Chapter 3

An Evolutionary Misclassification Cost Minimization Approach for Medical Diagnosis

Parag C. Pendharkar and Sudhir Nanda
Pennsylvania State University at Harrisburg, USA

James A. Rodger
Indiana University of Pennsylvania, USA

Rahul Bhaskar
Manugistics, Inc., USA

ABSTRACT

This chapter 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 (Doi et al., 1993; Kovalerchuck, Triantaphyllou, Ruiz & Clayton, 1997; Pendharkar, Rodger, Yaverbaum, Herman & Benner, 1999; Wu, Doi, Giger, Metz & Zhang, 1994). 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 nonparametric linear GA-based model that does not allow decision-makers to incorporate costs.

The rest of the chapter is organized as follows. In “Pure Frontier Models,” we provide an overview of linear discriminant analysis and genetic algorithm based models for classification. In “Contrasts of Meaning and Purpose,” we suggest modifications to the genetic algorithm based model that incorporates Type I and Type II cost based priorities. “Data Mining Uses and Suggested Guidelines” 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 “Other Models and Applications.”

OVERVIEW OF DISCRIMINANT ANALYSIS
AND GENETIC ALGORITHM APPROACHES
TO DISCRIMINANT ANALYSIS

Parametric linear discriminant analysis (LDA) was developed by Fisher (1936). 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:
Related Content

Periodic Streaming Data Reduction Using Flexible Adjustment of Time
Section Size
www.irma-international.org/article/periodic-streaming-data-reduction-using/1747/

Adaptive Scheduling for Real-Time Distributed Systems
Apurva Shah (2014). *Biologically-Inspired Techniques for Knowledge Discovery and Data Mining* (pp. 236-248).
www.irma-international.org/chapter/adaptive-scheduling-for-real-time-distributed-systems/110462/

Big Data Paradigm for Healthcare Sector
www.irma-international.org/chapter/big-data-paradigm-for-healthcare-sector/150182/

Data Mining and Explorative Multivariate Data Analysis for Customer Satisfaction Study
www.irma-international.org/chapter/data-mining-explorative-multivariate-data/73507/

A Perturbation Method Based on Singular Value Decomposition and Feature Selection for Privacy Preserving Data Mining