IDEA GROUP PUBLISHING

701 E. Chocolate Avenue, Hershey PA 17033-1117, USA
Tel: 717/533-8845; Fax 717/533-8661; URL-http://www.idea-group.com

ITP4251

An Evolutionary Misclassification Cost Minimization Approach for Medical Diagnosis

Parag C. Pendharkar
School of Business Administration, Penn State University at Harrisburg, paragp@psu.edu

James A. Rodger
Eberly College of Business and Economics, Indiana University at Pennsylvania, jrodger@grove.iup.edu

Sudhir Nanda
Quantitative Analyst, T Rowe Price, Maryland

ABSTRACT

This paper illustrates how a misclassification cost matrix can be incorporated into an evolutionary classification system for medical diagnosis. Most classification systems for medical diagnosis have attempted to minimize the misclassifications (or maximize correctly classified cases). The minimizing misclassification approach assumes that Type I and Type II error costs for misclassification are equal. There is evidence that these costs are not equal and incorporating costs into classification systems can lead to superior outcomes. We use principles of evolution to develop and test a genetic algorithm (GA) based approach that incorporates the asymmetric Type I and Type II error costs. Using simulated and real life medical data, we show that the proposed approach, incorporating Type I and Type II misclassification costs, results in lower misclassification costs than LDA and GA approaches that do not incorporate these costs.

INTRODUCTION

Current computer-based medical diagnostic methods use neural networks, discriminant analysis and other machine learning approaches for medical diagnosis [3], [10], [11], and [13]. Although somewhat useful these approaches do not incorporate the economic considerations of misclassification. There are two types of errors that are encountered in classification systems: False positive (Type I) and false negative (Type II) error. The costs of these errors are not equal. For example, predicting that a patient does not have heart disease when the patient has it is more costly than predicting that a patient has heart disease when he does not have it.

Traditional classification systems such as neural networks and linear discriminant analysis do not allow a user to incorporate asymmetric costs of misclassification. In fact, these costs are considered equal in most machine learning classification systems. In this chapter, we propose and implement a GA based classification model that allows the decision-maker to incorporate misclassification costs. Using simulated, real life heart disease, and liver disorder data sets, we show that the proposed GA model performs better than parametric linear discriminant analysis and a non-parametric linear GA based model that does not allow decision-makers to incorporate costs.

The rest of the paper is organized as follows. In section 2 we provide an overview of linear discriminant analysis and genetic algorithm based models for classification. In section 3 we suggest modifications to the genetic algorithm based model that incorporates Type I and Type II cost based priorities. Section 4 provides tests of the proposed genetic algorithm model using simulated and real life data sets. The summary of our research and directions for future work are in section 5.

OVERVIEW OF DISCRIMINANT ANALYSIS AND GENETIC ALGORITHM APPROACHES TO DISCRIMINANT ANALYSIS

Parametric linear discriminant analysis (LDA) was developed by Fisher [4]. The LDA procedure constructs a linear discriminant function by maximizing the ratio of between-groups variance to within-groups variances. For a binary classification problem, the discriminant function can be written as follows:

$$D(X) = X' \sum^{-1} (\mu_1 - \mu_2) - \frac{1}{2} (\mu_1 - \mu_2)^T \sum^{-1} (\mu_1 + \mu_2)$$

where μ_1, μ_2, and Σ^{-1} are mean vectors for group 1, group 2 and inverse of common covariance matrix respectively. LDA works well when normality and equal dispersion of two groups assumptions are met. However, in reality these assumptions are violated frequently and non-parametric approaches such as GA and neural networks are reported to perform better.

Heuristic genetic algorithms provide a popular non-parametric approach for classification when minimizing misclassification is considered as a performance metric. Genetic algorithms (GAs) use *survival of the fittest strategy* to learn coefficients of a linear discriminant function. GAs are parallel search techniques that start with a set of random potential solutions and use special search operators (evaluation, selection, crossover, mutation) to bias the search towards the promising solutions. At any given time, unlike any optimization approach, GA has several promising potential solutions (equal to population size) as opposed to one optimal solution. Each population member in a GA is a potential solution. A population member (P_l) used to learn the coefficients for a linear discriminant function will consists of a set of all the coefficients and the intercept. P_l can be represented as,

$$P_l = \langle w_1, w_2, w_3, w_4, c \rangle$$
$$Where, \quad (w_1, w_2, w_3, w_4, c) \in \Re$$

