지능정보연구 제14권 제2호 2008년 6월(pp.193~206)

Medical Diagnosis Problem Solving Based on the Combination of Genetic Algorithms and Local Adaptive Operations*

Ki-Kwang Lee
Institute of Industrial Management Research, School of Management, Inje University
(kiklee@inje.ac.kr)

Chang Hee Han
Division of Business Administration, Hanyang University
(chan@hanyang.ac.kr)

Medical diagnosis can be considered a classification task which classifies disease types from patient's condition data represented by a set of pre-defined attributes. This study proposes a hybrid genetic algorithm based classification method to develop classifiers for multidimensional pattern classification problems related with medical decision making. The classification problem can be solved by identifying separation boundaries which distinguish the various classes in the data pattern. The proposed method fits a finite number of regional agents to the data pattern by combining genetic algorithms and local adaptive operations. The local adaptive operations of an agent include expansion, avoidance and relocation, one of which is performed according to the agent's fitness value. The classifier system has been tested with well-known medical data sets from the UCI machine learning database, showing superior performance to other methods such as the nearest neighbor, decision tree, and neural networks.

Received : May 2008 Accepted : June 2008 Corresponding author : Chang Hee Han

1. Introduction

Classification learning systems are useful for decision making tasks in many diverse applications where classifying expertise is necessary (Tan et al., 2005; Weiss and Kulikowski, 1991). This wide range of applicability motivated many researchers to further refine classification methods in several research areas, such as statistical pattern recognition, machine learning and data mining (Tan et al., 2005; Jain, Duin and Mao, 2000; Michie et al., 1994; Simpson, 1992). The basic idea of classification is to assign a new data record represented as attributes to one of the possible classes with a minimal rate of misclassification. Solutions to a classification problem have been characterized in terms of parameterized or

* This work was supported by Korea Research Foundation Grant funded by Korea Government (MOEHRD, Basic Research Promotion Fund) (KRF-2005-003-D00464).

non-parameterized separation boundaries that could successfully distinguish the various classes in the attribute space (Pal et al., 1998).

Medical diagnosis can be considered a classification task : a record is a given patient's case, attributes are all patient's data including symptoms, signals, clinical history and results of laboratory tests etc., and the class is the diagnosis about disease or clinical condition that the physician has to discover, based on the patient's data (Bojarczuk et al., 2004). This paper proposes a new classification learning system to discover separation boundaries among various classes in the domain of medical diagnosis problems. A primary focus of the previous studies to build the separation boundaries has been on learning from examples, where a classifier system accepts case descriptions that are preclassified and then the system learns a set of separation surface that can classify new cases based on the preclassified cases (Nolan, 2002). Various learning techniques have been contrived to design the separation surfaces, depending on the kind of representation method, such as statistical functions (Duha and Hart, 1973), neural networks (McClellend and Rumelhart, 1988), and decision trees or production rules (Quinlan, 1986; Quinlan, 1993).

The method proposed in this paper to construct a classifier system for the medical diagnosis problem consists of two levels, i.e., the global search and the local improvement. The classifiers, which are delineated by geometrical ellipsoids and their variant, adjust their parameters to search the separation boundaries by using a hybrid method of a genetic algorithm (GA) and a heuristic local search algorithm.

In (Abe and Thawonmas, 1997), a classifier with ellipsoidal regions was shown to have the generalization ability comparable or superior to those of classifiers with the other shapes. Lee and Yoon (2005, 2006) and Lee et al. (2006) developed the examplar-based learning algorithm for the ellipsoids to fit the nonlinear separation surfaces of given data sets. Motivated by the result of (Abe and Thawonmas, 1997; Lee and Yoon, 2005; Lee et al., 2006; Lee and Yoon, 2008), the ellipsoids and their variants are adopted to fit the usual non-linear boundaries pattern inherent in the medical domain. The ellipsoid's variants are introduced to cover various nonlinear medical data patterns. While an ellipsoid is defined with two foci, the variants are supposed to have one or three foci. The specific ellipsoid with one central point is known as a sphere, and the other with three foci is called as a hyper-ellipsoid.

