Compact Fuzzy Rules Induction and Feature Extraction Using SVM with Particle Swarms for Breast Cancer Treatments

Shang-Ming Zhou, Robert I. John, Xiao-Ying Wang, Jonathan M. Garibaldi, and Ian O. Ellis

Abstract— **Developing a treatment plan for breast cancer patient is a very complex process. In this paper, we propose a scheme of inducing fuzzy rules that characterise breast caner treatment knowledge from data. These fuzzy rules can augment the human experts in the process of medical diagnosis to select optimal treatment for patients. The proposed machine learning scheme applies the particle swarm optimisation technique (PSO) to the construction of an optimal support vector machine (SVM) model for the sake of inducing accurate and parsimonious fuzzy rules and simultaneously reducing input space dimensions, in which a new fittness function that regularises the importance ranks of features with misclassification rate is suggested. The SVM-based fuzzy classifier evades the curse of dimensionality in high-dimensional breast cancer data space in the sense that the number of support vectors, which equals the number of induced fuzzy rules, is not related to the dimensionality. The experiments have shown that not only the classification performance achieved by the proposed fuzzy classifier outperforms the ones achieved by other methods in the literature, but also the input space dimension has been reduced greatly.**

I. Introduction

Breast cancer is a malignant neoplasm most frequently occurring to girls and women in western countries [1], [2]. Among women worldwide, breast cancer is the most common cause of cancer death [3]. The outcomes for patients depend critically on the timely diagnosis and quality of treatment given. It is widely accepted that developing a treatment plan for a breast cancer patient is a very complex process, many factors from patient and medical diagnosis need to be considered, like age, lymph node stage, tumour size, tumour grade, patient's preferences, and many more besides. So breast cancer treatment decision support systems (BCTDSSs) has emerged to augment, not replace, the doctors in the process of medical diagnosis to select optimal treatment for breast cancer [4]. The heart of BCTDSS technology is a knowledge base that works as an inductive engine to diagnose patients given breast cancer states. On the other hand, real world medical decision making is innately uncertain due to inherent subjectivity, vagueness in the articulation of human opinions, and inaccurate and imprecise medical measurements etc.. Hence, a promising framework of developing a knowledge base for BCTDSS is to model the uncertainty in the decision making process utilising a fuzzy logic methodology. The advantage of fuzzy logic systems is that the solutions to the problems are casted into fuzzy *if-then* rules with linguistic terms that human operators can understand.

Currently, one widely used strategy for constructing fuzzy logic systems is to induce fuzzy rules from data [5], [6], [7], [8]. This is particularly suitable for breast cancer treatment because. Due to the advanced development of modern information technology, plenty of breast cancer data, which records the historical treatments of breast cancer for patients, is available in hospitals. However, one challenge in inducing fuzzy rules from breast cancer data is that the complexity of the breast cancer treatment process leads to high-dimensional dataset. As a result, fuzzy rule induction would face the curse of dimensionality [9], [10]. The purpose of this paper is to generate an accurate and parsimonious fuzzy rule base from breast cancer data and automatically perform dimensionality reduction and extract important features in one model structure for determinant of therapy.

In traditional statistical system modelling, it is known that the support vector machine (SVM) approach [11], [12] has been widely regarded as the state-of-the-art parsimonious modelling technique for regression and classification with successful applications in many domains. Recently, some research has indicated that the advantage of the SVM in yielding parsimonious solutions can be exploited in fuzzy system modelling, so that an accurate and compact fuzzy rule base can be induced [13], [14]. What's more, as prototype-based classifiers, the SVM-based fuzzy classifiers evade the curse of dimensionality in high-dimensional space in the sense that the number of support vectors, which equals the number of induced fuzzy rules, is not related to the dimensionality. However, for most kernel machine models, how to select optimal parameters for the kernel functions remains open [15], [16], [17]. A gradient based algorithm with radius-margin as objective function has been proposed for L2-SVM-based fuzzy classifier [14] to train hyper-parameters from data, but optimal hyper-parameters selection is still an open problem in L1-SVM based fuzzy classification [13], because radius-margin bound does not hold in L1-SVM. Hence in this paper, we propose a new method of constructing a parsimonious fuzzy classifier by (L1-) SVM technique, in which the particle swarms optimisation (PSO) method [18], [19], [20] is used to train the hyper-parameters for SVM, and at the same time, automatically choose the number of fuzzy

Shang-Ming Zhou and Robert I. John are with the Centre for Computational Intelligence, School of Computing, De Montfort University, Leicester, LE1 9BH, UK (email: smzhou@ieee.org; rij@dmu.ac.uk)

Xiao-Ying Wang and Jonathan M. Garibaldi are with the School of Computer Science, University of Nottingham, Nottingham, NG8 1BB, UK (email: xyw@cs.nott.ac.uk; jmg@cs.nott.ac.uk),

Ian O. Ellis is with School of Molecular Medical Sciences, Nottingham University Hospitals and University of Nottingham, Queens Medical Centre, Derby Road, Nottingham, NG7 2UH, UK (email:Ian.Ellis@nottingham.ac.uk).

