decision boundary H in (10) can be written as (12).

$$H : \chi(X) = \Psi^T M(X) = 0 \tag{12}$$

After this transfer, decision boundary is a discriminant hyper-plane crossing the origin (having no bias). The Figure 3 depicts decision boundaries of the rules specified in Table 1, for both original feature space and compatibility grade space. The rules of Table 1 are based on sample fuzzy sets proposed in Figure 1.

As shown in Figure 3, decision boundary in compatibility grade space is a linear func-

Rule No.	Rule Description
1	If x^1 is High then class +1 with $\omega_1 = 0.5$
2	If x^1 is Low and x^2 is High then class -1 with $\omega^2 = 1$

TABLE 1. Two Sample Fuzzy Rules

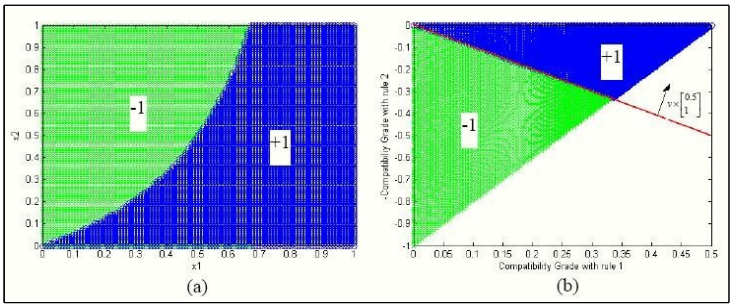

FIGURE 3. Decision Spaces of Each Rule Specified in Table 1: (a) in Original Feature Space. (b) in Compatibility Grade Space

tion of feature values even if it would not be linear in the original space. In addition, the perpendicular vector of discriminant hyper-plane in compatibility grade space (3.b) is the same as the vector of rule weights utilized in original feature space.

Inversely, this transfer function can be utilized to weighting some predefined fuzzy rules. First, transfer the training patterns to the compatibility grade space, and then, find a proper discriminant hyper-plane in compatibility grade space (by any linear classifier i.e. SVM here). Perpendicular vector of this hyper-plane can be finally used as fuzzy rule weights. It should be underlined that this discriminant hyper-plane is forced to contain the origin. Hence, the rule weighting problem can be presented as (13).

$$\begin{array}{ll} Minimize & \frac{1}{2}\Psi^2 + C\sum_p \epsilon_p s.t. \\ \forall X_p : & \Psi^T M(X_p) y_p \geq 1 - \epsilon_p \\ & \epsilon_p \geq 0 \\ \forall r_k \in R : & \omega_k \geq 0 \end{array} \tag{13}$$

This problem becomes so much similar to (9) by equalizing ψ and $M(X)$ to W and $\phi(X)$, respectively. There is a significant difference:

- Although w_k in SVM plays the same roles as ω_k in our proposed rule-weighing method,ω_k should be non-negative as a predefined condition in state of the art.

It should be commented that, in (13), the same as SVM, the margin is to be maximized.

6. **Rule-weighting Kernel for SVM**

Until now, all optimization problems are presented in primal form but, investigation of dual forms are useful from some perspectives e.g. process and memory space complexities

or similarity of different optimization problems. The dual problem of unbiased SVM with transfer function $\phi(X)$ which is presented in (9), is shown in (14).

$$\begin{array}{ll} Maximize & -\frac{1}{2}\sum_p \sum_q \alpha_p \alpha_q y_p y_q \phi(X_p)^T \phi(X_q) + \sum_p \alpha_p \\ s.t. & \\ & W = \sum_p \alpha_p y_p \phi(X_p) \\ & 0 \leq \alpha_p \leq C \end{array} \tag{14}$$

Considering equality in (14), a new representation of decision function in (5) is shown in (15).

$$f(X) = sgn\left(\sum_p \alpha_p y_p \phi(X_p)^T \phi(X_p)\right) \tag{15}$$

Based on (15), it is enough to store only nonzero ? values called support vectors in SVM. As demonstrated, both (14) and (15) use dot product of two data patterns after transferring to new feature space by $\phi(X)$. This is the most important reason to define a Kernel $K\langle Xp, Xq\rangle$, as presented in (16).

