Expert Systems with Applications 42 (2015) 6844–6852

Contents lists available at ScienceDirect

Expert Systems with Applications

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

# A fuzzy-rough nearest neighbor classifier combined with consistency-based subset evaluation and instance selection for automated diagnosis of breast cancer

Aytuğ Onan *

*Celal Bayar University, Department of Computer Engineering, 45140 Muradiye, Manisa, Turkey*

ARTICLE INFO

*Article history:*
Available online 11 May 2015

*Keywords:*
Fuzzy-rough set
Nearest neighbor classifier
Consistency-based subset evaluation
Breast cancer
Instance selection

ABSTRACT

Breast cancer is one of the most common and deadly cancer for women. Early diagnosis and treatment of breast cancer can enhance the outcome of the patients. The development of classification models with high accuracy is an essential task in medical informatics. Machine learning algorithms have been widely employed to build robust and efficient classification models. In this paper, we present a hybrid intelligent classification model for breast cancer diagnosis. The proposed classification model consists of three phases: instance selection, feature selection and classification. In instance selection, the fuzzy-rough instance selection method based on weak gamma evaluator is utilized to remove useless or erroneous instances. In feature selection, the consistency-based feature selection method is used in conjunction with a re-ranking algorithm, owing to its efficiency in searching the possible enumerations in the search space. In the classification phase of the model, the fuzzy-rough nearest neighbor algorithm is utilized. Since this classifier does not require the optimal value for K neighbors and has richer class confidence values, this approach is utilized for the classification task. To test the efficacy of the proposed classification model we used the Wisconsin Breast Cancer Dataset (WBCD). The performance is evaluated using classification accuracy, sensitivity, specificity, F-measure, area under curve, and Kappa statistics. The obtained classification accuracy of 99.7151% is a very promising result compared to the existing works in this area reporting the results for the same data set.



## 1. Introduction

The development of an effective computer-aided diagnosis and classification system for diseases is of great importance for medical informatics. Breast cancer is one of the most common and deadly cancer for women. In 2008, the number of new cancer cases is approximately 1.4 million and the number of cancer deaths is around 460,000 (Ma & Jemal, 2013). Although breast cancer is still a major cause for death from cancer among women, the cancer mortality exhibits a decreasing pattern with the help of early detection, appropriate therapy and treatment (Hayat, 2008). The diagnosis and treatment of breast cancer in its earlier possible phases can enhance the outcome of breast cancer patients (Tot & Dean, 2004). The early detection and accurate diagnosis of the disease are the two key factors of improved outcomes for breast cancer. Besides, women with not metastasized breast cancer can exhibit long-terms survival (West, Mangiameli, Rampal, & West, 2005). Moreover, common diagnostic techniques, such as mammography and fine needle aspiration cytology exhibit relatively low reliability in diagnosing malignancy (Chen, Yang, Liu, & Liu, 2011). All these factors intensify the motivations toward the development of reliable automated diagnosis systems that can facilitate the clinical decision making process and the early detection of breast cancer.

This study aims to build an automatic diagnostic system for breast cancer based on the fuzzy-rough nearest neighbor classifier. K-nearest neighbor method is one of the most widely employed algorithms for classification tasks in data mining and knowledge discovery. Pattern recognition, text categorization, ranking models, object recognition and event recognition are a few examples of application fields for K-nearest neighbor (Bhatia & Vandana, 2010). In K-nearest neighbor method, an object is assigned to the majority class among its K nearest neighbors based on a majority of its neighbors. K-nearest neighbor method has several advantages. First, it is very easy to understand and implement. Besides, the method performs well when the features are weighted carefully (Nisbet, Elder, & Miner, 2009). On the other hand, the method

* Tel.: +90 236 201 39 05/544 810 70 80; fax: +90 236 201 29 98.
*E-mail address:* aytug.onan@cbu.edu.tr

http://dx.doi.org/10.1016/j.eswa.2015.05.006
0957-4174/

suffers from some significant disadvantages. First of all, the classification performance of the method is highly dependent on the value of K parameter, which determines the space of the neighborhood. Besides, the method lacks possibilistic classification ability, namely it cannot make discrimination between equally close neighbors and equally far away neighbors (Frigui & Gader, 2009). Moreover, relative closeness of neighbors can be a problematic issue in the existence of noise or overlapping classes (Sarkar, 2007). The fuzzy-rough nearest neighbor classifier is based on the theory of fuzzy-rough sets. Rough set theory is a useful method to deal with vague and uncertain information, classical rough set model based on equivalence relation can only deal with complete and symbolic data sets. Fuzzy-rough sets can deal with numerical attributes. They can encapsulate concept of vagueness and indiscernibility owing to fuzzy sets and rough sets, respectively (Dai, 2013). Fuzzy-rough nearest neighbor classifier attempts to enhance the conventional K-nearest neighbor classifier by exploiting fuzzy-rough uncertainty. While preserving the advantages of the conventional K-nearest neighbor method, fuzzy-rough counterpart does not need to know the optimal value of K parameter; it has the possibilistic classification ability and has the worst-case time complexity which is the same as the conventional method. Moreover, it does not require any a priori information about the training data, though other parametric and semi-parametric classifiers involve that information (Sarkar, 2007). Owing to the aforementioned enhanced features, fuzzy-rough nearest neighbor method can be used as a viable tool for classification.

