Predicting Breast Cancer Survivability Using Data Mining Techniques

Abdelghani Bellaachia, Erhan Guven

Department of Computer Science
The George Washington University
Washington DC 20052
{bell, eguven}@gwu.edu

Abstract

In this paper we present an analysis of the prediction of survivability rate of breast cancer patients using data mining techniques. The data used is the SEER Public-Use Data. The preprocessed data set consists of 151,886 records, which have all the available 16 fields from the SEER database. We have investigated three data mining techniques: the Naïve Bayes, the back-propagated neural network, and the C4.5 decision tree algorithms. Several experiments were conducted using these algorithms. The achieved prediction performances are comparable to existing techniques. However, we found out that C4.5 algorithm has a much better performance than the other two techniques.

Keywords: Breast cancer survivability, data mining, SEER, Weka.

1. Introduction

Today, in the United States, approximately one in eight women over their lifetime has a risk of developing breast cancer. An analysis of the most recent data has shown that the survival rate is 88% after 5 years of diagnosis and 80% after 10 years of diagnosis [1].

The discovery of the survival rate or survivability of a certain disease is possible by extracting the knowledge from the data related to that disease. One of these data sources is SEER [2] (Surveillance Epidemiology and End Results), which is a unique, reliable and essential resource for investigating the different aspects of cancer. The SEER database combines patient-level information on cancer site, tumor pathology, stage, and cause of death [3, 4].

The characteristics of a population can be observed to establish the factors associated with a specific outcome. Observational studies, such as statistical learning and data mining, can establish the association of the variables to the outcome, but they do not always establish the cause-and-effect relationship of the association. Data driven statistical research is becoming a common complement to many scientific areas like medicine and biotechnology. This trend is becoming more and more visible as in the studies of Houston et al. [5] and Cios et al. [6].

In this paper, we present data mining techniques to predict the survivability rate of breast cancer patients. In our study, we have used the SEER data and have introduced a pre-classification approach that take into account three variables: Survival Time Recode (STR), Vital Status Recode (VSR), and Cause of Death (COD).

This paper is organized as follows. The next section reviews related work. Section 3 gives the methodology used to conduct the prediction analysis. Experimental results are presented in Section 4. Conclusion and future work are given in the last section.

2. Related Work

A literature survey showed that there have been several studies on the survivability prediction problem using statistical approaches and artificial neural networks. However, we could only find a few studies related to medical diagnosis and survivability using data mining approaches like decision trees [7, 8, 9].

In this work, we took the study of Delen et al. [9] as the starting point of our research. In his study, Delen et al. preprocessed the SEER data (period of 1973-2000 with 433,272 records named as breast.txt) for breast cancer to remove redundancies and missing information. The resulting data set had 202,932 records, which then pre-classified into two groups of "survived" (93,273) and "not survived" (109,659) depending on the Survival Time Recode (STR) field. The "survived" class is all records that have a value greater than or equal 60 months in the STR field and the "not survived" class represent the remaining records. After this step, the data mining algorithms are applied on these data sets to predict the dependent field from 16 predictor fields. The results of predicting the survivability were in the range of 93%

accuracy. After a careful analysis of the breast cancer data used in [9], we have noticed that the number of "not survived" patients used does not match the number of "not alive" (field VSR) patients in the first 60 months of survival time. As a matter of fact, the number of "not survived" patients is expected to be around 20% based on the breast cancer survival statistics of 80% [1]. In our discussion with the authors of [9], we found out that the pre-classification process was not accurate in determining the records of the "not survived" class. They did not take into consideration neither the Vital Status Recode (VSR), nor the Cause of Death (COD). They assume that all patients are dead with cancer, which is not always true.

In our study, we have used a newer version of SEER database (period of 1973-2002 with 482,052 records) and, unlike [9], we have included two other fields in the pre-classification process:

- Survival Time Recode (STR),
- Vital Status Recode (VSR),
- Cause of Death (COD)

The next section presents our pre-classification process.

3. Methodology

In this paper, we have investigated three data mining techniques: the Naïve Bayes, the back-propagated neural network, and the C4.5 decision tree algorithms. In this paper, we used these algorithms to predict the survivability rate of SEER breast cancer data set. We selected these three classification techniques to find the most suitable one for predicting cancer survivability rate.

The Naïve Bayes technique depends on the famous Bayesian approach following a simple, clear and fast classifier [10]. It has been called 'Naïve' due to the fact that it assumes mutually independent attributes. In practice, this is almost never true but is achievable by preprocessing the data to remove the dependent categories [10]. This method has been used in many areas to represent, utilize, and learn the probabilistic knowledge and significant results have been achieved in machine learning [10].

The second technique uses artificial neural networks. In this study, a multi-layer network with back-propagation (also known as a multi-layer perceptron) [10] is used.

The third technique is the C4.5 decision-tree generating algorithm [11]. C4.5 is based on the ID3 algorithm.

It has been shown that the last two techniques have better performance [7, 8, 9]. Therefore we have included them in our analysis.

