Ovarian cancer diagnosis using a hybrid intelligent system with simple yet convincing rules

Ng, Geok See; Wang, Di; Chai, Quek

2014

Wang, D., Chai, Q. , & Ng, G. S. (2014). Ovarian cancer diagnosis using a hybrid intelligent system with simple yet convincing rules. *Applied Soft Computing*, 2025-39.
doi:10.1016/j.asoc.2013.12.018

<https://hdl.handle.net/10356/89612>

<https://doi.org/10.1016/j.asoc.2013.12.018>

© 2014 Elsevier. This is the author created version of a work that has been peer reviewed and accepted for publication by *Applied Soft Computing*, Elsevier. It incorporates referee' s comments but changes resulting from the publishing process, such as copyediting, structural formatting, may not be reflected in this document. The published version is available at: [<http://dx.doi.org/10.1016/j.asoc.2013.12.018>].

Downloaded on 26 Apr 2023 18:34:49 SGT

Ovarian Cancer Diagnosis Using a Hybrid Intelligent System with Simple yet Convincing Rules

Di Wang^a, Chai Quek^{a,*}, Ng Geok See^b

^a*School of Computer Engineering, Nanyang Technological University, 639798, Singapore*

^b*School of Information Systems Technology & Design, Singapore University of Technology & Design, 138682, Singapore*

Abstract

Ovarian cancer is the ninth most common cancer among women and ranks fifth in cancer deaths. Statistics show that the five-year survival rate is greater than 75% if diagnosis occurs before the cancer cells have spread to other organs (stage I), but it drops to 20% when the cancer cells have spread to upper abdomen (stage III). Therefore, it is crucial to detect ovarian cancer as early as possible and to correctly identify the stage of the cancer to prevent any further delay of appropriate treatments. In this paper, we propose a novel self-organizing neural fuzzy inference system that functions as a reliable decision support system for ovarian cancer diagnoses. The system only requires a limited number of control parameters and constraints to derive simple yet convincing inference rules without human intervention and expert guidance. Because feature selection and attribute reduction are performed during training, the inference rules possess a great level of interpretability. Experiments are conducted on both established medical data sets and real-

*Corresponding author

Email address: ASHCQUEK@ntu.edu.sg (Chai Quek)

world cases collected from hospital. The experimental results of our proposed model in ovarian cancer diagnoses are encouraging because it achieves the most number of correct diagnoses when benchmarked against other computational intelligence based models. More importantly, its automatically derived rules are consistent with expert knowledge.

Keywords: Ovarian cancer diagnosis, Decision support system, Hybrid intelligent system, Interpretability, Feature selection

1. Introduction

Ovarian cancer begins with the formation of a tumor in a woman's ovary. It is classified as the epithelial cancer caused by genetic alterations that disrupts the regulation of proliferation, apoptosis, senescence, and DNA repair [1]. Ovarian cancer ranks fifth in cancer deaths among women and accounts for more deaths than any other cancer of the female reproductive system [2]. It is estimated that in United States, 2012, about 22,280 women receive new diagnoses of ovarian cancer and about 15,500 patients die from it [2]. Early detection of ovarian cancer is crucial because statistics [3] show that the five-year survival rate of ovarian cancer is greater than 75% if diagnosis occurs before the cancer cells have spread to other organs (stage I), but it drops to 20% when the cancer cells have spread to upper abdomen (stage III). However, early detection of ovarian cancer is not easy because it is rather hard to tell by the symptoms when a tumor in the ovary turning from benignancy to malignancy. Even annual routine gynecologic and pelvic examinations have only detected 3% of the early stage ovarian cancer cases [4]. Furthermore, the accurate staging of ovarian cancer is also important because the treatment

options depend on the type and the stage that ovarian cancer has advanced to [5]. Therefore, early detection and accurate staging of ovarian cancer are of great interests and importance because they are the keys to improve the survival rate of a patient [6].

There are a number of research works in literature focused on cancer diagnoses using various approaches, such as statistical methods, clustering techniques, decision trees, neural networks, and neural fuzzy inference systems. Valerio et al. [7] used χ^2 -test to identify unique protein peaks in pancreatic cancer diagnoses. Lee [8] applied ANOVA to select gene markers in ovarian cancer diagnoses. Generally speaking, statistical approach identifies important features and usually requires other methods or doctor expertise to perform actual cancer diagnoses. Petricoin et al. [9] proposed a self-organizing clustering technique to separate patients with or without prostate cancer. Poon et al. [10] applied a two-way hierarchical clustering algorithm to differentiate hepatocellular carcinoma from chronic liver disease. Generally speaking, classifications based on the clustering results are conceptually equivalent to similarity-based diagnoses, which do not provide concrete decisive statements. Adam et al. [11] developed a classification decision tree to identify better biomarkers for early detection of prostate cancer. Qu et al. [12] further improved the results on the same prostate cancer data set using boosted decision trees. Decision trees are commonly known as comprehensive inference tools, which employ crisp decision rules and perform feature selections. However, decision trees are sensitive to noisy inputs and the boosted decision trees are difficult to interpret although they increase diagnostic accuracies. Ball et al. [13] proposed a three-layered neural network, which

employs back propagation learning to analyze mass spectra for the prediction of astroglial tumor grades. Tsai et al. [14] used t -test to cull most of unconcerned genes and then applied the selected oncogenes to a three-layered neural network for ovarian cancer diagnoses. Although neural networks are accurate prediction tools, most of them function as black boxes [15] because users cannot make senses of the reasoning processes. Neural fuzzy inference system (or fuzzy neural network) combines the learning capabilities of neural networks and the transparent properties of fuzzy systems together by performing respective fuzzy or non-fuzzy operations in each layer of the network. Tan et al. [16] proposed neural fuzzy inference systems, which are inspired by the hippocampus and neocortex memory structures to diagnose ovarian cancer. When benchmarked against other models, their experimental results illustrated clear advantages of the proposed neural fuzzy inference systems that achieved higher accuracies and employed smaller number of interpretable fuzzy rules. Tung et al. [17] first used the Monte Carlo evaluative selection method [18] to select important features and subsequently trained the proposed neural fuzzy inference system for ovarian cancer diagnoses. However, their experimental results did not show superior performances even with feature selection applied. In recent years, there were studies focusing on obtaining highly compact fuzzy inference rules [19, 20]. The models were able to derive extraordinarily small numbers of rules through the systematic pruning process from both established medical data sets and real-world disease diagnoses. Furthermore, the hybridization of multiple techniques combines the advantages of each individual and alleviates certain limitations. Therefore, hybrid intelligent systems receive increasing attentions in the past few

years. For example, evolutionary algorithms are often combined with other techniques to solve optimization problems [21, 22] and to find optimal parameters [23]. A recent comparative study is reported in [24]. In this paper, we propose a novel neural fuzzy inference system, which self-organizes its network structure and performs feature selection and attribute reduction during training. Therefore, simple yet convincing fuzzy inference rules are systematically derived to diagnose both established medical data sets and real-world ovarian cancer cases collected from hospital for stage identifications.

The rest of this paper is organized as follows. Section 2 and 3 describe two techniques respectively to introduce the foundations of our proposed model. Section 4 provides the details of our proposed clustering method. Section 5 defines the system architecture of our proposed model, which employs the fuzzy rules automatically derived by the proposed clustering method. Section 6 benchmarks the performances of our proposed model on two established medical data sets. Section 7 presents the experimental results of applying our proposed model to diagnose ovarian cancer cases collected from hospital and analytically benchmarks the results against other computational intelligence based models. Section 8 concludes this paper.

2. Rough Set Theory for Knowledge Reduction

Rough sets [25] are always compared to fuzzy sets [26]. To describe belongingness, fuzzy sets use straightforwardly defined membership functions, while rough sets use relative relations named the lower and upper approximations. Both theories aim to achieve the same type of goal [27]. However, it is always better to have them both in one system to take advantages of

their complements [28].

2.1. Knowledge Representation System and Decision Table

Rough set theory is initiated to model relations in a given data set or knowledge base [25]. It is a formally defined methodology, which can be applied to reduce the dimensionality of a given data set [29], proceeding to train an inference system.

To express mathematically how rough set theory is applied in knowledge reduction, decision logic language is used to model the Knowledge Representation System (KRS). Such a system is represented in a pair $S = (U, A)$, where U is a nonempty and finite set named the universe of discourse and A is a nonempty and finite set of primitive attributes.

Decision tables can be defined in terms of KRS. If we have $S = (U, A)$ and $C, D \subset A$, which denote the condition and decision attributes respectively, then S with distinguished (C, D) pairs is essentially a decision table, i.e. $T = \{U, C, D\}$ or T is a CD-based decision table. Moreover, each element in U is a CD-based decision rule and it represents a cluster of data. Subsequently, U is the union of all clusters found in the data set. A simple example of decision table is given in Table 1 to illustrate the concept.

Table 1: A simple example to illustrate the concept of a decision table.

U	A		
	C		D
	height	weight	body size
1	tall	heavy	big
2	short	light	small

2.2. Indiscernible Relation and Approximations

Indiscernible relation over knowledge K , named $IND(K)$, is defined in Eq. (1). The family of all equivalence classes of the equivalence relation $IND(K)$ is named $U/IND(K)$.

$$IND(K) = \{(x, y) \in U^2 \mid \forall r \in K, r(x) = r(y)\}. \quad (1)$$

Rough set theory approximates knowledge using a pair of relational approximations. The lower and upper approximations of a set, when given an equivalence relation $IND(K)$, are defined in Eq. (2). By presenting knowledge K , lower approximation $\underline{K}Y$ is the set of elements that can be certainly classified by K and Y and upper approximation $\overline{K}Y$ is the set of elements that can be possibly classified by K and Y .

$$\underline{K}Y = \bigcup \{X : X \in U/IND(K), X \subseteq Y, Y \subseteq U\}, \quad (2a)$$

$$\overline{K}Y = \bigcup \{X : X \in U/IND(K), X \cap Y \neq \phi, X \subseteq Y, Y \subseteq U\}. \quad (2b)$$

2.3. Attribute Reduction and Feature Selection

Rough set theory performs knowledge reduction with two fundamental concepts named reduct and core. Intuitively, a reduct of knowledge is an essential subset of knowledge suffices to define all basic relations, whereas a core is the most fundamental subset of knowledge consists of the common attributes of all reducts.

