RESEARCH ARTICLE OPEN ACCESS

IDENTIFICATION WITH EXPRESSION LEVELS RELATED TO BREAST CANCER PROGNOSIS USING DATAMINING TECHNIQUES

[1]A. RAMYA, [2]G. SILAMBARASAN

[1]Research Scholar, Department of Computer Science, AnnaiVailankanni Arts and Science College, Thanjavur- 613007
[2]Assistant Professor, Department of Computer Science, AnnaiVailankanni Arts and Science College, Thanjavur- 613007

ABSTRACT

Providing clinical predictions for cancer patients by analyzing their genetic make-up is a difficult and very important issue. With the target of identifying genes more interrelated with the diagnosis of breast cancer, we used data mining techniques to study the gene expression values of breast cancer patients with known clinical outcome. Focus of our work was the creation of a arrangement model to be used in the clinical practice to support therapy prescription. We randomly subdivided a gene expression dataset of 112 samples into a training set to learn the model and a test set to validate the model and assess its performance. We evaluated several learning algorithms in their not weighted and weighted form, which we distinct to take into account the different clinical importance of false positive and false negative classifications. To develop in used the results, these final, especially when used in their combined form, appear to provide better executions of the results.

Key words: Data Mining, Gene expression; Breast Cancer Prognosis

1. INTRODUCTION

This paper was to understanding what portions of the genome are occupied in the development of cancer cells is a difficult and currently very important issue in medicine. as long as clinical predictions for cancer patients by analyzing their hereditary make-up is a central goal of many research groups. In this respect, our donation here illustrate regarded the use of

In the work was to appreciate which genes are more closely related to the classification of metastasis recovery patients information's. A gene expression dataset of 96 samples was obtained by merging two published works of breast cancer microarray analysis. It was then randomly subdivided into a 39-sample training set and a 57-sample test set. The initial step has been to reduce the datasets to study through a process of reduction of the unnecessary or redundant features for classification **(features selection phase).** This paper purpose we broken the potential of different data mining techniques, implemented in available software tools such as WEKA(Waikato Environment for Knowledge Analysis) and Yale(Yet another Learning Environment). After early analyses, we obtained reduced data sets(data samples with a smaller number of genes) and we confirmed whether the achieved data reduction increased the ability in prediction of metastasis. With the reduced datasets we were able to create great classification models by using five classification algorithms known in the literature, which represent a wide range of calculation technique.

The main aim of our work was the identification of genes with levels of expression associated with a clinical prognosis for breast cancer patients.

2. DATA MINING TECHNIQUES

In this work the advantage taken to several families of data mining techniques, including feature selection and classification methods, such as decision trees and bagging, bootstrapping and random forest ensemble algorithms.

2.1 Feature selection algorithms

We seen for patients in a collection of information with hundreds of features is a multifaceted challenge because of the idleness and noise in the raw guidance data. In our work we used a class of purpose made algorithms, known as feature selection algorithms. Using such methods let us increase the prediction accuracy as well as to get a greater firmness and a better perceptive of the examined concepts.

2.2 Decision Trees

From a mathematical portion a decision tree is a connected graph not containing closed loops. In machine learning it becomes a forecast model with outstanding properties, able to manage a vast deal of data. For our analyses we used various learning algorithms: single algorithms as decision trees and NaiveBayes, and ensemble technique as *AdaBoost M1, Bagging and Random Forests.*

2.3 Ensemble methods

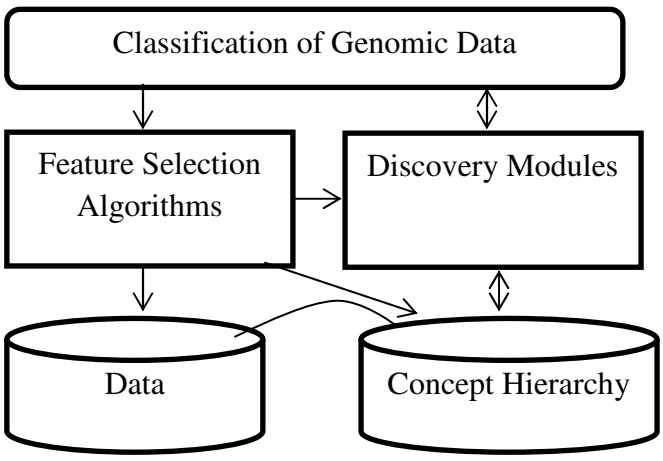
Ensemble methods (also known as Committee methods or model combiners) are aggregates of classifiers whose single predictions are combined with vote or weighted average approaches in order to build a unique classifier. Typically the classifiers composing one ensemble predictor are all of a single family, but ensemble predictors consisting of classifiers of different types were built as well. In this work three different ensemble methods were used: *bagging , boosting and random forests*.

