Chapter 7
Hybrid Metaheuristics for Medical Data Classification

Sarab Al-Muhaideb and Mohamed El Bachir Menai

Abstract. Medical data exhibit certain features that make their classification stand out as a distinct field of research. Several medical classification tasks exist, among which medical diagnosis and prognosis are most common. Deriving a medical classification is a complex task. In particular, the rule–discovery problem is NP-hard. Identifying the most suitable strategy for a particular medical classification problem along with its optimal parameters is no less difficult. Heuristics and meta-heuristics are normally applied to approximate its solution. This chapter reviews hybrid metaheuristics for medical data classification task, particularly diagnosis and prognosis, and their application to model selection, including parameter optimization and feature subset selection.

Keywords: Medical data classification, medical data complexity, evolutionary computation, swarm intelligence, model selection, model optimization, hybrid metaheuristics, artificial neural networks.

7.1 Introduction

Modern clinical information systems store extensive amount of data in medical databases. This encourages the extraction of useful knowledge from these databases providing valuable insight for medical decision support. A branch of data mining, known as medical data mining, is currently considered one of the most popular research subjects in the data mining community [68]. This, in part, is due to the

Sarab Al-Muhaideb · Mohamed El-Bachir Menai
Department of Computer Science, College of Computer and Information Sciences,
King Saud University, P.O. Box 51178, Riyadh 11543, Saudi Arabia
e-mail: salmuhaideb@acm.org,
menai@ksu.edu.sa

E.-G. Talbi (Ed.): Hybrid Metaheuristics, SCI 434, pp. 187–217.
springerlink.com

societal significance of the subject and also to the computational challenge it presents. Normally, there exist a dataset of historic data describing a particular medical disorder. Such datasets consist of records of patients' data relating to demographic, clinical and pathological data, along with results of particular investigations that were collected for the diagnosis and prognosis of a particular medical disorder. These medical datasets are typically incomplete, noisy, imbalanced and inexact [55]. Developing a computational diagnostic or a prognostic system is thus a challenging task.

This chapter is not intended to present a review of medical data classification techniques, but rather to introduce a snapshot of data mining techniques used to aid medical decision making. Several computational techniques have been proposed including machine learning, evolutionary computation and statistical techniques. Since each of these techniques have their own advantages and drawbacks, they are commonly hybridized in search of a more robust solution. Metaheuristics can be effective and efficient tools. They are well known for solving various optimization problems, for their adaptation ability and for their ability to produce good solutions in reasonable time and memory requirements. The chapter starts with a brief introduction of the classification problem in general, followed by medical data classification in particular. Next, features and challenges of medical datasets that make their classification stand out as a separate domain are explored. Based on that, the computational complexity of medical data classification is analyzed. Next, light is shed on some state-of-the-art solutions for medical data classification, in particular, hybrid meta-heuristics. It is possible to classify the hybrid metaheuristic techniques used for medical data classification into two broad categories according to their purpose:

1. Model learning and optimization; where the objective is to learn the classification hypothesis.
2. Model selection; that is selecting the model that best describes a dataset. This may include parameter and hyper-parameter optimization, neural network weight optimization, or feature subset selection, etc.

Each of these categories is illustrated by published work.

7.2 The Classification Problem

Classification aims at capturing hidden regulations and/or relations between the attributes (predictor features) in a set of class-labeled instances. These relations and/or regulations are modeled producing a general hypothesis. The resulting hypothesis is next applied to unseen future instances, with known predictor features and unknown class labels.The goal is to automatically make predictions about the class of those future instances [49]. Formally, given a set of training instances $\mathbf{D}_n$ with the form $\{(x_1,y_1),(x_2,y_2),\ldots,(x_n,y_n)\}$, the task is to approximate or project a function; $f(x)$, where $x \in \Re^m$ is a vector of attributes or predictor features of the form $\langle x_{i1}, x_{i2}, \ldots, x_{im} \rangle$, and y is the expected output (i.e. class) for the given x vector. Normally, y is drawn from a discrete set of classes [71]. The discovered model can be represented in different forms. Production rules in the form of $(IF\langle condition\rangle THEN\langle class\rangle)$ are often used. Other forms include decision trees (DTs) and artificial neural networks (ANNs).

Errors in classification may be in one of three cases [68]. Type-I error (false-positive) occurs when the system erroneously classifies a case as positive when in fact, it is not. For example, in a diagnosis scenario, a patient is wrongly labeled with a certain disease. Type-II error (false-negative), on the other hand, describes missing an existent positive. For example, a patient who is affected by a certain disease is diagnosed as disease-free. Usually, improving one type of error comes on the expense of the other [16]. In practice, the significance of these error costs vary with the application itself. For example, in life threatening medical conditions that require prompt intervention, missing a diagnosis (a false-negative) might result in a waste of time that may lead to life losses or at least cases that are not treated properly. On the other hand, a false-positive may result in unnecessary procedures, anxiety and financial costs [68]. The last type of error is the unclassifiable error. In this case, the system is unable to classify a case, possibly due to the lack of historic data.

There are many approaches to estimate the expected error of the classification model. Computing the error on the training set itself is an optimistic estimator of the true error [54]. In the training–testing method, the data set is normally split into two partitions called training and testing sets respectively. The most common technique is called the k-fold cross-validation [83]. Here, the whole data set $\mathbf{D}_n$ is partitioned into k disjoint folds, each of size k/n. Cross-validation is done k times each using $k-1$ folds for training the model and the one fold left out of the training phase is used as a test set. Each time a different fold is used as a test set. Results are then averaged over the k iterations.

Different performance metrics are used to measure the effectiveness of a classifier with respect to a given data set. The prior and posterior probabilities, also known as the Sensitivity (Sn), Specificity (Sp) and Precision (P) [71] are among the most commonly used.

Let the number of positive instances correctly classified be denoted TP, the number of positive instances incorrectly classified into negative FN. Similarly, the number of negative instances correctly classified as negative TN and those falsely classified into positive as FP. Sensitivity measures the proportion of positive samples being correctly classified as positive (7.1).

$$Sn = \frac{TP}{TP + FN} \tag{7.1}$$

Specificity, on the other hand, measures the proportion of negative samples being correctly recognized as negative (7.2).

$$Sp = \frac{TN}{TN + FP} \tag{7.2}$$

Precision (classification accuracy [26, 55] measures the proportion of samples being correctly classified (7.3).

$$P = \frac{TP + TN}{TP + TN + FP + FN} \tag{7.3}$$

Using Bayes theorem , it can be shown that P is entirely dependent on the values of Sp and Sn only if the data set is balanced [71]. A tradeoff between the hit rate (Sn, plotted on the Y-axis) and false alarm rate ($1 - Sp$, plotted on the X-axis) can be illustrated by the receiver operating characteristic curve. Each classification algorithm has a parameter, for instance, a threshold of decision, which can be fine-tuned to balance the tradeoff between hit rates and false alarms. Increasing the hit rate leads to an increase in false alarms as well. Different applications exhibit different significance levels of these two factors leading to the selection of a different point on the curve. Another performance measure used by classification algorithms is the area under the ROC curve (AUC). AUC index values range from 0.5 (random behavior) to 1.0 (perfect classification performance). For more detail see [12, 45].

7.3 Features and Challenges of Medical Data Classification

Several medical classification tasks exist, among which diagnosis and prognosis are most common. Other medical classification tasks include medical imaging, signal processing and scheduling [65]. In a diagnosis process, the patient's information is selectively collected and interpreted based on previous knowledge as evidence for or against the existence or nonexistence of disorders [58]. In the case of prognosis, the patient's information is selectively gathered and analyzed to predict the "course and outcome of disease process" [59]. Prognosis is considered an important instrument for medical management [59, 65]. For example, in the case of cancer prognosis, the intention is to predict cancer susceptibility, recurrence or survivability [19].

Medical diagnosis and prognosis can be modeled as classification problems. An instance is a patient's case. The predictor features are the patient's medical data. These might include demographic, clinical and pathological data. The class in case of diagnosis is the medical disorder. In case of prognosis, the class is the course and outcome of disease process. Production rules and decision trees are particularly attractive representation forms for the classification model in the medical field due to their comprehensibility. Using these forms, extracted models can be verified by medical experts and can enhance understanding the problem in-hand [84]. For example, in [47], a consultant pathologist in the domain of primary breast cancer evaluated the resulting rules for primary breast cancer diagnosis and classified them into three types; interesting new knowledge that could be further investigated, rules that are useful for the diagnosis and confirm medical knowledge, and those that contradict existing medical knowledge.

A physician relies on medical knowledge and personal experience to perform the desired classification task (diagnosis and/or prognosis). In many cases, physicians find difficulty in deciding the correct diagnosis or prognosis of a patient [11]. Patient presentation of disease varies significantly. It is a qualitative perception of symptoms that is difficult to quantify. The fluid representation is also perceived by qualitative receptors, i.e. physicians. These two factors; patient presentation and physician reception, are usually variable which participate significantly in understanding the medical case. The subjective interpretation results in variable output in terms of diagnosis and/or prognosis. For example, heart attack may be represented with pain in both arms; that can be interpreted as different diagnosis, some of which are not cardiac [76]. In addition, medical field experts are scarce and do not cooperate in converting their unique knowledge and art into a practical decision tool [65]. Also, the medical literature grows at a speed the physicians cannot cope with. Computer-aided diagnosis and/or prognosis systems bridge the knowledge gap in the era of evidence-based medicine [76]. The development of an adaptive model that learns from experience is more desirable than a best-fit solution for inherently complex and non-linear systems like the human body [92].

There are difficulties associated with medical data as well. Medical data includes demographic data, clinical observations, laboratory tests and radiology exams. Medical decisions are based on patient's medical records. Health care institutions are maintaining permanent patient medical records. Modern medical screening and diagnostic methods generate high volume of heterogeneous data. This data is continually accumulating. Mining such data requires intelligent methods [65, 88].

In addition to the high dimensionality, medical data exhibit unique features including noise resulting from human as well as systematic errors, missing values and even sparseness [88]. To illustrate, Table 7.1 presents medical data set examples. Most of these datasets are obtained from the UCI repository of machine learning databases, University of California-Irvine, Department of Information and Computer Science[1]. For example, some datasets like Dermatology, consist of different types of attributes. The high dimensionality is a feature of the Ovarian 8-7-02 dataset. Thyroid dataset contains more than 7000 instances. The Hepatitis dataset is imbalanced. The percentage of missing values in the Hungarian Heart dataset exceeds 20%. Finally, the Chest Pain dataset exhibits the multiclass problem featuring 12 different classes. Due to this nature, Tanwani et al. [88] calls for the classification of medical data as a separate domain, of which is currently considered one of the most popular research subjects in the data mining community [68]. This, in part, is due to the societal significance of the subject and also to the computational challenge it possess.

