Bio-Medical Materials and Engineering 24 (2014) 3397–3404
DOI 10.3233/BME-141163
IOS Press

Breast cancer early diagnosis based on hybrid strategy

Peng Li[a,*], Tingting Bi[b] , Jiuling Huang[b] and Siben Li[b]
[a]*School of Software, Harbin University of Science and Technology, Harbin 150080, China*
[b]*School of Computer and Science and Technology, Harbin University of Science and Technology, Harbin 150080, China*

Abstract. The frequent occurrence of breast cancer and its serious consequences have attracted worldwide attention in recent years. Problems such as low rate of accuracy and poor self-adaptability still exist in traditional diagnosis. In order to solve these problems, an AdaBoost-SVM classification algorithm, combined with the cluster boundary sampling preprocessing techniques (CBS-AdaBoost-SVM), is proposed in this paper for the early diagnosis of breast cancer. The algorithm uses machine learning method to diagnose the unknown image data. Moreover, not all of the characteristics play positive roles for classification. To address this issue the paper delete redundant features by using Rough set attribute reduction algorithm based on the genetic algorithm (GA). The effectiveness of the proposed methods are examined on DDSM by calculating its accuracy, confusion matrix, and receiver operating characteristic curves, which give important clues to the physicians for early diagnosis of breast cancer.

Keywords: Computer-aided diagnosis, image data mining, support vector machine, clustering sampling

1. Introduction

Breast cancer is one of the most common cancers throughout the world. Although experts said that many factors might increase the risk of breast cancer, the main pathogenesis has not been yet confirmed [1]. Furthermore, due to the limitation of medical technology and equipment, there is no effective method can eradicate breast cancer [2]. Unfortunately, 10%-30% breast cancer cannot be diagnosed by looking at the mammography, because the features in the early images are not obvious enough and doctors' judgments also affect the detection [3]. With the emerging and development of computer aided diagnosis (CAD) and data mining and discovery technology have applied extensively in the medicine field [4]. The methods are the foundations of the medical image automatic detection and they also promote the constant development of the clinical medicine [5].

As the rapid development of artificial intelligence and computer vision, more new algorithms and ideas have been used in CAD system for breast cancer [6–8]. However, diagnosis of breast cancer is a typical imbalanced data problem, meaning that a majority of people are in good health in the real data, but what we are often concerned about is the minority of patient who suffered from breast can-

*Corresponding author: Peng Li, School of Software, Harbin University of Science and Technology, Harbin 150080, China. Tel.: +0 086 0451 86397003; Fax: + 0 086 0451 57812666; E-mail: pli@hrbust.edu.cn.

cer. Studies have shown that conventional methods are not effective in solving the problem of imbalanced data classification [9].

In light of the existing problems, a method based on AdaBoost- SVM classification algorithm, combined with the cluster boundary sampling preprocessing technique (CBS-Adaboost-SVM), is proposed in this paper. Moreover, diagnoses are made by using a genetic algorithm (GA) based rough set attribute reduction algorithm to achieve better results.

2. Image processing and features selection

There are many factors may influence breast cancer early diagnosis in the process of data preprocessing and they form a multi attribute dimensional dataset. These attributes are called characteristics. Since not all aspects of a mammogram are valuable and some of them are even redundant, it is necessary to select those important features.

2.1. Pre-process of mammograms

In this paper, all mammograms are selected from the Digital Database for Screening Mammography (DDSM). The image quality of mammograms is directly related to many factors, but the resolutions and image grayscales are different from cases to cases. Therefore, all mammograms need to be normalized. Eq. (1) adjusts image gray of mammograms to a 0-to-255 scale.

$$G_1(x,y) = (G_0(x,y) - \min(G_0)) \times \frac{255-0}{\max(G_0) - \min(G_0)} \tag{1}$$

Where $G_0(x,y)$ is gray level of original X-ray image at point (x,y), $G_1(x,y)$ is gray level after scaling at point (x,y). All subsequent experiments are based on $G_1(x,y)$. Eq. (2) adjusts the resolutions of mammograms.

$$GDDSM_1 = \frac{GDDSM_0}{2^{12} / 2^8} \tag{2}$$

Eq. (3) performs normalizations to mammograms.

$$C(P_0) = \frac{area(p_1 \cap p_0)}{area(P_1)} \tag{3}$$

Where $C(P_0)$ is weight of P_0 relative to P_1, and P_0, P_1 are pixel dots.