$$K\langle X_p, X_q\rangle = \phi(X_p)^T \phi(X_q) \tag{16}$$

Similarities of the proposed method of rule weighting and SVM classifier can be touched in dual forms more than primal forms as defined by (17).

$$\begin{array}{ll} Maximize & -\frac{1}{2}\sum_p \sum_q \alpha_p \alpha_q y_p y_q M(X_p)^T M(X_q) + \sum_p \alpha_p - G \\ s.t. & \\ & G = \sum_{r_k \in R}\left(\beta_k \sum_p \alpha_p y_p h_k \mu_k(X_p) + \frac{\beta_k^2}{2}\right) \\ & 0 \leq \alpha_p \leq C,\ \beta_k \geq 0 \end{array} \tag{17}$$

Note that in this equation, G is a cost function which is the result of the constraint $\omega_k \geq 0$. Without this non-negativity inequality, G and β_k can be totally ignored and (17) will be equal to (14). In this paper, it is claimed that the proposed rule-weighting is a special case of unbiased SVM classifiers with a novel kernel defined in (18) and, of course, some conditions ($\omega_k \geq 0$).

$$K^*\langle X_p, X_q\rangle = M(X_p)^T M(X_q) = \sum_{r_k \in R} \mu_k(X_p)\mu_k(X_q) \tag{18}$$

We make it a convention now to call the SVM classifier with this novel kernel "Fuzzy Kernel SVM" (FKSVM). Considering some conditions, the proposed rule weighting method is the same as FKSVM:

(1) FKSVM would be unbiased implying that b would be set to zero
(2) There would be no restriction on rule weights (rule weights can be negative)

But in this paper, it is assumed that the weights are constrained to be nonnegative. In the next section, a multi-class FRBCS and associated rule-weighting method are presented.

7. Multi-class Rule-weighting

SVM classifiers are primarily proposed for binary class datasets in [7]. But so far, too many approaches are proposed to extend SVM classifiers for multi-class datasets [13]. Initially, some general methods for multi-class classification using binary classifiers are proposed such as "one vs. rest" and "one vs. one" [13]. Weston and Watkins [33] propose

a multi-class SVM, however, which optimizes a single function in order to adjust optimal hyper-planes. The primal problem of multi-class SVM has been shown in (19).

$$\begin{array}{ll} Minimize & \frac{1}{2}\sum_{c_i\in CL}\left(W^{c_i}\right)^2 + C\sum_{X_p}\sum_{c_i\in CL-y_p}\epsilon_p^{c_i} \\ s.t. & \\ \forall X_p: & \\ \forall c_i\in CL-y_p: & \left[\left(W^{y_p}\right)^T\phi(X_p)-b^{y_p}\right]-\left[\left(W^{c_i}\right)^T\phi(X_p)-b^{c_i}\right]\geq 2-\epsilon_p^{c_i} \\ & \epsilon_p^{c_i}\geq 0 \end{array} \tag{19}$$

whereupon W^c and b^c are, respectively, the perpendicular vector and the bias of the hyper-plane discriminating class c from the rest of classes. By this optimization, the classification function of X is defined as (20).

$$g(X) = arg\max_{c_i\in CL}\left(\left(W^{c_i}\right)^T\phi(X_p)-b^{c_i}\right) \tag{20}$$

Where, $g(X)$ is the class label predicted for pattern X. For binary datasets, this optimization problem and associated classification function is similar to binary SVM [33].
From another perspective, FRBCSs are not limited to binary datasets as can be concluded from section 2. But the rule-weighting proposed in sections 5 and 6 is a model just applicable to binary datasets. For multi-class datasets, each pattern should be voted by rules of correct class more than accumulated votes of associated rules for each one of other classes. Yet, minimizing 2^{nd} norm of the vector of rule weights, increases the margin and some patterns may also have some errors. Therefore, (13) is extended to (21) for multi-class labels.