We have used the Weka toolkit to experiment with these three data mining algorithms [12]. The Weka is an ensemble of tools for data classification, regression, clustering, association rules, and visualization. The toolkit is developed in Java and is an open source software issued under the GNU General Public License [10].

Preprocessing the input data set for a knowledge discovery goal using a data mining approach usually consumes the biggest portion of the effort devoted in the entire work [10]. We have developed a set of tools to extract and cleanup the raw SEER data.

A simple analysis shows that the SEER data has missing information in the fields of Extent of Disease (EOD) and Site Specific Surgery (SSS) fields for almost half of the records. Most of the missing information is in the records, which are gathered prior to 1988. Since we wanted to use all the available fields in the SEER database, we removed these records from the test data set. These records have Coding System for EOD coded as '4'. The SSS field usage has changed after 1998. Instead of the regular field, the information is split in five other fields. A mapping scheme from new SSS to old SSS is developed to fill the missing SSS fields. After this step, the records with missing information are removed from the data set.

The EOD field is composed of five fields including the EOD code. These fields (size of tumor, number of positive nodes, number of nodes, and number of primaries) contain missing information coded such as '999', '99' or '9' representing the 'unknown' information. Please note that, the statistics in Table 1 do not contain fields with 'unknown' values. The table also shows the fields used in our analysis.

Nominal variable name	**Number of distinct values**		
Race	19		
Marital status	6		
Primary site code	9		
Histologic type	48		
Behavior code	2		
Grade	5		
Extension of tumor	23		
Lymph node involvement	10		
Site specific surgery code	19		
Radiation	9		
Stage of cancer	5		
Numeric variable name	**Mean**	**Std. Dev.**	**Range**
Age	58	13	10-110
Tumor size	20	16	0-200
No of positive nodes	1.5	3.7	0-50
Number of nodes	15	6.8	0-95
Number of primaries	1.25	0.5	1-8

Table 1: Survivability Attributes

As stated in the previous section, we have adopted a different approach in the pre-classification process. Unlike [9], we have included three fields: STR, VSR, and COD. The STR field ranges from 0 to 180 months in the SEER database. The pre-classification process is outlined as follows.

```
// Setting the survivability dependent variable for 60
// months threshold
if STR ≥ 60 months and VSR is alive then
        the record is pre-classified as "survived"
else if STR < 60 months and COD is breast cancer, then
        the record is pre-classified as "not survived"
else
        Ignore the record
end if
```

In the above approach, the ignored records correspond to those patients that have an STR less than 60 months and are still alive, or those patients that have an STR less than 60 months but the cause of their death is not breast cancer.

Table 2 and Table 3 show the classes of our pre-classification process and the approach used in [9], respectively.

Class	No of instances	Percentage
0: not survived	35,148	23.2
1: survived	116,738	76.8
Total	151,886	100

Table 2: Proposed Survivability Class Instances

Class	No of instances	Percentage
0: not survived	162,381	58.3
1: survived	116,282	41.7
Total	278,663	100

Table 3: Survivability Class Instances based on the Previous Work (study [9])

After the preprocessing step, a common analysis would be determining the effect of the attributes on the prediction, or attribute selection. We used the information gain measure [10] to rank the attributes due to the fact that it is a common method and the C4.5 decision tree technique utilizes this measure. Information gain (IG) is measured as the amount of the entropy (H) difference when an attribute contributes the additional information about the class. The following is the information gain and the entropy before and after observing the attribute X_i for the class C:

$$\mathbf{H}(C) = -\sum p(c)\boldsymbol{log}p(c) \qquad , c \in C$$

$$\mathbf{H}(C|X_i) = -\sum p(x) \sum p(c|x)\boldsymbol{log}p(c|x) \qquad , x \in X_i, c \in C$$

$$\mathbf{IG}_i = \mathbf{H}(C) - \mathbf{H}(C|X_i)$$

Figure 1 shows the ranked survivability attributes of our data as calculated by the Weka toolkit. It clearly shows that Extension of Tumor has a higher rank than the Tumor Size.

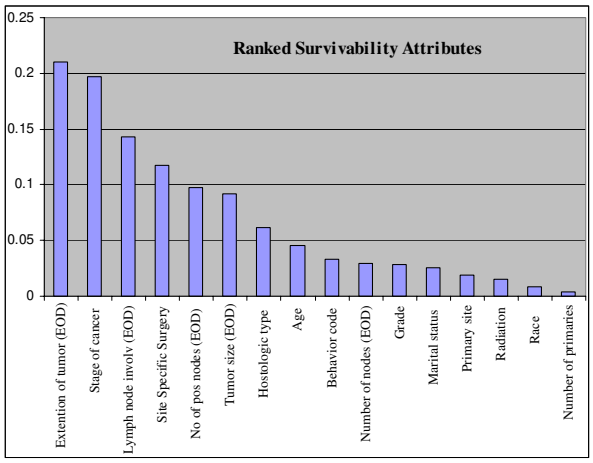

Figure 1: Ranked Survivability Attributes

We will use the performance metrics of accuracy, precision and recall to compare the three techniques. In order to have a fair measure of the performance of the classifier; we used a cross-validation with 10 folds. In its most elementary form, cross-validation consists of dividing the data into k subgroups. Each subgroup is predicted via the classification rule constructed from the remaining (k-1) subgroups, and the estimated error rate is the average error rate from these k subgroups. In this way, the error rate is estimated in an unbiased way. The final classifier rule is calculated from the entire data set. After running the classifier 10 times with 10 folds, we obtain the metrics of precision, recall, accuracy A_i and the Cross Validation Accuracy (CVA) to represent a classifier performance:

$$CVA = (1/10)\sum A_i \qquad i = 1, 2, \ldots, 10$$

A_i = # records correctly classified / total # records

The Weka toolkit can calculate all these performance metrics after running a specified k-fold cross-validation.