The advantage of the hybrid GAs is exploited in searching for a finite number of ellipsoids which can approximate the training data pattern while providing minimum misclassification of training sample points. GAs are iteratively run with a population consisting of individuals where the parameters of the ellipsoids are encoded. A local improvement procedure is incorporated into the GAs in order to complement a population-wide global search through an individual adaptive operation (Lee et al., 2006; Renders and Flasse, 1996). Each ellipsoid, i.e., individual in the population performs the adaptation procedure by appropriately moving, rotating, extending, or shrinking the ellipsoid itself according to the current state of the ellipsoid. The state of an ellipsoid

is represented by a fitness value, which is calculated from the class distribution of data located in the ellipsoid region and the number of attributes used. The details about the fitness value are explained in later section. The main idea of the adaptive procedure is to have the ellipsoid with high fitness values expanded in expectation of increase in the fitness value. On the other hand, the ellipsoid with low fitness value, due to existence of misclassified data, tries to avoid the misclassified data by relocation, rotation, or shrinkage.

The rest of this paper is organized as follows. Section 2 presents modeling of the general pattern classification problem in terms of three types of ellipsoidal individuals used in the hybrid GA. Section 3 describes the details about a proposed hybrid GA learning classifier system. Section 4 investigates the performance of the proposed method by applying it to well-known medical diagnosis data sets. Finally, conclusions are stated in section 5.

2. Modeling pattern classification problem with ellipsoidal regions

2.1 Pattern Classification Problem

Based on Lee and Yoon (2005, 2008) and Lee at al. (2006), our pattern classification problem assumes c classes in the n-dimensional pattern space $[0, 1]^n$ with continuous attributes. It is also supposed that a finite set of points $X = \{\mathbf{x}_p, p = 1, 2, \cdots, m\}$ are given as training data. Suppose that each point of X, $\mathbf{x}_p = (x_{p1}, x_{p2}, \cdots, x_{pn})$, is assigned to one of c classes, and let the corresponding subsets of X, having $N_1, N_2, \cdots, N_c$ points, respectively, be denoted $X_1, X_2, \cdots, X_c$. Because the pattern space is $[0, 1]^n$, values of attributes are $x_{pj} \in [0, 1]$ for $p = 1, 2, \cdots, m$ and $j = 1, 2, \cdots, n$. It is desired to cover the subset X_i ($i = 1, 2, \cdots, c$) by ellipsoidal regions labeled L_{ij} ($j = 1, \cdots$), so that new points can be assigned to one of c classes.

2.2 Classification rule representation with ellipsoidal regions

Assume that the data subset X_i for class C_i, where $i = 1, \cdots, c$, is covered by several ellipsoidal regions L_{ij}^K ($j = 1, \cdots$), where L_{ij}^K denotes the jth regional agent for class C_i. The regional agent L_{ij}^K is defined as three types according to the number of foci $\mathbf{f}_{ij}^{(k)}$, i.e., K and a constant, i.e., size factor, D_{ij} as follows :

$$L_{ij}^K : \sum_{k=1}^{K} dist(x, f_{ij}^{(k)}) \le D_{ij}, \text{ where } K = 1,$$
$$2 \text{ or } 3, \; dist(x, y) = \sqrt{(x-y)^t (x-y)} \qquad (1)$$

In Eq. (1), a regional agent L_{ij}^K can be interpreted as a sphere, ellipsoid, or hyper-ellipsoid, each of which has value of 1, 2, or 3 for the parameter K.

For each regional agent L_{ij}^K, we define the following classification rule where R_{ij} denotes the label of the j th rule for class C_i.

R_{ij} : If $\mathbf{x}$ is in L_{ij}^K then $\mathbf{x}$ belongs to class C_i. (2)

2.3 Classification rule strength and determination of class

For the pattern classification, it is reasonable to assume that the degree of membership of **x** for classification rule (2) increases as **x** moves toward the center of the ellipsoid L_{ij}^{K}, and decreases as **x** moves away from the center. To realize this characteristic, the degree of membership of **x** for a rule R_{ij} is defined as follows.

$$d_{ij}^{K}(x) = \frac{D_{ij}}{\sum_{k=1}^{K} dist(x, f_{ij}^{(k)})} \quad (3)$$

If the value of d_{ij}^{K}(x) in (3) is larger than 1, it indicates that point **x** is located within the ellipsoid L_{ij}^{K}. The value of (3) is less than 1 when **x** lies out of the boundary of the ellipsoid. Now the degree of membership of x for class C_i, denoted as d_i(**x**), is given by $d_i(x) = \max_{\forall j,K}\{d_{ij}^{K}(\mathbf{x})\}$. The class of input **x** is then determined as class C_{i^*} such that d_{i^*}(**x**) is the maximum among d_i(**x**), i = 1, ⋯, c.