Given a decision table $T = \{U, C, D\}$, an attribute a is dispensable if and only if $IND(C) = IND(C - \{a\})$. Otherwise, a is indispensable. The family C is independent if $\forall a \in C$ is indispensable in C . Attribute reduction is performed during the process of finding independent C with minimum

cardinality. If for all elements in U , an attribute b is dispensable, then b can be removed from the training data set. This process of eliminating attributes that do not contribute any information for inferences is considered as feature selection, which is extremely useful to handle problems with high dimensionality.

Based on indispensable relations, reduct and core are defined as follows. $Q \subseteq P$ is a reduct of P , if Q is independent and $IND(P) = IND(Q)$. The core of P is defined as the common parts of all reducts, i.e. $CORE(P) = \bigcap REDUCT(P)$. It is easy to infer that P can have many reducts, but only has one core.

2.4. Knowledge Reduction Process

When a training data set is given, continuous data can be represented with categorical values if separation boundaries in every dimension are set. After removing all dispensable attributes and merging all duplicates, a simplified decision rule base is obtained. This knowledge reduction process is actually the process of finding a reduct of the knowledge representation system constructed from the given data set.

3. Genetic Algorithm for Optimal Separation

Genetic Algorithm (GA) [30] is in the field of evolutionary algorithms. The idea of GA is inspired by the survival of the fittest theory proposed by Charles Darwin. GAs are searching algorithms based on the mechanics of natural selection and natural genetics [31]. GA normally starts with a randomly initialized population, which consists of artificial creatures named chromosomes. Based on their fitness values, some of them are selected in pairs

as parents and granted the opportunities to produce offspring by means of crossover operators. Subsequently, some chromosomes are randomly selected for mutation, which means their genes are to be varied. The chromosomes in the next generation are expected to perform better and the process goes on iteratively until any termination criterion is met. Although randomized, GA is not a random walk. It efficiently exploits historical information to speculate on new search points with expected improvements [31].

Because rough set theory only applies to categorical values, discretization of the training data set is required. GA is employed by our proposed model to search for optimal or at least satisfactory suboptimal separation boundaries in every input dimension. Different strategies that can be applied to each step of GA is not covered in this paper. However, all strategies applied to our proposed model is introduced with details in the following section.

4. Proposed Clustering Technique

Genetic Algorithm based Rough Set Clustering (GARSC) is a clustering technique that integrates genetic algorithm and rough set theory together. Genetic algorithm is applied to determine optimal or at least satisfactory sub-optimal solutions. Rough set theory is incorporated to alleviate the curse of dimensionality problem [32], which leads to unnecessarily large network sizes of many established inference systems. By applying rough set approximations, the original knowledge base is greatly reduced without losing essential information. This wonderful characteristic of rough set theory is extremely helpful to improve the interpretability [33] of an existing inference rule base, i.e. reduce the number of features needed for reasoning, the number of rules

employed, and the number of arguments stated in each rule. Therefore, the overall proposed system achieves a high level of interpretability without sacrificing accuracy. The overall GARSC process is illustrated in Fig. 1.

Figure 1: Flowchart of genetic algorithm based rough set clustering process.

Please note that up to here the inference rules are crisp decision rules and we need to transform them into fuzzy rules by generating Gaussian type fuzzy membership functions based on the clustering results and assigning corresponding linguistic terms. Subsequently, the transformed fuzzy rules are used to evaluate the performance of the current solution. This knowledge transfer process is illustrated in Fig. 2.

This particular process of knowledge transfer from crisp to fuzzy cannot be omitted because the crispness of separation adopted in rough set theory does not tolerant overlapping. Fuzzy membership functions are employed to represent the derived clusters to deal with inexact information and unfore-

Figure 2: Illustration of knowledge transfer from crisp rules to fuzzy ones.

seen circumstances. This kind of knowledge transfer has a great advantage because it naturally prevents the fuzzy membership functions from overlapping or separating too much with adjacent ones, which is another important aspect of interpretability in fuzzy modeling. Furthermore, because clustering is performed in each individual feature, no transformation or normalization is required and more importantly, semantic meanings of the assigned linguistic labels are preserved.

4.1. Constraints on Data Discretization

Before the detailed introduction of GARSC, there are certain constraints on data discretization to be defined. One is the maximum number of separation boundaries allowed in each dimension. This constraint straightforwardly determines the maximum number of fuzzy membership functions allowed in each dimension. However, the actual number of fuzzy membership functions formulated in each dimension is also determined by the knowledge reduction process. The smallest number of separation boundaries actually in use is zero, which means that particular feature is not considered in the simplified knowledge base. This constraint should not be set to a large value because employing a huge number of fuzzy membership functions degrades interpretability.

The other constraint is on the minimum distance between any pair of adjacent separation boundaries in the same dimension, which is termed *mindis*. This constraint is imposed to make sure that the actual constructed fuzzy membership functions have high level of generalization such that any adjacent pair of them should not be merged into one. The minimum distance constraint *mindis* is defined in Eq. (3). The *max* function in the denominator defines the level of generalization.

$$mindis_i = \frac{ub_i - lb_i}{\max(nop_i, M)}, \quad (3)$$

where i denotes the i th dimension; ub_i and lb_i denote the upper and lower boundaries of values in the i th dimension respectively; nop_i denotes the total number of possible separation boundaries in the i th dimension, which if not specifically stated otherwise, is assigned to the total number of different values in the i th dimension where corresponding values in the decision attribute are changed across; and M denotes the user specified minimum number of separation boundaries in every dimension, which has a default value of 10.

4.2. Knowledge Reduction and Removal of Inconsistent Rules

It has been discussed earlier in this paper that after the removal of all dispensable attributes, the finalized decision table is independent, which is actually one of the many possible reducts of the originally constructed decision table. Therefore, attribute reduction is performed during the process of constructing the independent decision table with minimum cardinality. If an attribute is dispensable in all rules, then that attribute can be removed from the given data set. This process of removing attributes that do not contribute to the essential knowledge base is referred as feature selection.

Decision rule reduction is conceptually similar to attribute reduction. Other than merging each set of duplicate rules into one single rule, a decision rule in the rule set is dispensable if and only if the performance of the rule base does not decrease with the removal of that rule. This is also referred as the pruning process of decision rules. Moreover, the removal of inconsistent rules is necessary to maintain the integrity of the knowledge base.

Inconsistent rules, which are the rules with the same conditional attributes but different decision attribute, do often exist in real-world applications. Only one rule from each inconsistent rule set should be preserved to remove ambiguities. To determine which rules should be preserved in the simplified decision table, we propose Eq. (4) (in rough set theory terms) to compute the confidence of the k th rule. The *min* function is applied to penalize information incompatibility.

$$conf(k) = \min \left(\frac{card(U_i(k_i) \cap d_k)}{card(U_i(k_i))} \right), \forall i \in C, \quad (4)$$

where *card* denotes the function to compute cardinality; U_i denotes the function to generate a union of decision attributes of every rule in the decision table that shares the same value on the i th attribute; k_i denotes the value of the i th attribute of the k th rule; and d_k denotes the decision attribute of the k th rule.

Based on the confidence evaluation function, the criteria to remove inconsistency rules from the decision table are defined as follows. (1) The rule with the maximum confidence value will be preserved in the decision table while all the other rules from the same inconsistent rule set will be removed. (2) If multiple rules are tied at the maximum confidence value in the same inconsistent rule set, the rule covers the most number of data elements is

preserved. (3) If multiple rules still tie in this situation, a random rule will be selected as the winner with equal probability.

4.3. Proposed Genetic Algorithm Strategies

4.3.1. Commonly Adopted Strategies

In genetic algorithm, the population size defines the total number of chromosomes existing in each generation. Increasing the population size is equivalent to increasing the number of possible solutions. Therefore, more candidates will be examined and a wider range of area in the universe of discourse will be explored.

In GARSC, real number coding strategy is employed to construct chromosomes. Each gene used in the chromosome represents a separation boundary in its respective dimension. Because GARSC technique only constrains the maximum number of partitions allowed in each dimension, the actual number of partitions varies, i.e. chromosomes in GARSC have different lengths.

Elitism replacement strategy is applied in GARSC. The elitism ratio μ , which is in the interval $[0, 1)$, defines how many percentage of highly fit individuals in the current generation $P(t)$ will be directly brought into the next generation $P(t+1)$. Normally μ is assigned to a relatively small number to prevent the domination of species on local optimums.

The overall clustering process stops when genetic algorithm reaches the predefined number of generations, which should sufficiently ensure the convergence of the genetic algorithm.

4.3.2. Fitness Evaluation Function

Fitness function evaluates the quality of each chromosome. It is probably the most important component in genetic algorithm, because it directly defines the performance of each chromosome and intimately characterizes the ideal solution that the user attempts to search for. Because the intension of the proposed soft computing model is to produce highly interpretable rules without sacrificing accuracy, fitness function f is defined in Eq. (5), where performance evaluations of both accuracy and interpretability are incorporated. We use capital letters to represent constants and small letters to represent variables. Term-1 and term-5 in Eq. (5) represent the accuracy of the model and the rest three terms represent the interpretability of the model because they are the scores on the number of selected features, the number of derived fuzzy rules, and the total number of arguments in the antecedent part of all rules respectively. This fitness function is to be minimized by the genetic algorithm.

$$f(x) = \underbrace{\tau_1(1-a)\frac{NOD}{NOF}}_1 + \underbrace{\tau_2\frac{nof}{NOF}}_2 + \underbrace{\tau_3\frac{nor}{NOD}}_3 + \underbrace{\tau_4\frac{noa}{NOF \cdot NOD}}_4 + \underbrace{\tau_5\frac{mse}{NOF}}_5, \quad (5)$$

where x denotes the chromosome to be evaluated; $\tau_1, \dots, \tau_5$ denote associated coefficient parameters for the user to specify; a denotes the accuracy of applying data set to the model derived using x ; NOD is the total number of data elements in the data set; NOF denotes the total number of input features in the data set; nof denotes the number of features actually used to construct the model; nor denotes the number of rules actually used to construct the model; noa denotes the total number of arguments in the antecedent part of all rules in the rule set; and mse denotes mean squared error

computation, which is defined in Eq. (6).

$$mse = \frac{1}{NOD} \sum_{i=1}^{NOD} (y_i - \hat{y}_i)^2, \quad (6)$$

where y_i denotes the predicted value and $\hat{y}_i$ denotes the expected value.