4. Experimental Results

In this study, the accuracy of three data mining techniques is compared. The goal is to have high accuracy, besides high precision and recall metrics. Although these metrics are used more often in the field of information retrieval, here we have considered them as they are related to the other existing metrics such as specificity and sensitivity. These metrics can be derived from the confusion

matrix and can be easily converted to true-positive (TP) and false-positive (FP) metrics [10].

The experimental results of our approach as presented in Table 4.

Classification Technique	Accuracy (%)	Class	Precision	Recall
Naïve Bayes	84.5	0	0.70	0.57
		1	0.88	0.93
Artificial Neural Net	86.5	0	0.83	0.52
		1	0.87	0.97
C4.5	86.7	0	0.80	0.56
		1	0.88	0.96

Table 4: Combined Results (our study)

Classification Technique	Accuracy (%)	Class	Precision	Recall
C4.5	81.3	0	0.86	0.81
		1	0.76	0.81

Table 5: Results for C4.5 (dataset as in Table 3)

As can be seen in Table 4, neural net and decision tree have comparable performances.

Table 5 shows the experimental results using the pre-classification approach used in [9] and the same dataset used in our approach. The results clearly show that the classification rate (81%) is much lower than the classification rate of our approach (~87%).

It may be worth noting that the computation times of the algorithms Naïve Bayes, neural net and C4.5 (on an AMD Athlon 64 4000+ machine) were in the ranges of 1 minute, 12 hours and 1 hour, respectively.

These obtained results in this work differ from the study of Delen et al. [9] due to the facts that we used a newer database (2000 vs. 2002), a different pre-classification (109,659 and 93,273 vs. 35,148 and 116,738) and different toolkits (industrial grade tools vs. Weka).

5. Conclusions and Future Work

This paper has outlined, discussed and resolved the issues, algorithms, and techniques for the problem of breast cancer survivability prediction in SEER database. Unlike the pre-classification process used in [9], our approach takes into consideration, besides the Survival Time Recode (STR), the Vital Status Recode (VSR) and Cause of Death (COD). The experimental results show that our approach outperforms the approach used in [9].

This study clearly shows that the preliminary results are promising for the application of the data mining methods into the survivability prediction problem in medical databases.

Our analysis does not include records with missing data; future work will include the missing data in the EOD field from the old EOD fields prior to 1988. This might increase the performance as the size of the data set will increase considerably.

Finally, we would like to try survival time prediction of certain cancer data such as respiratory cancer where the survivability is seriously low. We think of discretizing the survival time in terms of one year and then classifying using the aforementioned data mining algorithms.

References

[1] American Cancer Society. Breast Cancer Facts & Figures 2005-2006. Atlanta: American Cancer Society, Inc. (http://www.cancer.org/).

[2] Surveillance, Epidemiology, and End Results (SEER) Program (www.seer.cancer.gov) Public-Use Data (1973-2002), National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, released April 2005, based on the November 2004 submission.

[3] Cox DR. Analysis of survival data. London: Chapman & Hall; 1984.

[4] Benjamin F. Hankey, et. al. The Surveillance, Epidemiology, and End Results Program: A National Resource. Cancer Epidemiology Biomarkers & Prevention 1999; 8:1117-1121.

[5] Houston, Andrea L. and Chen, et. al.. Medical Data Mining on the Internet: Research on a Cancer Information System. Artificial Intelligence Review 1999; 13:437-466.

[6] Cios KJ, Moore GW. Uniqueness of medical data mining. Artificial Intelligence in Medicine 2002; 26:1-24.

[7] Zhou ZH, Jiang Y. Medical diagnosis with C4.5 Rule preceded by artificial neural network ensemble. IEEE Trans Inf Technol Biomed. 2003 Mar; 7(1):37-42.

[8] Lundin M, Lundin J, Burke HB, Toikkanen S, Pylkkanen L, Joensuu H. Artificial neural networks applied to survival prediction in breast cancer. Oncology 1999; 57:281-6.

[9] Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artificial Intelligence in Medicine. 2005 Jun; 34(2):113-27.

[10] Ian H. Witten and Eibe Frank. Data Mining: Practical machine learning tools and techniques, 2nd Edition. San Fransisco:Morgan Kaufmann; 2005.

[11] J. R. Quinlan, C4.5: Programs for Machine Learning. San Mateo, CA:Morgan Kaufmann; 1993.

[12] Weka: Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/