In this paper, fuzzy-rough nearest neighbor classifier is combined with consistency-based subset evaluation and fuzzy-rough instance selection method. Feature selection plays an important role for building a classification model with high predictive accuracy. In the consistency-based subset evaluation, attribute subset selection is applied based on the consistency-based subset evaluation metric. In this method, the worth of a subset of attributes is evaluated by the level of consistency in the class values to identify useful features and eliminate the irrelevant ones. Re-ranking algorithm, that is an approach analyzing only a few block of variables to decrease the number of wrapper evaluations, is utilized in consistency-based feature selection to search the space of feature subsets effectively. Moreover, a fuzzy-rough set based instance selection method is applied in order to significantly reduce the number of instances, but maintaining high classification accuracy. This instance selection approach uses the weak gamma evaluator.

The rest of the paper is organized as follows: Section 2 briefly reviews the existing machine learning related researches on the diagnosis of breast cancer. Section 3 gives a brief explanation for the methods used in the classification model, i.e. the consistency-based subset evaluation, the re-ranking algorithm, the fuzzy-rough instance selection, the fuzzy-rough nearest neighbor classifier are explained briefly. Section 4 presents the proposed classification model in detail. In Section 5, data set, evaluation metrics and experimental results are given. Finally, Section 6 presents the discussion and concluding remarks.

## 2. Related works

There has been a lot of research on the diagnosis of breast cancer with the WBCD data set in the literature with a relatively high predictive classification performance. Pena-Reyes and Sipper (1999) reached 97.80% classification accuracy using an approach that integrates the fuzzy systems and evolutionary algorithms. Chou, Lee, Shao, and Chen (2004) obtained 98.25% classification accuracy with the artificial neural networks and the multivariate adaptive regression splines. Übeyli (2007) compared the classification performance of a multilayer perceptron neural network, a combined neural network, a probabilistic neural network, a recurrent neural network and a support vector machine and the highest classification accuracy was achieved by support vector machine with 99.54% classification accuracy. Polat and Güneş (2007) reported 98.53% classification accuracy with the least square support vector machine classifier algorithm. Şahan, Polat, Kodaz, and Güneş (2007) achieved a classification accuracy of 99.14% with the hybridization of a fuzzy-artificial immune system with K-nearest neighbor classifier. Mu and Nandi (2007) evaluated the benefits of applying support vector machines, radial basis function networks and self-organizing maps and they reported an accuracy of 98.6%. Ryu, Chandrasekaran, and Jacob (2007) applied isotonic separation technique to the breast cancer prediction and the experimental results indicated that the method can be used as a viable tool for the problem. Karabatak and Ince (2009) utilized an association rule for dimension reduction and the neural network for performing classification. They reported a classification accuracy of 97.4%. Akay (2009) presented a support vector machine based diagnosis system combined with F-score based feature selection and the obtained classification accuracy was 99.51%. Hassan, Hossain, Begg, Ramamohanarao, and Morsi (2010) obtained 98.89% classification accuracy with an ensemble of area under ROC curves based feature selection and a hybrid hidden Markov model-fuzzy approach. Chen et al. (2011) presented a rough set based support vector machine classifier and obtained a classification accuracy of 99.41% for 50-50% of training-test partition. Marcano-Cedeno, Quintanilla-Dominguez, and Andina (2011) achieved a classification accuracy of 99.26% with the training of neural network by an artificial metaplasticity multilayer perceptron algorithm. Uzer, Inan, and Yılmaz (2013) integrated the principal component analysis with a sequential forward selection and sequential backward selection based feature selection method and reported 98.57% classification accuracy. Inan, Uzer, and Yılmaz (2013) presented an integrated model of association rule mining based feature selection, the principal component analysis and a neural network classifier and reported 98.29% classification accuracy. Li, Peng, and Liu (2013) applied the quasiformal kernel common locality discriminant analysis for dimensionality reduction and reported a classification accuracy of 97.26%. Zheng, Yoon, and Lam (2014) presented a K-means algorithm and support vector machine based model and reported a classification accuracy of 97.38%.

Seera and Lim (2014) presented a hybrid intelligent classification model for medical data. The model consists of the Fuzzy Min–Max neural network, the classification and regression tree and the Random Forest algorithm. The Fuzzy Min–Max neural network is responsible for incremental learning, the classification and regression tree is responsible for enhancing understandability and the Random Forest algorithm is utilized to enhance the predictive performance. The WBCD data set is among the medical datasets used in experimental evaluations. They reported a classification accuracy of 98.84% for breast cancer diagnosis.

Chen (2014) presented a hybrid intelligent model for breast cancer diagnoses that can work in the absence of labeled training data. Hence, this work studies the feature selection methods in unsupervised learning models. The model integrates clustering and feature selection. The study indicates that selecting a subset of relevant features instead of using all the features in the original data set can enhance the interpretability of clustering results.

Özşen and Ceylan (2014) examined the performance of artificial immune system as a data reduction algorithm. In order to evaluate the data reduction performance of artificial immune system, it is compared to the fuzzy c-means clustering algorithm. Both data reduction methods are combined with the artificial neural network classifier to obtain the classification results. They obtained a