We propose five terms to evaluate two different types of performance in the fitness function. The employment of this type of fitness function is one way to implement a multi-objective genetic algorithm [34]. If $\tau_1, \dots, \tau_5$ are assigned equally to a positive number, the five terms in Eq. (5) are placed in descending order of importance and the proof is given as follows.

Term-1 of Eq. (5) can be expanded into Eq. (7).

$$(1 - a) \frac{NOD}{NOF} = (1 - \frac{noc}{NOD}) \frac{NOD}{NOF} = \frac{NOD - noc}{NOF}, \quad (7)$$

where noc denotes the number of correctly classified data elements.

If the number of correctly classified data elements decreases by one, i.e. $noc \rightarrow noc - 1$, then we can rewrite Eq. (7) into Eq. (8).

$$\frac{NOD - (noc - 1)}{NOF} = \frac{NOD - noc}{NOF} + \frac{1}{NOF}. \quad (8)$$

Comparing Eq. (7) to Eq. (8), if all the other terms in Eq. (5) remain unchanged, then fitness value is increased by $\frac{1}{NOF}$, which denotes a lesser fit of the chromosome.

If term-2 of Eq. (5) changes in the slightest amount, i.e. $nof \rightarrow nof + 1$, which means one more feature is used to construct the model, then the fitness value is increased by $\frac{1}{NOF}$, which also means a lesser fit of the chromosome. The derivation steps are similar to the transformation from Eq. (7) to Eq. (8). Therefore, the derivation steps will not be repeated hereinafter.

Up to here, we show that the first two terms in Eq. (5) have the same amount of effect for the slightest changes. In other words, the amount of effect on accuracy is on the same level of feature selection. Comparing to many other neural fuzzy inference systems that only concern accuracy, our proposed model prefers more generalized solutions that utilize lesser amount of inference knowledge if accuracy remains unchanged.

If one more rule is employed in the inference knowledge base, i.e. $nor \rightarrow nor+1$, then the fitness value is increased by $\frac{1}{NOD}$. The number of dimensions in the given data set is normally smaller than the number of data elements (except for gene data), i.e. $NOF < NOD \Rightarrow \frac{1}{NOF} > \frac{1}{NOD}$, which implies that the effect of the slightest change in term-2 of Eq. (5) is greater than that of term-3.

If one more argument is employed in the rule set, i.e. $noa \rightarrow noa + 1$, then the fitness value is increased by $\frac{1}{NOF \cdot NOD}$. Because $NOF > 1$ in all real-world problems, the effect of the slightest change in term-3 of Eq. (5) is greater than that of term-4.

If all the first four terms of Eq. (5) remain the same, then the classification accuracy and interpretability of the constructed model do not vary. If there is a small change in term-5, then the model is slightly lesser fit for the given data set, which leads to an increase in the mean squared error. By substituting Eq. (6) into term-5 of Eq. (5), we can derive Eq. (9).

$$\frac{mse}{NOF} = \sum \frac{(y_i - \hat{y}_i)^2}{NOF \cdot NOD}. \quad (9)$$

According to Eq. (9), the change in the sum of squared error has the same level of effect as the change in the number of arguments employed in the rule set because they have the same denominator. The slightest change in

sum of squared error is assumed to be smaller than one, if only one predicted value among all the data elements is computed differently without changing the value of term-1 in Eq. (5). Therefore, term-4 takes on a greater effect than that of term-5 in a rather stable solution, where predictions do not vary drastically.

Thus, the complete proof has been given that the amount of effect decreases along the five terms of Eq. (5) if their coefficient parameters are set to the same positive value. Term-1, which concerns about accuracy, guarantees the correctness of the constructed model. Term-2, -3, and -4, which concern about interpretability, encourage the constructed model to achieve the same level of accuracy with lesser amount of inference knowledge. According to Occam’s razor that all other things being equal, the simplest solution is the best [35], a higher level of interpretability with the same level of accuracy constructs better model in terms of better generalization, which prevents the model from over-fitting. Term-5, which concerns about the mean squared error, is employed to refine the constructed model to achieve a better fitting on the given data set as long as accuracy and the amount of information employed are kept unchanged. The tradeoff between accuracy and interpretability should vary in different application domains and under different user requirements. However, tuning the coefficients, $\tau_1, \dots, \tau_5$, is sufficient to leverage the balance between accuracy and interpretability.

4.3.3. Tournament Selection Strategy

To produce new chromosomes, parents with high fitness values are selected from the current generation to produce offspring in the next generation. Tournament selection strategy [36] is employed in GARSC because we

can easily control the selection stress by adjusting only the tournament size m and the selection probability p .

For each chromosome to be selected for crossover, m candidates are randomly selected from the current generation for consideration and they are sorted in descending order based on their fitness values. The selection process starts with the first candidate and the probability of selecting the n th candidate $p(n)$ is defined by Eq. (10).

$$p(n) = p(1 - p)^{n-1}, \quad 0.5 < p \leq 1. \quad (10)$$

If m is large, it is more stressful for lesser fit candidates to be selected. On the other hand, a small tournament size increases the probability of lesser fit candidates being selected as they are competing with a lesser number of the others. In general, a large value of m is used in simple and unimodal application domains to accelerate the converging process and a small value of m is used in complex and multimodal application domains to better explore the universe of discourse [36].

In the early generations of GA, p should be set to a small value to give more chances to those lesser fit candidates to get selected. In this way, the search is prevented from premature convergence because more possible candidates are considered even if they have smaller fitness values. However, in the late generations, p should set to a large value because only those highly fit candidates are expected to lead the selection towards the best solution in more in-depth exploitation. Tournament selection probability p in GARSC is defined by Eq. (11).

$$p = 0.5 \left(1 + \frac{cgi}{NOG} \right), \quad (11)$$

where cgi denotes the current generation index and NOG denotes the maxi-

mum number of generations.

Because cgi is in the interval $[1, NOG]$, p is equally distributed in the interval $[0.5 + \frac{0.5}{NOG}, 1]$, which exactly fulfils the constraint defined in Eq. (10).

4.3.4. Modified Uniform Crossover Operator

When sufficient number of parents has been selected, the crossover operator is applied to each pair of parents to produce offspring. The probability for a pair of selected parents eventually mate is defined as the crossover rate. However, because elitism replacement strategy is applied in GARSC, the crossover rate is set to one, i.e. every pair of parents is crossed over to produce offspring. Because chromosomes in GARSC consist of real number coded genes to represent sets of separation boundaries that are different in length, there is no simple crossover operator feasible to perform the proposed information exchange between parents.

A modified uniform crossover operator is proposed to deal with chromosomes in different lengths. Similar to the conventional uniform crossover operator, a binary string is randomly created to control in each position, from which parent the child should inherit the gene. The length of the control string is assigned to the number of conditional attributes in the given data set. Therefore, there will be no risk taken on the misunderstanding of the dimensionality during the creation process of offspring. The modified uniform crossover operator is illustrated in Fig. 3.

Because feature selection is applied in GARSC, many chromosomes have no gene in certain input dimension as indicated with the empty brackets “[]” in Fig. 3. Therefore, when a pair of chromosomes is uniformly crossed over,

Figure 3: Illustration of the proposed modified uniform crossover operator.

there is a chance that one of the offspring has no gene in every input dimension. In such case, the “empty” chromosome will be replaced with a randomly constructed non-empty one.

4.3.5. Modified Mutation Operators

Unlike conventional mutation operators that simply vary the values of the selected genes, a set of three different mutation operators is proposed in GARSC. Whenever a gene is selected to be mutated, one of the following three operators are performed with equal probability: (1) Add one separation boundary to the gene, if applicable; (2) Remove one separation boundary from the gene, if applicable; (3) Vary the value of a randomly selected separation boundary in the gene, if applicable.

Similar to tournament selection probability p , the mutation rate mr_{rate} , which defines the probability for each gene to be mutated, should increase from a small value in the early generations to a large value in the late generations. Based on this policy, mr_{rate} is defined in Eq. (12).

$$mr_{rate} = \frac{1}{NOF} + \frac{(NOF - 1) \cdot cgi}{NOF \cdot NOG}. \quad (12)$$

4.4. Overall Proposed Clustering Algorithm

Up to here, we have introduced all the details of our proposed Genetic Algorithm based Rough Set Clustering (GARSC) method. The overall GARSC

algorithm is summarized in Table 2.

Table 2: The proposed clustering algorithm.

Step 1	Based on the given training data set, find all possible separation boundaries in all input dimensions; initialize the look-up table of fuzzy membership functions; initialize the first population of GA.
Step 2	Based on the chromosome being analyzed, construct corresponding decision table; merge duplicate rows so that each row represents a cluster of data.
Step 3	Find the reduct of the decision table to perform feature selection and attribute reduction; merge again if duplicate rows exist.
Step 4	If conflict exists, remove inconsistent rules and go back to Step 3.
Step 5	Convert the crisp decision rules into fuzzy rules; if corresponding fuzzy membership functions are not available in the look-up table, compute the membership functions and update the look-up table.
Step 6	Based on the fuzzy rules, construct the neural fuzzy inference system and subsequently apply the training data set to compute the fitness value.
Step 7	Go back to Step 2 until all chromosomes are evaluated.
Step 8	If the maximum number of generations is reached, terminate the clustering algorithm and output solutions.
Step 9	Select elites from the current generation and directly put them into the next generation without any changes.
Step 10	Select parents and subsequently apply crossovers and mutations on them.
Step 11	Go back to Step 2 with the new population of chromosomes.

5. Proposed System Architecture

The architecture of Genetic Algorithm and Rough Set Incorporated Neural Fuzzy Inference System (GARSINFIS) is illustrated in Fig. 4. GARSINFIS is a six-layered, feed-forward, and partially connected architecture. In each layer, neurons are not connected to each other but only connected to neurons in the adjacent layers.

Each layer of GARSINFIS performs respective fuzzy or non-fuzzy operations that are described as follows. Input layer receives the input data

Figure 4: The network architecture of GARSINFIS.

vector and translates it into fuzzy singletons in each dimension. Because feature selection is applied, not all given linguistic variables are presented to the connected neurons in the next layer. Condition layer stores respective fuzzy membership functions with respect to each selected linguistic variable and subsequently presents the activation values to the connected neurons in the next layer. Rule-base layer performs fuzzy reasoning and subsequently presents the activation value of the corresponding fuzzy rule to all the neurons in the next layer. Normalization layer scales the activation values of all fuzzy rules to the same level of reference and subsequently presents the normalized values to the connected neuron in the next layer. Consequence layer computes the prediction of each rule and subsequently present it to the only neuron in the next layer. Output layer consists of only one neuron, which aggregates all the inputs received and presents the result as the overall inference output.