3. Hybrid genetic algorithms for evolution of regional agents

This section describes how to evolve three types of ellipsoids based on a set of training data X_i for class C_i, where i = 1, 2, ⋯, c, to classify an n-dimensional input vector **x** into one of c classes for medical diagnosis problems. Each of a population of regional individuals is assigned to one of three types of ellipsoids and one of c classes. The proposed hybrid GA method is adopted to make coverage for given training data set. The hybrid GA interweaves the two evolutionary methods, relating the use of GA to the concept of "evolution" of a population of individuals and that of individual adaptation to the concept of "life" for each individual. In general, GA takes a population and makes it evolve in such a way that most of the population reaches the global optimization. The traditional GA makes the transition from one generation to another through crossover operations for the individuals selected by evaluating instantaneous fitness values at their "birth." The hybrid method, then, lets GA make its selections based on the fitness at the end of the individual life, resulting from the local adaptive operations as shown in <Figure 1>. The outline of the proposed hybrid GA procedure is as follows.

Step 1 : Generate an initial population of ellipsoidal individuals.

Step 2 : Evaluate each individual in the current population.

Step 3 : Perform the general GA operations : selection, crossover, and mutation.

Step 4 : Improve the newly generated offspring by the adaptation procedure.

Step 5 : Update the current population using an elit-

ist model.

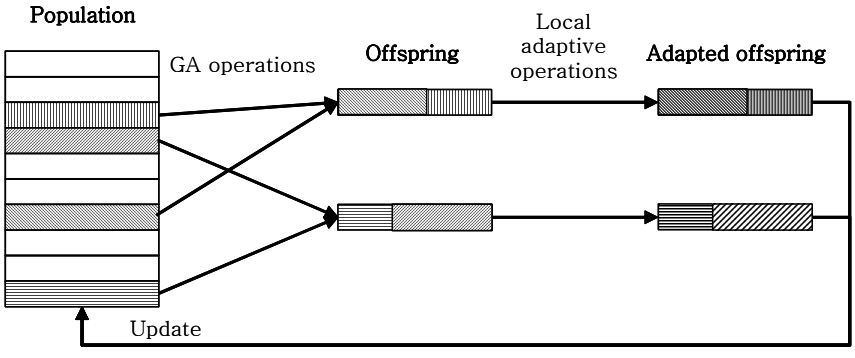
Step 6 : Repeat the **Step 2** to **Step 5** until termination criterion is achieved.

Each step of the procedure will be explained in detail in the following sections.

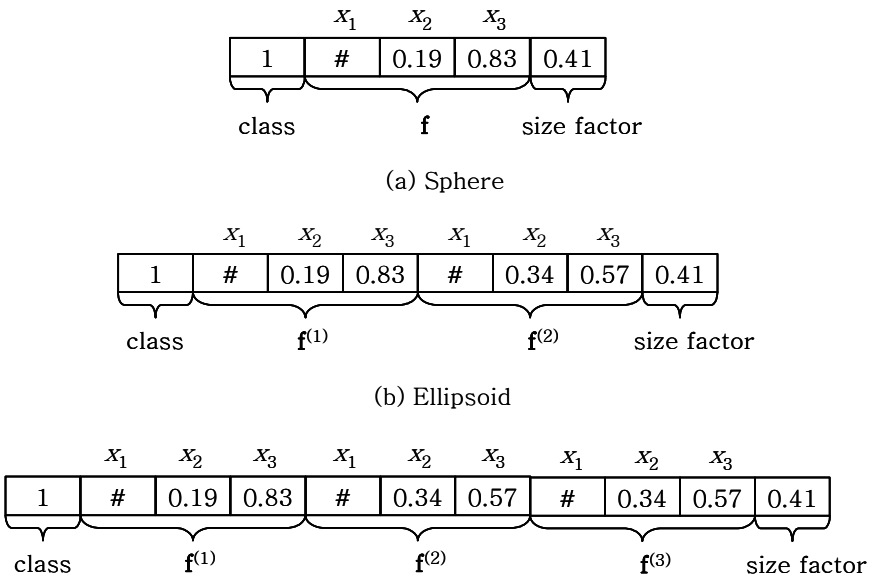
3.1 Chromosome representation and population initialization

