International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:10, No:5, 2016

Breast Cancer Survivability Prediction via Classifier Ensemble

Mohamed Al-Badrashiny, Abdelghani Bellaachia

Abstract—This paper presents a classifier ensemble approach for predicting the survivability of the breast cancer patients using the latest database version of the Surveillance, Epidemiology, and End Results (SEER) Program of the National Cancer Institute. The system consists of two main components; features selection and classifier ensemble components. The features selection component divides the features in SEER database into four groups. After that it tries to find the most important features among the four groups that maximizes the weighted average F-score of a certain classification algorithm. The ensemble component uses three different classifiers, each of which models different set of features from SEER through the features selection module. On top of them, another classifier is used to give the final decision based on the output decisions and confidence scores from each of the underlying classifiers. Different classification algorithms have been examined; the best setup found is by using the decision tree, Bayesian network, and Naïve Bayes algorithms for the underlying classifiers and Naïve Bayes for the classifier ensemble step. The system outperforms all published systems to date when evaluated against the exact same data of SEER (period of 1973-2002). It gives 87.39% weighted average F-score compared to 85.82% and 81.34% of the other published systems. By increasing the data size to cover the whole database (period of 1973-2014), the overall weighted average F-score jumps to 92.4% on the held out unseen test set.

Keywords—Classifier ensemble, breast cancer survivability, data mining, SEER.

I. Introduction

BREAST cancer is the leading type of cancer in women, accounting for 25% of all cases [1]. Survival rates in the developed world are high [2], with between 80% and 90% of the cases in England and the United States survive for at least 5 years [3]. However, it is poorer in the developing countries [1]. Early detection of the cases that have a low survivability probability, helps the doctors to to take the required precautions to help them.

This paper addresses the problem of predicting the survivability rate of the breast cancer patients using a classifier ensemble approach. The performance of the proposed approach is then examined against the state of the art systems using the same dataset.

II. Related Work

Survivability prediction of the patients has been studied by many researchers. Zhou et al. [4] presented a C4.5 rule preceded by artificial neural network ensemble for medical diagnosis. They first trained an artificial neural network. Then, a new training data set is generated by feeding the feature vectors of original training instances to the trained classifier and replacing the expected class labels of original training instances with the new class labels. Finally, a specific rule induction approach is used to learn rules from the new training data set. They showed that their system is able to give 97% 10-folds cross validation accuracy on the breast cancer patients. However, their dataset was very small. It was only 683 case for both training and testing.

Lundin M et al. [5] applied the artificial neural networks to survival prediction in breast cancer patients. They used 651 cases for training and 300 for testing. The best accuracy they got is 90.9% for 5 years of survivability prediction.

Delen et al. [6] made a comparison of three data mining methods for breast cancer survivability; artificial neural networks, decision trees, and logistic regression. They used SEER data (period of 1973-2000) [7] with 202,932 records after removing redundancies and missing information; 93,273 are tagged as "survived" and 109,659 as "not survived". The "survived" class is all records that have a value greater than or equal 60 months in the survival time recode (STR) field while the "not survived" class represent the remaining records. The reported accuracy was 93.6%, 91.2%, and 89.2% 10-folds cross validation on decision trees, ANN, and logistic regression receptively.

Bellaachia et al. [8] took the study of Delen et al. [6] as the starting point. They found that the pre-classification process of Delen et al. [6] was not accurate in determining the records of the "not survived" class. They did not take into consideration either the vita status recode (VSR) filed that show the status of the patient (alive or not) or the cause of death (COD). This reduced the best results of Delen et al. [6] to 81.3%. Therefore, Bellaachia et al. [8] defined a new pre-classification approach to fix the issue of Delen et al. [6]. They used the STR, VSR, and COD to define the "survived" and "not survived" classes and filter out any unrelated records. If the STR is greater than or equals to 60 months and the VSR is alive then the record is pre-classified as "survived". But if the STR is lower then 60 months and the COD is breast cancer, then the record is pre-classified as "not survived". Otherwise the record is being ignored. Bellaachia et al. [8] applied this pre-classification approach on a newer version of SEER data (period of 1973-2002) [9]. They found that 116,738 records can be tagged as "survived" and 35,148 as "not survived". They compared three classification methods; Naïve Bayes, the back-propagated neural network, and the C4.5 decision tree. The reported accuracy was 86.7%, 86.5%, and 84% 10-folds cross validation on the decision trees, ANN, and Naïve Bayes classifiers respectively.