The discriminant function takes the following form,

$$w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 + c = 0$$

The classification heuristic can be represented as,

IF $w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 + c \geq 0$ Then class=G_1

ELSE class = G_2

Any $w \in P_l$ is called a gene (coefficient) of a given population member P_l. A set of several population members is called the population Ω. The cardinality of the set of population members Ω (number of population members) is called population size. The cardinality of a population member (number of genes) is called the defining length of the population member ζ. The defining length for population member P_l, ζ=5. The defining length of all the population members in a given population is constant.

GA starts with a random set of the population. An evaluation operator is then applied to evaluate the fitness of each individual. In

This paper appears in *Issues and Trends of Information Technology Management in Contemporary Organizations,* the proceedings of the Information Resources Management Association International Conference.

case of learning coefficients for a discriminant function, the evaluation function is the number of correctly classified cases. A selection operator is then applied to select the population members with higher fitness (so that they can be assigned a higher probability of survival). Under the selection operator, individual population members may be born, allowed to live, or to die. Several selection operators are reported in the literature; they are proportionate reproduction, ranking selection, tournament selection, and steady state selection [5]. Among the popular selection operators are ranking and tournament selection. Goldberg and Deb [5] show that both ranking and tournament selection maintain strong population fitness growth potential under normal conditions. However, the tournament selection operator requires lower computational overhead. The time complexity of ranking selection is O (n log n), whereas the time complexity of tournament selection is O(n), where n is the number of members in a population. In tournament selection two random individuals are selected and the member with the better fitness of the two is admitted to the pool of individuals for further genetic processing. The process is repeated in a way such that the population size remains constant and the best individual in the population always survives. We use a tournament selection operator in this research.

After the selection operator is applied, the *new* population special operator called crossover and mutation is applied with a certain probability. For applying the crossover operator, the status of each population member is determined and each is assigned a status as a survivor or a non-survivor. The number of population members in survivor status is approximately equal to population size*(1 – probability of crossover). The number of non-surviving members is approximately equal to population size*probability of crossover. The non-surviving members in a population are then replaced by applying crossover operators to randomly selected surviving members. Though several crossover operators exist we describe and use a one-point crossover operator in our research.

In one-point crossover, two surviving parents and a crossover point are randomly selected. For each parent, the genes in the right hand side of the crossover point are exchanged to produce two children. Let P_1 and P_2 be two parents and the crossover point be denoted by "|". The two children C_1 and C_2 are produced as follows (we use the bold font to simplify the understanding),

$P_1 = < w_1, w_2, | w_3 , w_4, c >$
$\mathbf{P_2 = < w_1, w_2, | w_3 , w_4, c >}$
$\mathbf{C_1 = < w_1, w_2, | w_3 , w_4, c >}$
$\mathbf{C_2 = < w_1, w_2, | w_3 , w_4, c >}$

The mutation operator randomly picks a gene in a surviving population member (with the probability equal to probability of mutation) and replaces it with a real random number.

INTEGRATING TYPE I AND TYPE II COST PREFERENCES IN GENETIC ALGORITHM BASED CLASSIFICATION SYSTEMS

We use the GA model described in section 2 and incorporate Type I and Type II error costs. We name this model Integrated Cost Preference Based-GA Model (ICPB-GA). In the ICPB-GA model, we first calculate the ratio (preference) of Type I and Type II error costs as follows,

$$P_{TypeI} = \frac{\text{Cost of Type I Error}}{\text{Cost of Type I Error} + \text{Cost of Type II Error}} \quad \text{and}$$

$$P_{TypeII} = \frac{\text{Cost of Type II Error}}{\text{Cost of Type I Error} + \text{Cost of Type II Error}}$$

where P_{TypeI} is the preference for minimization of Type I error and $P_{Type\ II}$ is the preference for minimization of Type II error. The cost preferences can be directly incorporated into the fitness function of the genetic algorithm model. Since GAs use survival of the fittest strategy to evolve fit population members, we use the following fitness function to minimize priority based Type I and Type II errors of misclassification.