2.2. Genetic algorithm based rough set attribute reduction algorithm

Attribute reduction is one of important techniques for feature selection [10]. This paper proposes a rough set attribute reduction algorithm based on generic algorithm, which is able to deal with problems that heuristic algorithm can't solve is the procedure of the method is described as follows:

Set $P \subseteq C$, for $P \subseteq C, \{Y_1, Y_2 \ldots, Y_k\}$, the approximation accuracy of P is

$$\gamma_p = \sum_{i=1}^{k} card(P - Y_i) / card(U) \tag{4}$$

where P is condition attribute, Y_i is constraint condition, and $card(U)$ means the cardinal number of the set.

Fitness function:

$$F(x) = (1 - card(x)/n) + k \tag{5}$$

Where, n is the length of condition attributes, k is dependence of decision attribute to condition to attribute.

Input: decision table $S = (U, A, V, F)$, where $A = C \cup D$, C is condition attribute, and D is decision attribute.

Output: attribute reduction R of this decision table.

- The degree of dependency $\gamma_c(D)$ of decision attribute D on condition attribute is calculated by Eq. (4).
- Let $Core(C) = \varphi$ and attribute $c \in C$. If $\gamma_{C-c} \neq \gamma_c$, then $Core(C) = Core(C) \cup \{c\}$; If $\gamma_{core}(D) = \gamma_c(D)$, then $Core$ is the minimum relative reduction. Otherwise implements the next step.
- Randomly generates m binary strings with initial population length of n. The bit corresponds to the attribute in the nuclear is 1. The others are set to 1or 0 randomly.
- Calculate the dependency degree of decision attribute on contains condition attribute for each individual by Eq. (1). Calculate the adaptive value of each individual by Eq. (5). Then calculate the selected probability of each individual. At last using simulation bets (mean a random number in 0-1) to select the individual.
- Single-point crossover is used for the crossover operation with a (user-defined) crossover probability pc.
- A basic mutation method, where corresponding position of attributes in the nuclear always remain the same, is used in the mutation operation, with probability pm.
- Copy the optimal individual into the next generation group by using the optimal preservation strategy.
- If the best individual adaptive value of the Kth generation does increase, the algorithm stops; otherwise return to step 4.

3. Cluster boundary sampling (CBS) and adaboost-SVM algorithm

3.1. Cluster boundary sampling based on density clustering

Information is not uniformly distributed among samples and there should exist a kind of core information which influences the efficiency of classification. Therefore, this paper proposes an assumption that the boundary of cluster contains similarity information in a same cluster because there are significant differences among clusters.

The data elements are more intensive in the same cluster in the vector space by density clustering. This paper represents any data element x in the form of feature vector, and use the standard Euclidean distance as the distance between two vectors, is shown in Eqs. (6) and (7).

$$\langle \alpha_1(x),\ \alpha_2(x),\ \ldots,\alpha_n(x) \rangle \tag{6}$$

Where $\alpha_k(x)$ represent the k-th attribution of x. The Euclidean distance between x_i and x_j is:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{n} (\alpha_k(x_i) - \alpha_k(x_j))^2} \tag{7}$$