Tanwani and Farooq [85, 86, 87] performed an extensive study to present the challenges associated with biomedical data and approximate the classification potential of a biomedical dataset using qualitative measure of this complexity. The

[1] Address="`http://archive.ics.uci.edu/ml/datasets.html`"

complexity of biomedical datasets was found to be highly associated with a new factor; the correlation-based feature selection subset merit. This factor measures the quality of attributes in terms of how much they are correlated with the outcome class and not correlated with each other. Several empirical studies involving various evolutionary computing and machine learning classification algorithms were performed on UCI biomedical datasets. The classification accuracy was found to be dependent on the complexity of the biomedical dataset - not on the classifier choice. The two main effectors are noise and correlation-based feature selection subset merit. Second, the number and type of attributes has no noticeable effect on the classification accuracy as compared to the quality of the attributes. It is shown that biomedical datasets are noisy and that noise is the dominant factor that affects the resulting classification accuracy. Only high percentages of missing values severely degrade the classification accuracy. Third, evolutionary algorithms tend to overfit for small-sized datasets and are not much affected by the imbalanced classes' problem. A meta-study was performed consisting of the complexity measures as attributes. Using a decision tree and rule learner classifiers, the datasets were categorized into having good, satisfactory, or bad classification potential, according to their complexity factors. An equation is presented to find the classification potential of a dataset based on the level of its' noise and correlation-based feature selection subset merit.

Table 7.1 Example medical data sets and their associated complexity

Data set	Source	No. Instances	No. Attributes	No. Classes	Missing Values	Input Data Type
Chest Pain	[10]	138	165	12	No	Binary
Hungarian Heart	UCI; [88]	294	13	5	20.46%	3 Binary, 10 real
Dermatology	UCI; [88]	366	34	6	0.06%	1 Categorical, 1 binary, 32 integer
Wisconsin breast cancer (WDBC)	UCI; [84]	569	32	2[a]	No	Real
Hepatitis	UCI; [84]; [88]	155	19	2[b]	5.67%	13 Integer,6 real
Ovarian 8-7-02	CCR[c]; [88]	253	15, 154	2	No	Real
Thyroid	UCI; [88]	7200	21	3	No	15 Binary, 6 real

[a] Benign (62.7%)/Malignant (37.3%)

[b] Live (79.35)/Die (20.65%)

[c] Ovarian cancer studies, Center for Cancer Research, National Cancer Institute, USA, address="`http://home.ccr.cancer.gov/ncifdaproteomics/ppatterns.asp`"

In light of all of this, deriving a medical classification is a complex task [11, 65]. In particular, the rule-discovery problem is NP-hard [18]. This task involves searching for the hypothesis that models the diagnosis and/or prognosis concept, over all possible patient instances, in the space of all possible hypotheses. Penã-Reyes and Sipper [65] state "the medical search space is usually very large and complex" The Chest pain dataset [10] is a simple example to show the complexity of the search space. It consists of 165 binary attributes. The instance space $|X|$ contains exactly $2^{165} = 4.6768E49$ distinct instances. Therefore, the target hypothesis space includes $2^{|X|}$ possible hypothesis. That is, the target space includes $2^{|4.6768E49|}$ possible hypothesis. Execution time and memory demands grow rapidly with the number of instances and attributes of the problem at hand. Exact methods cannot be applied in this case.

Various classification paradigms exist, each with a related decision surface that decides the type of problems the classifier is suitable for. Machine learning algorithms like decision trees (DTs) suffer from trapping in local optima for a problem with a large number of attributes [18]. The back propagation algorithm (BP) [74] for training ANNs exhibit local search ability and can similarly get trapped into the nearest local optima [43]. A single run of BP is normally unrepeatable, unreliable and suboptimal, particularly on multi-local optima decision surfaces [43]. The main problem with machine learning methods is scalability especially when dealing with huge data [75]. In this respect, Provost and Kolluri [70] present a survey of methods for scaling up these algorithms. Statistical methods such as logistic regression (LR) and linear discriminant analysis (LDA) are widely used for classification. However, they do not produce accurate models when the relationship between the inputs and outputs of the dataset are non-linear and/or complex [25]. There exists no best classifier over all possible problem types [54]. Each technique has its own set of capabilities and limitations.

One way to deal with this shortage is to combine the properties of intelligent techniques so that each technique complements the capabilities and covers the limitations of the other. Combining or hybridizing various methods including heuristics and metaheuristics such as soft computing methods can significantly improve an analysis in terms of tractability, robustness, solution cost, and accuracy [25]. Metaheuristics in particular such as genetic algorithms (GA) [36], tabu search (TS) [29, 30], memetic algorithms (MA) [62], and simulated annealing (SA) [48], perform heuristic local search rather than exhaustive search producing good solutions within reasonable time and memory requirements [18]. Early hybrid systems, like evolutionary–neural hybrid systems [65] appeared in the early 90's. For instance, GAs were used to select predictor variables for the neural network [63] used to predict patient's response to Warafin. GAs were also used to optimize weights of ANNs in the prognosis for ICU patients [23]. Penã-Reyes and Sipper [64] used an evolutionary–fuzzy hybrid system for breast cancer diagnosis. In this study, a rule-based classifier that uses fuzzy logic called a 'fuzzy inference system' is used for the medical classification model learning. GAs are used to search for the parameters of the fuzzy inference system. A similar evolutionary–fuzzy hybrid was used by Jain et al. [44] for the diagnosis of coronary artery disease and breast cancer.

The next sections provide a snapshot of the state-of-the-art approaches in medical data classification. Section 7.4 demonstrates a sample of the literature that applies hybrid Metaheuristics for the problem of learning and optimizing medical data classification models. Section 7.5 illustrates the use of hybrid metaheuristics for model selection in medical data classification.

7.4 Hybrid Metaheuristics for Model Learning and Optimization in Medical Data Classification

This section starts with the use of learning classifier systems and their variants for model learning [5, 28, 38, 39, 40, 41, 47, 66, 77, 78, 84, 89, 91, 94, 97]. Other hybrid systems for model learning are next exemplified including the combination of genetic programming (GP) [51] with genetic algorithms [84], the blending of self-organized maps (SOMs) with ANNs and sUpervised Classifier Systems (UCSs) [73], the combination of TS with SA [18], and MAs [9]. Finally, two examples illustrate the use of metaheuristics for enhancing classifier accuracy as in the use of GA to enhance the classifier model generated by a decision tree classifier [75] and the use of homogeneity-based algorithm (HGA) [67] for optimizing the classifier models generated by support vector machines (SVMs), DTs and ANNs [68]. Table 7.2 presents a summary of these systems.

7.4.1 Learning Classifier Systems

Learning Classifier Systems (LCSs) [37, 95, 96] represent the merger of different fields of research including evolutionary computing and machine learning (reinforcement and supervised learning). They are adaptive systems that learn rules to direct their performance in a certain environment. In these rule-based systems, evolutionary methods (mainly GAs) are used to search the solution space while the reinforcement part from machine learning is used to guide the search to improved results. Their first appearance, Cognitive System One (CS-1) [37] seemed to be "complex and difficult to realize" [14]. The mid-1990s witnessed the birth of new models and new applications which revived this area. The 'zeroth-level' classifier system, ZCS [95] is a striped-down version of Holland's LCS that has better performance and comprehensibility. Wilson's ZCS was parameter-sensitive but has demonstrated optimal performance on several well-known test problems [13]. Not much later, Wilson introduced a variant of LCS with a new fitness measure, XCS [96]. Wilson's XCS has obtained more success and acceptance in the LCS community [14]. Stolzmann introduced a new line in the LCS research that stems from the theory of anticipatory behavioral control and cognitive psychology; Anticipatory Classifier Systems (ACSs) [82]. Rules in ACS aim at predicting action consequences in all possible cases in an environment. The models evolved in ACS direct the system to the most promoted action and also provide anticipation on what will happen next. On the other extreme, sUpervised Classifier Systems (UCSs) [8] replace the reinforcement learning component that was basic in all previous systems by supervised learning. That is, immediate reward system is used as the correct action is known in advance.

There are two styles of learning classifier systems. The first follows Holland's original model [37] developed at the University of Michigan and is thus termed 'Michigan-Style'. The solution is represented by the whole population. Rules

Table 7.2 Hybrid Metaheuristics for Model Learning and Optimization

System	Medical Purpose	Method	Rival Algorithms	Performance Metric
		Learning Classifier Systems		
EpiCS [38, 40, 41]	Epidemiologic surveillance	LCS	C4.5, LR	$S_n, S_p, P, AUC, \ldots$ etc.
EpiXCS [39]	Epidemiologic surveillance	XCS	See5 DT	$S_n, S_p, P, AUC, \ldots$ etc.
ClaDia [94]	Breast cancer diagnosis	LCS (fuzzy)	—	P
[47]	Breast cancer diagnosis	XCS	Bayesian, SVM, C4.5	P, medical expert
[77, 78]	EEG signal classification	XCS	NB, SMO, k-NN, PART	S_n, S_p, P
XCSI [97]	Breast cancer diagnosis	XCSI	Best on UCI cite, other published work	P, rule quality
LCSE [28]	Diabetes classification	LCS ensemble	LCS, DT, ANN	P
[91]	Mixed	ACS	XCS, XCSL, C4.5	P, no. rules
ZCS-DM [89]	Mixed	ZCS	DT, C4.5, XCS, HIDAR	P
[66]	Mixed	Pitt-style LCS	—	P, rule quality
		Other Hybrid Metaheuristics		
[57]	Breast cancer diagnosis	ensemble: SVM, AdaBoost, and GA	—	AUC
[3]	Mixed	Hybrid BN–k-NN–GA	Bayesian (EM)	P
[9]	Cancer cell diagnosis	MA-optimized cell graph coloring	—	No. colors (cell graph)
[18]	Mixed	Hybrid TS–SA	Ant Miner, CN2	$S_n \times S_p$, P, no. rules
[73]	Mixed	SOANN, SOUCS	UCS, ANN, other published work	P, computational time
[84]	Mixed	Hybrid GA–GP	C4.5, PART, NB, other published work	P
		Enhancing Classification Accuracy		
[75]	Mixed	GA	UCS, C4.5	P
[68]	Mixed	HBA and GA	DT, SVM, ANN, other published work	P

compete under GA which operates at the individual rule level. In Smith's Pitt-style [79] developed at the University of Pittsburg, each individual in the population represents a complete solution. In Pitt-style, GA operates at the rule-set level. Both styles have their own advantages and shortcomings. However, since entire solutions are simultaneously being evolved and compared in pitt-style, it is computationally heavier than Michigan-style LCS. This favors Michigan-style LCS in terms of popularity in the LCS community [92]. For interested readers, the survey by Urbanowicz and Moore [92] is recommended.