2.4 Classification algorithms

For our analyses we used various learning algorithms: single algorithms as decision trees(J48) and Naïve Bayes, and ensemble techniques as AdaBoost M1, Bagging and Random Forests; we also tested other techniques, as Neural Network, RVM Learner, Pso SVM and Evo SVM, In order to build our classification model to produce a valid prognosis for unclassified breast cancer patients. All our analyses have been run with in the WEKA software environment.

2.5 DB Miner

DBMiner, a data mining system for interactive mining of multiple point familiarity in big relational databases, has been developed based on our years-of-research. The system implements a wide spectrum of data mining functions, including simplification, classification, inequity, association, classification and prediction. By incorporation of several interesting data mining techniques, including attribute-oriented induction, progressive deepening for mining multiple-level rules, and meta-rule guided knowledge mining, the system provides a user-friendly, interactive data mining environment with nice presentation.



3. GENE SELECTION IN THE TRAINING SAMPLE SET

The onset and development of a complex disease, such as breast cancer, cannot be attributed to a single gene. Generally, more DNA portions are involved and related to the possibility of an individual to develop a pathology. Aiming at identifying the genes that are more correlated with the prognosis of breast cancer, we considered as first case study a group of 166 genes selected as differentially expressed in the 39-sample training set of breast cancer patients with different clinical outcomes. The time of survival without metastasis after surgery was considered for clinical classification.

Thus, the class attribute was set to Class 0 if the patient lived more than five years without metastasis(from the day the disease was first diagnosed), while it was set to Class 1 for patients who developed metastasis within five years. The genes selected as differentially expressed in the considered 39 patients(24 of Class 0,15 of Class 1) were used as classification attributes.

3.1 Features Selection

We extracted the most important attributes for the forecast of the class attribute; by apply nine feature selection algorithms, so as to obtain nine reduced datasets. To identify which of the attributes (genes) selected by the nine algorithms better describe the starting data set, we sorted in descending order each of the selected genes according to the number of algorithms selected it. Then we extracted the ten most selected genes. it was given the name of " The best >=2".

Selected Genes were characterized by a greater accuracy compared to those belonging to the original dataset and were used to build a prediction model for each of the five considered learning algorithms implemented in WEKA, i.e., Adaboost, Bagging, J48, NaiveBayes and Random Forests.

TABLE : 1 THE BEST >=2 REDUCED GENE SET

SELECTED GENE
Gene 483
Gene 510
Gene 202
Gene 322
Gene 515
Gene 286
Gene 453
Gene 159
Gene 505
Gene 518

4. TRAINING DATA SET AND PERFORMANCE EVALUATION

The expression data of the ten genes selected in the feature extraction step for the considered 39 patients constituted our training dataset. These data were used to create the model to be used for the class attribute prediction of the test dataset. To get more comprehensive and comparative accuracy of the results, we did not use only a single learning algorithm to create the model, but we used five algorithms among those most suited to the specific structure of data(i.e. Adaboost, Bagging, J48, Naïve Bayes and Random Forest).

The evaluation of the performance of the different algorithms was derived mainly from two software tools; the buffer output of WEKA and the Performance Vector of YaLE. The latter one, taking in input the training dataset and a learning algorithm, produces a range of statistical measures to access the quality of the learning performance. Each classification was evaluated by using the final confusion matrix of the classification results.

The arrangement values were two(0 and 1), the resulting 2 x 2 Confusion Matrix reports on the main crosswise the number of instances classified correctly(i.e., the true negatives(TN) and the true positives(TP)), and on secondary diagonal the

number of misclassified instances(i.e., the false positives(FP) and the false negative(FN)).

By observing the distribution of TN,TP,FN and FP values within the matrix, it is possible to derive estimates of the performance of the considered classification algorithms to be used for comparison purpose. quite a few important measures can be extracted from the perplexity Matrix of the classification results to calculate the obtained tagging quality:

- Accuracy, as the percentage of instances classified correctly out of the total instances.
- Recall, as the percentage of positive instances classified correctly out of all positive instances:
- Precision as the percentage of instances correctly classified positive out of all instances classified positive:
- F-measure, as the harmonic mean of precision and recall:

For our analysis and its diagnostic implications, the two types of errors FN and FP should be considered differently. In fact, the FP error indicates patients classified as 1 (metastasis within five years) when their true classification is 0 (no metastasis within five years); while the FN error indicates patients classified as 0(no metastasis within five years) when their true classification is 1(metastasis within five years).

