Expert Systems with Applications 36 (2009) 9240–9249

Contents lists available at ScienceDirect

Expert Systems with Applications

journal homepage: www.elsevier.com/locate/eswa

An application of a new meta-heuristic for optimizing the classification accuracy when analyzing some medical datasets

Huy Nguyen Anh Pham, Evangelos Triantaphyllou *

Department of Computer Science, Louisiana State University, 298 Coates Hall, Baton Rouge, LA 70803, United States

ARTICLE INFO

Keywords:
Optimization
Medical data mining
HBA
Genetic algorithms
Classification errors

ABSTRACT

Medical data mining has recently become one of the most popular topics in the data mining community. This is due to the societal importance of the field and also the particular computational challenges posed in this domain of data mining. However, current medical data mining approaches oftentimes use identical costs or just ignore them for the different cases of classification errors. Thus, their outcome may be unexpected. This paper applies a new meta-heuristic approach, called the Homogeneity-Based Algorithm (or HBA), for optimizing the classification accuracy when analyzing some medical datasets. The HBA first expresses the objective as an optimization problem in terms of the error rates and the associated penalty costs. These costs may be dramatically different in medical applications as the implications of having a false-positive and a false-negative case may be tremendously different. When the HBA is combined with traditional classification algorithms, it enhances their prediction accuracy. It does so by using the concept of homogenous sets. Five medical datasets, obtained from the machine learning data repository at the University of California, Irvine (UCI), USA, were tested. Some computational results indicate that the HBA, when it is combined with traditional methods, can significantly outperform current stand-alone data mining approaches.

1. Introduction

Increasing powerful mechanisms for storing data has made available lots of datasets related to medicine in recent decades. A motivation for extracting useful knowledge from such datasets and thus discovering decision-making insights for the diagnosis and treatment of diseases, is also increasingly recognized. In the typical setting a dataset of historic data, which describe some type of disease or a medical disorder, is assumed to be available. Such datasets consist of records of patients describing physical and laboratory examinations related to that type of disease or medical disorder. Then, the computational challenge is how to develop a diagnostic system, which could assist in diagnosing this type of ailment based on the knowledge extracted from the historic dataset. At this point, human analysts need special computational tools to process and comprehend such large and complex datasets.

Medical data mining can assist in addressing such challenges. Data mining analysts can extract decision regions from a given historic dataset related to a medical condition or disease. Usually, such decision regions consist of medical indicators, which could be used to diagnose the condition or disease. In medical diagnosis (as in most other domains), usually there are three different cases of possible errors:

- The false-negative case in which a patient, who in reality has the disease, is diagnosed as disease free.
- The false-positive case in which a patient, who in reality does not have the disease, is diagnosed as having the disease.
- The unclassifiable case in which the prediction system cannot diagnose a given case. This happens due to insufficient knowledge extracted from the historic data.

Under the above considerations, current medical data mining approaches oftentimes assign identical penalty costs for the false-positive and the false-negative cases or just ignore the penalty cost for the unclassifiable cases. Such approaches will be discussed in Section 2. Thus, their outcome may be unexpected or even unacceptable.

The two penalty costs for the false-positive and the false-negative cases could be dramatically different in a medical application. For instance, in the case of a life threatening condition where time is of essence, if one diagnoses a given case as false-negative, then his/her medical condition goes untreated or is treated inadequately. Thus precious time may be wasted and the situation may turn out to be eventually fatal to the patient. On the other hand, for the same situation, a false-positive diagnosis may just

* Corresponding author.
E-mail addresses: hpham15@lsu.edu (H.N.A. Pham), trianta@lsu.edu (E. Triantaphyllou).

0957-4174/$ - see front matter © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2008.12.007

add some financial costs and anxiety to the patient but not result in a life threatening condition.

A penalty cost for unclassifiable cases in medical data mining is needed as well. A diagnosis of a patient as an unclassifiable case may require additional medical examinations and involve some costs. However, that particular case may not necessarily result in a wrong diagnosis.

For the above reasons, this paper applies a new meta-heuristic approach, called the Homogeneity-Based Algorithm (or HBA) as developed by Pham and Triantaphyllou (2007, part 4, chap. 5) and Pham and Triantaphyllou (2008, chap. 2), on some well-known medical datasets. The HBA first defines the total misclassification cost of models extracted from classification algorithms as an optimization problem in terms of the false-positive, the false-negative, and the unclassifiable rates along with their penalty costs. The HBA then organizes the extracted models as mutually exclusive decision regions represented by homogeneous sets. These decision regions are refined based on their density by employing a genetic algorithm (GA) approach. This is done in order to minimize the total misclassification cost. The HBA is motivated by the large discrepancy in the previous three penalty costs.

The next section provides a literature review of some related developments. The third section has a brief description of the HBA as adopted from Pham and Triantaphyllou (2007) and Pham and Triantaphyllou (2008). That section shows how the HBA can yield an optimal or near optimal misclassification total cost. The fourth section discusses some computational results from the medical domain. These results give an indication of how this methodology may improve the prediction accuracy in computerized medical diagnosis. The paper ends with some conclusions and an appendix, which describes the key algorithmic aspects of the HBA.