In the data set D, the field of x instance can be defined as:

$$EPS(x) = \{y \in D \mid d(x, y) \le EPS\} \tag{8}$$

This paper uses the notation $|EPS(x)|$ to represent the number of the element x 's field. Two density thresholds are used to identify the boundaries of clusters in a two-step faction. The first one is called the density threshold of clusters. It is used to estimate and divide the whole data set to several clusters according to features and average distance. The other one is the density threshold of boundary. It is used to estimate the scales of clusters and find the boundary data elements of clusters. It es EPS_1 and $MINP_1$ to divide the data set into several clusters named C in the first step. In the second step, it find the boundary of each C_i using EPS_{ci} and $MINP_{ci}$. In this paper, it uses D to represent the training set, C_i to represent the n-th cluster, and B_i to represent the boundary of the n-th cluster, which is:

$$D = \{C_1, C_2, C_3, \ldots, C_n, C_{noise}\} \tag{9}$$

$$C_i = \{x \in D \mid |EPS(x)| \ge MINP_1\} \tag{10}$$

$$B_i = \{x \in C_i \mid |EPS(x)| \ge MINP_{ci}\} \tag{11}$$

The procedures of the algorithm are:

– Traverse the data elements of D and calculate the distance of elements;
– Estimate $MINP_1$;
– Cluster the data set using the first density threshold;
– Assign data elements to different clusters, C_i *or* C_{noise};
– Calculate the number of elements N_{ci} of C_i;
– Calculate $MINP_{ci}$ of C_i according to N_{ci};
– Calculate $MINP_{ci}$ of the second step and select B_i from C_i;
– Repeat step 4 until all data elements (not noise) are traversed
– Getting all B_i.

Due to the uneven distribution of positive and negative samples, there is a significant difference between the numbers of the two. Therefore, it retains all the positive samples and only identifies the boundaries of clusters on negative samples. Finally, the classifier is trained based on all the positive samples and the boundaries of negative sample clusters.

3.2. Adaboost-SVM algorithm

Support vector machine (SVM) is originally developed by Boser and Vapnik. It was developed based on the Vapnik-Chervonekis (VC) theory and structural risk minimization (SRM) principleby trying to find the trade-off between minimizing the training set error and maximizing the margin in order to achieve the best generalization ability and remains resistant to over fitting. In addition, a major advantage of SVM is the use of convex quadratic programming that provides a unique global minimum and hence avoids being trapped in local minima. In this section it will be concentrated on the basic SVM concepts for typical binary-classification problems.

The basic idea of the boosting algorithm is to produce several weak classifiers which are slightly better than random guessing and then incorporate them into estimations with high accuracy [11]. Adaboost has the following advantages: (1) Its training process focuses on the data that are difficult to be classified; (2) weighted voting is adopted during the weak classifiers integration instead of average voting mechanism; (3) Features are selected contingent on the features

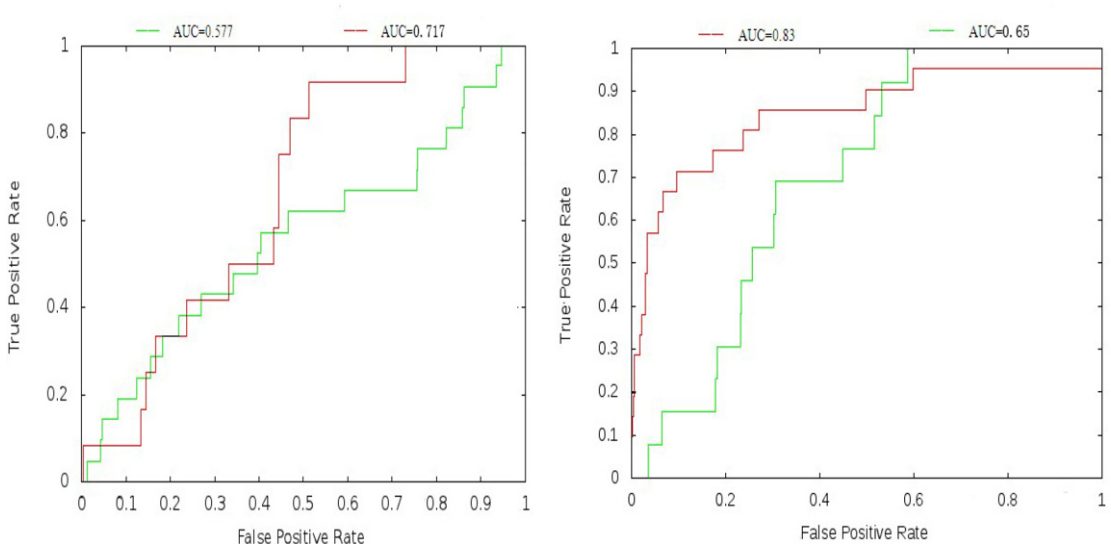
This paper uses SVM as the base classifier in the framework of Adaboost for reasons that, first, SVM can deal with a wide range of data sets including data sets with small sample sizes, nonlinearity or high dimensions, and second, there exists mature and convenient software package of SVM, such an LIBSVM.