978-1-4244-1823-7/08/$25.00

rules and identify the important input features.

The PSO, first suggested by Kennedy and Eberhart [18], [19], is a population based stochastic optimization technique that can mimic the sociological behaviours of bird flocks, fish schooling or a group of people. Although PSO is a relatively new optimization approach, it has received much attention in the recent period [20]. The advantage of PSO over gradient based optimisation algorithm is that it can be used in cases where the system function is non-differentiable or no gradient information is available. This is also much faster. In this paper, not only are the fuzzy rules generated optimally from data via the SVM with PSO technique, but also the dimensionality of input space is simultaneously reduced via feature ranking based on a proposed modulator function. To this end,we propose to regularise the importance ranks with the misclassification rate in the fitness for PSO training. As a result, the importance ranks associated with less influential features would be forced to approach to zero while the classifier maintains good classification performance.

The organisation of this paper is as follows. Section 2 describes the breast cancer treatment data collected from hospital. Section 3 proposes scheme of constructing fuzzy classifier and feature extraction by PSO technique. Section 4 evaluates the performance of the proposed scheme with high-dimensional breast cancer treatment data. Then Section 5 concludes the paper.

II. Clinical Data of Breast Cancer Treatment and its Pre-processing

The clinical data about breast cancer treatment used in this paper was collected from Nottingham Breast Institute at Nottingham City Hospital in the UK. This is a post-operative dataset recording the historial breast cancer treatements for the patients who had all undergone certain form of breast cancer operation (e.g. wide local excision, axillary node clearance or sample). The data is comprised of 17 attributes examined on each patient's post-operative visit and a follow up treatment decision. Table I illustrates the attributes involved in the data.

The clinical procedure employed for recording the data is summarised as follows:

1) The attribute information and additional comments related to each patient's treatment are recorded on a form;
2) The forms are discussed during the multi-disciplinary meeting and an agreement of a further course of action is reached;
3) After the meeting, all forms are collected and sent to a data analyst for entry into a computer database.

In the multi-disciplinary meeting, the care plans regarding the recommended course of follow up treatments are normally provided by a multi-disciplinary team from surgeons, pathologists and oncologists for the patients. In recent years, there have been various of life-saving treatment advances against breast cancer, which brings new hope and excitement.

TABLE I
Attrbitues for Breast Cancer Treatment Clinical Samples

Series No.	Attribute Meanings
1	Age
2	Invasive carcinoma size in mm
3	Invasive carcinoma grade
4	Invasive carcinoma type
5	Invasive carcinoma margins in mm
6	Lymph node stage
7	Number of positive lymph node spread when examined by a pathologist
8	Number of axillary lymph nodes taken for pathology examination
9	Number of apical lymph nodes taken for pathology examination
10	Nottingham prognostic index (NPI)
11	Estrogen receptor (ER) test result
12	Ductal carcinoma in situ (DCIS) size in mm
13	DCIS grade
14	DCIS type
15	Vascular invasion examination result
16	DCIS margins in mm
17	Whole tumour size in mm

The menu of treatment choices recommended by the multi-disciplinary team that fight the complex mix of cells in each individual cancer include:

- Radiation therapy/radiotherapy
- Hormone (anti-estrogen) therapy
- Chemotherapy
- Further operation
- Follow up
- Combinations of the above

but in this paper we only focus on the classification of the treatment decisions: *radiotherapy* and *hormone-therapy* according to the data availability.

It is worth noting that some issues arising from this clinical dataset itself make the task of contructing accurate treatment model tougher. First the many factors from patient and medical diagnosis as indicated in the Table I lead to high dimensional data, so it is impractical to induce fuzzy rules in grid partitions by traditional approaches. The data consists of the inputs from a number of different data analysts over a period of twenty-five years, and there is no standardised format for data entry, inconsistencies in data formats often occur, so it is inevitable to induce errors during standardising the data in a common format. Additionally, the treatment decisions lack clear indications in the raw records, they are hidden within some comments, rather than separately recorded in their own data field. Different treatment decisions may be identified through different notations or descriptions, which makes the data pre-processing more complicated and impact model classification accuracy. Moreover, the measurement variability in some attributes leads to the data with lots of noise. What's more, there exist many missing values for some attributes due to medical practice.

In this paper two additional pre-processing steps are conducted as follows:

- To discard the samples with missing attribute values. Among the available attributes, only the three: *age*,

invasive carcinoma type, and *NPI* have the full samples without missing values. After the samples with missing attribute values are removed, 330 samples associated with the treatment decisions- *radiotherapy* and *hormone-therapy* are obtained.

