Breast Cancer Diagnosis via the hybrid of genetic algorithm and support vector machine

Tianming Zhao

Research School of Computer Science, Australian National University,
Canberra, ACT, 2601, Australia
E-Mail: u5961519@anu.edu.au

Abstract. This research aims to explore the feasibility of using the hybrid of genetic algorithm (GA) and support vector machine (SVM) to diagnose breast cancer. The data set used in this research is the Breast Cancer Wisconsin(Diagnostic) data set which contains 30 features for classification. By applying the hybrid of GA and SVM, the highest classification accuracy achieved is 98.07%, which is higher than previous work did by Wolberg, Street and Mangasarian [1] who trained a classifier with accuracy 97.3%. Besides, in this research, a risk fitness value is introduced to GA to select out features that can train a SVM classifier making less false negative mistakes while diagnosing breast cancer. With the risk fitness value, 205 true breast cancer instances out of 212 are diagnosed correctly.

Keywords: Breast Cancer Diagnosis, Support Vector Machine, Genetic Algorithm, Feature Selection

1 Introduction

Nowadays, breast cancer is still a big issue that challenges women's health. As a stuff who used to work in a cancer hospital, I saw many women suffering from this disease and some of them even died on it. To reduce the harm caused by breast cancer, the best way is to diagnose it as soon as possible and do immediately treatment. Therefore, the accuracy of breast cancer diagnosis plays a key point and the price for making a false negative mistake (misdiagnose a true malignant breast cancer) can be extremely high.

To assist doctors to do more accurate breast cancer diagnosis, this research provides a machine learning approach to predict if the breast cancer is malignant or not via 30 Fine-needle aspiration (FNA) features. Exactly, this binary classifier is implemented by support vector machine (SVM) with kernel trick. Besides, a feature selection technique delivered by genetic algorithm (GA) is also applied to trained the optimal SVM classifier.

1.1 Data set

In this research, the data set used for training the SVM classifier is Breast Cancer Wisconsin (Diagnostic) Data Set. There are 569 instances in this data set and each of them contains 32 attributes. No value absents in the dataset. Besides, the balance of dataset is acceptable, in which 212 instances are malignant breast cancers and the other 357 instances are benign breast cancers.

For the 32 attributes in the data set, the first is the ID number of the instance, which is no help for training the SVM classifier. The second attribute is the label which indicates the instance is benign or malignant breast cancer. The remaining 30 attributes are computed from a digital image of a Fine-needle aspiration (FNA) of a breast mass [2], which are main features to train the SVM classifier. Besides, an important point is that these 30 attributes come from three individual cell nucleuses who share the same 10-dimension feature space. In other words, the 30 attributes of an instance are a combination of attributes from 3 individual cell nucleuses.

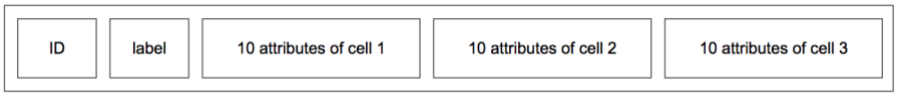

Fig.1. the components of the 32 attributes of an instance

1.2 Performance evaluation

In this research, the trained SVM classifier is evaluated by 10-fold cross validation. The data set is shuffled first and then split into 10 folds with the same size. Each time the classifier is trained with 9 folds and the remaining fold is used for testing. By this way, each instance will be used for validation once. The evaluation result is the average of 10 validation results. An advantage of training the SVM classifier with 10-fold cross validation is that it will utilize the limited data sufficiently and alleviates the problem of overfitting.

Confusion matrix is used to summary the performance of the SVM classifier. The confusion matrix is helpful for reporting the number of false positive and false negative mistakes that occur during the prediction phrase.

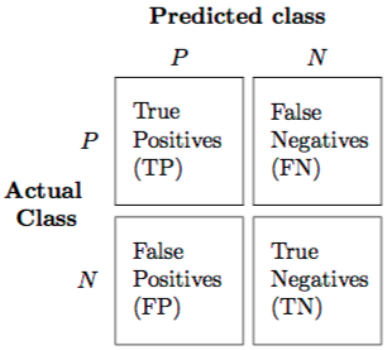

Fig.2. confusion matrix *[3]*

Another indicator that evaluates the performance of the trained classifier is the classification accuracy, which is calculated by the following equation

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

In this research, a self-defined indicator 'risk' is introduced to measure the performance of the classifier either, which is calculated by the following equation

$$Risk = 100 * FN + 1 * FP \tag{2}$$

where FN is the number of the cases that the 'Malignant' breast cancers are diagnosed as 'Benign' breast cancers, on the contrary, FP is the number of the cases that the 'Benign' breast cancers are diagnosed as 'Malignant' breast cancers. In this research, I allocate much more weight on the first case to calculate the risk, since false negative mistakes are much deadlier. If a real malignant breast cancer was diagnosed as a benign breast cancer, the patient would miss the best opportunity for treatment, which would lead the patient to death as the worst result. The risk indicator aims to select a classifier which has better performance on diagnosing malignant breast cancer.