$$\begin{array}{ll} Minimize & \frac{1}{2}\Psi^2 + C\sum_{p}\sum_{c_i\in CL-y_p}\epsilon_p^{c_i} \\ s.t. & \\ \forall X_p, c_i\in CL-y_p: & \sum_{r_k\in R^{y_p}}\omega_k\mu_k(X_p)-\sum_{r_k\in R^{c_i}}\omega_k\mu_k(X_p)\geq 1-\epsilon_p^{c_i} \\ & \epsilon_p^{c_i}\geq 0 \\ \forall r_k\in R: & \omega_k\geq 0 \end{array} \tag{21}$$

By defining $\Psi^c = [\omega_k|h_k = c]$, (21) can be written similar to multi-class SVM problem (19). Although there are some differences between Fuzzy Kernel SVM and our proposed rule weighting (especially in having no bias and no negative weights), from now on, we call this method of rule weighting as FKSVM.

8. **Rule Reduction**

Decreasing the number of the rules is not the direct goal of a rule weighting method but it may be achived as a side effect by setting some weights to zero. For example, IGARW follows a procedure to weighting the rules such that both of irrelevant and redundant rules can be removed due to having zero weights. In this paper, rule reduction is never considered directly in the objective function. But since accuracy of classification on training data is desired to be maximized, the weights of irrelevant rules are expected to be set to zero after minimizing $\|W\|$.
As a general trick in many rule weighting methods, the rules with weights less than a threshold can be removed. In this paper, after the process of weighting the rules, all of them are sorted according to their weights. Then a threshold for removing the low weighted rules will be chosen such that classification rate on training data would be maximized. The pseudo-code of proposed rule reduction is shown in Table 2.
It experimentally seems that, removing low weighted rules by this procedure, not only

results in a more compact rule-base but also decreases the complexity of the classification system which leads to have more generalization capability in classifying unseen data patterns [17].

function RULE_REDUCTION(R,Ψ)
    ▷ The fuzzy rules $R = \{r_1, r_2, \ldots, r_{|}R|\}$ ▷ and their associated weight $\Psi = \{\omega_1, \omega_2, \ldots, \omega|R|\}$
    $R^s \leftarrow \{r_{a_1}, r_{a_2}, \ldots, r_{a_{|R|}}\}$ where $\forall i < j \Rightarrow \omega_{a_i} \leq \omega_{a_j \neq a_i}$ ▷ Sort all the rules according to their weights in ascending order as R^s.
    $\Psi' \leftarrow \Psi$, $bestIndex \leftarrow 0$, $bestAccuracy \leftarrow acc(\Psi)$
    for $k = 1 \rightarrow |R|$ **do**
        $\omega_{a_k} \leftarrow 0$
        $tempAccuracy \leftarrow acc(\Psi)$
        if $tempAccuracy \geq bestAccuracy$ **then**
            $bestAccuracy \leftarrow tempAccuracy$
            $bestIndex \leftarrow k$
        end if
    end for
    $\Psi \leftarrow \Psi'$
    $\forall k \leq bestIndex : \ \omega_{a_k} \leftarrow 0$
    return Ψ
end function

TABLE 2. Pseudo-Code of Rule Reduction

9. **Experimental Results**

In this paper, 13 datasets from UCI repository, which are described in Table 3, are considered for classification tests. These datasets are selected from different cases with various numbers of class labels, features, patterns and variety in their complexities. Any missing value is replaced by zero in these datasets if there is. Although there are many methods to generate a good rule base [28, 27], since generation of the rule-base is not important, it is assumed that a pre-generated rule-base has been provided similar to [36, 25]. After weighting these rules, the ones with zero weights are removed from the rule base.

In following experiments, each rule-base contains exactly 30 rules from each class. To measure the power of each method to learn training data, "Full train-Full Test" validation is shown in Table 4. In this validation, all training data is used to train and all of them are tested by the trained classifier. Here the classification rates on training data, using 7 various methods of rule weighting, are compared with the case of having equal weights for all the rules (no weighting). These 7 methods are 4 greedy methods of rule-weighting proposed by Ishibuchi [24], called G1-4, a reward and punishment procedure (R& P) [31], our Iterative Greedy Accuracy-based Rule-Weighting (IGARW) [36] and the method of this paper (FKSVM). Although IGARW is an iterative method, due to the danger of over-fitting on training data, this method has run for only one iteration, in this paper. FKSVM has no parameter except of C which is 10 here.