- To initially remove redundant input attributes. In the original clinical dataset, in addition to the 17 attributes illustrated in Table I, the following two attributes:
  - Progesterone receptor (ProRec) test results
  - Number of apical lymph nodes spread (NoALNS) when examined by a pathologist

  were also used, but there are only a few samples whose *progesterone receptor* attribute have values. So in order to obtain plenty of samples as it can be, the attribute-ProRec has to be discarded. As for the NoALNS, after the samples with missing attribute values were removed, all the remaining samples have the same attribute value for the NoALNS, which does not play a role in distinguishing the treatments given the available patients in the dataset, so this attribute will not be used in this paper.

III. The Proposed SVM Based Fuzzy Classifier with Feature Extraction

A. Structure of (L1-)SVM based Fuzzy Classifier

In this paper, we consider the construction of a parsimonious zero-order Takagi-Sugeno (TS) fuzzy model [25] with L fuzzy rules in the following form:

$$R_i : if\ x_1\ is\ A_i^1\ and\ \cdots\ and\ x_n\ is\ A_i^n\ then\ y_i = b_i \quad (1)$$

where $i = 1, 2, \ldots, L$, x_j and y_i are the input and output variables of the ith rule R_i respectively, and A_i^j are the linguistic labels expressed as fuzzy sets with specific semantic meanings of behaviors of the system being modeled, which are characterized by membership functions $A_i^j(x_j)$ generated by expert knowledge or from data, b_i is the consequent parameter of the ith rule. Additionally the following auxiliary rule is added into the rule base as suggested in [13] for the sake of preventing the fuzzy system from breaking inference:

$$R_0 : if\ x_1\ is\ A_0^1\ and\ \cdots\ and\ x_n\ is\ A_0^n\ then\ y_0 = b_0 \quad (2)$$

where A_0^j denotes the domain of x_j and $A_0^j(x_j) \equiv 1$, and $b_0 \in \Re$. So the input space is thoroughly covered by the fuzzy rule "patches", and for any possible input vector, at least one rule should be fired.

Then the overall output of the system is obtained as follows:

$$F(x) = \frac{b_0 + \sum_{i=1}^{L} \tau_i(x) b_i}{1 + \sum_{i=1}^{L} \tau_i(x)} \quad (3)$$

where τ_i is the *firing strength* of the ith rule and usually calculated by an *T-norm* operator. In this paper the *T-norm* operator *product* is used, so

$$\tau_i(x) = \prod_j A_i^j(x_j) \quad (4)$$

The decision function (3) leads to a binary fuzzy classifier defined as follows:

$$f(x) = sgn\left(F(x) + t_f\right) \quad (5)$$

where t_f is a threshold parameter. Without loss of generality, t_f can be assumed to be 0. Since $1 + \sum_{i=1}^{L} \tau_i(x) > 0$, which does not impact the signs of the output (5), so the classifier can be simplified to

$$f(x) = sgn\left(\sum_{i=1}^{L} \tau_i(x) b_i + b_0\right), \quad (6)$$

which is in the form of a SVM classifier. So we only guarantee the $\tau_i(x)$ is a Mercer kernel, then the SVM algorithm can be applied to (6), as a result, a parsimonious model structure will be obtained. Interestingly, Chen and Wang [13] indicated that if the MFs $A_i^j(x_j)$ associated with the same input variable x_j are generated from location transformation of a reference function $a^j(\cdot)$ [26], i.e., $A_i^j(x_j) = a^j\left(x_j - m_i^j\right)$, then the *if-part* in each fuzzy rule defined as the *t-norm* of every variable's MF, i.e., the $\tau_i(x)$, is proven to be a Mercer kernel under the condition that the Fourier transform of the reference function is non-negative [27]. In this paper, we choose the following reference function $a^j(\cdot)$:

$$a^j(r) = e^{-\eta r^2}\ (\eta > 0) \quad (7)$$

whose Fourier transform is non-negative, hence $\tau_i(x) = \tau(x, m_i) = \prod_{j=1}^{n} a^j(x_j - m_i^j)$ is a Mercer kernel, where $m_i = \left(m_i^1, \cdots, m_i^n\right)^T$ is called prototype or kennel centre. The parameter η in the reference functions is the kernel parameter, but the hyper-parameters were manually selected in the modelling scheme used in [13]. However, it is impractical to manually choose different values for hyper-parameters in a high-dimensional input space in order to obtain a classification system with good generalization performance. This paper proposes to use a learning scheme based on PSO technique to automatically update the hyper-parameters, at the same time, the goodness of features is learned optimally from data through a special modular function.