Learning Classifier Systems exhibit several attractive features. First of all, their rule-based nature leads to comprehensible hypothesis, as opposed to black-box solutions presented by ANNs for example. This implies that physicians can validate if the resulting classification hypothesis is clinically plausible. This also means that there is room for discovering new interesting relations. Second, LCSs tackle complex learning problems [42], and this is particularly important when dealing with the medical domain. LCSs are also on-line learners that avoid local minima due to the EC component [6]. Different kinds of representation can be used for LCSs [6, 52]. Other advantages include adaptability, robustness [6], and good generalization ability [96]. These features are especially interesting when dealing with medical data.

The main weaknesses of LCSs include overfitting for small data [5, 28] and difficulty with imbalanced classes [6]; as they tend to bias towards the majority class.

The use of learning classifier systems and their variants for the purpose of model learning in medical data classification has been well established. LCSs were applied with considerable results in medical data classification field. For example, learning classifier systems for epidemiologic surveillance EpiCS [38, 40, 41], EpiXCS [39],LCS with fuzzy rule representation [94], XCS [5, 47, 77, 78, 97], learning classifier system ensembles [28], ACS [91], ZCS for data mining (ZCS-DM) [89], and Pitt-style LCS [66]. Below a summary is presented for medical data classification solutions that are based on learning classifier systems.

EpiCS [38] was the first specialized LCS in the medical field; a learning classifier system for epidemiologic surveillance data. EpiCS predicts risk of disease; the probability of developing a disease. The estimate is given by the proportion of the matching classifiers that classify the case as positive. Using synthetic epidemiologic data generated such that one variable is associated with the outcome, EpiCS was compared to logistic regression-derived probability of disease and has shown significant advantage in terms of classification performance measured using the area under the receiver-operating characteristic curve (*AUC*).

The study by Holmes et al. [40, 41] was performed on epidemiologic surveillance data obtained from the Partners for Child Passenger Safety (PCPS). The aim of the study was reducing child automobile crash-associated morbidity and mortality through discovering patterns associated with head injury (head-injury/no-head injury classification task) [40], with inappropriate child restraint (appropriate/inappropriate child restraint classification), and the associated risk analysis [41]. Epidemiologic data are characterized by their large size and number of features that may result in huge number of relations. These relations can be modeled using the IF-THEN format. 47 numeric features [40] were selected out of over 500 available variables. Data were equally partitioned into testing and training sets with positive and negative classes equally distributed. Missing data were treated as don't-cares. Performance was evaluated in terms of sensitivity, specificity and *AUC*. All of these evaluation metrics were modified by the indeterminate rate (IR); cases that the model could not classify. EpiCS significantly outperformed the decision tree classifier algorithm C4.5 and LR in terms of *AUC* (0.97%) [40]. The number of rules produced by C4.5 was significantly lower. Based on that, the authors suggest that the use of C4.5 to initialize the EpiCS population might be advantageous. The authors also point out the need to improve LCS in terms of macrostate reduction, dealing with numeric data in native form and dealing with noisy data. The paper addresses the limitations of decision trees and linear regression models with respect to clinical and epidemiologic data.

In 2005, Holmes and Sager [39] introduced a new LCS for the epidemiologic community, EpiXCS. EpiXCS is an XCS classifier application tailored to the needs of epidemiologic research. The main feature is an interface workbench that allows researchers to set different parameters in a user-friendly manner. Using EpiXCS, researchers can watch the performance in terms of parameters like sensitivity, specificity, *AUC*, learning rate, and indeterminate rate. These parameters are updated in frames of 100 iterations. In addition, EpiXCS views resulting rules both in textual (IF–THEN format) and graphical forms. The graphical rule display option enables researchers to see possible clustering of features and their values forming a certain outcome. EpiXCS was compared to the See5 decision tree classifier in forming rules that discover the features associated with teenage automobile fatality in the census of all fatal United States and Puerto Rico automobile crashes ;FARS database. Results show that while classification accuracy of both classifiers was comparable, EpiXCS produced far fewer rules making the analysis much more manageable. Also, EpiXCS has discovered several features that were missed by See5.

For the diagnosis of breast cancer, Walter and Mohan present a classifier system for disease diagnosis, ClaDia [94]. A fuzzy rule representation was used where the attribute values were mapped to the ranges (low, medium and high). Instance-rule match degree correspond to the median membership degree of the instance's constituent attributes. Rule fitness was computed as the difference between the number of correctly classified instances and those incorrectly classified. Rule fitness was later reinforced by correct classifications and penalized otherwise. Niching was applied such that recombination is only allowed among individuals in the same niche (benign/malignant). Unlike the original LCS, mutation is performed on rule antecedent as well as consequent. That is, the rule consequent of weak rules may be mutated (reversed) as these may result in good rules for the opposite class. ClaDia was applied to Wisconsin Breast Cancer (WBC) database from the UCI repository and achieved over 90% accuracy.

Bacardit and Butz [5] compared the performance and generality level of two LCS classifiers, namely the on-line XCS and the off-line GAssist [4]. The comparison is done over thirteen different data sets. While GAssist is a Pitt-style classifier, XCS is a Michigan-style classifier that basis fitness on rule accuracy and applies GA selection to the currently active classifier subsets. Six types of problem difficulty are considered in this study. These include the input data volume, size and type of the search space, concept complexity, input noise and missing data in addition to the overfitting problem. The goal is to achieve a maximum level of generality. Results show that while both systems perform well on all data sets, the produced solutions are quite different. XCS has a weaker strategy in handling missing data and tends to over-fit training data especially in small data sets. XCS thus requires a large training set. GAssist on the other hand tends to ignore additional complexity and struggles when facing problems with multiple classes or those featuring large search spaces. Two conclusions are drawn: XCS needs to address its generality difficulty and GAssist needs to address its problem handling data sets with multiple classes.

In Kharbat et al. [47] primary breast cancer data from the Franchay Breast Cancer (FBC) data set is mined using XCS. Results are compared to other classifying techniques including Bayesian network classifier, SVM and C4.5 decision trees. As a preprocessing step, numeric values were normalized and data in nominal and Boolean attributes were decoded. The imbalance problem is handled by random over-sampling. Missing data were treated with Wild-to-Wild method; in which missing values are replaced with don't-cares for nominal data and general intervals for numerical data. Results showed that XCS outperformed other methods. The number of rules produced with XCS was much more than those produced by the C4.5 algorithm. However, these rules were described by a medical expert to be more informative and useful. Clustering and rule compaction were applied on the resulting rules.

Skinner et al. [77, 78] also use XCS but for EEG signal classification. EEG signals are characterized by their high dimensionality and noisy nature. This study investigates the efficacy of XCS in the classification of mental tasks based on human multi-channel EEG signals. In particular, the binary classification of four diverse mental tasks for three individuals. The significance of this investigation lies in the potential to use EEG classification results to control wheelchairs or similar devices for paralyzed individuals. The novelty of the approach is in the investigation of using XCS to process large and noisy condition strings. EEG signals were preprocessed to reduce the number of channels and their associated frequencies. Data was then segmented. The results were compared with four ML classification methods; naïve Bayes (NB), SMO, *k*-nearest-neighbor (*k*-NN with *k*=3), and PART which combines the learning strategies of decision trees and rule learners. Results were compared in terms of classification accuracy and showed that XCS significantly outperformed PART and *k*-NN. XCS was comparable to the SMO but inferior to naïve Bayes.

The study by Skinner et al. [78] investigates the effect of different migration policies on distributed and parallel XCS classifier population with different topologies and parameters. The study was performed on the single-step classification for human EEG signals associated with two mental tasks; Mental Counting and Figure Rotation for two persons. Three topologies were examined; fully connected, and uni- and bi-directional rings with different number of demes (2, 4 and 8). Migration policies are based on the selection and replacement criteria for the immigrant rules; based on their fitness, numerosity, or random. The study concludes that lower migration frequencies and rates produce better classification performance. High degree of connectivity speeds up the learning process. All policies result in a significant classification accuracy improvement with respect to XCS alone. Random immigrant selection results in a slower learning. Also, fitness-based migration selection increases the selection pressure and thus degrades the classification performance. As for population size, it is entirely dependent on the immigrant selection policy. Fitness-based immigrant selection policy gives better results for the fully connected topology while random-based immigrant selection is more beneficial for the uni- and bi-directional rings.

Also in the field of breast cancer is the study by Wilson [97]. The classifier predicates of XCS describe logical problems by defining hyper-rectangles in the decision space. In [97], XCS was modified to handle integer input spaces. The modified version, XCSI, was tested on oblique data. The study started with a simple

2-dimensional synthetic oblique data for which XCS achieved 100% (training) classification accuracy. A second experiment was conducted also with synthetic oblique data that resembles the UCI WBC dataset in terms of the number and type of attributes, their data ranges and the number of instances. Again 100% (training) performance was achieved although slightly slower. The final experiment used the UCI WBC dataset with a 10-fold cross-validation technique. Accuracies averaging 95.56% were reached. Results further show that the hyper-rectangles modeled by XCS predicates were good at approximating the oblique discrimination surface of the data. Also, results suggest the presence of logical patterns in WBC dataset that is evident by the presence of several accurate classifiers showing logical dependencies on one or a few attributes. Classifiers describing regions close to the discrimination surface feature a match set with strong evidence on both directions that can be used for risk of disease analysis instead of a concrete diagnosis.

The first LCS ensemble was introduced by Gao et al. [28]. The Learning Classifier System Ensemble (LCSE) [28] is an extension of LCS that aims at achieving better generality through using several sub-LCSs. Diabetes data input is distributed over these sub-LCSs. Each sub-LCS may then produce different rules even for the same input data. Results are then aggregated by means of a popularity voting method. Overfitting problem is managed with a 10-fold cross-validation approach. Results of LCSE outperform LCS, DTs and ANNs as well. Experiments also show that the accuracy of results increases with the number of sub-LCSs.

Unold and Tuszynski [91] applied ACS to three data mining data sets from the UCI repository; Monks, Voting-record and WBC. Results show that ACS achieve results no less than 97% except for the Monk's 2 data set, were the accuracy was limited to 75%. A comparison with XCS, XCS with s-expression (XCSL) [52, 53] and C4.5 shows that overall; XCS and XCSL achieved best results. XCSL have succeeded in producing the least rule set size. C4.5 was far behind. Future research aims at developing ACS to enable the handling of attributes of continuous type.