2. Previous work

This paper studies five medical datasets, which current data mining approaches have often used for their analyses. The main characteristics of these datasets are depicted in Table 1. These datasets were selected because the number of attributes were in the range of values that the HBA can handle easily (i.e. approximately less than 9 or 10). Other reasons for selecting these datasets were that traditional approaches have analyzed them with variable success and these datasets represent a variety of important medical diseases and disorders.

The first dataset is the Pima Indian diabetes (PID) as described in Asuncion and Newman (2007). Attributes of 768 female patients of Pima Indian heritage were recorded in this dataset. The class variable denotes whether a person has diabetes or not. Smith, Everhart, Dickson, Knowler, and Johannes (1998) achieved 76% accuracy by using an Early Neural Network (ENN). Jankowski and Kadirkamanathan (1997) obtained 77.6% accuracy by using a radial basis function network suite, called IncNet. Au and Chan (2001) improved the correct classification percentage by using a fuzzy approach. Their approach achieved 77.6% accuracy. Rutkowski and Cpalka (2003) obtained 78.6% accuracy by introducing a new neural-fuzzy structure, called a flexible neural-fuzzy inference system (FLEXNFIS). Leon (2006) obtained 81.8% accuracy by using a Fuzzy Neural Network (FNN) associated with the BK-Square products. Different classification algorithms in the StatLog project in Michie, Spiegelhalter, and Taylor (1994, chap. 9) obtained less than 78% accuracy. Pham and Triantaphyllou (2008) applied the HBA in conjunction with some Support Vector Machine (SVM), Artificial Neural Network (ANN), and Decision Tree (DT) algorithms. Their accuracy reached about 93.8%.

The second medical dataset is the Haberman Surgery Survival (HSS) as described in Asuncion and Newman (2007). This is one of the most difficult datasets for classification algorithms. The dataset contains records which describe 306 patients who have undergone surgery for breast cancer. Kecman and Arthanari (2001) proposed an SVM approach using linear terms in the objective function for analyzing the HSS dataset. Their approach yielded 71.2% accuracy. Fung and Mangasarian (2001) reformulated Kecman's approach to decrease its complexity. Their approach is called the Proximal Support Vector Machine (PSVM) classifier and it uses a purely quadratic objective function with equality constraints. Their approach yielded 72.5% accuracy. Domm, Engel, Louis, and Goldberg (2005) proposed the Integer Support Vector Machine (ISVM) classifier, which used binary indicator error variables in order to directly minimize the number of potential errors. Their accuracy was 62.7%. Shevked and Dakovski (2007) represented sets of positive and negative training points as logical functions. These logical functions were then minimized in order to find the target functions, which were prime implicants. Their approach yielded 66.2% accuracy.

Some classification approaches used the breast cancer (BC) dataset as described in Asuncion and Newman (2007). This dataset contains records which describe 286 patients who had either breast cancer or no cancer. One of the tested algorithms on this dataset was C4.5 as developed by Quinlan (1996). Quinlan's approach reached 94.7% accuracy by using 10-fold cross-validation. Hamilton, Shan, and Cercone (1996) used the Rule Induction (RI) approach based on approximation of classification to enhance the accuracy. Their approach obtained 96% accuracy. Similarly, Ster and Dobnikar (1996) achieved 96.8% accuracy with the Linear Discriminant Analysis (LDA) approach. Bennet and Blue (1997) used an SVM approach. Their accuracy was 97.2%. In the following two years, Nauck and Kruse (1999) achieved 95.1% accuracy by using a Neuro-Fuzzy approach. At the same time, Pena-Reyes and Sipper (1999) developed a Fuzzy-GA approach, which yielded 97.5% accuracy. Furthermore, Setiono's approach (2000) reached 98.1% accuracy by using a Neuro-Rule approach. Abonyi and Szeifert (2003) applied the Supervised Fuzzy Clustering (SFC) approach and achieved 95.6% accuracy. Polat, Sahan, Kodaz, and Gunes (2007) applied the Fuzzy Artificial Immune Recognition System (FAIRS) to form fuzzy-logic rules. Their approach reached 98.5% accuracy.

The fourth medical dataset is the Liver Disorders (LD) as described in Asuncion and Newman (2007) that many classification approaches have used for their analyses in recent years. This dataset contains records which describe 345 patients who had con-

Table 1
Characteristics of the five medical datasets.

Dataset	No. attributes	No. records	No. positive records	No. negative records	No. records in the training dataset T_1	No. records in the testing dataset
Pima Indian diabetes (PID)	8	768	268	500	576	192
Haberman Surgery Survival (HSS)	3	306	225	81	230	76
Breast cancer (BC)	9	286	85	201	214	72
Liver disorders (LD)	6	345	145	200	276	69
Appendicitis (AP)	7	106	85	21	85	21