To this end, the input variables are scaled by the following modulator function:

$$\hat{x}_j = x_j \theta_j \quad (8)$$

where $\theta_j \in (0,\ 1)$ indicates the importance of the input variable x_j to the classification task and is defined as

$$\theta_j = 1 - \frac{1}{1 + e^{-\varphi_j}} \quad (9)$$

where $\varphi_j \in \Re$. In the following, the input variables are scaled, but in order to avoid the confusion of notations, we still use x_j rather than $\hat{x}_j$ to represent the input variables, unless otherwise stated. Then a SVM based fuzzy classifier can be expressed as

$$f(x) = sgn\left(\sum_{i=1}^{L} \tau(x, m_i) b_i + b_0\right) \quad (10)$$

where $\tau(x, m_i) = \prod_{j=1}^{n} a^j(x_j - m_i^j) = \prod_{j=1}^{n} e^{-\eta(x_j - m_i^j)^2}$. It can be seen from (22), (4), and (10) that each $\tau(x, m_i)$ is associated with a fuzzy rule. To construct the SVM based fuzzy classifier described by (10), the following parameters should be determined: the number of rules L, prototypes m_i, weights b_i, bias b_0, and scaling parameters θ_j.

Given a dataset $\left\{ x_s^{(l)}, y_s^{(l)} \right\}_{l=1}^{N_s}$ for constructing SVM, where $y_{svm}^{(l)} \in \{-1,\ 1\}$, by solving the quadratic programming problems involved in SVM, one obtains the optimal Langranrian coefficient vector $\alpha_0 = \left[\alpha_0^{(1)}, \cdots, \alpha_0^{(N_s)}\right]^T$, in which there would be many zero coefficients, and only those samples that correspond to non-zero coefficients will play a role in the determination of model parameter values and are called *support vectors*. Let L be the number of non-zero coefficients which are denoted as $\tilde{\alpha}_0^{(i)}$. Then the output of the ith fuzzy rule can be calculated as

$$b_i = \tilde{\alpha}_0^{(i)} \tilde{y}_s^{(i)} \tag{11}$$

where $\tilde{y}_s^{(i)}$, i=1, 2, …, L, are the class labels of the corresponding support vectors. Hence, the non-linear decision function (10) becomes

$$f(x) = sgn\left(\sum_{i=1}^{L} \tau(x, \tilde{x}_s^{(i)}) \tilde{\alpha}_0^{(i)} \tilde{y}_s^{(i)} + b_0 \right) \tag{12}$$

where $\tilde{x}_s^{(i)}$ represent support vectors which will be set as prototypes m_i in fuzzy rule induction, and the bias term b_0 can be computed as follows:

$$b_0 = \frac{1}{L} \sum_{j=1}^{L} \left(\tilde{y}_s^{(j)} - \sum_{i=1}^{L} \tilde{\alpha}_0^{(i)} \tilde{y}_s^{(i)} \tau(\tilde{x}_s^{(j)}, \tilde{x}_s^{(i)}) \right) \tag{13}$$

In the following, a PSO based learning scheme is used to optimally train the hyper-parameters, θ_j, η, and C. C is the regularization parameter penalizing the classification error in SVM.

B. PSO for Automatic Model Selection and Feature Extraction

The PSO is a stochastic optimization technique that exploits a population of individuals to iteratively explore promising regions of multidimensional search space for a global minimum (or maximum). In this context, the population is called a *swarm*, and the individuals are referred to as particles. The particles have memory and are able to keep track of previous best positions and corresponding fitness. The PSO works as follows: each particle represents a potential solution to the optimization task at hand, it is given a random velocity and "flies" through the problem space according to its adaptable velocity. In each iteration, a fitness value representing a quality measure for every particle is calculated, and every particle accelerates towards its own personal best position, which is associated with the best fitness it has achieved so far, as well as towards the best position discovered so far by its neighbours/entire population. Hence, if a particle finds a promising new solution, all the other particles/neighbours will move closer to it and explore the region more thoroughly in the iterative process. Generally, according to the modes of regulating how the "social" information is exchanged among particles, there are three versions of PSO [28]: *individual version, local version,* and *global version.* This paper considers the global PSO only, in which the 'social' knowledge used to drive the movements of particles includes the position of the best particle from the entire swarm.

First we need to encode the particles for solving problem. In this paper, the hyper-parameters to be considered include: $\{\varphi_j\}_{j=1}^{n}$, η and C. Because the parameters η and C are required to be $\eta > 0$ and $C > 0$, so we use the transforms:

$$\beta = \log(\eta), \quad u = \log(C) \tag{14}$$

to meet the requirements, where $\beta \in \Re$, $u \in \Re$. In such a way, each particle will be a group of $(n + 2)$ parameters: $\{\varphi_j\}_{j=1}^{n}$, β and u. Then we need to determine a fitness function f as a quality measure for each particle to evaluate the fitness of solution achieved so far.