In Fig. 4, the rectangular boxes used to represent neurons in condition layer and consequence layer denote the antecedent and consequent parts of

the employed fuzzy rules respectively. GARSINFIS employs first-order TSK type of fuzzy rules [37, 38], which are defined in Eq. (13).

$$\begin{aligned} &\text{IF } x_1 \text{ is } A_{(1,i)} \text{ and } \cdots \text{ and } x_n \text{ is } A_{(n,i)} \text{ and } \cdots \text{ and } x_N \text{ is } A_{(N,i)}, \\ &\text{THEN } y_i = \alpha_{(0,i)} + \alpha_{(1,i)}x_1 + \cdots + \alpha_{(n,i)}x_n + \cdots + \alpha_{(N,i)}x_N, \end{aligned} \quad (13)$$

where x_n denotes the n th input linguistic variable; $A_{(n,i)}$ denotes the fuzzy label defined in the i th rule on the n th input linguistic variable; N denotes the total number of linguistic variables; y_i denotes the output of the i th rule; and $\alpha_{(n,i)}$ denotes the coefficient of x_n in the i th rule.

In GARSINFIS, fuzzy membership functions of Gaussian type are employed. The input function f and output function o of neurons in the condition layer are defined in Eq. (14).

$$f_{ij} = -\frac{(x_i - c_{ij})^2}{2\sigma_{ij}^2}, \quad (14a)$$

$$o_{ij} = \exp(f_{ij}), \quad (14b)$$

where f_{ij} denotes the input of the j th neuron in the i th dimension; x_i denotes the i th element of the input data vector; c_{ij} denotes the centroid of the j th cluster in the i th dimension; σ_{ij} denotes the standard deviation of the j th cluster in the i th dimension; and o_{ij} denotes the output of the j th neuron in the i th dimension.

During the iterative fitness function optimization process, different sets of fuzzy membership functions are evaluated. Therefore, when the clustering process terminates, the derived membership functions of the best solution are empirically proven to be accurate and precise. However, at this time, the derived fuzzy rules are zero-order ones and tuning of their consequent parts is desired if user wants to further increase the accuracy (at the same

time decrease interpretability because first-order rules are difficult to comprehend). In GARSINFIS, we apply the Recursive Least Squares (RLS) algorithm, which is an adaptive version of the Kalman Filter (KF) algorithm [39], to estimate the optimal coefficient matrix W^* . Assume there are P training samples and let a_p denotes the p th row of the weighted input matrix A , then W^* can be recursively estimated using Eq. (15).

$$S_0 = \gamma I, \quad (15a)$$

$$W_0 = [c_{10} \cdots c_{M0} \overbrace{0 \cdots 0}^{MN}]^T, \quad (15b)$$

$$S_p = S_{p-1} - \frac{S_{p-1} a_p^T a_p S_{p-1}}{1 + a_p S_{p-1} a_p^T}, \quad p = 1, \dots, P, \quad (15c)$$

$$W_p = W_{p-1} + S_p a_p^T \underbrace{(D_p - a_p W_{p-1})}_{\text{prediction error}}, \quad (15d)$$

$$W^* = W_P, \quad (15e)$$

where γ is a large positive number; I is the identity matrix; c_{m0} denotes the value of the consequent part of the m th zero-order fuzzy rule; M is the total number of derived fuzzy rules; N is the total number of selected features; S_p denotes the error covariance matrix for the p th input vector; and D is the matrix of desired values.

The general function of mapping input $X = (x_1, \dots, x_i, \dots, x_I)$ into output y using M number of first-order TSK type of fuzzy rules is defined in Eq. (16).

$$y = \frac{\sum_{m=1}^M \alpha_m (c_{m0} + c_{m1}x_1 + \cdots + c_{mi}x_i + \cdots + c_{MI}x_I)}{\sum_{m=1}^M \alpha_m}, \quad (16)$$

where α_m is the firing strength of the m th rule.

When the implication operator is selected as the *min* operator, then α_m is defined by Eq. (17).

$$\alpha_m = \min(\mu_{A_{m1}}(x_1), \dots, \mu_{A_{mi}}(x_i), \dots, \mu_{A_{mI}}(x_I)), \quad (17)$$

where $\mu_{A_{mi}}(x_i)$ denotes the membership value of x_i according to the membership function A_{mi} of the m th rule in the i th dimension.

Because GARSINFIS directly employs the fuzzy rules derived by Genetic Algorithm based Rough Set Clustering (GARSC) technique, GARSINFIS self-organizes its network structure. Moreover, because GARSC systematically derives simplified fuzzy inference rules, the network size of GARSINFIS is smaller than that of the other similar models. Because feature selection is performed, not all input dimensions in the given data set are utilized. Because the maximum number of fuzzy membership functions in each dimension is constrained by GARSC, the number of neurons employed to represent the derived membership functions is also constrained. Because attribute reduction in the antecedent part of fuzzy rules is performed, hence the derived fuzzy rules may not necessarily employ all the fuzzy membership functions in every selected dimension. Because rule pruning is performed during the clustering process, the number of neurons used to represent the fuzzy rule set is also minimized. The compact GARSINFIS architecture is now ready for performance evaluations.

6. Benchmarks on Established Medical Data Sets

Before moving on to ovarian cancer diagnoses, in this section, we present the experimental results of our proposed model on two established medical

data sets [40], namely Pima Indian Diabetes (PID) and Wisconsin Breast Cancer (WBC). In order to benchmark the results against other established neural fuzzy inference systems, training and testing data sets are organized in accordance with [19, 20].

6.1. Pima Indian Diabetes Data Set

Pima Indian Diabetes (PID) data set consists of 768 samples, in which 268 (35%) are diagnosed as diabetic and the reminder as healthy. In accordance with [20], in each experiment, 75% of the samples are randomly selected for training and the remainder 25% for testing. The performance of GARSIN-FIS is taken the average of ten experiments to remove randomness and the same training and testing data sets are applied to C4.5 [41]. The experimental results reported in [19, 20] are included as benchmarks because both established models focus on the compactness of the derived rules through systematic pruning. MFMM-FIS [19] constructs hyper-boxes to differentiate data samples of each class in the feature space and subsequently translates them into fuzzy inference rules. After pruning, the expressiveness of the rules is significantly improved because the number of rules is kept extraordinarily small. EGART-FIS [20] applies adaptive resonance theory [42] to incrementally learn associations in the input and output spaces respectively. Again, after pruning the extracted fuzzy rules, the derived rule size is kept small.

There are three common parameters of GARSC to be defined. (1) Elitism ratio of GA is set to 0.1, i.e. in every generation, the top 10% chromosomes with the highest fitness values are directly brought into the next generation. (2) Tournament size is set to two to ease the stress during selection. (3) The maximum number of separation boundaries allowed in any input dimension is

set to two. Therefore, the maximum number of fuzzy membership functions in any input dimension is three. This guarantees a great level of interpretability because there are at most three linguistic labels, namely small, medium, and large, assigned in any input dimension.

To ensure convergence, the population size of GARSC is set to 100 and the number of generations is set to 10. Because it is difficult to achieve relatively high accuracy on PID data set, GARSINFIS employs first-order TSK type of fuzzy rules and GARSC employs the fitness function to maximize accuracy only ($\tau_1 = \tau_5 = 1$ & $\tau_2 = \tau_3 = \tau_4 = 0$ in Eq. (5)). The benchmarks on PID data set are shown in Table 3.

Table 3: Benchmarks on Pima Indian diabetes data set.

Model	nor	nof	noa	Accuracy (%)
GARSINFIS	38.1	7.6	130.2	74.36
C4.5	17.9	6.4	69.6	73.96
EGART-FIS [20]	6	8	48	73.05
MFMM-FIS [19]	8	8	64	72.93

nor: number of rules; nof: number of features; noa: number of (total) arguments

The number of total arguments (noa) listed in Table 3 refers to the accumulated number of arguments of every fuzzy inference rule. E.g., EGART-FIS derives 6 rules and employs all 8 features for PID data set without attribute reduction, the number of total arguments is then yield as 48.

In Table 3, it is noticeable that although GARSINFIS achieves the highest accuracy on the PID testing data sets, it constructs the largest inference rule set. Although each rule derived by GARSINFIS only utilizes $130.2/38.1 = 3.42$ arguments on average, the number of employed rules is

at least twice larger than those of the other models. One possible reason of why GARSINFIS derives larger inference rule set is that there are missing values (recorded as 0) in the PID data set, which may cause difficulties when GARSINFIS tries to merge inference knowledge. To show GARSINFIS can do better, we apply it to another established data set in the following subsection.

6.2. Wisconsin Breast Cancer Data Set

Wisconsin Breast Cancer (WBC) data set consists of 699 samples, in which 458 (65.52%) are diagnosed as benign and the remainder as malignant. In accordance with [19, 20], in each experiment, 80% of the samples are randomly selected for training and the remainder 20% for testing. The performance of GARSINFIS is taken the average of ten experiments to remove randomness and the same training and testing data sets are applied to C4.5. The results of MFMM-FIS and EGART-FIS are taken from [20]. The parameter values and constraints remain the same as reported in Section 6.1, except that GARSINFIS employs zero-order rules and GARSC employs the fitness function to maximize both accuracy and interpretability ($\tau_1 = \tau_2 = \tau_3 = \tau_4 = \tau_5 = 1$ in Eq. (5)). Because WBC data set is relatively easy to achieve high diagnosis accuracy, we aim to maximize the interpretability of the derived rule set in this subsection. The benchmarks on WBC data set are shown in Table 4.

It is shown in Table 4 that although MFMM-FIS and EGART-FIS employ extraordinarily small numbers of fuzzy inference rules, their overall rule sizes (if measured by noa) are not smaller than the one derived by GARSINFIS because they do not perform feature selection and attribute reduction.

Table 4: Benchmarks on Wisconsin breast cancer data set.

Model	nor	nof	noa	Accuracy (%)
GARSINFIS	10.1	6.2	27.0	95.33
C4.5	11.5	6.5	44.0	95.69
EGART-FIS [20]	3	9	27	93.56
MFMM-FIS [19]	4	9	36	92.56

nor: number of rules; nof: number of features; noa: number of (total) arguments

Furthermore, although C4.5 employs more inference knowledge, it achieves a slightly higher accuracy on the WBC testing data sets than GARSINFIS does. One possible reason of why C4.5 achieves the highest accuracy is that every feature of the WBC data set consists of categorical values, which may favor crisp decision techniques such as C4.5. To show GARSINFIS achieves its best performance when applied to complex and continuous data sets, we apply it to real-world cases collected from hospital in the next section.