Mohamed Al-Badrashiny and Abdelghani Bellaachia are with the Department of Computer Science, The George Washington University, Washington DC, 20052 USA (e-mail: badrashiny@gwu.edu, bell@gwu.edu).

III. Approach

We present a classifier ensemble system that uses the output from three different classifiers (Decision tree, Bayesian network, and Naïve Bayes) to predict wither the patient of the input feature vector will survive or not. The underlying classifiers are trained on gold labeled data with binary decisions of either being "survived" or "not survived". The system consists of two main components; features selection and classifier ensemble components. The features selection component is responsible for finding the most important features among a group of features that maximizes the weighted average F-score of a certain classification algorithm. This component is used with our three classifiers to maximize the performance of each one of them independently. The classifier ensemble component is then uses the output decisions and the confidences scores generated by the three classifiers to train another classifier that is used after that to produce the final output of the system. Each of the two components is described below:

A. Features Selection Component

This component is intended to find the best features from the whole available set of features that maximize the weighted average F-score of the whole system. In this paper we are using the latest version of SEER data (period of 1973-2014) [10]. The data set provides a lot of features that covers many aspects of different kinds on cancer. The same features that were used by Bellaachia et al. [8] are being used here in addition to two extra features (Standard Survival Age and Sex). In this component a heuristic method for features selection [11] is used. Fig. 1 shows that all the features are split into four different groups sets. Then two stages feature exploration are perform, where the features selection component exhaustively searches over all features in each group in the first phase, and then exhaustively searchs over all retained feature to find the best combination of features that maximizes the weighted average F-score. The features are listed below:

- **Patient-Info-Features-Group**: This group contains the basic information about the patient; Sex, Race, Marital Status, and the standard survival age of similar cases (StSurvAge).
- **Diagnose-Features-Group**: All features that are related to the diagnose are included in this group; primary site of the cancer, histologic type, behavior code of the cancer, and the age at diagnosis.
- **Tumor-Features-Group**: This group contains the tumor related features; grade of the tumor, tumor extension, tumor size, the number of nodes, the number of positive nodes, the number of primaries, and the stage of the cancer.
- **Therapy-Features-Group**: This group contains information about the methods that are used for treatment; radiation status, and the site surgery code.

The features selection component is run using decision tree, Bayesian network, and Naïve Bayes classifiers. This gives three different classifiers; each of them is trained using the best selected features that maximize its performance.

B. Classifier Ensemble Component

The system is based on the idea of classifier ensemble. It has been shown that such systems are less sensitive to inaccuracies in the data and to their internal parameter tuning settings [12]. Furthermore, these systems tend to be more accurate than a single classifier alone, especially if each underlying classifier has a different error distribution than the others.

Fig. 2 shows the system architecture of the classifier ensemble component. The ensemble uses the classes and the confidence scores of the preceding three classifiers on the training data to train a fourth one. Accordingly an input test feature vector goes through each of the three underlying classifiers, where each classifier gives a label and a confidence score. Then, fourth classifier uses the three labels and the three confidence scores to provide its final classification for the input feature vector.

IV. Experimental Setup

A. Datasets

The latest version of SEER data (period of 1973-2014) [10] is used in this paper after applying the same pre-processing and pre-classification approach that is used by Bellaachia et al. [8]. The data size after the pre-processing step is 639,532 records; 577,518 of them are tagged as "survived" and 62,014 as "not survived". This means that 90.3% of the data are "survived" and 9.7% are "not survived". We split the data into:

1) **Training dataset** (*TrDB*): Represents 80% of the data and used for system training;
2) **Development dataset** (*DevDB*): Represents 10% of the data and used for system tuning;
3) **Held out test dataset** (*TestDB*): Represents 10% of the data and used for system testing.