For the classification problem in this paper, the SVM model itself is constructed by the Vapnik-Chervonenkis (VC) dimension theory following the principle of structural risk minimisation [11], [12], all the patterns are fed to the SVM whose hyper-parameters are determined by the particle, we get the outputs and compare them with the standard outputs. So the misclassification rate can be used as the fitness value for the particle. Because our additional task in this research is to reduce dimensionality simultaneously via the proposed modular function, we propose to regularise the importance ranks θ_j with the misclassification rate as the final fitness value for the particles. Specifically, given a data set $\left\{ x_p^{(l)}, y_p^{(l)} \right\}_{l=1}^{N_p}$ used for performing model selection by the PSO, where $y_p^{(l)} \in \{-1,\ 1\}$, the proposed fitness function is as follows:

$$f = \frac{1}{N} \sum_{k=1}^{N} \left(\bar{y}_p^{(k)} - y_p^{(k)} \right)^2 + \lambda \sum_{i=1}^{n} \theta_i \tag{15}$$

where λ (> 0) is the regularization coefficient, $\bar{y}_p^{(k)}$ is the output achieved by the SVM model. Now we can exploit the PSO technique to train the hyper-parameters for the SVM by getting the lowest number of misclassified patters as possible, at the same time decreasing the influences of less important features as much as possible.

Let p denote the swarm size. Each particle $1 \le i \le p$ is characterised by its current position in the search space s_i, current velocity v_i and its personal best position in the search space g_i. In the iterative process, each particle is updated using (16) and (17) to minimise the function (15),

$$\begin{aligned} v_{i,j}(t+1) = w v_{i,j}(t) + c_1 r_{1,i}(t) \left[g_{i,j}(t) - s_{i,j}(t) \right] \\ + c_2 r_{2,i}(t) \left[\hat{g}_j(t) - s_{i,j}(t) \right] \end{aligned} \tag{16}$$

where $v_{i,j}$ is the velocity of the jth dimension of the ith particle, $j = 1, \cdots,\ n+2$, the w is called the inertia weight,

the c_1 and c_2 are the acceleration coefficients that control how far a particle will move in a single iteration, and $r_1 \sim U(0,1)$, $r_2 \sim U(0,1)$ are the elements from two uniform random sequences in the interval (0, 1). The new position of a particle is updated by

$$s_i(t+1) = s_i(t) + v_i(t+1) \tag{17}$$

while the personal best position of the ith particle is calculated using

$$g_i(t+1) = \begin{cases} g_i(t) & if\ f(s_i(t+1)) \geq f(g_i(t)) \\ s_i(t+1) & if\ f(s_i(t+1)) < f(g_i(t)) \end{cases} \tag{18}$$

and the global best position discovered by all particles during all previous iterations, $\hat{g}$, is obtained by

$$\hat{g} = \arg\min_{s_l} f(s_l(t+1)) \tag{19}$$

As a result, an SVM model with optimal hyper-parameters η and C for classification can be obtained, at the same time, the goodness of features can also be identified by the parameters θ_j.

C. Reduction of input space dimensionality

After the PSO optimises the SVM model, we can extract the features based on the values of the parameters θ_j. A larger value of θ_j indicates that feature x_j is more important. In the proposed scheme, the θ_j are regularised with misclassification performance measure in the form of (15) during model selection process. So, it is expected that the values of the θ_j associated with less important features will approach to zero. Although the parameters θ_j in (8) and (9) are designed as $\theta_j > 0$, through our experiments it can be found that some θ_j associated with less important features becomes very small like 1×10^{-40}, so these features whose θ_j values are small enough can be discarded in the fuzzy rule base, while the classification performance will remain unaltered or be not impacted much. In such a way, the input space dimension can be reduced greatly.

D. Induction of fuzzy rules

After the important features are extracted in terms of the values of the parameters θ_j, we can now induce the parsimonious fuzzy rules in the form of (22) based on the support vectors $\{x_s^{(l)}\}_{l=1}^{N_s}$ discovered by the SVM. Specifically speaking, the induction process is performed as follows:

1) Each support vector corresponds to a fuzzy rule. The number of fuzzy rules equals to the number of support vectors;
2) Given the ith support vector $\tilde{x}_s^{(i)} (i = 1, \cdots, L)$,
   a) The premise part of the ith fuzzy rule is evaluated as follows: the MF of fuzzy set for the jth input variable in the ith rule is
   $$A_i^j(x_j) = a^j(x_j - m_i^j) \tag{20}$$
   where $m_i^j$is the jth element of the ith support vector $\tilde{x}_s^{(i)}$.
   b) The consequent part of the ith fuzzy rule is induced from α_0 and class labels, i.e., the consequent value of the ith rule is
   $$b_i = \tilde{\alpha}_0^{(i)} \tilde{y}_s^{(i)} \tag{21}$$
   where $\tilde{\alpha}_0^{(i)}$ represents non-zero $\alpha_0^{(i)}$, and $\tilde{y}_s^{(i)}$ is the class label corresponding to the ith support vector $\tilde{x}_s^{(i)}$.

IV. Experimental Results

In this section, we apply the proposed scheme to the breast cancer clinical data collected from Nottingham City Hospital in the interests of inducing accurate and compact fuzzy rules for a breast cancer decision support system. The following experiments only consider the breast cancer cases associated with the treatments of *radiotherapy* and *hormone-therapy*, in which there exist 330 samples without missing values in the attributes.