This study adopts a real-coded GA in which the parameters of regional agents to be optimized are represented by a direct floating-point. The real-coded GA is expected to offer the advantages of being better adapted to numerical optimization for problems with continuous spaces, of speeding up the search and of making easier the development of approaches "hybridized" with the individual adaptation method to be described later. The chromosomes are represented by strings of a floating-point value in [0, 1] and # (don't care), encoding the parameters of ellipsoids. In order to represent a general medical diagnosis problem as a classification problem in a space of $[0, 1]^n$, every value of the attributes used in the medical diagnosis problem should be normalized into [0, 1]. <Figure 2> shows the structures for three types of chromosomes in three-dimensional pattern space as an example. The "don't care" bit in the string means that the corresponding attribute is not necessitated, accordingly it will be excluded from the feature space. Hence, an automatic attribute selection can be realized in the evolution process by utilizing the "don't care" bits. It should be noted that the "don't care" bits of two or three foci are located in the same feature positions because the foci must be defined by the same dimensional space.

An initial population is generated in such a way that each individual assigned to one of the classes and the types of ellipsoids is encoded in terms of foci, $\mathbf{f}^{(K)}$ and size factor D, which are randomly allocated in pattern space $[0, 1]^n$ to ensure sufficient diversity in the population. For each of



<Figure 1> Framework of the hybrid GA



<Figure 2> An example of chromosomes in three-dimensional attribute space

half individuals in the population, then, one of the foci is seeded with randomly selected training sample point for providing a good starting solution. The number of individuals with a certain class C_i, denoted by *Pop (i)*, in the population is determined in proportion to the number of training data with the same class. Consequently, the size of the population, denoted by *Pop_size*, is defined as the sum of *Pop (i)*, i =1, ···, c.

3.2 Fitness computation

Two measures are considered to evaluate a regional agent based on the classification result of the corresponding classifier : generalization ability and classification rate. In order to obtain good generalization ability of a regional agent, the region that the agent covers needs to grow as large as possible. Therefore, when the training data is divided by the agents, the number of data belonging to a regional agent should not be too small. For the high classification rate of a regional agent, the number of correctly classified data should be large relative to the number of incorrectly classified data among data belonging to the agent.

Considering the two measures, the fitness value of each agent is defined as follows, where *fitness* (L_{ij}^K) is the fitness value of the ellipsoid L_{ij}^K, *NC* (L_{ij}^K) is the number of training data that are correctly

classified by L_{ij}^K, $NI(L_{ij}^K)$ is the number of training data that are incorrectly classified by L_{ij}^K, and $weight(L_{ij}^K)$ is the weight value that multiplies $NI(L_{ij}^K)$.

$$fitness(L_{ij}^K) = NC(L_{ij}^K) - weight(L_{ij}^K) \times NI(L_{ij}^K) \quad (4)$$

The weight value for an ellipsoid L_{ij}^K is used to determine the tradeoff between the generalization ability and the expected classification rate of the ellipsoid on the basis of the ratio of the number of data with same class C_i to the total number of remaining data. Given a training data set with a large value of the ratio the ellipsoids is apt to be large with the large fitness value caused by large expected value of $NC(\ )$ and small expected value of $NI(\ )$. This will over-emphasize the generalization power relative to the classification rate. On the other hand, a small value of the ratio has the ellipsoids be small with the small fitness value caused by small $NC(\ )$ and large $NI(\ )$, which leads to low generalization ability. If the ratio has a large value, the expected classification rate should be emphasized with a large value of weight. Otherwise the generalization ability should be emphasized with a small value of weight. Based on the above relation, the weight value of each ellipsoid is calculated as follows, where N_i is the number of data of which class is C_i among remaining training data, N_{remain} is the number of total remaining training data, and $\alpha(\alpha > 1)$ and $\beta(0 < \beta < 1)$ is constant.

$$weight(L_{ij}^K) = \exp\left\{\alpha \times \left(\frac{N_i}{N_{remain}} - \beta\right)\right\} \quad (5)$$