= = = Cost Matrix = = =

a b ← classified as

0 1 | a = 0

FN 0 | b = 1

Cost Matrix structure

Since the clinical and therapeutic importance of the correct prognosis of metastasis, the two types of errors have a totally different practical aspect.

Classify a patient as FP means to predict the patient developing a metastasis within five years when she will not; thus it means to provide the patient with an unnecessary treatment. Classify a patient as FN means to predict that the patient will not develop a metastasis within five years when she will; thus it means not providing the patient with the treatment necessary for her health. Compared to the former, the latter case has therefore a higher cost; not treating a sick patient who will encounter a worsening of the disease, with the consequent risk of death.

= = = Confusion Matrix = = =

a b ← classified as

TN FP | a = 0

FN TP | b = 1

Confusion Matrix structure

Usually a good measure of the obtained classification quality is given by a high F-measure of class 1, with a recall of Class 1 higher than 0.5(i.e., when the number of FN is less than the number of TP). However, in our scenario this is not sufficient since the number of FN considered acceptable would result to be still too high, taking into account the high cost of misclassifying a patient that develops a metastasis within five years.

Thus, in evaluating the classification performance, we consideredthe important classification difference of our considered scenario by weighting the cost of a FN classification error more than that of a FP error. For this aim, we decided to use the cost matrix. Weighting more than 1 the FN, in order to unbalance the classification and obtain very low FN.

5. CLASSIFIER VALIDATION IN THE TEST DATASET

To test the classification model built on the training dataset, we considered the expression values of the differently expressed genes in the 57-sample test set, with the clinical outcomes classified as in the

training set(i.e., presence(Class 1, 38 patients) or not (Class 0, 19 patients) of metastasis within five years from surgery). The reduced set of genes used for the classification of the patient in the test setwas composed of the same 10 genes of the best>=2 reduced gene set, selected from the training dataset through the feature selection process previously described. Thus, in order to test the classification models built with the five learning algorithms considered (i.e., Adaboost, Bagging, J48, NaiveBayes and Random Forests), the expression values of such genes were used as input of each of these classification models defined on the training dataset as previously described.

5.1. Not weighted vs. weighted classification Analysis

Having defined, for each considered learning algorithm, a not weighted model, a weighted one, and a method to calculate a class probability, we were able to generate different results and analyze them. Initially, we produced a set of classifications by applying the learning algorithms without considering the different seriousness of a FP error with respectto a FN one. Obtained results show high accuracy, but an inadequate value of the recall index, because they do not give the correct significance to a FN error. Then, we emphasized the importance of an error type(FN) with respect to the other one(FP).

The used heuristic methods to look for different weights for the prediction errors obtained, thus trying to voluntarily produce a highly unbalanced prediction. A notable outcome was the bad performance of the Adaboost algorithm in the weighted classification. It resulted the less efficient classifiers for the question we faced in our study; it needed a heavy displacement weight in order to decrease the false negative classified patients, and even so it did not provide the desired results. The weighted average results of the unbalanced classification, either with or without the Adaboost predictor is presented here.

Not weighted Adaboost				
Confusion Matrix:				
Classified as				
A	B			
108	23	a		
49	17	b	real label	
Accuracy				63.45%
Recall(Class 1)				0.2576
Precision(Class1)				0.425
F-measure(Class 1)				0.328

Not weighted Bagging				
Confusion Matrix:				
Classified as				
A	b			
110	21	a		
44	22	b	real label	
Accuracy				67.01%
Recall(Class 1)				0.333
Precision(Class1)				0.5116
F-measure(Class 1)				0.4037

Based on obtained results, we can assert that a weighted unbalanced classification is significantly better than a balanced one, because it provides a reduced number of false negative prognoses. Although it generates an increased number of false positive parents, this latter error is less important from the clinical point of view.

5.2. Alternative Classifier

We used a neural network implementation in WEKA and the following four algorithms implemented in the YaLE software: RVM Learner, PsoSVM,Evo SVM and Perceptron(i.e.,a network of neurons in which the output(s) of some neurons are connected through weighted connections to the input(s) of other neurons). The Figuredisplays the results obtained with these alternative classifiers.

while the order two classifiers perform slightly better. With regard to recall and F-measure, only Pso SVM and Evo SVM have performances comparable with the previous considered weighted classifiers.

Looking for the best neural network topology, we tested networks with different numbers of neurons, hidden layers and neurons for each layer, obtaining three structures: the first was a Perceptron, the second had one hidden layer containing 15 neurons and the last had two hidden layers having 30 and 12 neurons, respectively.