These 330 samples are randomly separated into training subset with 60 samples, testing subset with 140 samples, and validation subset with 130 samples. The training samples are used to construct the SVM model, the validation samples to perform optimal model selection by PSO algorithm, whilst the testing samples to evaluate the generalisation performance of the optimal SVM model selected by the PSO. In the course of a training run by the PSO algorithm, the inertia weight w is typically setup to vary linearly from near 0 to 1. The acceleration coefficients c_1 and c_2 controlling how far the particles moving in the iteration are typically both set to a value of 2.0 [28]. The initial values of $\{\varphi_j\}_{j=1}^{17}$ are set to be 0, i.e., let initial $\{\theta_j\}_{j=1}^{17}$ be 0.5, while the initial η and C are chosen to be 0.1 and 1 individually. The regularisation parameter λ in the overall fitness function is 0.1. Then 15 particles in the swarm are used to identify the optimal hyper-parameters for the sake of model selection and dimensionality reduction.

After the PSO learning process, 35 support vectors are generated, which means 35 fuzzy rules are induced for the SVM based fuzzy classifier. The induced fuzzy classifier achieves 14 misclassifications on the testing samples with a classification rate (CR) of 90.0%. This indicates that the $(L1-)$ SVM based fuzzy classifier possesses good generalisation performance on the breast cancer clinical data. The importance of features obtained by the values of $\{\theta_j\}_{j=1}^{17}$ is illustrated in the Table II. It can be seen that the importance ranks associated with less important features approach to zeros. As a result, three most important features identified by the PSO technique come up, which are *estrogen receptor* test results, *DCIS margins* and *NPI* working as important determinant of therapy. That is to say, the fuzzy rules can be fully characterised by the three most important features with the scaling values $\{\theta_j\}$ shown in Table II while keeping the classification performance. For example, one of such rules is illustrated as follows:

$$R_2 : if\ x_{10}\ is\ A_2^{10}\ and\ x_{11}\ is\ A_2^{11}\ and\ x_{16}\ is\ A_2^{16}\ then\ y_2 = 95.0216 \tag{22}$$

where the fuzzy sets A_2^{10}, A_2^{11} and A_2^{16} are induced to characterise the attributes *NPI*, *estrogen receptor* and *DCIS margins* individually, and their corresponding membership functions are $A_2^{10}(x_{10}) = exp\big(-0.027(x_{10} - 3.834)\big)$, $A_2^{11}(x_{11}) = exp\big(-0.027(x_{11} - 1.0)\big)$, and $A_2^{16}(x_{16}) = exp\big(-0.027(x_{16} - 3.0)\big)$ respectively.

TABLE II
THE GOODNESS OF 17 FEATURES IDENTIFIED BY THE PSO FOR BREAST CANCER DATA

Features	1	2	3	4	5	6
Goodness of feature by θ_j	0.000	0.000	0.000	0.000	0.000	0.000
Features	7	8	9	10	11	12
Goodness of feature by θ_j	0.000	0.000	0.000	0.887	0.999	0.000
Features	13	14	15	16	17	
Goodness of feature by θ_j	0.000	0.000	0.000	0.999	0.000	

Interestingly, these three features discovered from data are clinically important factors for breast cancer treatment. It is clinically known that the choice of hormone therapy heavily depends on the test result of the *estrogen receptor*. The DCIS margin width, the distance between the boundary of the lesion and the edge of the excised specimen, is important for treatment decisions, such as surgery, radiotherapy and/or hormone therapy. Some research has shown that there is no significant benefit from postoperative radiotherapy therapy among patients with margin widths of 1 to $<$ 10mm, but patients in whom the margin width is less than 1 mm can benefit from postoperative radiotherapy [29]. The NPI is a well-established prognostic model for primary operable breast cancer in the adjuvant setting and has been validated repeatedly [30], [31], [32], [33], [34].

For the purpose of comparison, one linear classification method-the linear discriminant analysis (LDA), and one nonlinear one-the MLP, are applied to the breast cancer data with the same training set, test set, and validation set. First the 17 features are used to identify the treatment decisions. The LDA method does not need the validation samples, the validation data is only used in the MLP approach. The LDA misclassifies 48 samples with a CR of 65.7% on the test samples, which seemingly implies that the breast cancer treatment data is not linearly separable. The MLP consists of 17 input nodes, 25 hidden neurons and 2 output neurons. The generalized delta rule [35] is used to train the network, in which the momentum parameter and the learning rate are set as 0.01 and 0.1 separately, whilst the validation data is used to avoid overfitting of the network. The trained MLP misclassified 27 test samples with a CR of 80.7%. Furthermore, we use the 3 most important features extracted by the PSO, i.e., *estrogen receptor*, *DCIS margins* and *NPI*, to determine the therapy decisions for breast cancer. The LDA misclassifies 62 test samples with a CR of 55.7%. The MLP with the same structure achieves the CR 85.0% by misclassifying 21 testing samples, whilst the proposed scheme classifies the testing samples with the CR 90.0%. It can be seen that the extracted features are helpful in improving the generalisation performances for nonlinear classifiers. The above classification results are summarized in Table III.