To illustrate the fuzzy rules systematically derived and optimized by GARSC are easy to comprehend by human, all the eight rules derived in one experiment are shown in Table 5 (the numbers shown in the brackets are mean and variance respectively to represent corresponding fuzzy membership functions). The shown set of rules achieves 96.35% accuracy on the testing data set and 98.1% on the whole data set. It is obvious to notice from Table 5 that although eight rules are employed (four positive ones and four negative ones), they only utilize seven features (x_2 and x_3 are excluded) and most of them (seven out of eight) only consist two or three terms in the antecedent part. It is shown that GARSINFIS employs highly interpretable rules to achieve relatively high accuracy.

Table 5: One set of the derived fuzzy rules of WBC data set.

IF	x_1 is low (1.92, 0.79) and x_4 is low (1.44, 0.77) and x_6 is low (1.27, 0.51) x_4 is low (1.44, 0.77) and x_6 is low (1.27, 0.51)	THEN diagnose as benign
IF	and x_7 is low (2.06, 0.84) and x_8 is low (1.28, 0.66) and x_9 is low (1.07, 0.10)	THEN diagnose as benign
IF	x_4 is high (6.00, 4.74) and x_5 is low (1.99, 0.18) and x_8 is low (1.28, 0.66)	THEN diagnose as benign
IF	x_6 is low (1.27, 0.51) and x_7 is low (2.06, 0.84) and x_8 is medium (6.11 0.11)	THEN diagnose as benign
IF	x_1 is high (8.77, 1.50) and x_6 is high (6.14, 3.31)	THEN diagnose as malignant
IF	x_1 is high (8.77, 1.50) and x_7 is high (6.52, 3.68)	THEN diagnose as malignant
IF	x_1 is high (8.77, 1.50) and x_9 is high (5.29, 5.91)	THEN diagnose as malignant
IF	x_4 is high (6.00, 4.74) and x_6 is high (6.14, 3.31)	THEN diagnose as malignant

7. Experimental Results on Ovarian Cancer Diagnoses

Other than benchmarking the performances of our proposed model on the established medical data sets, the experimental results on the ovarian cancer cases collected from hospital are benchmarked and analyzed in this section.

7.1. Ovarian Cancer Diagnoses Data Set

The ovarian cancer diagnoses data set is based on the blood test results of 171 patients collected from Department of Oncology and Gynecology, National University Hospital, Singapore. All samples in the data set are categorized into one of the five groups according to the staging system defined by International Federation of Gynecology and Obstetrics [43]. The distribution of the data set is shown in Table 6. The seriousness of every patient is

determined by 28 features that consist of the age of the patient and certain blood test results. The list of all features is given in Table 7.

Table 6: Distribution of the ovarian cancer diagnoses data set.

Staging of ovarian cancer	Fraction
Normal	22 (12.87%)
Benign (benign cyst)	78 (45.61%)
Borderline (proliferating tumor)	10 (5.85%)
Stage I & II	19 (11.11%)
Stage III & Iv	42 (24.56%)

Table 7: Input features of the ovarian cancer diagnoses data set.

1	Age	15	Antithrombin III Antigen (ATAG)
2	Blood Platelet	16	Plasminogen
3	Packed Cell Volume (PCV)	17	D-Dimer
4	Hemoglobin (Hgb)	18	Tissue Plasminogen Activator (tPA) Activity
5	β -Thromboglobulin (BTG)	19	Tissue Plasminogen Activator (tPA) Antigen
6	Reaction Time of Thrombelastography (TEG-R)	20	Urokinase-like Tissue Plasminogen Activator (uPA) Activity
7	Coagulation Time of Thrombelastography (TEG-K)	21	Urokinase-like Tissue Plasminogen Activator (uPA) Antigen
8	Maximum Amplitude of Thrombelastography (TEG-MA)	22	Urokinase-like Tissue Plasminogen Activator (uPA) Receptor
9	Fibrinogen	23	Plasminogen Activator Inhibitor (PAI) Activity
10	Factor VII	24	Plasminogen Activator Inhibitor (PAI) Antigen
11	Von Willebrand-Factor (VWF)	25	Protein C Antigen (PCAg)
12	Thrombin-Antithrombin Complex (T/AT)	26	Tissue Polypeptide Specific (TPS) Antigen
13	Prothrombin Fragment 1 & 2 (F1+2)	27	Tissue Factor Pathway Inhibitor (TFPI)
14	Antithrombin III Activity (ATAC)	28	Cancer Antigen-125 (CA-125)

7.2. Design of Experiments

Experiments on the ovarian cancer diagnoses data set evaluate the capability of GARSINFIS to automatically derive human-like reasoning rules and subsequently perform accurate inferences without human intervention and expert guidance. To resemble the diagnosis processes of doctors, a systematic decision support system to diagnose ovarian cancer is constructed and shown in Fig. 5.

Figure 5: Illustration of the ovarian cancer diagnoses system.

When a new instance arrives, it is sequentially processed until the diagnoses system reports the final result. In such a way, at each stage, the two important measurements in medical diagnosis, namely sensitivity and specificity, are reported. The two measurements are defined in Eq. (18) and they assess the degree of true positive and true negative respectively. In ovarian cancer diagnoses, sensitivity is critical because early treatment is crucial and beneficial to the patients and specificity is critical because unnecessary treatment is harmful to the patients both mentally and physically.

$$\text{Sensitivity} = \frac{\text{Number of positive instances correctly diagnosed}}{\text{Total number of positive instances}}, \quad (18a)$$

$$\text{Specificity} = \frac{\text{Number of negative instances correctly diagnosed}}{\text{Total number of negative instances}}. \quad (18b)$$

There are four diagnoses in the overall system. In each diagnosis, two thirds of randomly selected data samples are used for training and the remaining one third for testing. Furthermore, to remove randomness and show robustness, ten experiments are conducted in each diagnosis. In this way, each diagnosis is evaluated as a three-fold cross validation problem and runs for ten times to average the performance.

For all diagnoses and all experiments, the three common parameters of GARSC use the same values as introduced in Section 6.1. The population size and the maximum number of generations vary from different diagnoses. Therefore, their values are reported in the corresponding subsections. Diagnosis-1 is the simplest diagnosis to separate normal persons from patients at risks. Therefore, GARSINFIS employs zero-order TSK type of fuzzy rules (for better interpretability) and GARSC employs the fitness function to maximize both accuracy and interpretability ($\tau_1 = \tau_2 = \tau_3 = \tau_4 = \tau_5 = 1$ in Eq. (5)). In the other three diagnoses, GARSINFIS employs first-order TSK type of fuzzy rules and GARSC employs the fitness function to maximize accuracy only ($\tau_1 = \tau_5 = 1$ & $\tau_2 = \tau_3 = \tau_4 = 0$ in Eq. (5)). Please note that although evaluations on interpretability measurements are no longer included in the fitness function, GARSC still performs knowledge reduction (without keeping scores) to optimize the inference rule base for a higher level of interpretability.

7.3. Benchmarking Models

When benchmarking the experimental results, some well established models such as C4.5 decision tree [41], Naive Bayes classifier [44], Multi-Layer Perceptron (MLP) network [45], Radial Basis Function (RBF) network [45], Adaptive Network-based Fuzzy Inference System (ANFIS) [46] and Dynamic Evolving Neural-Fuzzy Inference System (DENFIS) [47], are used for comparisons. All these models are applied to the same pairs of training and testing data sets that are applied to GARSINFIS. In this way, their performances are fairly compared on the same references.

In terms of comparisons on the selected features, only GARSINFIS and

C4.5 are discussed, because all the other models utilize all given features without preferences. Even if the associated weights of certain input dimensions can be significantly small in some models, it is still not considered as performing feature selection.

In terms of the size of the employed rule set, the actual number of employed rules is recorded for GARSINFIS, ANFIS, and DENFIS. For C4.5, the number of tree leaves is recorded because it is equivalent to the number of crisp decision rules. For Naive Bayes classifier, the number of rules is not applicable. For MLP, the number of hidden neurons, which is systematically determined by the total number of input dimensions and the number of output classes, is recorded. For RBF, the number of radial basis neurons, which is pre-determined by the number of clusters in every output class, is recorded.

In the comparison tables, the winner of any particular column is highlighted in bold. The winner is either the largest accuracy value or the least amount of information employed. Because multiple experiments are conducted on each model for each diagnosis, stability (standard deviations) of every model is also recorded and compared. Winners in stability are also highlighted in bold.

In this paper, ANFIS utilizes clustering results derived by the employed Fuzzy C-Means (FCM) clustering technique [48]. In this way, ANFIS is more comparable to GARSINFIS because the number of rules employed by ANFIS is greatly reduced. However, extra efforts are taken to carefully select the optimal number of pre-defined clusters through trial-and-error. This is a great disadvantage of ANFIS, because its network structure is not self-

organized but has to be pre-defined by employing additional knowledge.

The Evolving Clustering Method (ECM) [49] employed by DENFIS constructs clusters based on the distances among data samples in the high-dimensional space. ECM often derives a large number of clusters and as a result, DENFIS employs a large number of rules. Generally speaking, DENFIS often fails to alleviate the curse of dimensionality problem [32].

7.4. Diagnosis-1: Normal vs. Abnormal

In diagnosis-1, 22 samples are normal and 149 are abnormal. The population size of GARSC is set to 100 and the number of generations is set to 10. Table 8 reports the performance of GARSINFIS on ovarian cancer diagnosis-1 data set. On average, GARSINFIS utilizes 1.7 input features and employs 2 rules (1 negative and 1 positive) to achieve 98.95% accuracy on the testing data sets. Each rule employs averagely $2.7/2 = 1.35$ arguments in the antecedent part. The average training time is less than 1.5 minutes.

Table 8 presents the astonishing performance of GARSINFIS on ovarian cancer diagnosis-1 data set. All ten experiments achieve 100% specificity, which means all normal patients are correctly diagnosed. Furthermore, all seven experiments that employ two input features achieve 100% sensitivity. The selected features among all ten experiments are highly concentrated that TPS antigen is selected for ten times and CA-125 for seven times.