The classes distribution are being preserved in the three datasets (i.e. 90.3% of each dataset are "survived" and 9.7% are "not survived"). Table I shows the statistics of the data.

TABLE I
Classes Distribution over the *TrDB*, *DevDB*, and *TestDB*

	Survived	Not survived	Total
TrDB	462,000	49,628	511,628
DevDB	57,759	6,193	63,952
TestDB	57,759	6,193	63,952

Table I shows that there is a minority class problem in the dataset because the "not survived" class represents only 9.7% of the data. Therefor, using the accuracy measure is not a good choice in our case. Instead, the weighted average F-score is used as the main metric in this paper.

B. Baselines

In this paper, three main sets of experiments are conducted; features selection experiments, ensemble experiments, and experiments to compare the overall system performance against the other published approaches. For each set of experiments, some baselines are defined to compare the presented approach with.

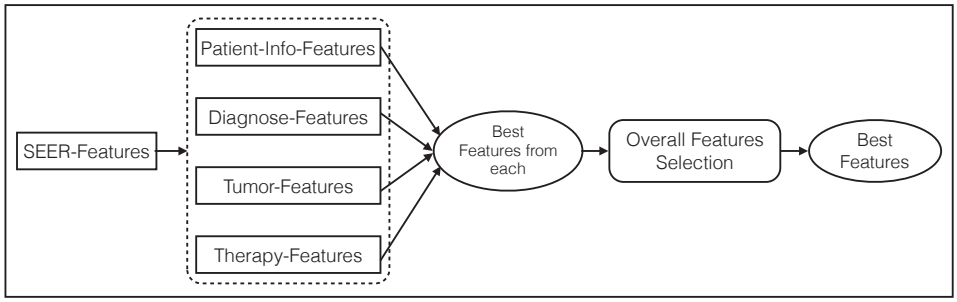

Fig. 1 Features selection architecture

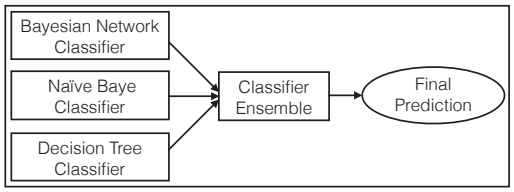

Fig. 2 Classifier ensemble architecture

1) Features Selection Baselines: In this set of experiments, the heuristic features selection method that is described in section III above is being examined using three different classification algorithms. The following six baselines are used to measure performance of this step:

- **All-Feat-DT-BL**: All features of SEER data [10] that are mentioned in section III above. They are put together in a single feature vector to train a decision Tree classifier.
- **All-Feat-NB-BL**: Same as (All-Feat-DT-BL) but using Naïve Bayes classifier.
- **All-Feat-BN-BL**: Same as (All-Feat-DT-BL) but using Bayesian network classifier.
- **WEKA-FS-DT-BL**: This baseline uses the automatic feature selection in *WEKA* [13] to select the best features from the training data instead of using all of them. *WEKA* selected (Sex-Stage-Race-HistologicType-Grade-PositiveNodes) as the best features. Then the selected features are used to train a decision Tree classifier.
- **WEKA-FS-NB-BL**: Same as (WEKA-FS-DT-BL) but using Naïve Bayes classifier.
- **WEKA-FS-BN-BL**: Same as (WEKA-FS-DT-BL) but using Bayesian network classifier.

2) Ensemble Experiments Baselines: Different classifiers are experimented in this step to combine the decisions of the underlying classifiers from the features selection step. One baseline is defined for this set of experiments:

- **Maj-BL** is the majority baseline. In this baseline, all records are assigned to the label of the most frequent class observed in the training data. In our case it is the "survived" class.

3) Overall System Performance Baselines: To be able to compare our approach with Bellaachia et al. [8], we removed all records that have the date field after 2002 to have the same data that have been used by Bellaachia et al. [8]. Furthermore, we used their evaluation metric by calculating the 10-folds cross validation accuracy on the whole data set. However, We also reports the 10-folds weighted average F-score because the reduced data still has the minority class problem (76.8% for the "survived" class and 23.8% for the "not survived" class). The baselines that are used in this step are:

- **Maj-BL**: Is the majority baseline. In this baseline, all records are assigned to the label of the most frequent class observed in reduced data. In our case it is the "survived" class.
- **Delen et al-BL**: The system performance of Delen et al. [6] after the correction that Bellaachia et al. [8] have made.
- **Bellaachia et al-BL**: The system performance of Bellaachia et al. [8].

