Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 127 (2018) 293–299

Procedia Computer Science

www.elsevier.com/locate/procedia

The First International Conference On Intelligent Computing in Data Sciences

Optimization of K-NN algorithm by clustering and reliability coefficients: application to breast-cancer diagnosis

Walid CHERIF[a,*]

[a]*Laboratory SI2M, Department of Computer Science, National Institute of Statistics and Applied Economics, B.P. 6217, Rabat, Morocco*

Abstract

There is a growing trend towards data mining applications in medicine. Different algorithms have been explored by medical practitioners in an attempt to assist their work; the diagnosis of breast cancer is one of those applications. Machine learning algorithms are of vital importance to many medical problems, they can help to diagnose a disease, to detect its causes, to predict the outcome of a treatment, etc. K-Nearest Neighbors algorithm (KNN) is one of the simplest algorithms; it is widely used in predictive analysis. To optimize its performance and to accelerate its process, this paper proposes a new solution to speed up KNN algorithm based on clustering and attributes filtering. It also includes another improvement based on reliability coefficients which insures a more accurate classification. Thus, the contributions of this paper are three-fold: (i) the clustering of class instances, (ii) the selection of most significant attributes, and (iii) the ponderation of similarities by reliability coefficients. Results of the proposed approach exceeded most known classification techniques with an average f-measure exceeding 94% on the considered breast-cancer Dataset.

Keywords: data mining; cancer diagnosis; supervised classification; unsupervised classification; k-nearest neighbors; k-means; similarity measurement.

* Corresponding author. Tel.: +212-672-277-806.
E-mail address: w.cherif@insea.ac.ma.

10.1016/j.procs.2018.01.125

1. Introduction

In medical domains, the volume and complexity of collected data is growing at a rapid pace. This includes besides the information coming from clinical studies, other data on patients [1]. The analysis of such data makes it possible to come up with new medical hypotheses or to confirm existing hypotheses. This partially overcomes the limitations of traditional medical studies which were restricted to small numbers of parameters and small numbers of instances.

In particular, breast-cancer is the most common type of cancerous disease among women of the western world. Approximately every tenth woman suffers from it in her lifetime, half of whom does not survive. Even though breast-cancer is such a severe disease, effective treatment is possible if it is detected at an early stage [2].

The social and economic values of breast-cancer diagnosis are very high [3]. As a result, the problem has attracted many researchers in the area of data mining recently [4, 5]. Their efforts generated various approaches for automatic diagnosis.

As breast-cancer diagnosis is an important and complicated task that needs to be extremely accurate and efficient. Its automation would be very advantageous. It may probably exceed traditional approaches [6]. However, standard algorithms are still limited and no algorithm has proved perfect diagnosis.

Technically, the problem of diagnosis belongs to binary classifications. It stays at the cross junction of statistics and artificial intelligence [7], it aims to classify instances into two groups on the basis of classification criteria.

Binary classification includes two types of models: predictive (supervised) such as KNN algorithm [8], in which the class of each instance is known, and exploratory (unsupervised) such as k-means algorithm whose task is the creation of clusters (classes of instances) [9]. Supervised classification models consist of two major steps: training and testing [10].

Therefore, this paper proposes a novel approach that extends the KNN algorithm by a normalization stage and by creating clusters inside each class. The second improvement is based on the elimination of insignificant attributes. Finally, obtained clusters are then compared to each new instance in terms of a weighted similarity measure. In experiments, each one of these improvements proved quite interesting and contributed into the overall performance.

The rest of this paper is structured as follows: Section 2 summarizes main approaches applied to breast-cancer datasets. Section 3 meticulously details the proposed approach; and in section 4, the obtained results are compared to those of most known classification techniques. Finally, the last section concludes this work.

2. Background

Several researchers in the literature have measured their performances on recognizing benign from malignant breast-tumors.

Sarkar and Leong [11] treated the breast-cancer diagnosis as a pattern classification problem. They used KNN algorithm as a nonparametric classifier to predict malignant samples. Their study included another enhancement: the fuzzy KNN.

Setanio proposed [12] a rule extraction algorithm based on artificial neural networks. The rules were extracted from the network. The pruning algorithm was used to remove redundant connections, and a clustering step was used to discretize the activation values of the input pattern. A similar rule extraction model was presented by Taha et al. [13]. It included three rule extraction algorithms based on artificial neural networks.

In the work of Reyes and Sipper [14], fuzzy logic and genetic algorithm were combined into a same classifier system which outperformed other artificial neural networks approaches.

Abbass [3] proposed an evolutionary artificial neural network approach based on the pareto-differential evolution algorithm augmented with local search for the prediction of breast-cancer.

Kuo at el. [15] opted for decision tree technique to classify breast cancers. Their work aimed to reduce the number of unnecessary biopsies and to increase the diagnosis confidence. 24 features were used to create a decision tree with the ability of recognizing malignant breast-cancers.

Sawarkar et al. [16] have used, in addition to artificial neural networks, support vector machines for breast cancers diagnosis. The implemented algorithm maps the input data into a high-dimensional space. Further, it associates the instances into their respective classes by separating formed hyperplanes.