Journal of Harbin Engineering University Vol 44 No. 7
ISSN: 1006-7043 July 2023

Classifying Seer Dataset for Breast Cancer Using C4.5 Algorithm

NISHARANI BHOI [1]
[1] Research Scholar, Department of Computer Application, Dr. A. P. J. Abdul Kalam University, Indore, Madhya Pradesh

DR. LOKENDRA SINGH SONGARE [2]
[2] Supervisor, Department of Computer Application, Dr. A. P. J. Abdul Kalam University, Indore, Madhya Pradesh

ABSTRACT
When it comes to cancer-related deaths among women, breast cancer is now the second largest cause. The chances of long-term survival for those with breast cancer are greatly improved if the illness is caught early. Patients in the SEER breast cancer dataset have been categorized using the C4.5 classification algorithm as having "Carcinoma in situ" (an early stage of cancer) or "Malignant potential." The raw dataset has been cleaned up by pre-processing methods, and the important attributes for classification have been extracted. Data that has already been cleaned and prepared for analysis has been randomly sampled for testing. The acquired rule set was put to the test on the remaining information. This section presents and discusses the obtained results.

Keywords: Breast Cancer, Diagnosis, Classification, Dataset, Data mining

I. INTRODUCTION

New medical diagnostic systems are being created as a result of the use of electronic and digital technology to address medical issues. In addition to standardizing data collecting, the created technologies are reducing diagnostic error. One of the most investigated areas in medicine is how to detect cancer at an early stage. New illness diagnosis methods were created so that doctors could employ electronic and digital technology to better treat their patients. Medically developed, these technologies streamline data collection while decreasing the likelihood of human mistake. The most researched medical issue is how to detect malignancies early on.

Cancer is a disease in which cells in the body cease performing their normal jobs and begin to divide and grow uncontrolled. When cancer cells multiply unchecked, tumours develop (bulks). There are two types of tumours: benign (mild) and malignant (ill-tempered). The development of malignant tumours is the major cause of mortality worldwide. According to statistics, 7.9 million individuals succumbed to cancer in 2007. When diagnosing cancer, doctors first look at the affected organ to determine the proper diagnosis. This malignant proliferation of cells in the breasts is known as breast cancer. Breast cancer is the second leading cause of cancer-related mortality in women, behind only lung cancer. The World Health Organization estimates that in 2008, around 460,000 women lost their lives to breast cancer. One in eight women will develop breast cancer in their lifetimes, although men are extremely unlikely to be affected. Also, a recent study conducted in Canada pegs this figure at a third. Researchers have identified risk factors including genetics, obesity, and age but have not been able to produce an effective therapy to prevent illness. Diagnosis at an early stage is the best therapy option here. In many cases, the condition can be prevented or the patient's life can be prolonged by receiving an early diagnosis. Cancer is usually diagnosed by using the findings of medical tests performed on people who are thought to be carriers of the illness. These studies can be broken down into demographic data, biometric measurements, and medical imaging scans (Roentgen, MR, Doppler and Mammography). After medical tests are completed, a plethora of data becomes available for study. In order to use the data for medical diagnosis, it must be evaluated quickly and effectively. The use of rapidly developing information technology, and

in particular data mining techniques, for analyzing these data, has become commonplace in recent years. Data mining is a popular method for collecting, validating, and estimating many types of information. Data mining techniques, which were previously known as data classification methods, are increasingly employed in pattern recognition and have been extensively applied to the problem of cancer diagnosis.

II. PROPOSED METHOD

Figure 1 detail the procedures used to process SEER data.