A modified version of ZCS for data mining applications named ZCS-DM is presented and applied to several UCI repository datasets including WBC, Hepatitis, Pima Indiana Diabetes and Bupa Liver Disorder benchmark sets [89]. The main changes include evolving the action part as well as the condition part and using user-tunable reward/penalty for the different class combinations (predicted and actual classes). In their model, users decide the number of individuals in the population and whether the action selection mechanism is deterministic or stochastic. A preprocessing step involves removing all duplicate rules after adjusting their strength. Rules are ordered based on their strength and either the first matching rule is selected or a voting scheme is employed. Missing values were treated by setting their corresponding predicates in the rule's condition part to true. Experiments were done using a 10-fold cross-validation and results show the classification accuracy advantage of the proposed model in 11 out of 12 UCI datasets over the DT algorithm C4.5, XCS and HIDAR [2]; which is a hierarchical decision rule that uses a sequential covering GA. C4.5 was the fastest algorithm.

An improved Pitt-style LCS is introduced by Peroumalnaik and Enee [66]. The training set is equally partitioned among the individual classifiers following the divide-and-conquer approach. Since prediction is performed by the whole population, and GA combines the genetic material of different individuals, the authors argue that the partitioning would eventually lead to a segmentation of the cognitive space. The proposed algorithm was tested on four medical UCI data sets with various parameters for the population size, number of rules per individual and rule selection strategy (random, most general or most specific). The proposed method has produced good results for the Wisconsin Diagnostic Breast Cancer (WDBC) dataset using a 10-fold cross-validation. No comparison with other methods was performed.

7.4.2 Other Hybrid Metaheuristics

Applying multiple classifiers is analogues to consulting a team of specialists. Each specialist considers the problem from a different perspective thus allowing the exploration of different regions of the search space. Multiple classifiers usually result in higher accuracy compared to a single classifier [88]. This is because the strengths of one method are utilized to complement the weaknesses of another [49]. The use of multiple classifiers is particularly useful for imbalanced data sets [88]. In addition, using multiple classifiers usually feature a strong generalization ability [28]. As classification model learners, there are endless possibilities for hybridizing metaheuristics together or with other classification methods in seek for a better classifier. However, using multiple classifiers comes in one of two forms. The first form is the ensemble classifier [28, 57]. Classifier ensembles achieve model diversification by using different subsets of training data with a single learning method, different training parameters with a single learning method, or using different learning methods [49].The second form for using multiple classifiers is employing a hybridization of metaheuristics with machine learning and/or evolutionary computing methods [3, 9, 18, 73, 84].

Like employing a team of specialists, the cost of using multiple classifiers is more than that of a single classifier. First, since all component classifiers need to be stored after training, the storage requirement increases accordingly. Second, all component classifiers need to be processed adding to the computational cost. Finally, it is more difficult when using multiple classifiers to comprehend the underlying reasoning and conclude a classification, particularly for non-experts [49].

An ensemble method for the detection of breast cancer from x-ray images is investigated by the authors in Lo et al. [57]. The proposed classifier was chosen as the joint winner in KDD Cup 2008. The data set is characterized by being highly imbalanced. The ratio of positive samples to negative samples is 163. Each patient is represented by a set of data points. The evaluation criterion was to minimize the AUC per patient rather than per data point. This was intended to minimize overfitting. The ensemble consisted of four classifiers; AdaBoost, Class-based SVM (CB-SVM), Patient-based SVM (PB-SVM), and GA. In CB-SVM, the intention was to balance the positive and negative classes. A class-sensitive loss function was employed where the weight of positive samples was 163 times more than that of negative ones. The problem faced was that patients with fewer positive instances were more difficult to identify than those with more positive instances. To resolve this problem, PB-SVM was designed such that the sum of weights of positive samples was equal to the sum of weights of negative ones. In addition, for each patient i, the sum of positive sample weights is equal to that of patient j. A slight improvement was obtained over the CB-SVM. AdaBoost was based on 50 weak learners, were Classification and Regression Tree (CART) was chosen as a weak learner. As for GA, the fitness was based on the AUC itself and resulted in better recognition of patients with fewer positive instances. The best ensemble outcome was obtained by averaging the two best classifiers.

The study by [3] focuses on randomly generating data sets based on the observed data and that will maximize classification accuracy. This technique is particularly useful in cases featuring missing data, small training data set size or noisy data. The study suggests an iterative hybrid model that starts with applying the Bayesian method based on the expected maximization algorithm (EM). Misclassifications are recorded. Next, a new data set twice as large as the observed data is randomly generated. A k-NN classifier is trained on this data and tested on the observed data set. This process is repeated until a lower misclassification rate is observed. Then, GA is used to further improve the generated instances. Bayesian classification based on EM is applied on the resulting data and new data generations are evolved and tested until an improved misclassification number is obtained. The algorithm was applied to five UCI data sets including Iris, Breast Cancer, Wine, Yeast and Glass. Results were compared against using the Bayesian classification based on EM alone. Improvements up to about 75% were recorded for the Breast cancer data set. On the other hand, Wine dataset resulted in a slight retrogression. The algorithm involves several iteration cycles resulting in an increased computational time. Also, results were not compared to other algorithms that are not based on data set generation.

Bhattacharyya et al. [9] present an introductory work for the diagnosis of cancerous cells from human-extracted low-resolution biopsy BMP images. Currently the diagnosis is based on the subjective pathologist evaluation of the tissue sample. The authors introduce a new automated diagnostic method that is based on the generic organizational structure of tissue cells. Two phases are implemented. The first is constructing the Cell Graph. This step transforms the BMP image into a monochrome graph; where nodes correspond to cells or cell clusters depending on the resolution used. Edges are assigned on a probability based on the Euclidian distance between the nodes. The second phase is graph coloring using the minimum number of colors such that nodes within the same range of Euclidian distance obtain the same color. Memetic algorithms (MAs) are used to optimize graph coloring. In this work, MAs are composed of a heuristic search; sequential graph coloring algorithm, and a genetic algorithm with a modified mutation operator. The output of the program was the number of colors used for the sample image. It was not clarified how cancerous cell diagnosis can be derived from this information. However, the work provides more formalism about the density/organizational characteristics of the tissue cells that aid in the diagnosis process.

A hybrid tabu search–simulated annealing rule induction algorithm for classification tasks is presented by Chorbev et al. [18]. Continuous attributes were discretized. Classification rules are created incrementally and pruned for better readability and higher predictive value. The probability of an addition of a term to a classification rule depends on the entropy of the attribute's value as opposed to entropy in decision trees that is computed for attributes as a whole. Tabu timeouts aim at reducing the probability that a particular attribute value is selected twice; therefore increasing the search diversity. The quality of a rule computed as the product of sensitivity and specificity serves as the energy parameter for SA. An initial high temperature in SA allows low quality solutions to be accepted in the beginning for better exploration of the search space. As the search proceeds and the temperature cools down, only high quality rules will be accepted into the final rule list thus intensifying the search in promising areas. The proposed classification algorithm was compared against Ant Miner and the rule induction algorithm CN2 on four UCI medical datasets. In terms of predictive accuracy, Ant Miner was in the lead. However, in terms of the number of rules and terms per rule, SA Tabu Miner achieved good results that outperformed CN2 and was highly comparable to that for Ant Miner.

Rojanavasu et al. [73] present the use of self-organized maps (SOMs) as a pregate. That is, the SOM is used to cluster the data on-line and thus decomposing the search space into smaller sub-problems that are conceptually simpler. Class labels for the data are masked in this phase. Separate classifiers are then used to learn each sub-problem. The paper investigates the utility of connecting the pregate to two different classifies; a set of sUpervised Classifier Systems (UCSs) thus forming Self-Organized UCS or SOUCS; and an artificial neural network (ANN) thus forming SOANN. ANN layout is fixed for each dataset. The authors experiment with three data sets; the first is a group of five synthetic problems of increasing complexity. The second is a set of UCI datasets, and the third is a large and complex Forest-Cover-Type dataset from the Roosevelt National Forest in northern Colorado. Experiments have been applied with varying number of SOM sizes $(2 \times 2,\ 3 \times 3, and\ 4 \times 4)$; which implies a different number of UCS classifiers, and varying the number of individual UCS populations. All experiments are done using 10-fold cross-validation. In comparison with UCS alone and other published work, SOUCS showed an equivalent or better results in terms of classification accuracy except for the Forest-Cover-Type data set. The complexity of this dataset was not properly addressed by the smaller population sizes in the individual constituent UCSs. The SOUCS was superior in terms of computational time. The reason is the smaller population size in each UCS. Experiments also show the high sensitivity of the outcome to the population size and number of SOM cells as well as the problem type. This was suggested as a future research line. The ANN/SOANN environment obtained better results for the Forest-Cover-Type dataset but no better in the rest.

Finally, in Tan et al. [84], a two-phase strategy is presented as follows: the first phase uses a hybrid Michigan GA and GP, to produce per-class single rule poles in the form of: $\langle IF\ X_1\ and\ X_2\ and\ \cdots\ X_n\ THEN\ class = Y\rangle$. Michigan GA is applied to numeric data while the GP is applied to nominal data sets. The second phase involves applying Pittsburgh GA to find an optimal combination of the resulting rules with the OR operator. The population is divided into a number of subpopulations evolving simultaneously and corresponding to different number of rules in a single solution (rule set). Results of applying the algorithm to hepatitis prognosis (live/die) and breast cancer diagnosis (benign/malignant) were very encouraging and outperformed other classification methods like the DT algorithm C4.5 and trained neural networks. However, the system is computationally expensive and is suited for off-line classification.

7.4.3 Hybrid Metaheuristics for Enhancing Classification Accuracy in Medical Data Classification

Metaheuristics like GA that are well known for solving complex optimization problems can be used to optimize classification models obtained using other data mining techniques. The idea is to evolve a population of individuals (classification models or classification model components) that compete on the basis of their fitness. This 'fitness' measure can be defined in terms of their classification accuracy, compactness, computational complexity, or some other similar or compound measure. For example, a two-stage hybrid machine learning classifier approach is proposed by [75]. The first phase involves creating an initial classification rule set by the C4.5 decision tree classifier. 3-Fold cross-validation is used and the best accuracy generated rule set out of the three is used for the second stage. In the second stage, a genetic algorithm with one-point cross-over and optional mutation is applied to improve the generated rule set. Rules having invalid class type and resulting from crossover operation are omitted. Invalid attribute values in rules resulting also from crossover operator are replaced with don't-cares. In comparison with accuracy scores of the C4.5 alone and the accuracy-based learning classifier UCL, the proposed approach produced better classification results over most of the eight UCI data sets used in the study.