3.3 Genetic operations

In order to generate new offspring for class C_i, a pair of individuals with the same class C_i is selected from the current population. Each individual is selected by the following selection probability based on the roulette wheel selection with the linear scaling, where $fitness_{\min}(S_i)$ is the minimum fitness value of the individuals in the current set S_i.

$$P(L_{ij}^K) = \frac{fitness(L_{ij}^K) - fitness_{\min}(S_i)}{\sum_{L_{ik}^K \in S_i}\left\{fitness(L_{ik}^K) - fitness_{\min}(S_i)\right\}} \quad (6)$$

From the selected pair of ellipsoids, the arithmetic crossover for randomly taken genes generates two offspring. For an example of the i-th genes, a_i and b_i of the selected pair of ellipsoids are replaced by $\lambda a_i + (1-\lambda) b_i$ and $(1-\lambda) a_i + \lambda b_i$ respectively, where $0 < \lambda < 1$. Note that the size factor is determined by a random number drawn from a uniform distribution $U(dist(\mathbf{f}^{(1)\prime}, \mathbf{f}^{(2)\prime}), 1)$ in order to keep the size of the ellipsoid greater than distance between its two modified foci.

Each parameter of ellipsoids generated by the crossover operation is randomly replaced using a random number from U (0, 1) at a prespecified mutation probability. As in the crossover operation, the size factor is recomputed with the modified distance between the two altered foci.

3.4 Local adaptive operations

Local adaptive operations are devised to make up for the weak point of GAs in finding the precise local solution in the pattern space where the algorithm converges. The adaptive operations have three operations, i.e., expansion, avoidance, and relocation, of which the most probable one is selected for each regional agent based on its fitness value. The agent with a positive value of fitness is expanded to have a chance to contain more data patterns. If an agent has a negative value of fitness (i.e., an agent includes at least one misclassified data), the agent avoids the misclassified examples by rotating or contracting itself. Finally if the fitness value of an ellipsoid is zero (i.e., an ellipsoid does not contain any training data), the agent moves to other location in the pattern space. The fitness value of an ellipsoid can be zero even though the ellipsoid contains data from (5). However, there is a bare possibility that the number of correctly classified data is same as the number of misclassified data multiplied by weight value because the value of weight in (6) is a real number calculated by an exponential function with a real number of parameter. Nevertheless, the overall performance does not take a sudden turn for the worse (Lee and Yoon, 2005; Lee et al., 2006; Lee and Yoon, 2008).

In summary, each agent in a pool is updated by iteratively executing one of three adaptive operations based on the fitness value of the ellipsoid. The adaptive operations can make the fitness values of the agent either larger or smaller than those of the agent before the adaptation. The update of the regional pool is performed to the only agents whose fitness values are increased after the local adaptation process. The three adaptive operations for three types of regional agents are described in <Table 1>, of which main idea comes from the adaptive methods for ellipsoids proposed in Lee et al. (2006).

<Table 1> Adaptive operations for three types of agents

Agent type	Operation	Description
Sphere	Avoidance	Moving the center to the opposite direction of the negative example
	Expansion	Increasing the size factor, i.e., radius
	Move	Randomly allocating the location of the center
Ellipsoid or Hyper-Ellipsoid	Avoidance	Moving the focus which is close to the negative example to the opposite direction of the example
	Expansion	Increasing the size factor
	Move	Randomly allocating the location of a focus selected from two or three foci at random

3.5 Update of the population

The proposed hybrid GA procedure applies genetic operations after population elitist selection (Eshelman, 1991). With the population elitist selection, predefined *Pop_size* individuals are selected from the current population and a set of the newly generated offspring. This updating method guarantees that the best *Pop_size* individuals seen so far always survived.

3.6 Termination test for GA

The termination criterion used in this paper is to terminate the iteration of the GA operations when either all the training samples are covered by the regional agents in the population or the specified maximum number of iterations is exceeded. The ultimate solution obtained by the GA procedure is not the final population itself but the best agents in the final population, which cover all the training samples contained by the final population. The selection of the best agents in the final population for the eventual output of the algorithm can eliminate the redundant agents whose removal does not change the recognition capability of the classifier.