4. Experiment verification and analysis

4.1. Evaluation index of breast cancer CAD

This paper conducted comparative experiments on DDSM in order to verify the effectiveness of our methods. Detailed analyses were carried out for three specific techniques: attribute reduction, data resampling, and classification algorithm.

In recent years, a new study has shown that Receiver Operating Characteristic (ROC) curve and Area Under ROC Curve (AUC) are better metrics to evaluate the efficiency of computer-aided detection



(a)ROC curve before and after CBS based on SVM (b) ROC curve before and after CBS based on AdaBoost-SVM

Fig. 1. ROC curves of all characters in DDSM.

Table 1

AUC values for different machine learning methdos before and after sampling

	SVM	Adaboost-SVM
Before sampling	0.577	0.65
After sampling	0.717	0.83

and diagnosis algorithms, because ROC curve and AUC will not be affected by imbalanced data or non-normally distributed data. It means that ROC curve and AUC will not change even if the numbers of positive and negative samples in test dataset change. Therefore the ROC curve evaluation index is more scientific and intuitive [12]. In this paper, it uses the true positive rate (sensitivity) as the vertical axis and the false positive rate (1- specificity) as the horizontal axis.

4.2. Experimental results and discussions

Experiment 1: The effectiveness of sampling method and ensemble learning method were tested and verified on all dataset without attribute reduction. The comparison ROC curves results of two techniques are piloted in Figure 1. The AUC values calculated for all scenarios are shown in Table 1.

The results for experiment 1 show that the CBS method can improve detection rate significantly for both classification algorithms. When using SVM as the classifier, the AUC value was increased from 0.577 to 0.717. When using AdaBoost-SVM as the classifier, the AUC value was increased from 0.65 to 0.83. Moreover, The AUC value was increased from 0.577 to 0.65 before sampling after using ensemble learning method. Similarly, the use of learning method improved AUC value from 0.717 to 0.83 after sampling. It can conclude from the results that the classification performance can be improved significantly by balancing the sizes of different types of samples, refining the core information

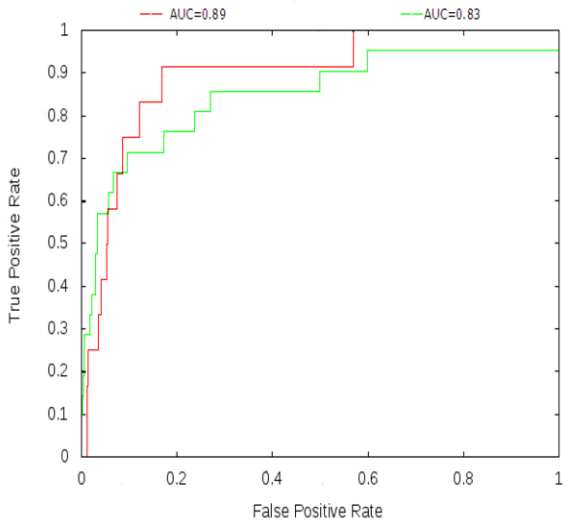


Fig. 2. ROC curves for detections using and not using GA based rough set attribute reduction algorithm

of dataset, and reducing the noise data and using ensemble learning base on SVM classification algorithm can enhance the stability in CAD for breast cancer.

Experiment 2: GA based rough set attribute reduction algorithm eliminates the attributes which have negative effect to the classification. The CBS-Adaboost-SVM method was then used. The ROC curves are shown in Figure 2.