$$Fitness = (\text{Total Cases}) - (P_{TypeI}\ Total\ Type\ \text{I}\ Errors) - (P_{TypeII} Total\ TypeII\ Errors)$$

The above fitness function is always positive since the total number of errors can never exceed total cases in the data set. Our model is different from the traditional classification model in which the fitness function maximizes correctly classified cases. The fitness function for the traditional model can be written as,

$$Fitness = (\text{Total Cases}) - (\text{Total Type I Errors}) - (\text{Total Type II Errors})$$

The genetic learning procedure begins with a population of random strings, and can be summarized as:

{
Randomly initialize coefficients of discriminant function ∈ [-1, 1]
While (not terminating-condition){
evaluate-fitness of population members
perform tournament selection
With probability *pcross*
perform single point crossover on two parents to get two new offsprings
With probability *pmutate*
perform mutation on a offspring
Replace parents with offsprings if offsprings have higher fitness
}
}

The values of population members for the classification model for ICPB-GA is restricted between –1 and +1 to improve the speed and solution accuracy.

EXPERIMENTS ON SIMULATED AND REAL LIFE DATA SETS

In this section we present the results of our experiments on three data sets. The first is a simulated data set, which incorporates a number of distribution assumptions. The others are real life heart disease and liver disorders data sets.

Simulated Data

Our experiments are based on data previously used for comparing a number of statistical and linear programming techniques for discrimination. Joachimsthaler and Stam [6] examined Fisher's linear discriminant function, the quadratic discriminant function, the logistic discriminant function, and a linear programming approach under varying group distribution characteristics. Koehler and Erenguc [9] and Abad and Banks [1] used the same data generator to establish identical experimental conditions to evaluate a number of other linear programming approaches. Koehler [8] used this data to determine the effectiveness of a genetic search approach for discrimination. Recently, Bhattacharyya and Pendharkar [2] used this data set to evaluate various induction, evolutionary and neural techniques for discrimination problem.

We use four simulated data sets for our research. The data varies with respect to the type of distribution, determined through the kurtosis. Four kurtosis values of -1, 0, 1, and 3 correspond approximately to samples drawn from uniform, normal, logistic and Laplace population distributions. Each data set consists of 20 data samples. Each sample has three attributes and has 100 observations equally split between two groups. In order to minimize the effect of group overlap, the group means are set as follows: the group 1 mean is =(0,0,0) throughout, and the group 2 mean was =(.5, .5, .5). The dispersions of the two groups were the kept same. A more detailed description of the data can be found in Joachimsthaler and Stam [6]. We use 40 data samples (10 from each of the 4 kurtosis values) for training and remaining 40 data samples for testing (holdout set). The cost preferences, based on the assumed cost matrix shown in Figure 1, are P_{TypeI} = 0.66 and P_{TypeII} = 0.33 respectively. Figures 2 through 5 illustrate the results of our experiments.

Figures 2 through 5 show that ICPB-GA minimizes Type I errors at the expense of total correct classification. In other words, when

Figure 1: The misclassification cost matrix

	Reject	Accept
Predicted Reject	0	2
Predicted Accept	1	0

Figure 2: Results for correct classification in the training sample for simulated data

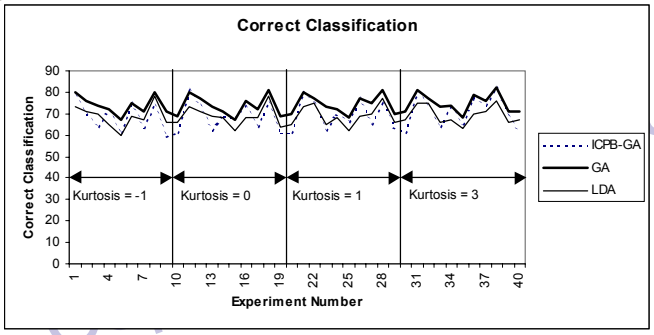

Figure 3: Type I error in the training sample for simulated data

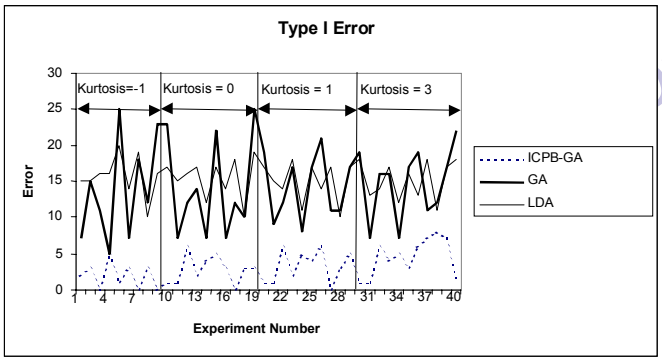

Figure 4: Results for correct classification in the holdout sample for simulated data