4. Experimental results

Three data sets, on liver, diabetes, and breast cancer, respectively, were analyzed to verify the effectiveness of the proposed hybrid GA-based classification methods. The data sets are available from the UCI machine learning repository (Blake and Merz, 2006) and have real values of attributes. As a preprocessing of the data for the proposed classifier system, the value of each attribute was normalized as having the maximum value of one and the minimum value of zero.

4.1 Medical data sets used for performance evaluation

The performance of the proposed classification method was evaluated on three data sets, i.e., liver, diabetes, and breast cancer, which are usually employed as benchmarks for medical decision making applications. The liver disorder data set consists of 345 examples with six attributes and three classes. The classification task is to predict whether a patient's liver is disordered from excessive alcohol consumption. The class is distributed with 200 (57.97%) liver disorder positive examples, 145 (42.03%) negative examples. The data set of diabetes has 768 examples belonging to two classes, i.e., positive or negative for diabetes, described by eight numeric-valued attributes. The diagnosis class is distributed with 500 (65.1%) negative examples and 268 (34.9%) positive examples. The Wisconsin diagnostic breast cancer (WDBC) data set was used for breast cancer diagnosis problem. The data set consists of 569 patient's records with 30 attributes and two classes of benign or malignant. Among them, 212 (37.26%) are reported to have breast cancers whereas the remaining 357 (62.74%) are not.

The summary of the data sets is described in <Table 2>. The data sets were normalized so that real-valued attributes ranged from [0, 1], and then each data set was partitioned into training set and test set for the tenfold cross validation.

<Table 2> Summary of data sets used for evaluation

Data set	No. of attributes	No. of classes	No. of examples
Liver	6	2	345
Diabetes	8	2	768
WDBC	30	2	569

4.2 Results of computational evaluation

Here we present the results achieved by the proposed hybrid GA approach and compare them with the performances of existing well-known classifiers, i.e., a *k*-nearest neighbor (Weiss et al., 1991), a decision tree with C4.5 (Quinlan, 1993), and a neural network with backpropagation (Quinlan, 1994). This paper tried to guarantee the proper prediction power of the classifiers even under insufficient training data for scarce classes by adopting rather simple structures such as $k = 3$ and one hidden layer. In order to evaluate the overall classification capability of the classifiers, the misclassification costs were taken into account. It is apparent that the cost associated with Type I error (i.e., a patient without disease is misclassified as a patient with disease) and Type II error (i.e., a patient with disease is misclassified as a patient without disease) are significantly different (Ross, 1987; Chou et al., 2004). In general, the misclassification costs related with Type Ⅱ errors are much higher than those of Type I errors. Hence, special attention should be paid to Type Ⅱ errors in so as to evaluate the overall diagnostic capability. <Table 3> shows the comparison results of the Type I and Type Ⅱ errors for the three data sets mentioned in the previous subsection. As the results revealed in <Table 3>, the proposed hybrid GA-based classification method with three types of ellipsoidal regions achieved the lowest Type Ⅱ error for test data in comparison with other popular methods such as the *k*-nearest neighbor, the C4.5 decision tree, and the neural network. Also the proposed method had a superior classification rate to those of the other methods. Therefore it can be concluded that the proposed classifier system not only has better average classification rate, but also has lower Type Ⅱ errors and hence can successfully reduce the possible risks due to the high misclassification costs associated with Type Ⅱ errors.

5. Conclusion

This paper proposes a hybrid GA-based classifier learning method for medical diagnosis problems which can be considered as multidimensional pattern classification problems. The method adds a local improvement loop to the traditional genetic algorithm loop. The traditional GA determines overall separation boundaries from the patterns of given

<Table 3> Comparison results on liver, diabetes, and WDBC

Classifier	Type I error (%)			Type II error (%)			Classification rate (%)		
	Liver	Diabetes	WDBC	Liver	Diabetes	WDBC	Liver	Diabetes	WDBC
k-NN[a](k = 3)	39.31	27.60	7.00	34.00	27.24	1.42	63.77	72.53	95.08
C4.5	36.55	29.00	1.96	32.00	27.99	2.36	66.09	71.35	97.89
Neural network	34.48	27.20	4.48	32.00	26.12	0.47	66.96	73.18	97.01
Hybrid GA	34.48	26.60	2.52	31.00	23.88	0.47	67.54	74.35	98.24

Note) [a] k-NN : k-nearest neighbor.

training data. The proposed adaptive operations, then, were applied to the individuals in the GA population to refine the overall separation boundaries generated by GA operations such as crossover.

