Expert Systems with Applications 39 (2012) 12371–12377

Contents lists available at SciVerse ScienceDirect

Expert Systems with Applications

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

Computational intelligence for microarray data and biomedical image analysis for the early diagnosis of breast cancer

Jesmin Nahar [a,*], Tasadduq Imam [a], Kevin S. Tickle [a], A.B.M. Shawkat Ali [a], Yi-Ping Phoebe Chen [b]

[a] Central Queensland University, Faculty of Arts, Business, Informatics and Education, Queensland, Australia
[b] Department of Computer Science and Computer Engineering, La Trobe University, Melbourne, Victoria 3086, Australia

ARTICLE INFO

Keywords:
Breast cancer
Microarray data
Image data
Computational intelligence
SMOTE
SMO

ABSTRACT

The objective of this paper was to perform a comparative analysis of the computational intelligence algorithms to identify breast cancer in its early stages. Two types of data representations were considered: microarray based and medical imaging based. In contrast to previous researches, this research also considered the imbalanced nature of these data. It was observed that the SMO algorithm performed better for the majority of the test data, especially for microarray based data when accuracy was used as performance measure. Considering the imbalanced characteristic of the data, the Naive Bayes algorithm was seen to perform highly in terms of true positive rate (*TPR*). Regarding the influence of SMOTE, a well-known imbalanced data classification technique, it was observed that there was a notable performance improvement for J48, while the performance of SMO remained comparable for the majority of the datasets. Overall, the results indicated SMO as the most potential candidate for the microarray and image dataset considered in this research.

1. Introduction

The key focus in this paper is to identify the best suited algorithms for early breast cancer detection considering different factors. Several publicly available databases have been used in different research studies and a number of algorithms have been trialed (Haibe-Kains et al., 2012; Keleş & Yavuz, 2011; Wang et al., 2011). However, a research that compares the effectiveness of the best suited algorithms and at the same time takes into account the imbalanced nature of these data is still missing. Two types of data are generally used in early breast cancer identification: image based and microarray based. Image based identification focuses on analyzing biomedical images to extract features, rules, similarities and dissimilarities using classification algorithms (Sherrow, 2006). In contrast, microarray based identification focuses on gene level structure for reliable classification (Berns, 2000; Campbell, Li, & Tipping, 2001; Lee & Lee, 2003; Nahar, Ali, & Chen, 2007; Peterson & Coleman, 2008; Pomeroy et al., 2002; Veer et al., 2002; Wong & Wang, 2008).

Early breast cancer detection is a difficult and important problem from clinical view, with recent research showing that the chance of a woman affected by invasive breast cancer at some point during her lifetime is about one in eight, and the chance of death is one in thirty five (ACS, 2010). As the causes of breast cancer are still not clear, accurate early detection is very important to reduce the high mortality rate (Cheng et al., 2006). This early detection requires accurate and reliable diagnostic processes, as well as the use of robust prediction techniques. The two identification processes mentioned earlier, image based identification and microarray based identification are popular, but it is unclear how the different classification techniques compare to each other in terms of performance. Further, surveys of existing studies indicate that researchers are yet to fully consider the imbalanced nature of the data used in this domain (detailed in the Section 3.1 and 3.2). Imbalanced nature of data can affect prediction accuracy and, therefore, needs further attention. This research, through analysing a number of classification techniques and also considering imbalanced nature of data, has attempted to provide direction towards the best suited algorithm for use in early breast cancer identification. Thus, this research addresses an important problem in this health informatics research field.

It is to be mentioned that this paper is an extension of the authors' previously published work (Nahar, Tickle, Ali, & Chen, 2009) in this field. But, this paper has been modified substantially from the previous paper to clearly focus on the main objective and also the technical process used varies considerably. Further, this paper has considered SMOTE (Chawla, Bowyer, Hall, & Kegelmeyer, 2002), an imbalanced data classification algorithm, an issue ignored in the previous articles.

* Corresponding author.
E-mail addresses: j.nahar@cqu.edu.au (J. Nahar), t.imam@cqu.edu.au (T. Imam), k.tickle@cqu.edu.au (K.S. Tickle), s.ali@cqu.edu.au (A.B.M. Shawkat Ali), phoebe.chen@latrobe.edu.au (Y.P.P. Chen).

0957-4174/$ - see front matter © 2012 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.eswa.2012.04.045

This paper is organised as follows: Section 2 briefly talks about breast cancer, its characteristics, symptoms, causes and treatment options. Section 3.0 will explain the image and microarray based existing identification process and limitations. Section 4 will describe the datasets that will be used for the experiment. Section 5 will then describe the experimental setup for the comparative research and the outcomes of the experiment. Section 6 will conclude the paper with a summary of the results and possible future research directions.