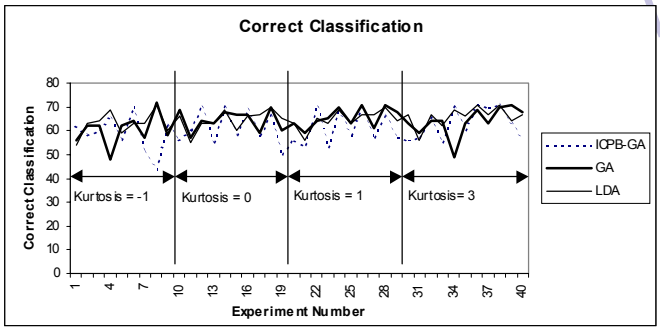

Figure 5: Type I error in holdout sample for simulated data

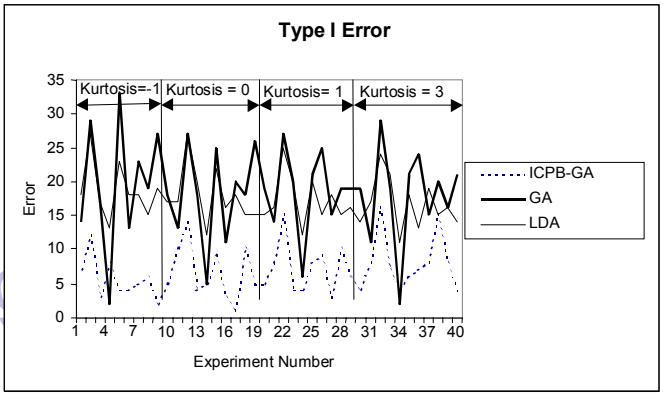

Table 1: Results of tests of difference in means for training and holdout samples

Type	Hypothesis	Mean (%)	Mean (%)	t-value	P>t
Training Sample					
Correct Classification	$\mu_{GA}=\mu_{LDA}$	μ_{GA} =74.2	μ_{LDA} = 68.9	18.56	0.000*
	$\mu_{LDA}=\mu_{ICPB-GA}$	$\mu_{ICPB-GA}$=69.4	μ_{LDA} = 68.9	0.63	0.266
	$\mu_{GA}=\mu_{ICPB-GA}$	μ_{GA} =74.2	$\mu_{ICPB-GA}$=69.4	7.62	0.000*
Type I Error	$\mu_{GA}=\mu_{LDA}$	μ_{GA} =14.17	μ_{LDA} = 15.32	1.52	0.0680
	$\mu_{LDA}=\mu_{ICPB-GA}$	$\mu_{ICPB-GA}$=5.2	μ_{LDA} = 15.32	18.92	0.000*
	$\mu_{GA}=\mu_{ICPB-GA}$	μ_{GA} =14.17	$\mu_{ICPB-GA}$=5.2	10.88	0.000*
Holdout Sample					
Correct Classification	$\mu_{GA}=\mu_{LDA}$	μ_{GA} =63.6	μ_{LDA} = 64.6	-1.16	0.126
	$\mu_{LDA}=\mu_{ICPB-GA}$	$\mu_{ICPB-GA}$=61.1	μ_{LDA} = 64.6	-3.205	0.001*
	$\mu_{GA}=\mu_{ICPB-GA}$	μ_{GA} =63.6	$\mu_{ICPB-GA}$=61.1	1.85	0.071
Type I Error	$\mu_{GA}=\mu_{LDA}$	μ_{GA} =18.5	μ_{LDA} = 17.52	-1.13	0.132
	$\mu_{DA}=\mu_{ICPB-GA}$	$\mu_{ICPB-GA}$=6.9	μ_{LDA} = 17.52	16.64	0.000*
	$\mu_{GA}=\mu_{ICPB-GA}$	μ_{GA} =18.5	$\mu_{ICPB-GA}$=6.9	10.50	0.000*

* significant at level of significance = 0.01

ICPB-GA minimizes Type I errors, Type II errors increase and overall correct classification goes down. Table 1 provides tests of difference in means for total classification and type I error for the three techniques.