The proposed three types of ellipsoids as the regional agents, whose parameters are foci and a size factor, have the advantage of interpretability, tractability and robust generalization ability. The GA procedure to fit the ellipsoids to the data patterns is expedited by three adaptive operations : expansion, avoidance, and relocation. The attribute selection is embedded in the loops of the hybrid GA procedure by introducing 'don't care' bits, which induce the robust performance for problems with a large number of features. The proposed hybrid GA-based classification method was applied to well-known medical data sets, i.e., liver, diabetes, and breast cancer diagnosis problems. The performance results showed the superiority of the proposed method to the traditional representative classification methods, i.e., the k-nearest neighbors, the neural networks, and the C4.5 algorithms. In particular, the outstanding performance for the type II error reveals that the proposed classification method guarantees against the risk of missing patients with critical diseases.

References

Abe, S., R. Thawonmas, "A fuzzy classifier with ellipsoidal regions", *IEEE Transactions on Fuzzy Systems*, Vol.5(1997), 358～368.

Blake, C. L., C. J. Merz, UCI Repository of Machine Learning Databases, available on-line : http://www.ics.uci.edu/~mlearn/MLRepository.html, 2006.

Bojarczuk, C. C., H. S. Lopes, A. A. Freitas, E. L. Michalkiewicz, "A constrained-syntax genetic programming system for discovering classification rules : application to medical data sets", *Artif. Intell. Med*. Vol.30(2004), 27～48.

Chou, S.-M., T.-S. Lee, Y.E. Shao, I.-F. Chen, "Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines", *Expert Systems with Applications*, Vol.27(2004), 133～142.

Duha, R., P. Hart, Pattern Classification and Scene Analysis, Wiley, New York, 1973.

Eshelman, L. J., The CHC adaptive search algorithm

: How to have safe search when engaging in nontraditional genetic recombination, in : G. J. E. Rawlins (Ed.), Foundations of Genetic Algorithms, Morgan Kaufman, San Mateo, CA, (1991), 265～283.

Haykin, S., Neural Networks : A Comprehensive Foundation, Prentice-Hall, New Jersey, 1994.

Jain, A. K., R. P. W. Duin, and J. Mao, "Statistical pattern recognition : a review", *IEEE Trans*, Pattern Anal, Mach. Intell, Vol.22(2000), 4～37.

Lee, K. K., W. C. Yoon, "A classifier learning system using a coevolution method for deflection yoke misconvergence pattern classification problem", *Information Sciences*, Vol.178(2008), 1372～1390.

Lee, K. K., W. C. Yoon, "Adaptive classification with ellipsoidal regions for multidimensional pattern classification problems", *Pattern Recognition Letters*, Vol.26(2005), 1232～1243.

Lee, K. K., W. C. Yoon, D. H. Baek, "A classification method using a hybrid genetic algorithm combined with an adaptive procedure for the pool of ellipsoids", *Applied Intelligence*, Vol.25(2006), 293～304.

Liu, H., R. Setiono, Incremental feature selection, Applied Intelligence, Vol.9(1998), 217～230.

Luukka, P., T. "Leppalampi, Similarity classifier with generalized mean applied to medical data," *Computers in Biology and Medicine*, Vol.36 (2006), 1026～1040.

McClellend, J., D. Rumelhart, Explorations in Parallel Distributed Processing, MIT Press, Cambridge, MA, 1988.

Michie, D., D. J. Spiegelhalter, and C. C. Taylor, "Machine Learning, Neural and Statistical Classification", Ellis Horwood, New York, 1994.

Nolan, J. R., "Computer systems that learn : an empirical study of the effect of noise on the performance of three classification methods", *Expert Syst*. Appl. Vol.23(2002), 39～47.

Pal, S. K., S. Bandyopadhyay, and C. A. Murthy, "Genetic algorithms for generation of class boundaries", *IEEE Trans. Syst. Man Cybern.-Part B*, Vol.28(1998), 816～828.

Quinlan, J. R., C4.5 : Programs for Machine Learning, Morgan Kaufmann, San Mateo, CA, 1993.