2. Breast cancer

Breast cancer is the situation when abnormal cells occur in the breast and then multiply uncontrollably (BCOSU, 2011; Caars., 2010; Setchi, Jordanov, Howlett, & Jain, 2010). Considering the size of tumor, there are four stages of breast cancer (Fallowfield & Clark, 1991): (i) T1: tumors are of size 2 cm or less, (ii) T2: tumours between 2 and 5 cm, (iii) T3: tumours having size greater than 5 cm and (iv) T4: tumours can be of any size. The majority of early breast cancer cases, however, have no early symptom. The particular causes of breast cancer are unknown, but major risk factors like obesity (Begum, Richardson, & Carmichael, 2009; Brown & Simpson, 2010), early menarche and late childbirth (Bombardieri, Bonadonna, & Gianni, 2008), breastfeeding status (de Silva, Senarath, Gunatilake, & Lokuhetty, 2010; Gajalakshmi et al., 2009), hormonal factors (Sieri et al., 2008; Woolcott et al., 2010) and family history (Mavaddat et al., 2010; Tao, Yu, Ross, & Xiu, 2006) have been claimed to increase the risk of developing breast cancer. There are some treatment options (Bombardieri et al., 2008; Healthi., 2010; NCI., 2010; Taghian, Smith, & Erban, 2009) like surgery, radiation therapy, chemotherapy and hormonal therapy. But treatment effectiveness depends on factors (Bombardieri et al., 2008; NCI, 2010) like the cancer's stage, growth rate, and the physical properties of the patient. As, it is a highly mortal disease and difficult to treat in advanced stages, early diagnosis is very important (Evans, 2012; Murukutla, Vivekanandaraja, Paramanathan, Kong, & Varma, 2012; Pieters et al., 2012; Ward & Smith, 2010).

3. Approaches in computer aided cancer identification

In recent times, computer aided early breast cancer detection has gained popularity (Alizart, Saunus, Cummings, & Lakhani, 2012; Gatza et al., 2010; Haibe-Kains et al., 2012; Molloy, Roepman, Naume, & Van't Veer, 2012; Tan, Quek, Ng, & Ng, 2007; Truong et al., 2010). Generally two identification processes are used in determining cancers: *Image based identification* and *Microarray based identification*. Image based cancer identification mainly rely on analyzing different types of images like mammography, ultrasound, x-ray, CT scan and MRI to diagnosis the disease (Orlov et al., 2009a; Sherrow, 2006). Microarray based cancer identification mainly uses gene information (Stillman & Stewart, 2005; Teicher, 2006) for this purpose. The following subsections detail existing computer aided diagnostics based on each of these data formats.

3.1. Image data based identification

Image-based identification is becoming popular in cancer studies due to advancement of image extraction and manipulation facilities, and also due to the easier understanding by the doctors. A number of researches have employed this approach. Antonie et al. (2001) for example, emphasised on using automatic reading as well as double reading of digital mammograms and observed mammography a very reliable method in the early detection of breast cancer. Wang, Zhou, and Geng (2005) used a decision tree algorithm based model for mammography classification and achieved 98% accuracy for breast cancer diagnosis. Another digital imaging and image analysis based system, using mammographic features and classification techniques, (Rangayyan, Ayres, & Desautels, 2007) also showed improvement in diagnosis. Some other similar works are: a novel image analysis method utilizing SVM classifier (Doyle, Agner, Madabhushi, Feldman, & Tomaszewski, 2008), pattern recognition processes for biomedical image-guided diagnosis (Orlov et al., 2009b), method to estimate breast tumour's growth rate from breast screening information (Michaelson et al., 2003), receiver operating curve methods for the analysis, classification of breast masses (Rangayyan & Nguyen, 2005), and breast cancer diagnosis through thermogram analysis of using Complementary Learning Fuzzy Neural network (Tan et al., 2007),. There are also other similar research studies (Nattkemper et al., 2005; Rodrigues, Ruey-Feng, & Suri, 2006; Sheshadri & Kandaswamy, 2007; Twellmann, Meyer-Baese, Lange, Foo, & Nattkemper, 2008). However, image-based identification has limitations, such as data storage and proper pre-processing for classification, as well as processing time.

3.2. Microarray based identification

In the microarray based identifications, tumour's gene expression profiles are used for early cancer diagnosis and have become popular due to reliable classification results. Lu and Han (2003), for example, revealed that cancer classification using microarray data has potentials in providing a robust approach in identifying tumour types. Berrar, Downes, and Dubitzky (2003) demonstrated the use of a probabilistic neural network (PNN) for multiclass data in cancer diagnosis. Sharma and Paliwal (2008) proposed the gradient LDA (linear discriminant analysis) technique to avoid the singularity problem in this field. Other similar works using microarray based cancer identification are: ant colony optimisation (ACO) algorithm and various machine learning procedures to select genes relevant to cancer (Chiang, Chiang, & Lin, 2008), nonlinear kernel SVM based diagnosis (Qizhong, 2007), gene markers identification through Neural Networks (Minn et al., 2005), chronological feature extraction approach through Naive Bayes (Fan, Poh, & Zhou, 2009), an ensemble method to choose multiple gene subsets (Liu et al., 2009) and a SVM based approach to improve breast-cancer diagnosis (Chen & Yang, 2012). There are also other studies that have used microarray data for cancer detection (Berns, 2000; Campbell et al., 2001; Chen & Hsu, 2006; Chen & Yang, 2012; Lee & Lee, 2003; Nahar et al., 2007; Peterson & Coleman, 2008; Pomeroy et al., 2002; Veer et al., 2002; Wong & Wang, 2008).

3.3. Overall limitations of existing research

As shown in the literature review, a number of researchers have developed diagnosis methods for the identification of cancers. But, a detailed comparative analysis of the suitable algorithms is missing and imbalanced nature of the data is yet to be well considered. Both of these issues are addressed in this research article for the diagnosis of early breast cancer.

4. Data collection and preparation

This research uses three image based and two microarray based breast cancer datasets, as also have been used in the authors' previously published work (Nahar et al., 2009). The image dataset was based on cancer affected and healthy breast images (IBC, 2010). The image manipulation and extraction facilities provided by Matlab, along with multidimensional image filtering, were applied to generate the datasets for the experiments.