The second example on the use of hybrid metaheuristics for enhancing classification accuracy in medical data classification is the work by Pham and Triantaphyllou [68]. While most studies assign equal weight to the different types of error for a classifier; FP, FN, and UC (un-classifiable), the study by Pham and Triantaphyllou [68] focuses on the optimization problem of the penalty costs for those three error types. The study investigated the use of three traditional machine learning classification algorithms; DT, SVM and ANN in combination with a metaheuristic termed Homogeneity Based Algorithm (HBA) [67], to optimize the penalty costs of the three error types. HBA works on defragmenting the decision surface space resulting from the classifiers into homogeneous regions according to their density. In addition, GA is also applied to optimize the parameters for HBA. The proposed hybrid system was tested on five medical datasets from the UCI repository. Results were compared to using the three classification algorithms alone and with other published studies. It is shown that using HBA significantly improves the classification accuracy for all five datasets. The shortage of HBA is its high complexity as it cannot deal with datasets with a high number of attributes (greater than 10). This is currently the research focus of the authors.

7.5 Hybrid Metaheuristics for Model Selection in Medical Data Classification

There is a recent trend to use optimization techniques, including mainly EC methods, for parameter selection, feature subset selection, class representative selection and even for preprocessing and classifier selection. These factors highly affect the quality of the resulting classifier. For example, there is no rule of thumb on how to guide parameter setting for ANNs. Usually these are determined experimentally for each problem. Also, the high dimensionality of the data set not only slows down the classification process, but also confuses the classification algorithm and may lead to poor results [33]. The main benefit is the achievement of competitive classifiers without using background knowledge, careful data analysis, long experimental trials, or even knowledge about the classification model being employed [24].

Model selection is defined as “estimating the performance of different models in order to choose the best one” in describing a dataset [35]. There are many interpretations for model selection, these include parameter selection and optimization [15, 17, 25, 60], feature subset selection [90, 93], artificial neural network modeling including learning the weights of the neural nets [61, 80, 81], or optimizing the architecture of the neural net [16, 43]. Reference [24] have extended this definition and introduced the so called full model selection (FMS). In this system, a pool of preprocessing methods, feature subset selection and learning algorithms is introduced and the task is to select the best combination that would yield the lowest classification error for a given problem. In addition, the parameters for these methods are being selected as well. Stochastic optimization algorithms are well suited for dealing with the vast search space introduced by such problems. This section samples published work that utilize hybrid metaheuristics for model selection in medical data classification. Section 7.5.1 focuses on their use for feature subset selection. Section 7.5.2 illustrates the use of hybrid metaheuristics for ANN model selection in medical data classification. Finally, sect. 7.5.3 exemplifies their use for FMS. Table 7.3 presents a summary of these systems.

The work by Candelieri [15] is derived from the author’s PhD thesis. The paper investigates and compares the hybridization of several metaheuristics including GA, TS and ACO to perform model selection for an SVM classifier both as a single classifier and as an ensemble. In this framework, SVM is used for learning the classification model. In the case of single classifier, the metaheuristic is either used to search for the best performing kernel function (linear, Normalized Polynomial, or Radial Basis Function) and its associated parameter(s); of which the author calls Model

Table 7.3 Hybrid Metaheuristics for Model Selection

System	Medical Purpose	Method	Rival Algorithms	Performance Metric
[15]	Mixed	SVM hyberdized with GA, TA, and ACO	—	Balanced classification accuracy
[17]	Mixed	Ensemble of BP-ANN, SVM, C4.5, parameter and FSS by SS	Other published work, individual ensemble components	*P*
[25]	Mixed	Clustering by CBR then fuzzy DT	*k*-NN, NB, SVM, fuzzy DT, other published work	*P*
[60]	Mixed	Class representatives and, parameters chosen by DE	*k*-NN, SVM, discriminant analysis, ANN	*P*
		Hybrid Metaheuristics for Feature Subset Selection		
[90]	Mixed	LR with PSO for FSS	Exhaustive search, TS, SS, random subset generation	*P*, computational time
[93]	Mixed	Fuzzy rule-based classifier and ACO for FSS	Fuzzy rule-based classifier, other published work	*P*, min no. features
		Hybrid Metaheuristics for Artificial Neural Network Model Selection		
[61]	Mixed	EDA for training ANN	BP-ANN, LM-ANN	*CEP*
[80]	Colorectal cancer prognosis	Step-wise regression, clustering, ANN ensemble with LS + BP training	—	*P*
[81]	Mixed	ANN trained by ACOR, ACOR–BP, and ACOR–LM	ANN trained with BP, LM, GA, GA–BP, GA–LM	CEP
[16]	Mixed	MG-Prob for MOP of MLP ensemble	—	minimize(FP, FN, network size)
[43]	Mixed	BP–ANN, PSO-trained ANN	BP vs. PSO, other published work	CEP, MSE, computational time
		Hybrid Metaheuristics for Full Model Selection		
PSMS [24]	Mixed	PSO, PS, other systems	BER	

Selection. The second feature, Multiple Kernel Learning, is to use the metaheuristic to optimize the n kernels along with their related parameters and coefficients. In the case of Ensemble Learning, the metaheuristic is used to find the best m SVM classifiers and their associated weights for combination. ACO was only applied to Multiple Kernel Learning. The framework was tested on 8 datasets of which several were medical. Balanced classification accuracy using 10-fold cross-validation was used for evaluation. Results show that the three models were highly competitive. GA was generally faster and more effective than TS. As for ACO, results were comparable to the others and promising as a new application.

In a study by Chen et al. [17]; classifier parameters and data features are stochastically chosen and evolved independently using scatter search [31]. The parameters were for an ensemble of three classifiers; SVM, BP-ANNs and DT (C4.5). Data instances with missing values were removed from all classifiers except for the DT. Each classifier was run three times and a majority voting was obtained to combine the 9 runs. Experiments were conducted on 18 UCI datasets and were compared to four similar studies of which some involve using ensembles of a larger number of classifiers. The proposed approach achieved the highest classification accuracy. Also, in comparison with results of individual component classifiers, the average performance of the ensemble outperformed the single classifiers.

Fan et al. [25] introduce a four-stage model for medical data classification that utilizes data preprocessing and clustering techniques for improved classification accuracy. Comprehensibility of the generated model was a main objective. First stage involved feature subset selection using step-wise regression. Selected features are then weighted by the gradient method. The next step involves case-based reasoning (CBR) clustering of the input data. Next, the fuzzy Triangle member ship function is applied to discretize the data and ID3 decision tree is applied to build a classification tree for each cluster. The last step involves evolving the fuzzy terms used in the decision tree by means of genetic algorithms to further improve the classification accuracy. The model was applied to two UCI datasets; WDBC and liver disorder. Comparisons against several ML classification methods including *k*-NN, naïve Bayes, SVM, fuzzy decision tree, and to other similar studies were done. Results show the consistent advantage of the proposed model. Average accuracy rate achieved was 99.5% and 85% for breast cancer and liver disorder respectively. The paper does not show clearly how GA encoding is done to allow for different number of fuzzy terms for each feature.

Also to demonstrate the advantage of using evolutionary computation methods to guide the parameter and class representatives' choice is the study by Luukka and Lampinen [60]. This study assess the effect on classification accuracy of adding noise to dataset features, adding extra noisy variables, and adding all two-component variables . A simple minimum distance classifier (instance based) was applied to four UCI data sets; New Thyroid, Hungarian Heart, Heart–Statlog and Lenses. Minkowsky distance metric parameter and individual class representatives were stochastically chosen and evolved by using differential evolution (DE) [69]. Three sets of experiments were conducted to study the effect of adding noise directly to data set feature values or as independent variables as well as adding all two-component terms on the classification accuracy. Noise variance and the number of noisy variables added were varied. Experimental results show improvements in classification accuracy up to 8% in the case of adding all component terms. Performance was degraded in the other cases. However, a comparison with k-NN, SVM, Discriminant analysis and BP-ANNs showed that the DE-enhanced classifier obtained higher classification accuracy in all experiments. Particularly, BP-ANNs and the DE-enhanced classifier showed the best results in the extra noise parameters and the two-component terms experiments.

7.5.1 Hybrid Metaheuristics for Feature Subset Selection in Medical Data Classification

Many of the feature attributes in a typical medical dataset are collected for reasons other than data classification. Some of the features are redundant while others are irrelevant adding more noise to the dataset. The Feature Subset Selection problem (FSS) consists in selecting the minimum subset of feature that represents the dataset without loss in classification accuracy [93]. FSS not only reduces storage and computational complexity, but also enhances comprehensibility and classification accuracy particularly in small sample size datasets. FSS also reduce the overfitting effect [50]. In medical diagnosis, it is desirable to select the clinical tests that have the least cost and risk and that are significantly important in determining the class of the disease. There are two approaches for solving the FSS problem. In the filter approach, features are selected independently of the classifier. In wrapper approach, a classifier is used to test each feature subset candidate and thus is classifier-dependent and computationally heavier than the filter approach. FSS problem is NP-hard [50]. Using exhaustive search to find all the possible feature subsets is computationally impractical, even for a medium sized feature set. This requires the use of heuristics and meta-heuristics. A recent survey of feature selection problem for machine learning classification can be found in [50]. What is needed is an algorithm with good global and local search abilities, that can converge to a near optimal solution in reasonable time, and that is computationally efficient [50]. Nature inspired methods like Particle Swarm Optimization (PSO) [46] Ant Colony Optimization (ACO) [22] have been successfully applied for many combinatorial optimization problems including FSS. These swarm intelligence search algorithms are based on the collective behavior of intelligent agents that use both direct and indirect interaction.

For example, in Unler and Murat [90], a Discrete PSO method is applied as wrapper feature subset selection methodology for a (binary) classification problem. Linear regression was chosen as the learning algorithm. Features are considered on an individual basis and the decision for inclusion combines the feature's predictive contribution, independent likelihood as well as the stochastic factor. The feature is added to the so-far collected feature subset if it results in improved classification accuracy. Computational complexity is managed by restricting the number of features considered in every iteration. PSO parameters are based on earlier empirical and theoretical research. Experiments were conducted on 10 UCI datasets and comparisons were made to exhaustive search and Random Subset Generation. Results show that PSO produced identical or near identical accuracy to exhaustive search accompanied with a significant time advantage. Results further show that PSO accuracy was competitive to Random Subset Generation. Other comparisons were done with other wrapper feature subset selection methods introduced in published research and using the same learning algorithm and data sets. These studies include tabu search and scatter search. Results show the superiority of the proposed algorithm in terms of both classification accuracy and computational cost.

The ACO metaheuristic was used by Vieira et al. [93] in conjunction with fuzzy rule-based classifiers. Fuzzy rule-based classifiers perform model learning while the ACO selects feature subsets. This specialized feature selection ACO is termed (AFS). It consists of two colonies; the first is assigned the determination of the number of features to be selected. The second, selects the features themselves. Choosing the cardinality of features is based on Fisher discriminant criterion. The fitness function is based on minimizing the number of selected features and the classification accuracy of the obtained model as evaluated using the fuzzy rule-based classifier. Experiments were conducted using 5 UCI benchmark datasets among which 2 are medical. Results for the medical data sets show a significant improvement over fuzzy rule-based classifiers alone and also over other published studies that use PSO and rough set-based feature selection.