The results in Table 1 support the observation that ICPB-GA lowers Type I error at the expense of lowering overall correct classification. For Type I error, the difference of means between ICPB-GA, and both GA and LDA is significant in both the training and holdout samples. For correct classification, the difference of means between ICPB-GA and GA is significant in the training sample, but not significant in the holdout sample. The test statistic for difference of means between ICPB-GA and LDA for correct classification is not significant in the training sample, but is significant in the holdout sample.

Heart Disease Data

We apply the three classification techniques on real life data that has been used in previous studies [7]. The dataset comes from the Cleveland Clinic Foundation and is now a part of the collection of machine learning databases at the University of California, Irvine. We use the three approaches (LDA, GA, ICPB-GA) for predicting heart disease. The data set consists of 270 total examples with two group, presence and absence of heart disease. The group covariances are equal. The kurtosis value for the data was 3.6.

There are 13 attributes with eight attributes being numerical continuous variables and 5 having categorical values. The data set also has a misclassification cost matrix, which was supplied by doctors in Leeds, Great Britain. The misclassification cost matrix is shown in Figure 6. We took the original data set of 270 examples and divided it into two sets. The training data contained 160 examples and the holdout data set contained 80 examples. Only 240 examples out of a total of 270

were used so that both training and test datasets contain 50% examples belonging to class 1 (presence of heart disease) and other 50% belonging to class 2 (absence of heart disease).

Figure 6: The misclassification cost matrix for heart disease data set

	Reject	Accept
Predicted Reject	0	5
Predicted Accept	1	0

Table 2 presents the results of our experiments on the heart disease data set. As expected, ICPB-GA lowered the Type I error when compared to LDA and GA in both the training and holdout samples. When the number of correct classifications was considered as the performance metric, LDA performed better than GA and ICPB-GA in the training sample, and GA performed better than LDA and ICPB-GA in the holdout sample. When we consider the misclassification cost as a performance metric, ICPB-GA provides the lowest misclassification cost. The misclassification cost is defined as follows, Misclassification Cost = Cost of Type I Error * (Total Type I Errors) + Cost of Type II Error * (Total Type II Errors).

Table 2: Results of tests in training and holdout samples of heart disease data

Type	LDA	GA	ICPB-GA
Training Sample			
Correct Classification (%)	85.0	82.5	70.6
Type I Error (%)	6.3	7.5	1.8
Misclassification Cost	10.62	12.62	9.79
Holdout Sample			
Correct Classification (%)	83.7	86.3	75
Type I Error (%)	7.5	7.5	1.3
Misclassification Cost	6.14	5.81	3.98

BUPA Liver Disorders Data

The BUPA liver disorders data set was created by BUPA Medical Research Limited. The data set was donated by Richard Forsyth and is available as a part of the UC Irvine Machine Learning Databases. There are 6 usable attributes in the data set. We use 5 of these attributes as predictor variables (results of the blood test) and one attribute as a class variable. The class variable is the number of alcoholic drinks, which is 0 if number of drinks is less than 3 and 1 otherwise. This data has been previously used by Turney [12]. The original data set contains 345 cases with no missing values. We divided the data randomly into 173 training cases and 172 test cases. Table 3 provides the results of our experiments on comparing the three techniques.

Table 3: Results for tests in training and holdout samples for BUPA dataset

Type	Training Sample		Holdout Sample	
	Correct Classification	Type I Error	Correct Classification	Type I Error
LDA	75.1	9.2	60.5	18.0
GA	73.9	9.2	56.9	22.6
ICPB-GA	72.2	1.1	59.9	6.9

The results of the liver disease data set are consistent with those of simulated and heart disease data. The cost based approach (ICPB-GA) performed better than the non-cost based approaches if Type I error is the performance metric. LDA performed best if correct classification is used as the performance metric

The results of our experiments on simulated and real life data sets illustrate the benefits of incorporating cost based preferences with GA classification systems. Thus, in classification problems, in which decision-makers tradeoff between misclassification costs, an integrated cost based preference classification approach such as ICPB-GA may be a promising approach when compared to traditional LDA or GA.