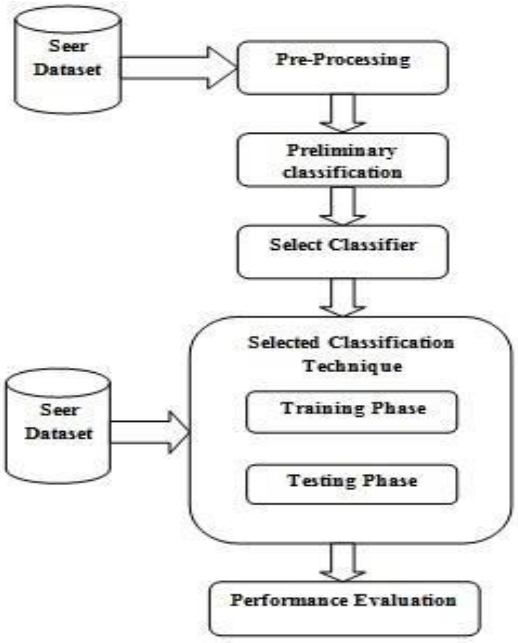

Figure 1: Processing steps

Data mining and categorization were practised on the SEER (Surveillance, Epidemiology, and End Results) dataset from the NCI's Cancer Control and Research Program. Today, SEER compiles and publishes cancer survival and incidence rates from community-based cancer registries that account for around 28% of the US population. The Surveillance, Epidemiology, and End Results (SEER) database is the gold standard for cancer statistics in the United States. It contains data on cancer incidence, prevalence, and survival rates for various regions of the country's population. During the time span covered (1973-2008), the dataset utilised included information on every conceivable kind of cancer. 1403 of these records were representative samples, and each had 124 characteristics relevant to breast cancer.

- **Pre-processing**

The raw SEER data was prepared by the application of data pre-processing. Data pre-processing is an essential procedure for preparing raw data for further analysis by data mining methods and for enhancing data quality. According to the literature, attribute selection is crucial for pinpointing the most crucial diagnostic factors in breast cancer. It was also shown that a minimal number of non-redundant characteristics was sufficient to preserve the prediction quality.

First, sociodemographic variables that were not directly connected to cancer were filtered out. For instance, race and ethnicity-related filters were disregarded. Eighteen traits were eliminated, bringing the total number of attributes down from 124 to 106. Next, we eliminated properties where more than 60% of records had missing data. For instance, in no record were any values for the parameter EOD TUMOR SIZE present. As a result, we were left with 72 traits after eliminating 34. Then, the characteristics that were redundant were either rewritten to remove the redundant data or removed altogether. To give just one example, the characteristic HISTOLOGIC TYPE was re-coded as HISTOLOGIC TYPE ICD-O-3. HISTOLOGIC TYPE was therefore disregarded.

As a result of this procedure, 56 qualities were chosen.

The next stage was to enter real data into the predetermined fields. All of this was accomplished by consulting the manual that came with the SEER database. For instance, the property VITAL STATUS RECODE was updated to reflect the correct values, such as "alive" for code 1, and "dead" for code 2. After all these steps, we had a total of 15 qualities. Five of the fifteen traits on the list were continuous, while the others may take on a range of discrete values. Table 1 provides a list of continuous properties and their respective explanations (taken from the SEER documentation).

Table 1: Seer Continuous Attributes After Pre-Processing

S. No.	Attribute	Description
1	AGE AT DIAGNOSIS	The age of the patient at diagnosis for this cancer which is coded as 1- 130 actual age and 999-unknown
2	REGIONAL NODES POSITIVE	Records the exact number of regional lymph nodes examined by the pathologist that were found to contain metastases.
3	SEQUENCE NUMBER—CENTRAL	This sequence number counts all tumors that were reportable in the year they were diagnosed even if the tumors occurred
4	CS TUMOR SIZE	Records the largest dimension or diameter of the primary tumor, and is always recorded in millimeters.
5	CS EXTENSION	Identifies contiguous growth (extension) of the primary tumor within the organ of origin or its direct extension into neighbouring organs.

As a last step, records were removed from consideration if they lacked information for any of these 5 criteria. Therefore, 1183 records (out of a total of 1403) were chosen since they had no missing information.

- **Preliminary Classification**

Using BEHAVIOR CODE ICD-O-3 as the target class and the aforementioned 5 continuous qualities as input attributes, we then performed a preliminary analysis on the 1183 data records with several classification approaches. BEHAVIOR CODE ICD-O-3 values of 2 indicate "Carcinoma in situ," whereas values of 1 indicate "Malignant Potential." The C4.5 algorithm's final set of rules is listed in figure 1.