7.5.2 Hybrid Metaheuristics for Artificial Neural Network Model Selection in Medical Data Classification

ANNs are capable of learning complex, non-linear decision surfaces with multiple classes [43]. A recent survey about machine learning in cancer prediction and prognosis [19] shows that more than half of the surveyed papers were using or referring to ANNs. The idea was derived from human biological neural system where multiple neurons are interconnected to each other. The basic unit is called a neuron. The simplest ANN is called a perceptron and is able to do a binary classification task that has a linear discriminate function. Neurons are organized in layers; producing a structure called a multilayer perceptron (MLP). The first layer is connected to the inputs. Each input predictor is normally connected to an input neuron. The last layer produces the output(s). There is usually one output neuron per class in the dataset [81]. With a certain precision, a two layer MLP can approximate any classification region [54]. In feed-forward ANNs, the most popular ANNs, there are no backward connections and no loops. Each node in a hidden layer has connections coming from the nodes in the previous layer, and others going to nodes in the next layer. Assuming that the weights of neuron connections are available, the features of an instance are fed to the input layer, and is propagated through the hidden layer(s), until it reaches the output layer. Each output neuron is associated with a class. The output neuron that generates the highest signal wins in determining the class of that instance.

There is no universal ANN architecture. The architectural design of ANNs should be optimized for each application [19]. In order to generate an ANN classifier, the weights of the network's connections needs to be determined. As these weights are real-valued, the problem of determining ANN weights (the optimization of the network training error, or in short training ANNs [61]) can be casted as a continuous optimization problem [81]. Back-propagation (BP) optimization algorithm [74] is normally used for tuning the values of the set of weights. It follows a gradient-decent technique on the error surface and exhibits a local search ability that causes it to get trapped into the nearest local minima [43].

The problem of simultaneous optimization of the network's training error and its architecture can be modeled as a multi-objective optimization problem [1]. Abbas [1] found that combining back-propagation algorithm with an evolutionary multi-objective optimization algorithm leads to a considerable drop in computational cost. In the following, an illustration is presented of literature that substitutes or combines traditional ANN training algorithms with hybrid metaheuristics. The last two examples use hybrid metaheuristics for the task of optimizing both ANN architecture and training error.

The use of Estimation of Distribution Algorithms (EDAs) [7] is investigated by Madera and Dorronsoro [61] for the training of ANNs. The ANNs are used for the medical classification task of four PROBIN1 benchmark datasets. EDAs are evolutionary stochastic search techniques that base the construction of a new generation on estimations of the probability distribution of current population, rather than by means of variation operators. In this study, six different EDAs are being tested. These algorithms cover discrete and continuous search spaces and in each type of search space, three different correlation types of input variables are tested: those without dependencies, with bivariate dependencies and multiple dependencies. Training–testing method was used rather than the cross-validation. The performance was compared against other famous ANN training techniques; namely BP and Levenberg–Marquardt (LM) [32] algorithm. Performance was also compared to few ANN training techniques based on EAs; including those using GAs, MAs, or evolutionary programming (EP) [26, 27]. In general, EDAs performance was comparable to the others. This initial result is promising as further parameter tuning might likely improve the findings. EDAs for discrete domains were generally slower and less accurate than those for continuous domains. This is due to the large search space resulting from the discretization of the input variables. In addition, EDAs based on higher degrees of dependencies are better suited for the more complex problems as they exhibit slow convergence.

The study by Smithies et al. [80] aims at predicting the recurrence of colorectal cancer. In particular, the study tests the efficacy of a type of chemotherapy termed FUFA in preventing cancer recurrence. The data set obtained from NHS hospitals in the UK features different types of attributes and a considerable amount of missing data. The classification process started with a relaxed linear regression stage to remove irrelevant attributes and those with markedly missing data. Next, data is clustered to better deal with the different data types and allow a better inference mechanism for missing data rather than statistical methods. A new clustering methodology is introduced that is tailored to data with mixed attribute types. Each cluster is then fed to a three-layer feed-forward ANN ensemble. The members of each ensemble differ only in the number of hidden nodes. For training the neural nets, local search is combined with a modified form of back propagation algorithm known as the batched error back propagation with an enhanced Resilient Propagation for learning rate adaptation (iRPROP). The combination of gradient-independent local search with the gradient-based enhanced iRPROP has shown to enhance the classification performance. In addition, the proposed search algorithm utilizes the forbidden neighborhood region idea from tabu search. Given that previous studies focus on statistical models, the 66% of patients being correctly predicted forms a promising result and encourages further enhancements.

Reference [81] extends the ACO algorithm to tackle unconstrained continuous optimization problems; ACOR. It is then used to optimize the weights for a feed-forward ANN used in a medical classification task. Classification is applied on three PROBIN benchmark medical datasets. The resulting performance was compared against two neural network training algorithms; BP and LM. Training–testing method was used rather than the cross-validation. As a general optimization algorithm and unlike BP and LM, ACOR does not require that the neuron function is known and is differentiable. However, ACOR does not exploit additional information, such as the gradient information. Results show that the performance of ACOR was inferior to the other two. A hybrid approach was also tested where ACOR was combined with BP (ACOR-BP) and with LM (ACOR-LM). In the hybrid approach, each solution of ACOR is enhanced by running a single iteration of the BP or LM methods respectively. The performance of the hybrid approached was comparable to BP and LM and in some cases outperformed them. The proposed algorithms where also compared to GA and its' hybrids (GA-BP and GA-LM) on the same data sets, and have significantly outperformed them.

Castillo et al. [16] and Ince et al. [43] investigate the optimization of ANN architecture and training error. Castillo et al. [16] used a multi-objective evolutionary algorithm called MG-Prob for the simultaneous optimization of three objectives; the reduction of type-I and type-II errors as well as minimizing the artificial neural network's size. MG-Prob is based on the Single Front Genetic Algorithm (FSGA) [20] that builds on the Pareto optimality principle. Elite set represents the non-dominated individuals and in FSGA only a diverse part of this set that is spread across the search space is copied into the next generation. Individuals are multi-layer perceptrons (MLPs). The resulting non-dominated individuals in the population are used as an ensemble to perform the classification. Three methods were used to combine the ensemble results; voting, average and largest activation among all outputs. Experiments on breast cancer dataset from the UCI repository demonstrate the effectiveness of this method as compared to other methods obtaining slightly better classification error with a minimal difference between the two type errors, and smaller network size for individual MLPs.

PSO was used by Ince et al. [43] to set the parameters for feed-forward fully-connected ANNs and compared the results with those obtained by the traditional BP training algorithm for different training depths (deep/shallow). Experiments were conducted on three UCI Probin1 medical datasets; breast cancer, heart disease and diabetes. Classification Error Percentage (*CEP*) was used to measure performance were $CEP = 1 - P$. Other performance measures were used including mean square error (*MSE*) and average processing time (ms). The study concluded that PSO has better generalization ability and more stable performance with respect to changing network architecture. On the other hand, BP resulted in better classification accuracy for smaller networks. The accuracy for PSO and BP is otherwise comparable. BP training however was consistently superior in terms of computational complexity.

7.5.3 Hybrid Metaheuristics for Full Model Selection in Medical Data Classification

PSO is used to perform a full model selection (FMS) for a classification task [24]. No background knowledge about the problem is required. Full model selection involves choosing and chaining preprocessing methods (zero or more), feature subset selection method (zero or one), a learning algorithm and a post processing method (zero or one). FMS includes the choice of all the associated parameters as well as the order of preprocessing and feature subset selection (i.e. to perform FSS first or else preprocessing first). The choice of these methods is made from objects available at the CLOP machine learning package. This package includes three preprocessing methods, twelve feature selection methods, ten ML classification algorithms and a single post processing method. Each individual is encoded such that it represents a definition to all the previous stages. Fitness is evaluated in terms of balanced error rate (BER). The advantage of this evaluation criterion is that it considers classification errors in both classes and thus avoids rewarding an algorithm that favors the majority class. Computational complexity is reduced by means of sub-sampling heuristic. The proposed PSO-based FMS was compared to another FMS method that is based on a simple direct search and optimization algorithm termed pattern search (PS) [21]. Results show that the PSO alternative consistently outperformed the PS-based FMS. The proposed algorithm was also challenged against other models that use background knowledge or are based on model selection for a single learning algorithm in the framework of a model selection competition named Agnostic Learning vs. Prior Knowledge Challenge. The proposed model has demonstrated comparable results.

7.6 Conclusion

Medical data classification is a new field of research that will improve the cost, accessibility, and quality of health care. The complexity associated with medical data classification prohibits the use of exact methods. This chapter overviews the state-of-the-art approaches in medical data classification. Studies suggest the use of fuzzy and hybrid meta-heuristic methods for model learning, selection, and optimization. Hybridizing different approximation algorithms, including metaheuristics, is a promising approach. However, most of the studies choose their system components arbitrarily, for example due to their success in other fields of study. Model comprehensibility is an important factor in the selection of these components. There exists a need for a meta-study that focuses on the basis of choosing hybrid system components for medical data classification. Perhaps the work by Tanwani and his team [85, 86, 87, 88] on formalizing medical data complexity forms a gateway to this area.

The use of transparent comprehensible models enables physicians to validate the clinical plausibility of the resulting classification hypothesis and allows discovering new interesting relations. In some cases, obtaining explanations and conclusions that enlighten and convince medical experts is more important than suggesting a particular class. XCS particularly showed comprehensible results with high classification accuracies. Several studies employing XCS were presented in this document. However, there is still room for research and improvement. For example, the use of XCS ensembles is an interesting line of research to be investigated.

Most of the studies highlighted in this document aim at the design of general classification systems rather than systems that are geared towards a specific medical disorder. It is true that medical data have many common characteristics. However, each dataset has its own character (for example, see Table 7.1). Therefore, classification models that work well with one dataset may not exhibit the same level of performance on another. Medical data classification may benefit more with focused research.

Another issue is the evaluation metric. Most studies limit the evaluation metric to precision (P). As this may be satisfactory in the machine learning and data mining community, it may be not for the medical community. The inclusion of other evaluation metrics that are routinely used in the medical field would certainly give more value and depth to the results. For example, Holmes EpiXCS tailor XCS to the epidemiologic community mainly by considering evaluation metrics needed by the epidemiologic field.

Due to the societal significance of the subject and also to the computational challenge it presents, more research in the field of medical data classification is needed. The papers introduced in this chapter only represent the tip of the iceberg in the medical data classification field. Despite this wealth in literature, very few systems are put into practical use. The inclusion of a medical professional in the study team is a valuable asset that is most often ignored. Several authors have addressed the clinical approval of intelligent systems before [34, 56, 72] and so they need to be considered seriously.