TABLE III
GENERALIZATION PERFORMANCES OF THE ALGORITHMS ON BREAST CANCER TREATMENT DATA

Methods	LDA	MLP	The proposed method
CRs of 17 features	65.7%	80.7%	90%
CRs of 3 features	55.7%	85%	90%

V. CONCLUSIONS

In this paper, we proposed a scheme of constructing accurate and parsimonious fuzzy classifiers based on the SVM for breast cancer treatment, in which model selection and feature extraction are performed simultaneously in an integrated manner by applying the PSO technique. Because the number of induced fuzzy rules is not related to the dimensionality of input space and a new modular function is applied to each attribute, the proposed scheme provides an efficient way of not only avoiding the "curse of dimensionality" during constructing fuzzy classifiers in high-dimensional space, but also reducing the input space dimensions. The experiments on breast cancer high-dimensional data have shown that the proposed fuzzy classifier achieves better classification performance than the well-known classifiers in the literature, and at the same time, the breast cancer data dimension was effectively reduced to 3 from 17.

Although the SVM learning algorithm can lead to parsimonious model structure for fuzzy classification, there exist potentially redundant support vectors, so there are maybe redundant fuzzy rules in the fuzzy classifier initially induced by the SVM. What's more, the number of variables used in each rule has certain influence on the number of fuzzy rules in rule base, feature extraction may lead to redundant fuzzy rules. We believe that the compactness of the fuzzy rule base generated by the proposed scheme has further room for improvement.

ACKNOWLEDGMENT

This work has been supported by the EPSRC Research Grant EP/C542215/1 and EP/C542207/1.

REFERENCES

[1] I. Jatoi and A.B. Miller, "Why is breast-cancer mortality declining?" *Lancet Oncology*, vol.4, no.4, 2003, pp. 251–254.
[2] J. L. Botha, F. Bray, R. Sankila and D.M. Parkin, "Breast cancer incidence and mortality trends in 16 European countries," *European Journal of Cancer*, vol.39, no.12, 2003, pp. 1718–1729.
[3] WHO, "Fact sheet No. 297: Cancer," *Media Centre for World Health Organization*, February, 2006 Retrieved on April, 2007.
[4] M. T. Skevofilakasa, and K. S. Nikitaa, "A decision support system for breast cancer treatment based on data mining technologies and clinical practice guidelines," *Proc. of the 27th IEEE Annual Conference on Engineering in Medicine and Biology*, Shanghai, China, September 1-4, 2005, pp.2429-2432.

[5] L.-X. Wang and J.M. Mendel, "Generating fuzzy rules by learning from examples," *IEEE Trans. on Systems, Man and Cybernetics*, vol.22, no.6, pp.1414-1427, 1992.

[6] J.-S. R. Jang, "ANFIS: adaptive-network-based fuzzy inference system," *IEEE Trans. on Systems, Man and Cybernetics*, Vol.23, no.3, pp.665-685, 1993.

[7] S. Guillaume, "Designing fuzzy inference systems from data: an interpretability oriented review," *IEEE Trans. on Fuzzy Systems*, vol.9, no.3, pp.426-443,2001.

[8] S. Guillaume, and B. Charnomordic, "Generating an interpretable family of fuzzy partitions from data," *IEEE Trans. on Fuzzy Systems*, vol.12, no.3, 2004, pp.324-335.

[9] M. Brown, K.M. Bossley, D.J. Mills and C.J. Harris, " High dimensional neurofuzzy systems: overcoming the curse ofdimensionality," *Proc. of International Joint Conference of the Fourth IEEE International Conference on Fuzzy Systems and The Second International Fuzzy Engineering Symposium*, 20-24 Mar, 1995, Yokohama, Japan, vol.4, pp.2139-2146.

[10] M.K. Giiven and K. M. Passino, "Avoiding exponential parameter growth in fuzzy systems," *IEEE Trans. on Fuzzy Systems,* vol.9, no.1, pp.194-199, 2001.

[11] V. Vapnik, *The Nature of Statistical Learning*, New York: Springer Verlag, 1995.

[12] V. Vapnik, *Statistical Learning Theory*, Wiley, New York, NY, 1998.

[13] Y. Chen, and J. Z.Wang, "Support vector learning for fuzzy rule-based classification systems," *IEEE Trans. Fuzzy Systems*, vol. 11, no. 6, pp.716–728, Dec. 2003.

[14] S. M. Zhou, and J. Q. Gan, "Constructing L2-SVM-based fuzzy classifiers in high-dimensional space with automatic model selection and fuzzy rule ranking," *IEEE Trans. Fuzzy Systems*, vol. 15, no. 3, pp. 398–409, 2007.

[15] S. M. Zhou and J. Q. Gan, "An unsupervised kernel based fuzzy c-means clustering algorithm with kernel normalization," *International Journal of Computational Intelligence and Applications*, vol. 4, no. 4, pp. 355-373, 2004.