The results of experiment 2 show that #subset after rough set attribute reduction algorithm based on GA, then classified by CBS-AdaBoost-SVM method in experimental 1. The AUC value was increased from 0.83 to 0.89. It can draw a conclusion: in the case of more attributes, some of them are played a negative role for classification.

In all, the experiments proved that the methods for breast cancer early detection are effective. The three specific techniques improved the detection performance gradually and it shows that three methods have a good supplementary effect to the overall performance.

5. Conclusion

The identification and diagnosis of early breast cancer using mammograms is currently a difficult problem. This paper proposed a classification method base on imbalanced data mining technology and provided an effective solution to the detection of early breast cancer. Our methods have three major contributions. First, a genetic algorithm based rough set algorithm was proposed for image preprocessing and attribute reduction optimization for mammograms where a multi attribute dimension vector dataset was constructed for subsequent calculations. Second, we developed a re-sampling method based on clustering boundary sample (CBS) that reduced the data imbalance ratio, refined the core information of data, and eliminated noisy data in order to improve the classification performance. Finally, classification for dataset based on Adaboost-SVM method, and ensemble learning is leaded into the standard classifier. It was also found that the classification performance could be improved by

gradually adding new features to the classifier. Through the contrast experiments, it showed that the detection performance for early stage breast cancer of the proposed approach is better than conventional approaches. Furthermore, the method can help doctors to improve the diagnostics accuracy and reduce false detection rate.

Acknowledgement

This paper is partially supported by National Natural Science Foundation of China (61103149), Postdoctoral Science Foundation (2011M500682, LBH-Z11106), Technological Innovation Foundation for Youth Scholars of Harbin (2012RFQXG093) and Natural Science Foundation of Province (QC2013C060).

References

[1] M.L. Irwin, C. Duggan, C.Y. Wang et al., Fasting c-peptide levels and death resulting from all causes and breast cancer: The health, eating, activity, and lifestyle study, Journal of Clinical Oncology **1** (2011), 47–53.
[2] C. DeSantis, R. Siegel, P. Bandi et al., Breast cancer statistics, A Cancer Journal for Clinicians **6** (2011), 408–418.
[3] A. Jemal, F. Bray, M.M. Center et al., Global cancer statistics, A Cancer Journal for Clinicians **2** (2011), 69–90.
[4] R. Wiemker, P. Rogalla, T. Blaffert et al., Aspects of computer-aided detection (CAD) and volumetry of pulmonary nodules using multislice CT, The British Journal of Radiology **1** (2014), 46–56.
[5] H.D. Cheng, J. Shan, W. Ju et al., Automated breast cancer detection and classification using ultrasound images: A survey, Pattern Recognition **1** (2010), 299–317.
[6] J.J. Fenton, L. Abraham, S.H. Taplin et al., Effectiveness of computer-aided detection in community mammography practice, Journal of the National Cancer Institute **7** (2011), 1152–1161.
[7] E. Alberdi, A.A. Povyakalo, L. Strigini et al., Use of computer-aided detection (CAD) tools in screening mammography: a multidisciplinary investigation, The British Journal of Radiology **1** (2014), 31–40.
[8] M.M. Eltoukhy, I. Faye and B.B. Samir, Breast cancer diagnosis in digital mammogram using multiscale curvelet transform, Computerized Medical Imaging and Graphics **4** (2010), 269–276.
[9] M. Khalilia, S. Chakraborty and M. Popescu, Predicting disease risks from highly imbalanced data using random forest, BMC Medical Informatics and Decision Making **7** (2011), 51–64.
[10] Y. Qian, J. Liang, W. Pedrycz et al., Positive approximation: an accelerator for attribute reduction in rough set theory, Artificial Intelligence **9** (2010), 597–618.
[11] J.H. Morra, Z. Tu, L.G. Apostolova et al., Comparison of AdaBoost and support vector machines for detecting Alzheimer's disease through automated hippocampal segmentation, IEEE Transactions on Medical Imaging **1** (2010), 30–43.
[12] X. Robin, N. Turck, A. Hainard et al., pROC: An open-source package for R and S+ to analyze and compare ROC curves, BMC bioinformatics **3** (2011), 77–85.