2 Methods

The general idea of this research is using GA to select out the most separable features first. Then, the SVM classifier will be trained with the selected features. This idea is inspired by [4]. In their research, hundreds of features were collected with sensors to recognize stress. Then, they did feature selection with GA. With the selected features, they trained models with artificial neural network (ANN) and SVM respectively. From their result, the hybrid of GA and SVM performed better on both prediction accuracy and execution time. Therefore, I will focus on the hybrid of GA and SVM in this research.

2.1 Support Vector Machine

SVM is a classification model which has been successfully used in many fields and it is quite good at solving binary classification problems. In a classification problem, a SVM classifier will generate an optimal decision hyperplane by maximizing the distance from the support vectors to the hyperplane, which can separate the training instances effectively. Besides, kernel trick can be used in a SVM classifier, which provides the SVM classifier the ability to perform decent non-linear classification. With the kernel trick, the original feature space can be transformed to high dimensional feature space implicitly. In the high dimensional feature space, the training instances can be separated much easier [5]. In this research, the classification performance of different kernels will be compared to select out the best SVM classifier for breast cancer diagnosis.

2.2 Genetic Algorithm

GA is a widely-used technique for searching the global optimal solution, which is inspired by the evolution of creatures. In GA, solutions are represented by chromosomes. The initial population of GA are formed with randomly generated

chromosomes. Then, crossover, mutation and selection methods will be applied onto the population over many generations until the stop condition is satisfied. In the last generation, the dominant chromosome in the population will be the fittest chromosome which is also the optimal solution returned by GA. In this process, crossover and mutation methods keep the diversity of the chromosomes in the population, which avoids the solution being caught in local optimal. Besides, the selection method selects out the best chromosomes as parents. These parents will generate children to form the population of next generation [4]. After many generations' evolution, the fittest chromosome will dominate the whole population, which is an analogy of the evolution process of natural creatures.

GA is one of the most effective algorithms for feature selection [6]. To apply GA for feature selection, the candidate solutions of the feature selection need to be represented by chromosomes (one example of the chromosome is showed in figure 2) to form the population. In my research, the chromosome of a solution of the feature selection is a binary list where 1 indicates the feature is selected and 0 indicates the feature is discarded. After many generations, the fittest features will be selected out.

feature 1	feature 2	feature 3	feature 4	feature 5	feature 6	feature 7	feature 8	feature 9	feature 10
0	1	1	0	0	1	1	0	0	1

Fig.3. a candidate chromosome of feature selection

3 Implementation

3.1 Dataset Preprocessing

Since the first attribute in the data set is the ID number of an instance, which is no help for training the SVM classifier. Therefore, it is save to drop the first column of the data set.

The second column of the data set is the label whose value is 'M' or 'B' which indicate the instance is a malignant or a benign breast cancer. For the convenience of training the SVM classifier, 'M' is converted into 1 and 'B' is converted into 0.

The remaining 30 attributes are useful features for training the SVM classifier. To get a classifier with decent performance, these attributes are standardized with equation

$$x' = \frac{x - \bar{x}}{\sigma} \tag{3}$$

where x is the feature vector, $\bar{x}$ is the mean of that feature vector, and σ is the standard deviation of the feature vector.

3.2 Implement GA feature selection

The workflow of feature selection with GA is showed in figure 4.

