Chapter 27
A Comparison of Open Source Data Mining Tools for Breast Cancer Classification

Ahmed AbdElhafeez Ibrahim
Arab Academy for Science, Technology, and Maritime Transport, Egypt

Atallah Ibrahin Hashad
Arab Academy for Science, Technology, and Maritime Transport, Egypt

Negm Eldin Mohamed Shawky
Arab Academy for Science, Technology, and Maritime Transport, Egypt

ABSTRACT

Data Mining is a field that interconnects areas from computer science, trying to discover knowledge from databases in order to simplify the decision making. Classification is a Data Mining chore that learns from a set of instances in order to precisely classify the target class for new instances. Open source Data Mining tools can be used to make classification. This paper compares four tools: KNIME, Orange, Tanagra and Weka. Our goal is to discover the most precise tool and technique for breast cancer classifications. The experimental results show that some tools achieve better results more than others. Also, using fusion classification task verified to be better than the single classification task over the four datasets have been used. Also, we present a comparison between using complete datasets by substituting missing feature values and incomplete ones. The experimental results show that some datasets have better accuracy when using complete datasets.

INTRODUCTION

Breast cancer has become the main reason of death in women in advanced countries. Many researchers have tried to apply machine learning algorithms for discovering survivability of cancers in human beings (Padmapriya & Velmurugan. 2014). Data mining has become one of the most explored tools for decision makings, where it discovers new forms within the data. The classification methods can

DOI: 10.4018/978-1-5225-2229-4.ch027

achieve high accuracy in classifying certain applications as it classifies a data item into one of several predefined categorical classes. The fusion classification task (H. Liu et.al 2005) is a set of classifiers that are combined in order to build new example. Combining classifiers shows good classification accuracy and produces more accurate results because diverse classifiers make different errors on different samples (Rosly et al. 2006). The datasets named Wisconsin Diagnostic Breast Cancer Database (WDBC) (Chien-Hsing Chen 2014), Wisconsin Breast Cancer Database Original (WBC), Wisconsin prognosis Breast Cancer Database (WPBC) and Ljubljana Breast Cancer Database University (LBCD) (J. Chhatwal et. al., 2009) are obtained from university of California Irvine (UCI) repository and The Wisconsin Madison University (UCI.2016).The four open source data mining tools KNIME, Orange, Tanagra and Weka are applied over different classification techniques (Wang et. al. 2007). Three datasets have missing feature values, hence substituting missing values by median value is applied. Furthermore, a comparison between using complete datasets by substituting missing feature values and incomplete ones by eliminating instances which have missing feature values is performed (M.A. Jayaram et. Al. 2010). In order to measure the performance, 10-fold cross validation technique is used on datasets (T. Kohonen et. Al., 2000). The paper is prearranged as follows; in the following section named proposed methodology presents the preprocessing steps, the proposed approach and the classification tasks. The results are discussed in the experimental results section. Finally, the latter section introduces the conclusion of this study (C.-H. Chen, 2011).

PROPOSED METHODOLOGY

Data Processing

Preprocessing steps are applied to the data before classification:

- **Data Cleaning:** There are 16 instances in WBC and 4 instances in WPBC that contain a single missing attribute value, denoted by "?"And there are 9 instances in LBCD that have two missing values which substituted by the median value for that feature built on statistics (M. Shah et.al. 2012).
- **Relevance Analysis:** The WBC, WPBC and WDBC have one irrelevant feature (D. Sun et.al., 2010) named 'Sample code number' which has no influence in the classification procedure; therefore, the feature is not considered.
- **Data Normalization:** The goal of normalization is to convert the feature values to a small-scale range (H. Yin et.al., 2002).

The Proposed Approach

We suggested a method for realizing breast cancer using four different data sets based on data mining as follows:

- Selection of Data Mining Tools to test.
- Import the Dataset.
- Discard the irrelevant features.