Hybrid Tolerance Rough Set:
PSO Based Supervised Feature Selection for Digital Mammogram Images

G. Jothi, Department of IT, Sona College of Technology (Autonomous), Salem, Tamil Nadu, India

H. Hannah Inbarani, Department of Computer Science, Periyar University, Salem, Tamil Nadu, India

Ahmad Taher Azar, Faculty of Computers and Information, Benha University, Banha, Egypt

ABSTRACT

Breast cancer is the most common malignant tumor found among young and middle aged women. Feature Selection is a process of selecting most enlightening features from the data set which preserves the original significance of the features following reduction. The traditional rough set method cannot be directly applied to deafening data. This is usually addressed by employing a discretization method, which can result in information loss. This paper proposes an approach based on the tolerance rough set model, which has the flair to deal with real-valued data whilst simultaneously retaining dataset semantics. In this paper, a novel supervised feature selection in mammogram images, using Tolerance Rough Set - PSO based Quick Reduct (STRSPSO-QR) and Tolerance Rough Set - PSO based Relative Reduct (STRSPSO-RR), is proposed. The results obtained using the proposed methods show an increase in the diagnostic accuracy.

Keywords: *Classification, Feature Extraction, Hybrid Intelligent Techniques, Mammogram Images, Particle Swarm Optimization, Segmentation (PSO), Supervised Feature Selection, Tolerance Rough Set (TRS)*

INTRODUCTION

Breast cancer is a kind of cancer that develops from breast cells. It is the most common invasive cancer in females worldwide. Breast cancer rates are much higher in developed nations compared to developing ones. There are several reasons for this, with possibly life suspense being one of the key factors. Breast cancer is more common in elderly women. Women in the richest countries live much longer than those in the poorest nations. Experts believe that different lifestyles and eating habits of females in rich and poor countries are also contributory factors (http://www.medicalnewstoday.com). There are an estimated 1, 00,000 - 1, 25,000 new breast cancer cases in India every year. An article in the Times of India reports that the number of breast cancer cases in India is estimated to double by 2025 (http://articles.timesofindia.indiatimes.com). Premature time diagnosis of the cancer's type could improve patients'

DOI: 10.4018/ijfsa.2013100102

treatment effect. Feature Selection (FS) plays an imperative role in cancer classification. Feature Selection is an important component of pattern recognition, data mining techniques, artificial intelligence and medical image processing. Feature selection, as a preprocessing step to machine learning, is effective in reducing dimensionality, removing irrelevant data, increasing predictive accuracy, and improving result comprehensibility. However, the recent increase of dimensionality of data poses a severe challenge to many existing feature selection methods with respect to efficiency and effectiveness. Rough set theory (RST) is an approach that can be used for dimensionality reduction, whilst simultaneously preserving the semantics of the features (Guyon & Elisseeff 2003). The main disadvantage of RST is its inability to deal with real-valued data. In order to tackle this problem, methods of discretizing the data were employed prior to the application of RST. The use of such methods can result in information loss, however, and a number of extensions to RST have emerged which have attempted to address this inability to operate on real-valued domains. Tolerance Rough Set model has the ability to operate effectively on real-valued (and crisp) data, thus minimising any information loss (Parthaláin & Shen, 2009). Particle Swarm Optimization (PSO) is an evolutionary computation technique which graphically simulates the movement of bird flocking behaviour (Hannah Inbarani et al., 2012a). Rough set is a powerful tool for data reduction based on dependency between attributes. The existing feature selection methods are not effective in selecting features which comprises real valued data. The proposed method combines the advantages of both Tolerance Rough Set (TRS) and PSO. Nineteen features are extracted from the segmented mammogram images using gray level co-occurrence matrix (GLCM). In this work, hybrid supervised Tolerance Rough Set and PSO based Relative Reduct (TRSPSO-RR) and hybrid Tolerance Rough Set and PSO based Quick Reduct (TRSPSO-QR) are proposed for feature reduction and the classification accuracy is measured. The main purpose of the proposed algorithm is to increase the efficiency of features selection method based on the strengths of PSO and Tolerance Rough set.

The rest of the paper is structured as follows: First section describes the related work of feature selection and classification of mammogram images using various techniques. Second section presents the research motivation behind this work. The methodology followed in this study for mammogram images is described in the next section. The features extracted from the mammogram image are explained. The proposed supervised feature selection methods are presented in the next section. Comparative analyses of the proposed algorithms are discussed in the experimental analysis section. Finally, conclusion is presented.

RELATED WORK

There are numerous studies reported focusing on the diagnosis of breast cancer using mammography. Chen et al., (2011) proposed rough set (RS) based supporting vector machine classifier (RS_SVM) for breast cancer diagnosis. The experimental results demonstrate that the proposed RS_SVM can not only achieve very high classification accuracy but also detect a combination of five informative features, which can give an important clue to the physicians for breast cancer diagnosis. Nithya and Santhi (2011) proposed Maximum Difference Feature Selection (MDFS). This neural network based model selects only essential features, eliminates the irrelevant features and improves the classification accuracy. Thangavel and Roselin (2012) have preprocessed the mammogram images, extracted the features and then identified the abnormality through the classification. Fuzzy-Rough feature selection with π membership function is proposed and it increases the classification accuracy. Luo and Cheng, (2010) used two feature selection methods, forward and backward selection, to remove irrelevant features. In addition, decision tree (DT), support vector machine sequential minimal optimization (SVM-SMO) and their ensembles were applied