Acknowledgements. The authors would like to thank Dr. Bryan Bergeron, a member of the Affiliate Faculty, Harvard-MIT Division of Health Sciences and Technology, for his valuable input in reviewing this document.

References

1. Abbas, H.A.: Speeding up Back-Propagation using Multiobjective Evolutionary Algorithms. Neural Computation 15(11), 2705–2726 (2003)
2. Aguilar-Ruiz, J., Riquelme, J., Toro, M.: Evolutionary Learning of Hierarchical Decision Rules. IEEE Transactions on Systems, Man, and Cybernetics, Part B 33(2), 324–331 (2003)
3. Aci, M., Inan, C., Avci, M.: A Hybrid Classification Method of K-Nearest Neighbor, Bayesian Methods and Genetic Algorithm. Expert Systems with Applications 30, 5061–5067 (2010)
4. Bacardit, J.: Pittsburgh Genetics-Based Machine Learning in the Data Mining era: Representations, generalization, and run-time. Ramon Llull University, Dissertation (2004)
5. Bacardit, J., Butz, M.V.: Data Mining in Learning Classifier Systems: Comparing XCS with GAssist. In: Kovacs, T., Llorà, X., Takadama, K., Lanzi, P.L., Stolzmann, W., Wilson, S.W. (eds.) IWLCS 2003. LNCS (LNAI), vol. 4399, pp. 282–290. Springer, Heidelberg (2007)
6. Bacardit, J., Bernadó-Mansilla, E., Butz, M.V.: Learning Classifier Systems: Looking Back and Glimpsing Ahead. In: Bacardit, J., Bernadó-Mansilla, E., Butz, M.V., Kovacs, T., Llorà, X., Takadama, K. (eds.) IWLCS 2006 and IWLCS 2007. LNCS (LNAI), vol. 4998, pp. 1–21. Springer, Heidelberg (2008)
7. Back, T., Fogel, D., Michalewicz, Z.: Handbook of Evolutionary Computation. Oxford University Press, London (1997)
8. Bernadó-Mansilla, E., Garrell-Guiu, J.M.: Accuracy based learning classifier systems: models, analysis and applications to classification tasks. Evolutionary Computation 11(3), 209–238 (2003)
9. Bhattacharyya, D., Pal, A.J., Kim, T.: Cell-graph coloring for cancerous tissue modeling and classification. Multimedia Tools and Applications (2011), doi:10.1007/s11042-011-0797-y
10. Bojarczuk, C., Lopes, H., Freitas, A.: Genetic Programming for Knowledge Discovery in Chest Pain Diagnosis. IEEE Engineering in Medicine and Biology Magazine 19(4), 38–44 (2000)
11. Bojarczuk, C., Lopes, H., Freitas, A., Michaliewicz, E.: A Constrained-syntax Genetic Programming System for Discovering Classification Rules: Application to Medical Data Sets. Artificial Intelligence in Medicine 30(1), 27–48 (2004)
12. Bradley, A.P.: The Use of the Area under the ROC Curve in the Evaluation of Machine Learning Algorithms. Pattern Recognition 30, 1145–1159 (1997)
13. Bull, L., Hurst, J.: ZCS Redux. Evolutionary Computation 10(2), 185–205 (2002)
14. Bull, L., Bernadó-Mansilla, E., Holmes, J.: Learning Classifier Systems in Data Mining: An Introduction. In: Bull, L., Bernadó-Mansilla, E., Holmes, J. (eds.) Learning Classifier Systems in Data Mining: Studies in Computational Intelligence, pp. 1–16. Springer (2008)
15. Candelieri, A.: A hyper-solution framework for classification problems via metaheuristic approaches. 4OR-Q J Oper Res (2010), doi:10.1007/s10288-011-0166-8
16. Castillo, P.A., Arenas, M., Merelo, J.J., Rivas, V.M., Romero, G.: Multiobjective Optimization of Ensembles of Multilayer Perceptrons for Pattern Classification. In: Runarsson, T.P., Beyer, H.-G., Burke, E.K., Merelo-Guervós, J.J., Whitley, L.D., Yao, X. (eds.) PPSN 2006. LNCS, vol. 4193, pp. 453–462. Springer, Heidelberg (2006)
17. Chen, S., Lin, S., Chou, S.: Enhancing the Classification Accuracy by Scatter Search-Based ensemble Approach. Applied Soft Computing 11, 1021–1028 (2011)
18. Chorbev, I., Mihajlov, D., Jolevski, I.: Web Based Medical Expert System with Self Training Heuristic Rule Induction Algorithm. In: Proceedings of the First International Conference on Advances in Databases, Knowledge, and Data Application (DBKDA 2009), pp. 143–148. IEEE Computer Society, Washington, DC (2009), doi:10.1109/DBKDA.2009.21

19. Cruz, J., Wishart, D.: Applications of Machine Learning in Cancer Prediction and Prognosis. Cancer Informatics 2006(2), 59–77 (2006)
20. de Toro, F., Ortega, J., Fernandez, J., Diaz, A.: Parallel genetic algorithm for multiobjective optimization. In: Proceedings of the 10th Euromicro Workshop on Parallel, Distributed and Network-based Processing, pp. 384–391. IEEE Computer Society (2002)
21. Dennis, J., Torczon, V.: Derivative-free pattern search methods for multidisciplinary design problems. In: Proceedings of the 5th AIAA/ USAF/NASA/ISSMO Symposium on Multidisciplinary Analysis and Optimization, Panama City, FL, pp. 922–932 (1994)
22. Dorigo, M.: Optimization, Learning and Natural Algorithms, Dissertation, Politecnico di Milano, Italie (1992)
23. Dybowski, R., Weller, P., Chang, R.: Gant V Prediction of outcome in critically ill patients using artificial neural network synthesized by genetic algorithm. Lancet 347(9009), 1146–1150 (1996)
24. Escalante, H., Montes, M., Sucar, L.: Particle Swarm Model Selection. J. Mach. Learn. Res. 10, 405–440 (2009)
25. Fan, C., Chang, P., Hsieh, J.L.: A Hybrid Model Combining Case-based Reasoning and Fuzzy Decision Tree for Medical Data Classification. Applied Soft Computing 11, 632–644 (2011)
26. Fogel, L.: Evolutionary Programming in Perspective: the Top-Down View. In: Zurada, J., Marks II, R., Robinson, C. (eds.) Computational Intelligence: Imitating Life, pp. 135–146. IEEE Press (1994)
27. Fogel, L., Owens, A., Walsh, M.: Artificial Intelligence through a Simulation of Evolution. In: Callahan, A., Maxfield, M., Fogel, L.J. (eds.) Biophysics and Cybernetic Systems, pp. 131–155. Spartan, Washington DC (1965)
28. Gao, Y., Huang, J., Rong, H.: Learning Classifier System Ensemble for Data Mining. In: Proceedings of the 2005 Genetic and Evolutionary Computation Conference IWLCS, pp. 63–66 (2005)
29. Glover, F.: Tabu Search - Part I. ORSA Journal on Computing 1(3), 190–206 (1989)
30. Glover, F.: Tabu Search - Part II. ORSA Journal on Computing 2(1), 4–32 (1990)
31. Glover, F.: A Template for Scatter Search and Path Relinking. In: Hao, J.-K., Lutton, E., Ronald, E., Schoenauer, M., Snyers, D. (eds.) AE 1997. LNCS, vol. 1363, p. 13. Springer, Heidelberg (1998)
32. Hagan, M.T., Menhaj, M.B.: Training Feedforward Networks with the Marquardt Algorithm. IEEE Transactions on Neural Networks 5, 989–999 (1994)
33. Han, J., Kamber, M.: Data Mining: Concepts and Techniques, 2nd edn., Jim Gray. The Morgan Kaufmann Series in Data Management Systems. Series Editor Morgan Kaufmann Publishers (2006) ISBN 1-55860-901-6
34. Hanson, C.W.: Marshall BEArtificial intelligence applications in the intensive care unit. Crit. Care Med. 29(2), 427–435 (2001)
35. Hastie, T., Tibshirani, R., Friedman, J.: The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd edn. Springer (2009) ISBN: 9780387848570
36. Holland, J.H.: Adaptation in Natural and Artificial Systems, 1st edn. The University of Michigan Press, Ann Arbor (1975); MIT Press, Cambridge, MA (1992)
37. Holland, J., Reitman, J.: Cognetive Systems based on Adaptive Agents. In: Waterman, D.A., Inand, F. (eds.) Pattern-Directed Inference Systems, Hayes-Roth (1978)
38. Holmes, J.: Discovering Risk of Disease with a Learning Classifier System. In: Proceedings of the 7th International Conference on Genetic Algorithms (ICGA 1997), pp. 426–433 (1997)
39. Holmes, J., Sager, J.: Rule Discovery in Epidemiologic Surveillance Data Using EpiXCS: An Evolutionary Computation Approach. In: Miksch, S., Hunter, J., Keravnou, E.T. (eds.) AIME 2005. LNCS (LNAI), vol. 3581, pp. 444–452. Springer, Heidelberg (2005)
40. Holmes, J., Durbin, D., Winston, F.: Discovery of Predictive Models in an Injury Surveillance Database: An Application of Data Mining in Clinical Research. In: Proceedings of AMIA Symposium, pp. 359–363 (2000a)