Quinlan, J. R., "Induction of decision trees", *Machine Learning*, Vol.1, No.1(1986), 81～86.

Renders, J.-M., S. Flasse, "Hybrid methods using genetic algorithm", *IEEE Transactions on System, Man, and Cybernetics*, Vol.26, No.2(1996), 243～258.

Ross, S. M., Introduction to Probability and Statistics for Engineering and Scientists, John Wiley and Sons, New York, NY, 1987.

Sima, C., S. Attoor, U. Brag-Neto, J. Lowey, E. Suh, E. R. Dougherty, "Impact of error estimation on feature selection", *Pattern Recognition*, Vol.38 (2005), 2472～2482.

Simpson, P. K., "Fuzzy min-max neural networks-part1 : classification", *IEEE Trans. Neural Networks*, Vol.3(1992), 776～786.

Tan, P., M. Steinbach and V. Kumar, "Introduction to Data Mining", Pearson Education, Boston, 2005.

Weiss, S. M. and C. A. Kulikowski, "Computer Systems That Learn", Morgan Kaufmann, San Francisco, CA, 1991.

Abstract

유전자 알고리즘 및 국소 적응 오퍼레이션 기반의 의료 진단 문제 자동화 기법 연구

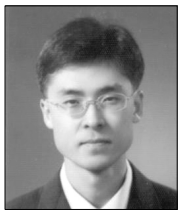
이기광* · 한창희**

의료 진단 문제는 기정의된 특성치들로 표현되는 환자의 상태 데이터로부터 병의 유무를 판단하는 일종의 분류 문제로 간주할 수 있다. 본 연구는 혼용 유전자 알고리즘 기반의 분류방법을 도입함으로써 의료 진단 문제와 같은 다차원의 패턴 분류 문제를 해결할 수 있는 방안을 제안하고 있다. 일반적으로 분류 문제는 데이터 패턴에 존재하는 여러 클래스 간 구분경계를 생성하는 접근방법을 사용하는데, 이를 위해 본 연구에서는 일단의 영역 에이전트들을 도입하여 이들을 유전자 알고리즘 및 국소 적응조작을 혼용함으로써 데이터 패턴에 적응하도록 유도하고 있다. 일반적인 유전자 알고리즘의 진화단계를 거친 에이전트들에 적용되는 국소 적응조작은 영역 에이전트의 확장, 회피 및 재배치로 이루어지며, 각 에이전트의 적합도에 따라 이들 중 하나가 선택되어 해당 에이전트에 적용된다. 제안된 의료 진단용 분류 방법은 UCI 데이터베이스에 있는 잘 알려진 의료 데이터, 즉 간, 당뇨, 유방암 관련 진단 문제에 적용하여 검증하였다. 그 결과, 기존의 대표적인 분류기법인 최단거리이웃방법(the nearest neighbor), C4.5 알고리즘에 의한 의사결정트리(decision tree) 및 신경망보다 우수한 진단 수행도를 나타내었다.

Keywords : 의료진단, 분류, 혼용 유전자 알고리즘, 적응 오퍼레이션, 영역 에이전트

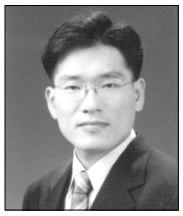
* 인제대학교 경영학부/산업경영연구원
** 한양대학교 경영학부

저 자 소 개



이기광

한양대 산업공학과에서 학사, 한국과학기술원 산업공학과에서 석사 및 박사 학위를 취득하였다. LG전자 UMTS시스템연구소 선임연구원 및 동사 마케팅전략그룹 과장을 거쳐 현재 인제대학교 경영학부 조교수로 재직하고 있다. 주요 관심분야는 데이터마이닝, 정보이론 및 지능형 의사결정지원시스템 등이다.



한창희

현재 한양대학교 경영학부 부교수로 재직 중이다. 한양대 산업공학과에서 학사, 한국과학기술원 산업공학과에서 석사, 한국과학기술원 테크노경영대학원에서 박사 학위를 취득하였다. Georgia Institute of Technology 초청연구원을 역임하였으며, 현대정보기술, 오픈타이드에서 컨설팅을 수행하였다. 주요 관심분야는 인터넷 서비스 설계 및 평가, 전략 의사결정 분석, 온라인 게임 등이다.