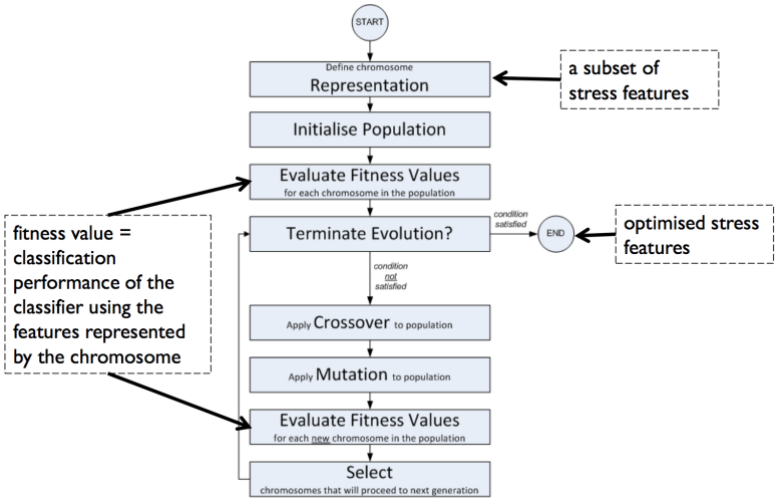

Fig.4. workflow of feature selection *[7]*

In this research, each instance has 30 attributes which can be used for training the SVM classification. However, instead of a list with 30 binary elements, in the step of representing solutions of feature selection with chromosomes, I decide to

encode the solution via a list with 10 binary elements. The reason is that the 30 attributes come from 3 individual cell nucleuses sharing the same 10-dimension feature space and we can do the feature selection on a single cell nucleus. When the fittest solution of feature selection of a cell nucleuses is found, we can use it to select features for an instance in the data set. In this way, computation capacity is saved in crossover stage and mutation stage. Then, the initial population is generated randomly with 1000 such 10-element binary chromosomes.

The fitness value of a chromosome is the 10-fold cross validation performance of the SVM classifier which is trained with the selected features by that chromosome. In this research, the performance of the SVM classifier is evaluated with two standards. The first is the classification accuracy, with which the selected features by the GA will train a SVM having highest classification accuracy. The second is the risk which has been listed in equation 2. With it, the trained SVM classifier will make false negative mistakes as less as possible. Furthermore, since the method to calculate the fitness value of a chromosome stays the same in all generations, the fitness values are saved in a python directory. Therefore, when processing an observed chromosome, instead of calculating again, the fitness value can be collected by looking up the dictionary, which will save execution time tremendously.

After calculating the fitness value of each chromosome in the population, a parent set with the same size of the population will be sampled out according to the distribution listed in equation 4, where $F(x_i)$ is the fitness value of a chromosome, N is the size of the population. In other words, the chromosomes with higher fitness value have higher probability to be selected out as parents.

$$P(\mathrm{x_i}) = \frac{\mathrm{F(x_i)}}{\sum_{\mathrm{j=1}}^{\mathrm{N}} \mathrm{F(x_j)}} \tag{4}$$

With setting the crossover rate to 0.8, one chromosome in the parents set has 80% probability to generate a child with a random chromosome in the parents set. Besides, the mutation rate is set to 0.005, which indicates 0.5% genes of the children's chromosome will mutate when they are generated. Then, the parents in the parents set will be replaced by the new born chromosomes, which will form a new population. Lastly, the total generation of the GA is set to 1000 where the GA will terminate.

3.3 Implement SVM

In this research, I use the third part package 'scikit-learn' to implement the SVM classifier. The SVM provided by 'scikit-learn' has 4 built-in kernel functions can be used, which are 'liner', 'polynomial', 'rbf' and 'sigmoid'. The kernel functions are listed as equations 5, 6, 7 and 8 [8].

$$liner: < x, x' > \tag{5}$$

$$polynomial: (\gamma < x, x' > + r)^d \tag{6}$$

$$rbf: \exp(-\gamma |x - x'|^2) \tag{7}$$

$$sigmoid: (\tanh(\gamma < x, x' > + r)) \tag{8}$$

4 Results & Discussion

Table 1 feature selection solutions with accuracy as fitness value

kernel	radius	texture	perimeter	area	smoothness	compactness	concavity	concave points	symmetry	fractal dimension
Linear	1	1	1	0	1	1	1	0	1	0
Polynomial	1	1	0	0	1	0	0	1	0	0
Sigmoid	1	1	1	1	1	0	1	0	1	0
RBF	1	1	1	1	1	1	1	1	1	0

Table 2 feature selection solutions with risk as fitness value

kernel	radius	texture	perimeter	area	smoothness	compactness	concavity	concave points	symmetry	fractal dimension

Linear	1	1	1	0	1	1	1	0	1	0
Polynomial	0	1	1	0	1	1	0	0	0	0
Sigmoid	1	1	1	1	0	1	0	1	0	0
RBF	1	1	0	1	1	0	0	1	1	1

Table 3 performance of SVM classifier without feature selection

Accuracy	Risk	False Negative	Feature Number
0.9789	804	0.0377	10

Table 4 performance of SVM classifier aiming at highest accuracy

Kernel	Accuracy	Risk	False Negative Rate	Feature Number
Linear	0.9807	803	0.0377	7
Polynomial	0.9244	4300	0.2028	4
Sigmoid	0.9649	1604	0.0754	7
RBF	0.9807	803	0.0377	9