As shown in Table 4, G1-4 cannot improve the classification rate even on training data. The method R&P is better than G1-4 but, on the one hand, it is highly time consuming, and on the other hand, it has been compared with IGARW in [36]. It was shown that IGARW improves the classification rate more than R&P. Indeed, none of G1-4 and R&P is aimed to improve an explicit objective function.

Dataset	Pattern No.	Feature No.	Class No.	Dataset	Pattern No.	Feature No.	Class No.
Haberman	306	3	2	Iris	150	4	3
Pima	768	8	2	Wine	178	13	3
Bcancer	684	10	2	Lung Cancer	32	56	3
WDBC5	569	30	2				
WPBC	198	33	2	Heart	287	13	5
Ionosphere	351	34	2	Glass	214	9	6
Sonar	208	60	2	Image	210	19	7

TABLE 3. List of UCI Datasets Used in This Paper

Dataset	No weight	G1	G2	G3	G4	R&P	IGARW	FKSVM
Haberman	73.86	73.53	73.86	=G2	=G2	75.49	79.41	78.10
Pima	71.61	72.14	72.40	=G2	=G2	74.87	77.21	77.34
BCancer	94.74	94.59	94.44	=G2	=G2	94.74	97.37	96.64
WDBC5	91.74	92.27	92.44	=G2	=G2	94.02	95.08	95.25
WPBC	77.27	77.27	76.77	=G2	=G2	83.84	84.34	85.35
Ionosphere	80.91	82.05	82.05	=G2	=G2	86.61	88.89	88.32
Sonar	75.96	76.44	75.96	=G2	=G2	77.40	80.29	78.85
Iris	96.00	96.00	96.00	96.00	96.00	95.33	98.00	96.00
Wine	96.07	96.07	96.07	96.07	96.07	95.51	98.88	98.31
Lung Cancer	84.38	84.38	87.5	84.38	84.38	93.75	93.75	93.75
Heart	60.63	60.63	60.63	60.63	60.28	64.11	67.25	63.76
Glass	64.49	64.49	64.49	65.42	64.49	65.89	70.56	68.69
Image	80.00	80.00	80.00	80.00	80.00	80.48	81.43	80.48
Average	80.59	80.76	80.97	80.80	80.75	83.23	85.57	84.68

TABLE 4. Classification Rate on Training Data After Rule-Weighting by G1-4 [24], R&P [31], IGARW [36] and FKSVM

In Table 5, full train- full test classification rate and the number of remained rules (rules with nonzero weights) for IGARW and FKSVM are compared. The proposed rule pruning in Table 2 is also applied after FKSVM and the results are reported in Table 5. As can be seen, it seems that the proposed rule pruning and IGARW, compete with each other to reduce the number of the rules. Based on our experiments, this pruning increases the generalization of FKSVM due to decreasing the system complexity. The classification rate of FKSVM is often less than IGARW on training data. This is due to considering generalization in FKSVM such that, in cases that IGARW over-fits on training data, the proposed rule weighting improves the accuracy of classification on unseen data.

In order to have a more statistical comparison of FKSVM and other methods, in Table 6, these methods are compared by Ten Fold-Ten Cross Validation (10CV). Here, C is found by cross-validation from the set of values {100, 10, 1}. In our experiments it is usually set to 10.

In this table, mean and standard deviation of classification rates of all 100 runs are reported for each pair of dataset and weighting method. In addition, for each dataset, paired T-test is used to compute statistical significance of the hypothesis that FKSVM is better than