V. Evaluation

A. Features Selection Experiments

The heuristic features selection method that is described in section III above has been used to select the best features the maximize the weighted average F-score on the *DevDB* using decision tree, Naïve Bayes, Bayesian network classifiers receptively. Table II compares the heuristic features selection method with the automatic features selection method by *WEKA* [13] and all features without any features selection using the same three classifiers. The bottom three rows in the table represents the results of the heuristic features selection method. We can see that each of them is better than its *WEKA* rival and also better than using the full features set. For example, the heuristic features selection method using Bayesian network classifier (Heuristic-FS-BN) gives 92.41% weighted average F-score, while its *WEKA* opponent (WEKA-FS-BN-BL) gives 90.62% and results without any features selection using the same classifier (All-Feat-BN-BL) is 92.28%.

B. Ensemble Experiments

The output decisions and confidence scores from the Heuristic-FS-DT, Heuristic-FS-BN, and Heuristic-FS-BN from the previous set of experiments are combined using an ensemble classifier. We experimented using different classification algorithms; Decision tree, Bayesian network, neural network, and Naïve Bayes classifiers. The weighted average F-score is calculated for each experiment on the *DevDB* and *TestDB* and then has been compared with the majority baseline. Table III

TABLE II
BEST SELECTED FEATURES AND THE WEIGHTED AVERAGE F-SCORE ON THE *DevDB*

Exp-Name	Weighted-Avg-Fscore	Best-Selected-Features
All-Feat-DT-BL	91.98%	All features
All-Feat-NB-BL	88.85%	All features
All-Feat-BN-BL	92.28%	All features
WEKA-FS-DT-BL	90.54%	Sex-Stage-Race-HistologicType-Grade-PositiveNodes
WEKA-FS-NB-BL	88.45%	Sex-Stage-Race-HistologicType-Grade-PositiveNodes
WEKA-FS-BN-BL	90.62%	Sex-Stage-Race-HistologicType-Grade-PositiveNodes
Heuristic-FS-DT	92.24%	StSurvAge-PrimarySite-Grade-Extension-TumorSze-NodesNum-PositiveNodes-PrimariesNum-Radiation-SiteSurgeryCode
Heuristic-FS-NB	89.38%	PrimarySite-HistologicType-Grade-Extension-PositiveNodes
Heuristic-FS-BN	92.41%	Sex-MaritalStatus-StSurvAge-HistologicType-AgeDiagnosis-Grade-Extension-TumorSze-NodesNum-PositiveNodes-PrimariesNum-Stage-SiteSurgeryCode

shows that combining the three underlaying classifiers using Naïve Bayes classifier (Ensemble-NB) gives the best results. It's better than using any of the underlying classifiers alone. The table shows that the weighted average F-score on the *DevDB* is 92.48% and on the heldout unseen test data *TestDB* is 92.42%.

TABLE III
CLASSIFIER ENSEMBLE RESULTS ON THE *DevDB* AND *TestDB* USING DIFFERENT CLASSIFICATION ALGORITHMS

Exp-Name	Dev-Weighted-Avg-Fscore	Test-Weighted-Avg-Fscore
Maj-BL	85.72%	85.72%
Ensemble-DT	92.41%	92.28%
Ensemble-BN	92.44%	92.27%
Ensemble-NT	92.46%	92.36%
Ensemble-NB	**92.48%**	**92.42%**

C. Overall System Performance Experiments

The presented approach is compared to best published systems using the same data (SEER data on the period of 1973-2002) and the same evaluation metrics (10-fold cross validation accuracy and the 10-folds cross validation weighted average F-score). Table IV shows that the presented approach outperforms all baselines using all evaluation metrics. However, we are more interested in the weighted average F-score more than the accuracy because of the minority class problem that is described in section IV above. The table shows that presented approach gives 87.65% weighted average F-score while the best baseline (Bellaachia et al-BL) gives 85.82%.