Table 5 performance of SVM classifier aiming at lowest risk

Kernel	Accuracy	Risk	False Negative Rate	Feature Number
Linear	0.9807	803	0.0377	7
Polynomial	0.9244	4201	0.2028	4
Sigmoid	0.9525	1116	0.0519	6
RBF	0.9754	707	0.0330	7

Table 1 shows the solutions of feature selection generated by GA when the classification accuracy of the SVM classifiers with different kernels is used as the fitness value. The performances of the SVM classifiers which are trained with these selected features are listed in Table 4. As a contrast, the performance of the SVM classifier which is trained without feature selection is show in Table 3. Comparing the results in Table 3 and Table 4, we can find that the highest classification accuracy achieved with applying GA feature selection is a little bit higher than without applying feature selection, which shows the hybrid of GA feature selection and SVM can generate better classifier. Besides, comparing the results within Table 4, we can find that linear kernel SVM classifier and RBF kernel SVM classifier result in the same performance. However, the feature spaces of these two cases are not the same. Exactly, the feature space of the linear kernel SVM classifier case is in lower dimension, which provides the classifier trained with this feature space the superiority while considering computing capacity and generalization.

Table 2 shows the solutions of feature selection generated by GA when the classification risk of the SVM classifiers with different kernels is used as the fitness value. The performance of each trained SVM classifier is showed in Table 5. Comparing results in Table 4 and Table 5, we can find that the classifier which trained with feature selection will make less false negative mistakes. This result further proves the effectiveness of the hybrid of GA feature selection and SVM. Besides, among the results in Table 5, We can find that the SVM classifier with RBF kernel can result in the lowest risk as well as a decent classification accuracy.

According to [1], the same dataset was used to trained the breast cancer classifier and the evaluation method for the classifier was 10-fold cross-validation either. In Wolberg, Street and Mangasarian's research, the classifier was trained basing on MSM-Tree method. The highest accuracy achieved by their classifier was 97.3%. Compared with their classifier, the highest classification accuracy achieved by my classifier is 98.07%, which improves the classification accuracy a little bit.

5 Conclusion & Future work

From the result and discussion above, we see the fact that the hybrid of GA feature selection and SVM can generate optimal classifier according to people's purpose by setting an appropriate fitness value. In the task of diagnosing breast cancer, we can generate a classifier that could achieve best classification accuracy. Besides, we can also train a classifier

which would make as less as possible false negative mistakes. These two cases show the feasibility of this method to train breast cancer classifiers with different preferences.

From the performance of different breast cancer classifiers showed in result section, we can find that the kernel of the SVM affects the performance significantly. Therefore, in the future, we can focus on constructing kernels ourselves to get better performance classifier. Moreover, we can combine the GA feature selection with other classification models instead of SVM and try to find the best combination.

Reference

[1] W. Wolberg, W. Street and O. Mangasarian, "machine learning techniques to diagnose breast cancer from find-needle aspirates," *canceer letters,* vol. 77, pp. 163-171, 1994.

[2] W. Wolberg, W. Street and O. Mangasarian, "Breast Cancer Wisconsin (Diagnostic) Data Set," 1995. [Online]. Available: http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29.

[3] S. Raschka, "Confusion Matrix," [Online]. Available: https://rasbt.github.io/mlxtend/user_guide/evaluate/confusion_matrix/.

[4] N. Sharma and G. Tom, "Hybrid Genetic Algorithms for Stress Recognition in Reading," *Evolutionary Computation, Machine Learning and Data Mining in Bioinformatics,* pp. 117-128, 2013.

[5] S. W. Noble, "What is a support vector machine?," *Nature Biotechnology,* vol. 24, no. 12, pp. 1565-1567, 2006.

[6] F. Gómez, A. Quesada and Artelnics, "Genetic algorithms for feature selection in Data Analytics," [Online]. Available: https://www.neuraldesigner.com/blog/genetic_algorithms_for_feature_selection.

[7] T. Gedeon, "Genetic Algorithms for Feature Selection," [Online]. Available: https://wattlecourses.anu.edu.au/pluginfile.php/1762198/mod_resource/content/0/IN1-%20GAs%20for%20Feature%20Selection.pdf.

[8] scikit learn, "Support Vector Machines documentation," [Online]. Available: http://scikit-learn.org/stable/modules/svm.html#kernel-functions.