41. Holmes, J., Durbin, D., Winston, F.: The learning classifier system: an evolutionary computation approach to knowledge discovery in epidemiologic surveillance. Artificial Intelligence in Medicine 19, 53–74 (2000b)
42. Holmes, J., Lanzi, P., Stolzmann, W., Wilson, S.: Learning Classifier Systems: New Models, Successful Applications. Information Processing Letters archive 82(1), 23–30 (2002)
43. Ince, T., Kiranyaz, S., Pulkkinen, J., Gabbouj, M.: Evaluation of global and local training techniques over feed-forward neural network architecture spaces for computer-aided medical diagnosis. Expert Systems with Applications 37, 8450–8461 (2010)
44. Jain, R., Mazumdar, J., Moran, W.: Application of fuzzy-classifier system to coronary artery disease and breast cancer. Australas Phys. Eng. Sci. Med. 21(3), 141–147 (1998)
45. Jiang, Y., Metz, C.E., Nishikawa, R.M.: A Receiver Operating Characteristic Partial Area Index for Highly Sensitive Diagnostic Tests. Radiology 201, 745–750 (1996)
46. Kennedy, J., Eberhart, R.C.: Particle swarm optimization. In: Proceedings of IEEE International Conference on Neural Networks, Piscataway, NJ, vol. 4, pp. 1942–1948 (1995)
47. Kharbat, F., Bull, L., Odeh, M.: Mining breast cancer data with XCS. In: Proceedings of the 9th Annual Conference on Genetic and Evolutionary Computation, pp. 2066–2073 (2007)
48. Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P.: Optimization by Simulated Annealing. Science. New Series 220(4598), 671–680 (1983)
49. Kotsiantis, S.B.: Supervised Machine Learning: A Review of Classification. Informatica Journal 31, 249–268 (2007)
50. Kotsiantis, S.B.: Feature selection for machine learning classification problems: a recent overview. Artif. Intell. Rev (2011), doi:10.1007/s10462-011-9230-1
51. Koza, J.R.: Genetic Programming. MIT Press, Cambridge (1992)
52. Lanzi, P.: Extending the Representations of Classifier Conditions. Part II: From Messy coding to S-expression. In: Banzhaf, W., et al. (eds.) Proceedings of the Genetic and Proceedings of the Genetic and Evolutionary Computation Conference, vol. 1, pp. 345–352 (1999)
53. Lanzi, P.: Mining interesting knowledge from data with the XCS classifier system. In: Proceedings of the Genetic and Evolutionary Computation Conference (GECCO 2001), pp. 958–965. Morgan Kaufmann, San Francisco (2001)
54. Larrañaga, P., Calvo, B., Santana, R., et al.: Machine Learnig in Bioinformatics. Brief Bioinform. 7(1), 86–112 (2006)
55. Lavrac, N.: Selected Techniques for Data Mining in Medicine. Artificial Intelligence in Medicine 16(1), 3–23 (1999)
56. Lisboa, P.J.: Taktak AFG The use of artificial neural networks in decision support in cancer: A systematic review. Neural Networks 19(4), 408–415 (2006)
57. Lo, H.-Y., Chang, C.-M., Chiang, T.-H., et al.: Learning to Improve Area-Under-FROC for Imbalanced Medical Data Classification Using an Ensemble Method. In: ACM SIGKDD Explorations Newsletter, vol. 10(2), ACM, New York (2008), doi:10.1145/1540276.1540290
58. Lucas, F.: Analysis of Notions of Diagnosis. Artificial Intelligence 105(12), 295–343 (1998)
59. Lucas, F., Abu-Hanna, A.: Prognosis Methods in Medicine. Artificial Intelligence in Medicine 15(2), 105–119 (1998)
60. Luukka, P., Lampinen, J.: Differential Evolutionary Classifier in Noisy Settings with Interactive Variables. Applied Soft Computing 1, 891–899 (2011)
61. Madera, J., Dorronsoro, B.: Estimation of Distribution Algorithms. In: Metaheuristic Procedures for Training Neutral Networks Operations Research/Computer Science Interfaces Series, Part III, vol. 36, pp. 87–108 (2006)
62. Moscato, P.: On Evolution, Search, Optimization, Genetic Algorithms and Martial Arts: Towards Memetic Algorithms - Caltech Concurrent Computation Program, C3P Report (1989)
63. Narayanan, M.N., Lucas, S.B.: A genetic algorithm to improve a neural network to predict a patient's response to warfarin. Methods Inf. Med. 32(1), 55–58 (1993)

64. Penã-Reyes, C., Sipper, M.: A fuzzy-genetic approach to breast cancer diagnosis. Artif. Intell. Med. 17(2), 131–135 (1999)
65. Penã-Reyes, C., Sipper, M.: Evolutionary Computation in Medicine: an Overview. Artif. Intell. Med. 19(1), 1–23 (2000)
66. Peroumalnaik, M., Enee, G.: Prediction using Pittsburgh Learning Classifier Systems: APCS use case. In: Proceedings of the 12th Annual Conference Companion on Genetic and Evolutionary Computation GECCO 2010, pp. 1901–1907 (2010)
67. Pham, H.N.A., Triantaphyllou, E.: The impact of overfitting and overgeneralization on the classification accuracy in data mining. In: Maimon, O., Rokach, L. (eds.) Soft Computing for Knowledge Discovery and Data Mining, Part 4, pp. 391–431. Springer, New York (2007)
68. Pham, H.N., Triantaphyllou, E.: An application of a new meta-heuristic for optimizing the classification accuracy when analyzing some medical datasets. Expert Systems with Applications 36, 9240–9249 (2009)
69. Price, K., Storn, R., Lampinen, J.: Differential Evolution - A Practical Approach to Global Optimization. Springer (2005)
70. Provost, F., Kolluri, V.: A Survey of Methods for Scaling up inductive Algorithms. Datamining and Knowledge Discovery 3(2), 131–169 (1999)
71. Ranawana, R., Palade, V.: Optimized Precision, A New Measure for Classifier Performance Evaluation. In: Proceedings of the IEEE Congress on Evolutionary Computation, Vancouver, Canada, pp. 2254–2261 (2006)
72. Rao, R.B., Bi, J., Fung, G., et al.: LungCAD: A Clinically Approved, Machine Learning System for Lung Cancer Detection. In: Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, ACM, New York (2008), doi:10.1145/1281192.1281306
73. Rojanavasu, P., Dam, H., Abbass, H., Lokan, C., Pinngern, O.: A Self-Organized, Distributed, and Adaptive Rule-Based Induction System. IEEE Transctions on Neural Networks 20(3), 446–495 (2009)
74. Rumelhart, D., Hinton, G., Williams, R.: Learning Representations by Backpropagation Errors. Nature 323, 533–536 (1986)
75. Sarkar, B.K., Sana, S.S.: A Hybrid Approach to Design Efficient Learning Classifiers. Computers and Mathematics with Applications 58, 65–73 (2009)
76. Shortliffe, E., Cimino, J.: Biomedical Informatics: Computer Applications in Health Care and Biomedicine. Springer, New York (2006)
77. Skinner, B., Nguyen, H., Liu, D.: Classification of EEG Signals Using a Genetic-Based Machine Learning Classifier. In: Proceedings of the 29th Annual International Conference of the IEEE EMBS, Lyon, France, pp. 3120–3123 (2007a)
78. Skinner, B., Nguyen, H., Liu, D.: Distributed Classifier Migration in XCS for Classification of Electroencephalographic Signals. In: Proceedings of the IEEE Congress on Evolutionary Computation CEC 2007, Singapore, pp. 2829–2836 (2007b)
79. Smith, S.: A learning system based on genetic adaptive algorithms. Dissertation, University of Pittsburgh, Pittsburgh (1980)
80. Smithies, R.G., Salhi, S., Queen, N.M.: Predicting colorectal cancer recurrence: A hybrid neural networks-based approach. In: Ibaraki, T., Nonobe, K., Yagiura, M. (eds.) Metaheuristics: Progress as Real Problem Solvers. Series: Operations Research/Computer Science Interfaces Series, vol. 32, pp. 259–285 (2005)
81. Socha, K., Blum, C.: Ant Colony Optimization. In: Metaheuristic Procedures for Training Neutral Networks, Part IV. Operations Research/Computer Science Interfaces Series, vol. 36, pp. 153–180 (2006)
82. Stolzmann, W.: Anticipatory Classifier Systems. In: Proceedings of the 3rd Annual Genetic Programming Conference, pp. 658–664 (1998)
83. Stone, M.: Cross-validatory choice and assessment of statistical predictions. Journal of the Royal Statistical Society Series B 36, 111–147 (1974)
84. Tan, K., Yu, Q., Heng, C., Lee, T.: Evolutionary Computing for Knowledge Discovery in Medical Diagnosis. Artificial Intelligence in Medicine 27(2), 129–154 (2003)

85. Tanwani, A., Farooq, M.: Performance Evaluation of Evolutionary Algorithms in Classification of Biomedical Datasets. In: Proceedings of the 11th Annual Conference Companion on Genetic and Evolutionary Computation: Late Breaking Papers, GECCO 2009, Canada, pp. 2617–2624 (2009a)
86. Tanwani, A., Farooq, M.: The Role of Biomedical Dataset in Classification. In: Combi, C., Shahar, Y., Abu-Hanna, A. (eds.) AIME 2009. Lecture Notes in Computer Science (LNAI), vol. 5651, pp. 370–374. Springer, Heidelberg (2009b)
87. Tanwani, A.K., Farooq, M.: Classification Potential vs. Classification Accuracy: A Comprehensive Study of Evolutionary Algorithms with Biomedical Datasets. In: Bacardit, J., Browne, W., Drugowitsch, J., Bernadó-Mansilla, E., Butz, M.V. (eds.) IWLCS 2008/2009. Lecture Notes in Computer Science (LNAI), vol. 6471, pp. 127–144. Springer, Heidelberg (2010)
88. Tanwani, A., Afridi, J., Shafiq, M., Farooq, M.: Guidelines to Select Machine Learning Scheme for Classification of Biomedical Datasets. In: Pizzuti, C., Ritchie, M.D., Giacobini, M. (eds.) EvoBIO 2009. LNCS, vol. 5483, pp. 128–139. Springer, Heidelberg (2009)
89. Tzima, F., Mitkas, P.: ZCS Revisited: Zeroth-level Classifier Systems for Data Mining. In: Proceedings of the 2008 IEEE International Conference on Data Mining Workshops, pp. 700–709 (2008)
90. Unler, A., Murat, A.: Discrete Optimization: A discrete particle swarm optimization method for feature selection in binary classification problems. European Journal of Operational Research 206(3), 528–539 (2010)
91. Unold, O., Tuszynski, K.: Mining Knowledge from Data using Anticipatory Classifier Systems. Knowledge-Based Systems 21(5), 363–370 (2008)
92. Urbanowicz, R., Moore, J.: Review Article: Learning Classifier Systems: A Complete Introduction, Review and Roadmap. Journal of Artificial Evolution and Applications, 1–25 (2009)
93. Vieira, S.M., Sousa, J., Runkler, T.A.: Multi-Criteria Ant Feature Selection Using Fuzzy Classifiers. In: Swarm Intelligence for Multi-objective Problems in Data Mining: Studies in Computational Intelligence, vol. 242, pp. 19–36 (2009)
94. Walter, D., Mohan, C.: ClaDia: A Fuzzy Classifier System for Disease Diagnosis. In: Proceedings of the 2000 Congress on Evolutionary Computation, CA, USA, vol. 2, pp. 1429–1435 (2000)
95. Wilson, S.W.: ZCS: A Zeroth-Level Learning Classifier System. Evolutionary Computation 2(1), 1–18 (1994)
96. Wilson, S.W.: Classifier Fitness Based on Accuracy. Evolutionary Computation 3(2), 149–175 (1995)
97. Wilson, S.W.: Mining Oblique Data with XCS. In: Lanzi, P.L., Stolzmann, W., Wilson, S.W. (eds.) IWLCS 2000. LNCS (LNAI), vol. 1996, pp. 158–290. Springer, Heidelberg (2001)