TABLE IV
PERFORMANCE OF OUR BEST SETUP (ENSEMBLE-BN) AGAINST ALL BASELINE USING THE 10-FOLD CROSS VALIDATION ACCURACY AND THE WEIGHTED AVERAGE F-SCORE ON THE SEER DATA (PERIOD OF 1973-2002)

Exp-Name	Accuracy	Weighted-Avg-Fscore
Maj-BL	76.86%	66.80%
Delen et al-BL	81.30%	81.34%
Bellaachia et al-BL	86.70%	85.82%
Our approach	**87.65%**	**87.39%**

VI. CONCLUSION AND FUTURE WORK

The paper presented a classifier ensemble system for breast cancer survivability prediction. The system performs a heuristic features selection approach to train three different classifiers by maximizing the weighted average F-score of each of them on the development set. The classifier ensemble uses the predictions and the confidence scores of the three underlying classifiers to predict the final label of any new input feature vector. The results show that the classifier ensemble yields higher performance than using either of the three classifiers separately. The results outperform all published systems on the same data set. We plan on exploring the deep learning approach for the classification part in addition to using log-linear models and feature-rich neural networks to perform automatic selection and tuning of the features.

REFERENCES

[1] "World health organization," in *World Cancer Report*, 2014, pp. Chapter 1.1, ISBN 92–832–0429–8.
[2] "International agency for research on cancer," in *World Cancer Report*, 2008.
[3] "Breast cancer. nci," in *SEER Stat Fact Sheets*, 2014.
[4] Z.-H. Zhou and Y. Jiang, "Medical diagnosis with c4.5 rule preceded by artificial neural network ensemble," *Information Technology in Biomedicine, IEEE Transactions on*, vol. 7, no. 1, pp. 37–42, March 2003.
[5] M. Lundin, J. Lundin, H. B. Burke, S. Toikkanen, L. Pylkkänen, and H. Joensuu, "Artificial neural networks applied to survival prediction in breast cancer," *Oncology*, vol. 57, no. 4, pp. 281–286, 1999. [Online]. Available: http://www.karger.com/DOI/10.1159/000012061
[6] D. Delen, G. Walker, and A. Kadam, "Predicting breast cancer survivability: a comparison of three data mining methods," *Artificial Intelligence in Medicine*, vol. 34, no. 2, pp. 113–127, Jun 2005. [Online]. Available: http://www.aiimjournal.com/article/S0933-3657(04)00101-0/abstract
[7] "Seer cancer statistics review. surveillance, epidemiology, and end results (seer) program (www.seer.cancer.gov) public-use data (1973-2000). national cancer institute, surveillance research program, cancer statistics branch, released april 2003. based on the november 2002 submission. diagnosis period 1973-2000, registries 1-9."
[8] A. Bellaachia and E. Guven, "Predicting breast cancer survivability using data mining techniques," in *Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth SIAM International Conference on Data Mining (SDM 2006)*, April 22 2006.
[9] "Seer cancer statistics review. surveillance, epidemiology, and end results (seer) program (www.seer.cancer.gov) public-use data (1973-2002). national cancer institute, surveillance research program, cancer statistics branch, released april 2005. based on the november 2004 submission."

[10] "Surveillance, epidemiology, and end results (seer) program (www.seer.cancer.gov) research data (1973-2011), national cancer institute, dccps, surveillance research program, surveillance systems branch, released april 2014, based on the november 2013 submission."
[11] R. Eskander, M. Al-Badrashiny, N. Habash, and O. Rambow, "Foreign words and the automatic processing of arabic social media text written in roman script," *In Proceedings of the First Workshop on Computational Approaches to Code-Switching. EMNLP 2014, Conference on Empirical Methods in Natural Language Processing, October, 2014, Doha, Qatar*, 2014.
[12] J. Kittler and F. Roli, Eds., *Multiple Classifier Systems, First International Workshop, MCS 2000, Cagliari, Italy, June 21-23, 2000, Proceedings*, ser. Lecture Notes in Computer Science, vol. 1857. Springer, 2000.
[13] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The weka data mining software: an update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, 2009.