To illustrate the derived simple yet convincing inference rules, fuzzy membership functions and the corresponding rules employed in experiment-1 are shown in Fig. 6 and Table 9 respectively.

The most simple diagnosis criterion adopted by doctors is that ovarian cancer usually happens in women over age 50, but it is not reliable be-

Table 8: Summary of applying GARSINFIS on ovarian cancer diagnosis-1 data set.

Experiment Index	Accuracy (%)	Sensitivity (%)	Specificity (%)	nof	nor	noa	Time (seconds)
1	100.0	100.0	100.0	2	2	3	62.76
2	100.0	100.0	100.0	2	2	3	84.58
3	100.0	100.0	100.0	2	2	3	71.50
4	100.0	100.0	100.0	2	2	3	66.75
5	94.74	94.00	100.0	1	2	2	75.05
6	100.0	100.0	100.0	2	2	3	74.36
7	98.25	97.96	100.0	1	2	2	68.28
8	100.0	100.0	100.0	2	2	3	75.92
9	100.0	100.0	100.0	2	2	3	72.54
10	96.49	95.92	100.0	1	2	2	80.93
mean	98.95	98.79	100.0	1.7	2	2.7	73.27
std	1.89	2.16	0.00	0.48	0.00	0.48	6.49

nof: number of features; nor: number of rules; noa: number of (total) arguments

Figure 6: One set of the derived fuzzy membership functions of ovarian cancer diagnosis-1 data set.

Table 9: One set of the derived fuzzy rules of ovarian cancer diagnosis-1 data set.

IF TPS antigen is low and CA-125 is low	THEN the sample is normal
IF TPS antigen is high	THEN the sample is abnormal

cause more deterministic tests are required to make a definite diagnosis [50]. GARSINFIS automatically discovers this knowledge because the input feature “age” is not utilized in all experiments. “A normal CA-125 level is less than 35. However, this test is not always an accurate way to tell whether a woman has ovarian cancer” [51] and “the combination of TPS and CA-125 will increase the accuracy” [52]. These statements validate the correctness of GARSINFIS in terms of feature selection because only the two doctor suggested features are utilized. Furthermore, in [52, 53], the cut-off values for TPS antigen and CA-125 are given as 97U/l and 35mU/ml respectively. These diagnosis criteria are fulfilled in Fig. 6. Therefore, the rules derived by GARSINFIS are consistent with those doctor adopted ones. This significant finding best illustrates the superior property of GARSINFIS, which systematically discovers intrinsic knowledge embedded in the data set without any prior knowledge or human intervention.

Table 10 reports the performance comparisons of GARSINFIS against other benchmarking models on ovarian cancer diagnosis-1 data set. Although GARSINFIS does not achieve the highest accuracy and sensitivity, the differences are as small as approximately 0.5 and 0.6 data sample respectively. However, only GARSINFIS and C4.5 perform feature selection and all the other models utilize all the 28 input features. Although C4.5 utilizes lesser number of input features, its accuracy is lower than that of GARSINFIS. More importantly, among all ten C4.5 decision trees, CA-125 is only utilized

Table 10: Benchmarks on ovarian cancer diagnosis-1 data set.

Model		Accuracy (%)	Sensitivity (%)	Specificity (%)	Features	Rules
GARSINFIS	mean	98.95	98.79	100.0	1.7	2.0
	std	1.89	2.16	0.00	0.48	0.00
C4.5	mean	98.07	97.79	100.0	1.1	2.1
	std	1.00	1.15	0.00	0.32	0.32
Naive Bayes	mean	98.25	100.0	86.30	28	N.A.
	std	1.43	0.00	11.69		
MLP	mean	99.82	99.80	100.0	28	17
	std	0.55	0.65	0.00		
RBF	mean	99.82	100.0	98.57	28	2
	std	0.55	0.00	4.52		
ANFIS	mean	99.30	99.20	100.0	28	2.0
	std	1.23	1.42	0.00		0.00
DENFIS	mean	99.82	99.80	100.0	28	39.3
	std	0.55	0.63	0.00		2.67

three times and TPS antigen is not utilized at all. This implies that although C4.5 utilizes lesser number of input features (1.1 comparing to 1.7 utilized by GARSINFIS) in this ovarian cancer diagnosis-1 application, it performs worse in terms of knowledge discovery.

7.5. Diagnosis-2: Benign vs. Malignant

In diagnosis-2, 78 samples are benign and 71 are malignant. The population size of GARSC is set to 100 and the number of generations is set to 20. Table 11 reports the performance of GARSINFIS on ovarian cancer diagnosis-2 data set. On average, GARSINFIS utilizes 5 input features and employs 8.5 rules (2.7 negative ones and 5.8 positive ones) to achieve 71.26% accuracy on the testing data sets. Each rule employs averagely $20.8/8.5 = 2.45$ arguments in the antecedent part. The average training time is less than 39.5 minutes.

Table 11: Summary of applying GARSINFIS on ovarian cancer diagnosis-2 data set.

Experiment Index	Accuracy (%)	Sensitivity (%)	Specificity (%)	nof	nor	noa	Time (seconds)
1	72.00	58.33	84.62	5	7	18	2186.12
2	68.00	70.83	65.38	4	6	13	2405.43
3	62.00	58.33	65.28	6	11	29	2258.81
4	74.00	70.83	76.92	6	12	33	2552.52
5	77.55	65.22	88.46	5	9	20	2185.78
6	72.00	70.83	73.08	6	10	22	2203.56
7	69.39	56.52	80.77	5	8	19	2618.34
8	68.00	66.67	69.23	4	7	16	2136.77
9	75.71	69.57	80.77	4	6	15	2287.63
10	74.00	75.00	73.08	5	9	23	2696.64
mean	71.26	66.21	75.77	5.0	8.5	20.8	2353.16
std	4.52	6.43	7.92	0.82	2.07	6.25	202.51

nof: number of features; nor: number of rules; noa: number of (total) arguments

Among all input features, D-Dimer is employed in nine experiments. This again proves the capability of GARSINFIS in terms of feature selection because D-Dimer is utilized by doctors. In [54], it is stated that “in patients with ovarian cancer before they receive chemotherapy, the level of fibrin degradation products (D-Dimer) is correlated with the tumor load.”

Table 12 reports the performance comparisons of GARSINFIS against other benchmarking models on ovarian cancer diagnosis-2 data set. Comparing Table 12 to Table 10, performances of all models drop significantly. However, this observation is within expectation because diagnosis-2 data set is the most complex one among all four diagnoses. In nature, benign and malignant tumors are easily separable using extra information such as ultrasound image or CT scan. However, these more distinctive features are not included in this data set, which leads to the highest accuracy of 71.26% on

Table 12: Benchmarks on ovarian cancer diagnosis-2 data set.

Model		Accuracy (%)	Sensitivity (%)	Specificity (%)	Features	Rules
GARSINFIS	mean	71.26	66.21	75.77	5.0	8.5
	std	4.52	6.43	7.92	0.82	2.07
C4.5	mean	63.89	63.29	64.62	7.7	10.6
	std	3.70	7.86	9.56	1.34	1.08
Naive Bayes	mean	68.41	55.27	80.38	28	N.A.
	std	3.81	6.30	8.59		
MLP	mean	63.78	59.92	67.31	28	17
	std	4.00	7.52	6.60		
RBF	mean	68.01	59.86	75.38	28	6
	std	5.94	10.15	7.52		
ANFIS	mean	68.41	59.07	76.92	28	5.8
	std	7.30	11.69	6.79		1.32
DENFIS	mean	66.80	62.87	70.38	28	37.9
	std	8.14	10.42	8.32		1.37

the testing data set.

GARSINFIS performs best in terms of both accuracy and sensitivity and second best in terms of specificity (the difference is as small as approximately 0.3 data sample). Furthermore, although GARSINFIS utilizes lesser number of input features and employs lesser number of rules, it achieves higher accuracy when comparing to C4.5. Among all models, ANFIS employs the least number of rules. However, it does not perform feature selection and requires extra efforts to determine the number of clusters for the employed FCM clustering technique through trial-and-error.

7.6. Diagnosis-3: Borderline vs. beyond Borderline

In diagnosis-3, 10 samples are borderline and 61 are beyond borderline. The population size of GARSC is set to 200 and the number of generations is set to 10. Table 13 reports the performance of GARSINFIS on ovarian cancer

diagnosis-3 data set. On average, GARSINFIS utilizes 2.4 input features and employs 2.7 rules (1.2 negative ones and 1.5 positive ones) to achieve 81.58% accuracy on the testing data sets. Each rule employs averagely $4.1/2.7 = 1.52$ arguments in the antecedent part. The average training time is less than 1.5 minutes.

Table 13: Summary of applying GARSINFIS on ovarian cancer diagnosis-3 data set.

Experiment Index	Accuracy (%)	Sensitivity (%)	Specificity (%)	nof	nor	noa	Time (seconds)
1	82.61	89.47	50.00	2	3	4	81.96
2	83.33	95.00	25.00	3	3	4	73.87
3	78.26	80.00	66.67	2	2	3	70.71
4	83.33	95.00	25.00	3	3	5	76.68
5	82.61	100.0	0.00	2	2	3	71.37
6	79.17	76.19	100.0	2	2	3	70.70
7	91.67	100.0	50.00	3	3	6	82.16
8	82.61	95.00	0.00	2	3	4	80.46
9	73.91	80.00	33.33	3	3	5	83.30
10	78.26	94.74	0.00	2	3	4	78.93
mean	81.58	90.54	35.00	2.4	2.7	4.1	77.01
std	4.69	8.73	32.58	0.52	0.48	0.99	5.03

nof: number of features; nor: number of rules; noa: number of (total) arguments

Sensitivity on the testing data set is 90.54% on average because ovarian cancer in late stages is easier to be discovered. Once the cancer grows beyond the borderline size, it will develop rapidly in ovary and may spread to other organs. At that time, the characteristics of ovarian cancer become increasingly distinctive. However, specificity on the testing data set varies drastically from 0% to 100% with the average below 50% because there are not enough data instances. In each experiment, only six or seven instances from the borderline class are available for training and their common char-

acteristics may not be distinctive enough. On the other hand, there are only three or four instances from the borderline class in the testing data set. Both of these lead to the drastic variations of specificities on the testing data set.

Among all input features, the most frequently selected feature is PAI activity, which is utilized in five experiments. PAI antigen and CA-125 are both utilized in four experiments respectively. These features selected by GARSINFIS once again inosculate with the ones that doctor uses. In [55], it is stated that PAI activity and antigen in blood and tumor fluids are examined to determine the seriousness of the tumor in ovary.