We could observe a good adaption of this family of algorithms to gene expression numerical data, but it was not enough to provide accurate prognosis for cancer patients because of the low recall and precision values, in spite of a high accuracy.

Adaboost Weight 200

Confusion Matrix:

Classified as

a	B	
85	46	a
24	42	b real label

Accuracy	64.07%
Recall(Class 1)	0.6364
Precision(Class1)	0.4773
F-measure(Class 1)	0.5455

Bagging Weight 12.5

Confusion Matrix:

Classified as

a	b	
62	69	a
18	48	b real label

Accuracy	55.84%
Recall(Class 1)	0.7273
Precision(Class1)	0.4103
F-measure(Class 1)	0.5246

NaiveBayesWeight 12

Confusion Matrix:

Classified as

a	b	
71	60	a
15	51	b real label

Accuracy	61.93%
Recall(Class 1)	0.7727
Precision(Class1)	0.4595
F-measure(Class 1)	0.5763

Random Forest Weight 12.8

Confusion Matrix:

Classified as

a	b	
72	59	a
20	46	b real label

Accuracy	59.90%
Recall(Class 1)	0.697
Precision(Class1)	0.4381
F-measure(Class 1)	0.538

5.3. Nearest Mean Classifier

We also considered the Nearest Mean Classifier (NMC) algorithm. It bases its classification on the genes with most different expression values in the two considered classes, as identified by their single to noise ratio(SNR) index calculated:

Feature selection phase of the NMC algorithm selects such genes performing the following steps:

- For each gene, calculation of its SNR index
- Ordering genes based on their SNR
- Selection of the genes

6. DISCUSSION AND CONCLUSIONS

In this paper, we applied several data-mining techniques in a biomedical scenario. A preliminary analysis, aimed at analyzing the distribution of data in the considered dataset, guaranteed applicability of each technique to the dataset considered. With nine algorithms of feature selection we extracted a group of subsamples of data, which was analyzed with different classification algorithms for comparison purpose. In our tests w used five learning algorithms, implemented in YaLE or WEKA. The latter was used for the opportunity to weight the classification in order to unbalance the prediction of class to the number of incorrectly classified patients predicted with metastasis within five years from surgery. This was made in order to decrease the occurrences of incorrectly classified patients predicted without metastasis within five years. Such points is very important in our study. In fact, due to the diagnostic and therapeutic consequences of the two classification, classifying a patient as " good prognosis" when she is in a state that will develop metastasis (i.e., a FN error) is much more serious than classifying a patient as " poor prognosis".

To identify additional methodologies to further improve classifications of our data, we took into account also other classifiers, specifically suited to numerical data, e.g., neural networks and support vector machine (SVM) classifiers achieved good results, but less satisfactory than the considered weighted classifiers. Of great importance was the Nearest Mean Classifier , a technique based on the distribution of k-means clustering, which assigns Class 0 to the majority class. With this classifier we obtained a classification with a high value of false positives, but a low value of false negative. The algorithm classified ill patients more accurately(lower FN and higher TP) at the expense of the classification of healthy patients(higher FP and lower TN), which was a major goal for our analysis.

7. REFERENCES

1.Data Mining Concepts by Arun K. Pujari

2.I.H. Witten, and E. Frank, " Data Mining: Practical Machine Learning" ,kaufmann, San Fransisco, 2005.

3. WEKA,http://www.cs.waikato.ac.nz/ml/weka

4. WEKAWiki, http://weka.wiki.sourceforge.net/

5. YaLE:Yet another Learning Environment, http://rapid-i.com/

6.https://pdfs.semanticscholar.org/4945/4263b6a75a87dbeb94dbe0ba418dba16f459.pdf Diagnosis of Breast Cancer using Clustering Data Mining Approach

7. Breast cancer diagnosis using machine learning algorithms –a survey, International Journal of Distributed and Parallel Systems (IJDPS) Vol.4, No.3, May 2013

8. N.Sureshbabu, S.Sivakumar , "A Study on Genes Identification with Expression Levels Related to Breast Cancer Prognosis Using Data Mining Techniques" International Journal of Computer Techniques – Volume 4 Issue 4, July – August 2017

9.Gabriele Giarratana ; Marco Pizzera ; Marco Masseroli ; Enzo Medico ; Pier Luca Lanzi "Data Mining Techniques for the Identification of Genes with Expression Levels Related to Breast Cancer Prognosis" 2009 Ninth IEEE International Conference on Bioinformatics and Bio Engineering DOI: 10.1109/BIBE.2009.37 , Print ISBN: 978-0-7695-3656-9