Dataset	IGARW						FKSVM			
	1 Iteration		2 Iteration		3 Iteration		No Rule Reduction		With Rule Reduction	
	Rate	Rule No.	Rate	Rule No.	Rate	Rule No.	Rate	Rule No.	Rate	Rule No.
Haberman	79.41	29	79.74	27	80.39	28	78.10	60	78.76	45
Pima	77.21	34	77.86	28	77.99	27	77.34	60	77.73	3
Cancer	97.37	20	97.37	19	97.37	19	96.64	60	97.22	8
WDBC5	95.08	52	95.25	46	95.25	45	95.25	60	95.96	15
WPBC	84.34	32	84.34	31	84.34	31	85.35	60	87.37	45
Ionosphere	88.89	24	89.17	18	89.17	13	88.32	60	88.32	28
Sonar	80.29	48	80.29	25	80.29	17	78.85	60	80.77	23
Iris	98.00	37	98.00	37	98.00	37	96.00	90	98.00	7
Wine	98.88	24	98.88	15	98.88	13	98.31	90	99.44	26
Lung Cancer	93.75	12	93.75	9	93.75	7	93.75	90	96.88	34
Heart	67.25	25	67.60	22	67.60	21	63.76	150	63.76	58
Glass	70.56	32	71.50	25	71.50	23	68.69	180	69.63	104
Image	81.43	9	81.43	7	81.43	7	80.48	210	80.48	137

TABLE 5. Comparison Between IGARW and FKSVM with Rule Reduction According to Classification Rate on Training Data and the Number of Remained Rules

Datasets	No Weight	G1	G2	G3	G4	R&P	IGARW with one iteration	FKSVM with Rule reduction	T-test
Haberman	71.80	71.96	73.00	=G2	=G2	73.54	72.43± 8.23	76.25± 7.09	+
Pima	71.38	71.59	72.19	=G2	=G2	74.76	74.97±4.21	77.13±3.83	+
Cancer	94.23	94.11	94.05	=G2	=G2	94.43	96.74±1.95	96.68±1.85	-
WDBC5	91.41	91.90	92.43	=G2	=G2	93.78	93.95±3.25	95.32±2.76	+
WPBC	76.30	76.31	76.29	=G2	=G2	77.83	78.85±8.31	82.19±7.57	+
Ionosphere	80.99	81.43	82.00	=G2	=G2	85.41	85.59±6.10	87.36±5.60	+
Sonar	74.50	73.59	72.88	=G2	=G2	75.75	77.81±9.48	80.45±8.19	+
Iris	95.93	95.93	96.00	96.00	96.00	95.47	95.13±5.32	96.33±4.47	+
Wine	94.99	94.84	95.02	95.29	94.96	95.10	96.13±4.58	98.04±3.01	+
Lung Cancer	49.33	54.33	52.42	49.50	51.50	57.16	54.5±24.82	66.58±19.81	+
Heart	56.83	56.47	56.28	56.48	56.38	53.86	54.01±8.50	57.94±8.60	+
Glass	54.71	57.58	56.07	56.45	57.37	58.77	58.78±10.89	62.31±10.14	+
Image	79.43	80.19	80.05	80.10	80.19	80.00	79.48±9.32	81.00±8.61	-

TABLE 6. Comparison of FKSVM with Other Rule Weighting Methods Using 10CV

IGARW (for significance level of 95%). Not only, in none of cases, FKSVM is significantly worse than IGARW, but also in the most of cases, which are marked by (+), FKSVM is significantly better than IGARW. The datasets, for which the hypothesis is rejected, are marked by (-).

Paired T-test can be used in order to statistically comparing two methods over only one dataset. Also, T-test have been used in the literature review for comparing two classifiers on multiple datasets, but it is not recommended in recent researches [8, 11] for this task. In this paper, two nonparametric methods are used, to compare performance of two classifiers over multiple datasets [8]: "Wilcoxon signed-ranks test" and "Counts of wins, losses and ties: Sign test". Using both of these tests, FKSVM is significantly

better than other methods for significance level of 95%, based on the classification results reported in Table 6.

10. **Conclusion**

In this paper, a novel quadratic model is proposed to tune the weights of fuzzy classification rules. Improving the classification rate on training data and increasing the generalization capability by a margin based approach, both of them, are considered explicitly in the objective function. A fuzzy kernel for Support Vector Machine (SVM) classifiers has been introduced here, and equality of the proposed method of rule-weighting, in special conditions, with nonbiased Fuzzy Kernel SVM (FKSVM) has been investigated here in order to use the deep background theory of SVM. In addition, experimental results show a significant improvement on classification rate of Fuzzy Rule-Based Classifier Systems (FRBCS) after weighting with FKSVM, in comparison with other methods of rule-weighting such as fast and powerful Iterative Greedy Accuracy-based Rule-Weighting (IGARW). Moreover, not only the proposed method has been extended for multi-class datasets, but also some other approaches such as having negative rules or bias rule in FRBCSes are touched for future works. Inspecting this quadratic programming model for rule-weighting with Single Winner reasoning, tracing the global solution by changing the value of C, local training, and reducing the complexity of this model respect to the special properties of the rule-weighting, are some of works which can be followed in future.