Table 14: Benchmarks on ovarian cancer diagnosis-3 data set.

Model		Accuracy (%)	Sensitivity (%)	Specificity (%)	Features	Rules
GARSINFIS	mean	81.58	90.54	35.00	2.4	2.7
	std	4.69	8.73	32.58	0.52	0.48
C4.5	mean	76.07	85.35	25.00	2.7	4.0
	std	6.96	7.39	19.02	0.68	0.82
Naive Bayes	mean	67.09	73.23	33.33	28	N.A.
	std	8.08	7.40	24.04		
MLP	mean	76.50	85.86	25.00	28	17
	std	6.07	8.72	22.03		
RBF	mean	75.27	86.54	14.17	28	2
	std	8.73	10.45	21.89		
ANFIS	mean	71.79	75.25	52.78	28	2.2
	std	10.05	11.62	30.44		0.42
DENFIS	mean	80.77	86.87	47.22	28	28.8
	std	5.97	7.99	31.54		1.23

Table 14 reports the performance comparisons of GARSINFIS against other benchmarking models on ovarian cancer diagnosis-3 data set. GARSINFIS performs best in terms of both accuracy and sensitivity and third best in terms of specificity (the difference is as small as approximately 0.6 data

sample). Furthermore, although GARSINFIS utilizes lesser number of input features and employs lesser number of rules, it achieves higher accuracy when comparing to C4.5. Among all models, ANFIS employs the least number of rules. However, the credit belongs to the employed FCM clustering technique, which determines the number of clusters through trial-and-error.

7.7. Diagnosis-4: Early Stage vs. Late Stage

In diagnosis-4, 19 samples are in early stage and 42 are in late stage. The population size of GARSC is set to 200 and the number of generations is set to 20. Table 15 reports the performance of GARSINFIS on ovarian cancer diagnosis-3 data set. On average, GARSINFIS utilizes 2.5 input features and employs 2.7 rules (1.5 negative ones and 1.2 positive ones) to achieve 83.31% accuracy on the testing data sets. Each rule employs averagely $3.9/2.7 = 1.44$ arguments in the antecedent part. The average training time is less than 4 minutes. Among all input features, D-Dimer is utilized in all ten experiments and CA-125 is utilized in eight. These two features are the ones suggested by doctors [54].

Table 16 reports the performance comparisons of GARSINFIS against other benchmarking models on ovarian cancer diagnosis-4 data set. GARSINFIS performs best in terms of both accuracy and sensitivity and second best in terms of specificity (the difference is as small as approximately 0.3 data sample). Furthermore, GARSINFIS utilizes the least number of input features and employs the least number of rules.

Table 15: Summary of applying GARSINFIS on ovarian cancer diagnosis-4 data set.

Experiment Index	Accuracy (%)	Sensitivity (%)	Specificity (%)	nof	nor	noa	Time (seconds)
1	80.95	80.00	83.33	3	3	4	206.57
2	75.00	84.62	57.14	2	2	3	176.61
3	85.00	92.86	66.67	3	3	5	187.80
4	71.43	71.43	71.43	2	3	4	221.09
5	90.00	85.71	100.0	3	3	4	235.06
6	85.00	84.62	85.71	3	3	4	237.07
7	85.71	86.67	83.33	2	3	4	174.93
8	85.00	85.71	83.33	2	2	3	223.05
9	90.00	100.0	66.67	3	3	5	232.96
10	85.00	85.71	83.33	2	2	3	241.67
mean	83.31	85.73	78.09	2.5	2.7	3.9	213.68
std	5.99	7.42	12.44	0.53	0.48	0.74	25.60

nof: number of features; nor: number of rules; noa: number of (total) arguments

Table 16: Benchmarks on ovarian cancer diagnosis-4 data set.

Model		Accuracy (%)	Sensitivity (%)	Specificity (%)	Features	Rules
GARSINFIS	mean	83.31	85.73	78.09	2.5	2.7
	std	5.99	7.42	12.44	0.53	0.48
C4.5	mean	69.95	77.14	53.97	3.6	4.6
	std	10.66	11.70	17.53	0.84	0.84
Naive Bayes	mean	77.83	80.00	73.02	28	N.A.
	std	6.17	8.85	16.30		
MLP	mean	78.33	82.86	68.25	28	17
	std	6.21	9.53	16.54		
RBF	mean	75.33	82.31	61.19	28	2
	std	5.89	7.49	22.16		
ANFIS	mean	75.86	80.71	65.08	28	2.9
	std	10.34	9.19	24.49		0.99
DENFIS	mean	78.33	68.25	82.86	28	26.7
	std	6.00	20.30	5.82		1.06

7.8. Summary on Experimental Results

The features selected by doctors, GARSINFIS, and C4.5 are summarized in Table 17. It is obvious that GARSINFIS performs better than C4.5 in terms of feature selection and more importantly, the features selected by GARSINFIS is more consistent with the ones suggested by doctors.

Table 17: Summary of selected features on ovarian cancer diagnosis data set.

Diagnosis	1	2	3	4
Doctor suggests	CA-125 [51–53] TPS antigen [52]	D-Dimer [46]	PAI activity [55] PAI antigen [55]	D-Dimer [54] CA-125 [54]
GARSINFIS selects	TPS antigen (10) CA-125 (7)	D-Dimer (9)	PAI activity (5) PAI antigen (4) CA-125 (4)	D-Dimer (10) CA-125 (8)
C4.5 selects	ATAC (7) CA-125 (3)	D-Dimer (8) CA-125 (8)	CA-125 (5)	D-Dimer (5) CA-125 (5)

the number in () denotes the number of occurrences among 10 experiments.

To systematically perform diagnoses on new input data instances, all the four diagnoses are integrated to form one decision support system, which is illustrated in Fig. 5. Table 18 summaries the accumulated accuracies of all models and Fig. 7 visualizes Table 18 with respect to GARSINFIS.

Table 18: Accumulated accuracies of different ovarian cancer diagnosis systems

Model	Normal	Benign	Borderline	Stage I & II	Stage III & IV	Overall
GARSINFIS	100.0%	74.85%	22.89%	50.77%	46.24%	111.74
C4.5	100.0%	63.19%	15.47%	28.51%	40.75%	95.37
Naive Bayes	86.30%	80.38%	18.42%	29.55%	32.38%	102.74
MLP	100.0%	67.18%	14.95%	35.04%	42.54%	100.42
RBF	98.57%	75.38%	8.48%	42.64%	31.70%	102.75
ANFIS	100.0%	76.30%	30.93%	28.70%	35.59%	105.01
DENFIS	100.0%	70.24%	29.63%	45.16%	37.20%	103.95
No. of samples	22	78	10	19	42	171

Figure 7: Visualization of accuracies on ovarian cancer diagnosis application.

It is shown in Fig. 7 that GARSINFIS is always one of the best performing models and it is stated in the last column of Table 18 that GARSINFIS achieves the most number of correct diagnoses if the testing data set has the same distribution presented in Table 6. In each diagnosis, GARSINFIS utilizes a small number of features that varies from 1.7 to 5 and employs a small number of rules that varies from 2 to 8.5. Therefore, GARSINFIS utilizes only the most essential knowledge to achieve a high level of accuracy. More importantly, it is shown in Table 17 that the rules employed by GARSINFIS are consistent with those suggested by doctors. This proves that GARSINFIS can automatically discover the intrinsic knowledge embedded in the ovarian cancer diagnoses data set without any prior information or human intervention.

8. Conclusions

Ovarian cancer diagnosis is an important study because early detection and accurate staging are the keys to significantly increase the survival rate of the patient. In this paper, we propose a novel hybrid intelligent system, which derives simple yet convincing fuzzy inference rules to diagnose ovarian cancer and identify its stage according to the level of seriousness. Our proposed self-organizing model is named Genetic Algorithm and Rough Set Incorporated Neural Fuzzy Inference System (GARSINFIS), which utilizes the inference rule base automatically derived by our proposed Genetic Algorithm based Rough Set Clustering (GARSC) technique. By fusing various soft computing techniques together, we combine the advantages of the individuals and alleviate certain limitations. Before GARSINFIS is applied to the real-world cases collected from hospital, two established medical data sets are benchmarked against other established models, which also focus on the compactness of the derived inference rules. All experimental results are encouraging, especially the ovarian cancer diagnoses. Because GARSINFIS only requires a limited number of control parameters and constraints and no human intervention and expert guidance to achieve the most number of correct diagnoses when benchmarked against other models. More importantly, it automatically selects features and derives rules that are suggested and applied by doctors.

Ovarian cancer data sets that consist of DNA microarray gene expressions [8, 14, 16] and actual images (CT, MRI, ultrasound, and etc.) [56–58] besides the blood test results are also studied in literature. However, these information corresponding to the samples in our data set is not available. We

will look into the possibilities to obtain an ovarian cancer diagnoses data set with richer distinctive information and we believe at that time our proposed model will increase its accuracy and become more reliable.