SUMMARY AND DIRECTIONS FOR FUTURE WORK

We have shown that certain medical diagnosis problems, in which it is important to lower the total cost of Type I and Type II errors, could be treated as cost minimization problems. We used the cost matrices to obtain cost tradeoffs and incorporated these tradeoffs into a linear GA based classification system. After incorporation of the cost preferences, the resulting classification system (ICPB-GA) performed better than classification systems that do not incorporate the cost preferences (LDA and GA).

In our research, we assumed that the Type I and Type II error costs are constant. In certain medical situations these costs may vary over time and may follow a statistical distribution. The current approach can be easily modified to incorporate costs that are not constant. It is likely that in the event costs are not constant, the performance of ICPB-GA could be impacted by the distribution of the costs of Type I and Type II errors. Future research should focus on investigating the impact of different cost distributions on the performance of ICPB-GA.

REFERENCES

[1] Abad, P.L. and Banks, W.J. (1993), "New LP based heuristics for the classification problem," *European Journal of Operations Research*, vol. 67, pp. 88-100.

[2] Bhattacharyya, S. and Pendharkar, P.C. (1998), "Inductive evolutionary and neural techniques for discrimination," *Decision Sciences,* vol. 29, pp. 871-899.

[3] Doi, K., Giger, M.L., Mishikawa, R.M., Hoffmann, K.R., Macmahon, H., Schmidt, R.A., and Chua, K.G. (1993), "Digital radiography: A useful clinical tool for computer-aided diagnosis by quantitative analysis of radiographic images," *Acta Radiologica*, vol. 34, pp. 426-439.

[4] Fisher, R.A. (1936), "The use of multiple measurements in taxonomic problems," *Annals of Eugenics*, vol. 7, pp. 179-188.

[5] Goldberg, D.E. and Deb, K. (1991), "A comparative analysis of selection schemes used in genetic algorithms," In G. Rawlins (Ed.), *Foundations of Genetic Algorithms,* pp. 69-93. San Mateo, CA: Morgan Kaufmann.

[6] Joachimsthaler, E.A. and Stam, A. (1988), "Four approaches to the classification problem in discriminant analysis: An experimental study," *Decision Sciences*, vol. 19, pp. 322-333.

[7] King, R.D., Henry, R., Feng, C., and Sutherland, A. (1994), "A Comparative Study of Classification Algorithms: Statistical, Machine Learning and Neural Network," *Machine Intelligence, 13: Machine Intelligence and Inductive Learning*, K. Furukawa, D. Michie and S. Muggleton (ed.), Clarendon Press, Oxford.

[8] Koehler, G.J. (1991), "Linear discriminant functions determined by genetic search," *ORSA Journal on Computing*, vol. 3**,** pp. 345-357.

[9] Koehler, G.J. and Erenguc, S.S. (1990), "Minimizing misclassifications in linear discriminant analysis," *Decision Sciences*, vol. 21, pp. 63-85.

[10] Kovalerchuck, B., Triantaphyllou, E., Ruiz, J.F., and Clayton, J. (1997), "Fuzzy logic in computer-aided breast cancer diagnosis: Analysis of lobulation," *Artificial Intelligence in Medicine*, vol. 11, pp. 75-85.

[11] Pendharkar, P.C. Rodger, J.A., Yaverbaum, G.J., Herman, N., and Benner, M. (1999), "Association, statistical, mathematical, and neural approaches for mining breast cancer patterns," *Expert Systems with Applications*, vol. 17, pp. 223-232.

[12] Turney, P.D. (1995), "Cost-sensitive classification: Empirical evaluation of a hybrid genetic decision tree induction algorithm," *Journal of Artificial Intelligence Research*, vol. 2, pp. 369-409.

[13] Wu, Y., Doi, K., Giger, M., Metz, C., and Zhang, W. (1994), "Reduction of false positive in computerized detection of lung nodules in chest radiographs using artificial neural networks, discriminant analysis and a rule-based scheme," *Journal of Digital Imaging*, vol. 17**,** pp.196-207.