References

[1] S. M. Chen, *Generating weighted fuzzy rules from relational database systems for estimating values using genetic algorithms*, IEEE Trans. on Fuzzy Systems, **11(4)** (2003), 495–506.

[2] S. M. Chen, *A new weighted fuzzy rule interpolation method based on GA-based weights-learning techniques*, proced. of ICMLC, **5** (2010), 2705-2711.

[3] S. M. Chen, *Weighted fuzzy rule interpolation based on GA-based weight-learning techniques*, IEEE Trans. on Fuzzy Systems, **19(4)** (2011),729–744.

[4] Z. Chi, H. Yan and T. Pham, *Fuzzy algorithms: with applications to image processing and pattern recognition*, World Scientific, Singapore, 1996.

[5] O. Cordon, M. J. Del Jesus and F. Herrera, *A proposal on reasoning methods in fuzzy rule-based classification systems*, Internat. J. Approx. Reason, **20** (1999), 21–45.

[6] O. Cordon, F. Gomide, F. Herrera, F. Hoffmann and L. Magdalena, *Ten years of genetic fuzzy systems: current framework and new trends*, Fuzzy Sets and Systems, **141** (2004), 5–31.

[7] C. Cortes and V. Vapnik, *Support vector networks*, Machine Learning, **20** (2004).

[8] J. Demsar, *Statistical comparisons of classifiers over multiple data sets*, Journal of Machine Learning Research, **7** (2006), 1–30.

[9] G. Forman and I. Cohen, *Learning from Little: Comparison of Classifiers Given Little Training*, Springer-Verlag, PKDD 2004, (2004), 161–172.

[10] L. Fu, *Rule generation from neural networks*, IEEE Transaction on systems, Man, and Cybernetics, **24(8)** (1994).

[11] S. Garcia and F. Herrera, *An extension on statistical comparison of classifiers over multiple data sets for all pair wise comparisons*, Journal of Machine Learning Research, **9** (2008), 2677–2694.

[12] H. B. Gurocak and A. de Sam Lazaro, *A fine tuning method for fuzzy logic rule bases*, Elsevier, Fuzzy Sets and Systems, **67** (1994), 147–161.

[13] C. W. Hsu and C. J. Lin, *A comparison of methods for multiclass support vector machines*, IEEE Transaction on neural networks, **13(2)** (2002).

[14] Q. Hu, P. Zhu, Y. Yang and D. Yu, *Large-margin nearest neighbor classifiers via sample weight learning*, Neurocomputing,**74(4)** (2011), 656–660.

[15] H. Ishibuchi, T. Murata and I. B. Turksen, *Single-objective and two-objective genetic algorithms for selecting linguistic rules for pattern classification problems*, Fuzzy Sets and Systems, **89(2)** (1997), 135–150.

[16] H. Ishibuchi and T. Nakashima, *Effect of Rule Weights in Fuzzy Rule-based Classification Systems*, IEEE Transactions on Fuzzy Systems, **9(4)** (2001), 506–515.

[17] H. Ishibuchi, T. Nakashima and T. Morisawa, *Simple fuzzy rule-based classification systems perform well on commonly used real-world data sets Fuzzy Information Processing Society*, NAFIPS '97., 1997 Annual Meeting of the North American, (1997), 251–256.

[18] H. Ishibuchi, T. Nakashima and T. Morisawa, *Voting in Fuzzy Rule-based Systems for Pattern Classification Problems*, Fuzzy Sets and Systems, **103(2)** (1999), 223–238.

[19] H. Ishibuchi, T. Nakashima and M. Nii, *Classification and modeling with linguistic information granules: advanced approaches to linguistic data mining*, Springer Verlag, 2004.

[20] H. Ishibuchi and M. Nii, *Techniques and applications of neural networks for fuzzy rule approximation*, Fuzzy Theory Systems, (1999), 1491–1519.