References

- [1] M. Schena, D. Shalon, R. W. Davis, P. O. Brown, Quantitative monitoring of gene expression patterns with a complementary DNA microarray, *Science* 270 (5235) (1995) 467–470.
- [2] ACS, American Cancer Society, Available online: <http://www.cancer.org/> (2013).
- [3] eMCH, e-Medicine Consumer Health, Available online: <http://www.emedicinehealth.com/> (2013).
- [4] P. Petignat, F. Joris, D. Faltin, Importance of the general practitioner in the early detection of ovarian cancer, *Gynecologic Oncology* 90 (2003) 491–500.
- [5] P. G. Anastasiadis, K. N. Romanidis, A. Polichronidis, N. G. Koutlaki, D. Tamiolakis, K. Simopoulos, The contribution of rapid intraoperative cytology to the improvement of ovarian cancer staging, *Gynecologic Oncology* 86 (3) (2002) 244–249.
- [6] A. Jermal, T. Murray, A. Samuels, A. Ghafoor, E. Ward, M. J. Thun, Cancer statistics, *Cancer Journal for Clinician* 53 (1) (2003) 5–26.
- [7] A. Valerio, D. Basso, S. Mazza, G. Baldo, A. Tiengo, S. Pedrazzoli, R. Seraglia, M. Plebani, Serum protein profiles of patients with pancre-

- atic cancer and chronic pancreatitis: searching for a diagnostic protein pattern, *Rapid Communications in Mass Spectrometry* 15 (24) (2001) 2420–2425.
- [8] Z.-J. Lee, An integrated algorithm for gene selection and classification applied to microarray data of ovarian cancer, *Artificial Intelligence in Medicine* 42 (1) (2008) 81–93.
- [9] E. F. Petricoin, D. K. Ornstein, C. P. Paweletz, A. Ardekani, P. S. Hackett, B. A. Hitt, A. Velasco, C. Trucco, L. Wiegand, K. Wood, et al., Serum proteomic patterns for detection of prostate cancer, *Journal of the National Cancer Institute* 94 (20) (2002) 1576–1578.
- [10] T. C. W. Poon, T. T. Yip, A. T. C. Chan, C. Yip, V. Yip, T. S. K. Mok, C. C. Y. Lee, T. W. T. Leung, S. K. W. Ho, P. J. Johnson, Comprehensive proteomic profiling identifies serum proteomic signatures for detection of hepatocellular carcinoma and its subtypes, *Clinical Chemistry* 49 (5) (2003) 752–760.
- [11] B.-L. Adam, Y.-S. Qu, J. W. Davis, M. D. Ward, M. A. Clements, L. H. Cazares, O. J. Semmes, P. F. Schellhammer, Y. Yasui, Z.-D. Feng, et al., Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men, *Cancer Research* 62 (13) (2002) 3609–3614.
- [12] Y.-S. Qu, B.-L. Adam, Y. Yasui, M. D. Ward, L. H. Cazares, P. F. Schellhammer, Z.-D. Feng, O. J. Semmes, G. L. Wright, Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral

- serum profiles discriminates prostate cancer from noncancer patients, *Clinical Chemistry* 48 (10) (2002) 1835–1843.
- [13] G. Ball, S. Mian, F. Holding, R. O. Allibone, J. Lowe, S. Ali, G. Li, S. McCardle, I. O. Ellis, C. Creaser, et al., An integrated approach utilizing artificial neural networks and SELDI mass spectrometry for the classification of human tumours and rapid identification of potential biomarkers, *Bioinformatics* 18 (3) (2002) 395–404.
 - [14] M.-H. Tsai, C.-H. Lai, S.-S. Yu, A statistical and learning based onco-gene detection and classification scheme using human cDNA expressions for ovarian carcinoma, *Expert Systems with Applications* 38 (8) (2011) 10066–10074.
 - [15] T. Bernd, M. Kleutges, A. Kroll, Nonlinear black box modelling - fuzzy networks versus neural networks, *Neural Computing and Applications* 8 (1999) 151–162.
 - [16] T. Z. Tan, C. Quek, G. S. Ng, Ovarian cancer diagnosis by hippocampus and neocortex-inspired learning memory structures, *Neural Networks* 18 (special issue) (2005) 818–825.
 - [17] S. W. Tung, C. Quek, C. Guan, SoHyFIS-Yager: A self-organizing Yager based hybrid neural fuzzy inference system, *Expert Systems with Applications* 39 (17) (2012) 12759–12771.
 - [18] K. H. Quah, C. Quek, Mces: a novel Monte Carlo evaluative selection approach for objective feature selections, *IEEE Transactions on Neural Networks* 18 (2) (2007) 431–448.

- [19] A. Quteishat, C. P. Lim, A modified fuzzy min-max neural network with rule extraction and its application to fault detection and classification, *Applied Soft Computing* 8 (2) (2008) 985–995.
- [20] K. S. Yap, C. P. Lim, J. Mohamad-Saleh, An enhanced generalized adaptive resonance theory neural network and its application to medical pattern classification, *Journal of Intelligent Fuzzy Systems* 21 (1, 2) (2010) 65–78.
- [21] A. R. Yildiz, Hybrid Taguchi-differential evolution algorithm for optimization of multi-pass turning operations, *Applied Soft Computing* 13 (3) (2013) 1433–1439.
- [22] A. R. Yildiz, A new hybrid artificial bee colony algorithm for robust optimal design and manufacturing, *Applied Soft Computing* 13 (5) (2013) 2906–2912.
- [23] A. R. Yildiz, A new hybrid differential evolution algorithm for the selection of optimal machining parameters in milling operations, *Applied Soft Computing* 13 (3) (2013) 1561–1566.
- [24] A. R. Yildiz, A comparative study of population-based optimization algorithms for turning operations, *Information Sciences* 210 (2012) 81–88.
- [25] Z. Pawlak, Rough sets, *International Journal of Information and Computer Science* 11 (5) (1982) 341–356.
- [26] L. A. Zadeh, Fuzzy sets, *Information Control* 8 (1965) 338–353.

- [27] Y. Y. Yao, A comparative study of rough sets and fuzzy sets, *Journal of Information Sciences* 109 (1998) 227–242.
- [28] Z. Pawlak, *Rough Sets*, Kluwer Academic Publishers, 1991.
- [29] Q. Shen, A. Chouchoulas, Rough set-based dimensionality reduction for supervised and unsupervised learning, *Applied Mathematics and Computer Science*, Special Issue on Rough Sets and Their Applications 11 (3) (2001) 583–601.
- [30] J. Holland, *Adaptation in Natural and Artificial Systems*, MIT Press, 1975.
- [31] D. E. Goldberg, *Genetic Algorithms in Search, Optimization, and Machine Learning*, Addison-Wesley Publishing, 1989.
- [32] I. Eccles, M. Su, Illustrating the curse of dimensionality numerically through different data distribution models, in: *Proceedings of International Symposium on Information and Communication Technologies*, 2004, pp. 232–237.
- [33] R. P. Paiva, A. Dourado, Interpretability and learning in neuro-fuzzy systems, *Fuzzy Sets and Systems* 147 (1) (2004) 17–38.
- [34] K. Deb, *Multi-Objective Optimization Using Evolutionary Algorithms*, John Wiley & Sons, 2001.
- [35] J. Zahalka, F. Zelezny, An experimental test of occams razor in classification, *Machine Learning* 82 (3) (2010) 475–481.

- [36] B. L. Miller, D. E. Goldberg, Genetic algorithms, tournament selection, and the effects of noise, *Complex Systems* 9 (1996) 193–212.
- [37] T. Takagi, M. Sugeno, Fuzzy identification of systems and its applications to modelling and control, *IEEE Transactions on Systems, Man and Cybernetics, Part B* 15 (1) (1985) 116–132.
- [38] M. Sugeno, G. T. Kang, Structure identification of fuzzy model, *Fuzzy Sets and Systems* 28 (1988) 13–33.
- [39] C. T. Lin, C. S. G. Lee, *Neural Fuzzy Systems*, Prentice-Hall, 1996.
- [40] K. Bache, M. Lichman, UCI machine learning repository (2013).
URL <http://archive.ics.uci.edu/ml>
- [41] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, 1993.
- [42] G. A. Carpenter, S. Grossberg, D. B. Rosenm, Fuzzy art: Fast stable learning and categorization of analog patterns by an adaptive resonance system, *Neural Networks* 4 (1991) 759–771.
- [43] FIGO, International Federation of Gynecology and Obstetrics, Available online: <http://www.figo.org/> (2013).
- [44] S. M. Ross, *Introduction to Probability Models*, Academic Press, 1985.
- [45] S. Haykin, *Neural networks: A Comprehensive Foundation*, Prentice Hall, 1998.

- [46] J. S. R. Jang, ANFIS: Adaptive network-based fuzzy inference systems, *IEEE Transactions on Systems, Man and Cybernetics, Part B* 23 (1993) 650–684.
- [47] N. K. Kasabov, Q. Song, Denfis: Dynamic evolving neural-fuzzy inference system and its application for time-series prediction, *IEEE Transactions on Fuzzy Systems* 10 (2002) 144–154.
- [48] J. C. Bezdek, *Pattern Recognition With Fuzzy Objective Function Algorithms*, Kluwer Academic Publishers, 1981.
- [49] Q. Song, N. Kasabov, ECM: A novel on-line, evolving clustering method and its applications, in: *Proceedings of Conference on Artificial Neural Networks and Expert Systems*, 2001, pp. 87–92.
- [50] NLM, U.S. National Library of Medicine, Available online: <http://www.nlm.nih.gov/> (2013).
- [51] AAFP, American Academy of Family Physicians, Available online: <http://www.familydoctor.org/> (2013).
- [52] G. Sliutz, C. Tempfer, C. Kainz, G. Mustafa, G. Gitsch, H. Koelbl, C. Biegelmayer, Tissue polypeptide specific antigen and cancer associated serum antigen in the follow-up of ovarian cancer, *Anticancer Research* 15 (3) (1995) 1127–1129.
- [53] L. R. Medeiros, D. D. Rosa, M. I. Da Rosa, M. C. Bozzetti, Accuracy of CA 125 in the diagnosis of ovarian tumors: a quantitative systematic review, *European Journal of Obstetrics Gynecology and Reproductive Biology* 142 (2) (2009) 99–105.

- [54] S. Mirshahi, E. Pujade-Lauraine, C. Soria, M. Mirshahi, J. Fretault, A. Bernadou, J. Soria, D-dimer and CA 125 levels in patients with ovarian cancer during antineoplastic therapy, *Cancer* 69 (9) (1992) 2289–2292.
- [55] B. Casslen, T. Bossmar, I. Lecander, B. Astedt, Plasminogen activators and plasminogen activator inhibitors in blood and tumour fluids of patients with ovarian cancer, *European Journal of Cancer* 30A (9) (1994) 1302–1309.
- [56] I. Jacobs, A. P. Davies, J. Bridges, I. Stabile, T. Fay, A. Lower, J. G. Grudzinskas, D. Oram, Prevalence screening for ovarian cancer in postmenopausal women by CA 125 measurement and ultrasonography, *British Medical Journal* 306 (1993) 1030–1034.
- [57] P. Gu, L.-L. Pan, S.-Q. Wu, L. Sun, G. Huang, CA 125, PET alone, PET-CT, CT and MRI in diagnosing recurrent ovarian carcinoma: a systematic review and meta-analysis, *European Journal of Radiology* 71 (1) (2009) 164–174.
- [58] Y. Yang, X. Li, T.-H. Wang, P. D. Kumavor, A. Aguirre, K. K. Shung, Q.-F. Zhou, M. Sanders, M. Brewer, Q. Zhu, Integrated optical coherence tomography, ultrasound and photoacoustic imaging for ovarian tissue characterization, *Biomedical Optics Express* 2 (9) (2011) 2551–2561.