[16] S. W. Kim and B. J. Oommen, "On utilizing search methods to select subspace dimensions for kernel-based nonlinear subspace classifiers," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 27, no. 1, pp. 136–141, Jan. 2005.

[17] F. Camastra and A. Verri, "A novel kernel method for clustering," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 27, no. 5, pp. 801–805, May 2005.

[18] J. Kennedy, and R. C. Eberhart, "Particle swarm optimization," *Proc. of the* 4^{th} *IEEE International Conf. on Neural Networks*, vol. 4, Perth, Australia, 27 November $\sim$1 December, 1995, pp. 1942-1948.

[19] R. C. Eberhart, and J. Kennedy, "A new optimizer using particle swarm theory," *Proc. of the Sixth International Symposium on Micro Machine and Human Science*, Nagoya, 4$\sim$6 October, 1995, Japan, pp. 39-43.

[20] R. C. Eberhart, and Y. Shi, "Special issue on particle swarm optimization," *IEEE Trans. on Evolutionary Computation*, vol. 8, no. 3, pp.201-202, 2004.

[21] R. De, N. R. Pal, and S. K. Pal, "Feature analysis: neural network and fuzzy set theoretic approaches," *Pattern Recognition*, vol. 30, no. 10, pp. 1579–1590, 1997.

[22] D. Tikk, T. D. Gedeon, and K. W. Wong, "A feature ranking algorithm for fuzzy modeling problems," In J. Casillas, O. Cordón, F. Herrera, L. Magdalena (eds.), *Interpretability Issues in Fuzzy Modeling,* number 128 in *Studies in Fuzziness and Soft Computing*, pp.176–192, Springer–Verlag, Heidelberg, 2003.

[23] D. Chakraborty, and N. R. Pal, "A neuro-fuzzy scheme for simultaneous feature selection and fuzzy rule-based classification," *IEEE Trans. on Neural Networks*, vol. 15, no.1, pp.110-123, 2004.

[24] D. Chakraborty, and N. R. Pal, "Integrated feature analysis and fuzzy rule-based system identification in a neuro-fuzzy paradigm," *IEEE Trans. on Syst., Man, Cybern. -Part B*, vol.31, no.3, pp. 391-400, 2001.

[25] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modelling and control," *IEEE Trans. on Syst., Man, Cybern.*, vol. 15, no. 1, pp. 116-132, 1985.

[26] D. Dubois and H. Prade, "Operations on fuzzy numbers," *Int. Journal of Syst. Sci.*, vol. 9, no. 6,pp. 613-626, 1978.

[27] A. J. Smola, B. Schölkopf, and K.-R. Müller, "The connection between regularization operators and support vector kernels," *Neural Networks*, vol. 11, no. 4, pp. 637–649, 1998.

[28] R. C. Eberhart, P. Simpson, and R. Dobbins, *Computational Intelligence PC Tools*. New York: Academic, 1996.

[29] M. J. Silverstein, M. D. Lagios, S Groshen, J. R. Waisman, B. S. Lewinsky, S. Martino, P. Gamagami, and W. J. Colburn, "The influence of margin width on local control of ductal carcinoma in situ of the breast**,**" *New England Journal of Medicine*, vol.340, no.19, pp.1455-1461, 1999.

[30] C. W. Elston, I. O. Ellis and S. E. Pinder, "Pathological prognostic factors in breast cancer," *Critical Reviews in Oncology/Hematology*, vol.31, no.3 1999, pp. 209-223.

[31] J. L. Haybittle, R. W. Blamey and C. W. Elston, J. Johnson , P. J. Doyle, F. C. Campbel, R. I. Nicholson, K. Griffiths, "A prognostic index in primary breast cancer," *British Journal of Cancer*, vol. 45, no.3, 1982, pp. 361–366.

[32] M. H. Galea, R. W. Blamey, C. E. Elston and I. O.Ellis, "The Nottingham prognostic index in primary breast cancer," *Breast Cancer Research and Treatment*, vol. 22, no. 3, 1992, pp. 207-219.

[33] I. Balslev, C. K. Axelsson, K. Zedeler, B. B. Rasmussen, B. Carstensen and H. T. Mouridsen, "The Nottingham Prognostic Index applied to 9,149 patients from the studies of the Danish Breast Cancer Cooperative Group (DBCG)," *Breast Cancer Res Treat*, vol.32, no.3, 1994, pp. 281–290.

[34] J. Kollias, C. A. Murphy, C. W. Elston, I. O. Ellis, J. F. Robertson and R. W. Blamey, "The prognosis of small primary breast cancers," *European Journal of Cancer*, vol.35, no.6, 1999, pp. 908-912(5).

[35] S. Haykin*, Neural Networks: A Comprehensive Foundation*, 2nd ed. Upper Saddle River, NJ: Prentice-Hall, Inc., 1999.