[21] H. Ishibuchi and Y. Nojima, *Analysis of interpretability-accuracy tradeoff of fuzzy systems by multiobjective fuzzy genetics-based machine learning*, International Journal of Approximate Reasoning, **44(1)** (2007), 4–31.

[22] H. Ishibuchi, K. Nozaki and H. Tanaka, *Distributed representation of fuzzy rules and its application to pattern classification*, Fuzzy Sets and Systems, **52(1)** (1992), 21–32.

[23] H. Ishibuchi, K. Nozaki, N. Yamamoto and H. Tanaka, *Selecting fuzzy if-then rules for classification problems using genetic algorithms*, IEEE Transactions on Fuzzy Systems, **3(3)** (1995), 260–270.

[24] H. Ishibuchi and T. Yamamoto, *Rule weight specification in fuzzy rule-based classification systems*, IEEE Trans. on Fuzzy Systems, **13(4)** (2005), 428–435.

[25] H. Ishibuchi and T. Yamamoto, *Fuzzy rule selection by multi-objective genetic local search algorithms and rule evaluation measures in data mining*, Fuzzy Sets and Systems, **141(1)** (2004), 59–88.

[26] J. Langford, *Tutorial on practical prediction theory for classification*, Journal of Machine Learning Research, **6** (2005), 273–306.

[27] E. G. Mansoori, M. J. Zolghadri and S. D. Katebi, *Using distribution of data to enhance performance of fuzzy classification systems* Iranian Journal of Fuzzy Systems, **4(1)** (2007), 21–36.

[28] E. G. Mansoori, M. J. Zolghadri, S. D. Katebi and H. Mohabatkar, *Generating fuzzy rules for protein classification*, Iranian Journal of Fuzzy Systems, **5(2)** (2008), 21–33.

[29] R. Mikut, J. Jakel and L. Groll, *Interpretability issues in data-based learning of fuzzy systems*, Elsevier, Fuzzy Sets and Systems, **150** (2005), 179–197.

[30] T. Nakashima, G. Schaefer, Y. Yokota and H. Ishibuchi, *A weighted fuzzy classifier and its application to image processing tasks*, Fuzzy Sets and Systems, **158(3)** (2007), 284–294.

[31] K. Nozaki, H. Ishibuchi and H. Tanaka, *Adaptive fuzzy rule-based classification systems*, IEEE Transactions on Fuzzy Systems, **4(3)** (1996), 238–250.

[32] V. Vapnik, *The nature of statistical learning theory*, Springer Verlag, New York, 1995.

[33] J. Weston and C. Watkins, *Support vector machines for multi-class pattern recognition*, ESANN'1999 proceedings - European Symposium on Artificial Neural Networks, Bruges (Belgium), isbn:2-600049-9-X, (1999), 219–224.

[34] L. Yu and J. Xiao, *Trade-off between accuracy and interpretability: experience-oriented fuzzy modeling via reduced-set vectors*, Elsevier, Computers and Mathematics with Applications, **57** (2009), 885–895.

[35] M. J. Zolghadri and E. G. Mansoori, *Weighting fuzzy classification rules using receiver operating characteristics (ROC) analysis*, Information Sciences, **177(11)** (2007), 2307–2296.

[36] M. J. Zolghadri and M. Taheri, *A proposed method for learning rule weights in fuzzy rule-based classification systems*, Fuzzy Sets and Systems, **159** (2008), 449–459.

Mohammad Taheri*, Department of Computer Science & Engineering & IT, Shiraz University, Shiraz, Fars, Iran
E-mail address: mtaheri@cse.shirazu.ac.ir

Hamid Azad, Department of Electrical Engineering, Science & Research Branch, Islamic Azad University, Marvdasht, Fars, Iran
E-mail address: azad@shirazu.ac.ir

Koorush Ziarati, Department of Computer Science & Engineering & IT, Shiraz University, Shiraz, Fars, Iran
E-mail address: koorush@ziarati.net

Reza Sanaye, Department of Computer Science & Engineering & IT, Shiraz University, Shiraz, Fars, Iran
E-mail address: reza_sanaye@cse.shirazu.ac.ir

*Corresponding author