CS EXTENSION < 35.0000
then BEHAVIOR = Malignant Potential
CS EXTENSION >= 35.0000
then BEHAVIOR = Carcinoma In Situ

Figure 2: Classification Rules for All 1183 Records

When compared to other features, CS EXTENSION was deemed to be the most important, as is evident. The remaining features were disregarded if CS EXTENSION was included in the set of categorization attributes. Due to its glaring lack of consideration for attributes such CS TUMOR SIZE, which signals the development of tumors and is thus crucial to illness categorization, this rule set is plainly not useful. This resulted in the removal of the CS EXTENSION characteristic.

The remaining four continuous characteristics were also put through categorization procedures. Table 2 provides a comparison of the error rates derived from the various categorization methods.

Table 2: Comparison of Classification Techniques

S. No	Technique	Error Rate
1	C-RT	0.1014
2	CS-MC4	0.0803
3	C 4.5	0.0761
4	ID3	0.1014
5	K-NN	0.0752
6	LDA	0.1074
7	NAÏVE BAYES	0.1183
8	PLS-LDA	0.1183
9	RND TREE	0.0414
10	SVM	0.1014

As can be seen from the table, the RND TREE method has the lowest error rate, at 0.0414, or around 4%. However, it was discovered that the rule set was excessively big and cumbersome, making it hard to apply to any dataset. The combined use of the KNN and C4.5 Classification Algorithms yields an accuracy of around 92%.

KNN or C4.5 are both viable classification algorithms that may have been used moving forward. C4.5 was selected since it is an established decision tree induction learning approach that has been widely and fruitfully used for medical data.

- **Classification using C4.5**

For those problems that ID3 did not fully solve, Quinlan developed a software add-on called C4.5. Over fitting the data should be avoided, and other important considerations include minimizing pruning errors, performing post-pruning on rules, managing continuous attributes, and dealing with missing attribute values. Entropy and information gain are employed by the C4.5 classification method during the tree-splitting process. During the testing phase, we made use of the previously compiled training data. In order to derive the rule set, the C4.5 method was used.

III. DATA ANALYSIS

The generated classification rules were then applied to the entire set of preprocessed data in the testing phase. Analyses are performed on the gathered data.

- **Training Phase**

Out of the original 1183 records in the pre-processed data, we picked three random groupings of 500 records each. This served as the training data fed into C4.5 to generate the sets of classification rules. Table 4 displays the percentage of incorrect classifications for each of the three sample groups. As can be seen in table 3, random Sample 2 had the lowest error rate of 0.599.

Table 3: C4.5 Training Phase Error Rates

Sample	Sample I	Sample II	Sample III
Error rate	0.0640	0.0599	0.0719

We used the other classification algorithms on the second set of 500 records in Sample 2 as an additional verification step. Table 4 displays the collected data.

Table 4: Comparison of Classification Techniques

S. No	Classification Technique	Error rate for Sample set 2
1.	C-RT	0.0918

2.	CS-MC4	0.0858
3.	C 4.5	0.0599
4.	ID3	0.0918
5.	K-NN	0.0739
6.	LDA	0.0918
7.	NAÏVE BAYES	0.0938
8.	PLS-LDA	0.0938
9.	RND TREE	0.0918
10.	SVM	0.0918

The table and the graph both show that the best outcomes are achieved by using the C4.5 method. Because of this, we know that using C4.5 to categorize SEER data was the right decision.

- **Testing Phase**

We used the C4.5 rules for random Sample 2 to classify all 1183 records in the testing phase. Table 5 shows the confusion matrix based on the actual and predicted values from the classification exercise.

Table 5: Confusion Matrix for Seer

	Malignant Potential	**Carcinoma In Situ**	**Total**
Malignant potential	56 (TP)	64 (FN)	120
Carcinoma In Situ	28 (FP)	1035 (TN)	1063
Total	84	1099	1183

An examination of the confusion matrix reveals that ninety-two of 1183 records have uncertain classifications. Sixty-four of the cases with "Malignant potential" were found to be Carcinoma in situ. As of now, 28 of the Carcinoma cases have been determined to be Malignant.

IV. CONCLUSION

We have utilized the C4.5 algorithm to try to categorize the SEER breast cancer data into Carcinoma in situ and malignant potential categories. We utilized a subset of the breast cancer dataset (500 records) as a training set and then applied the resulting classification rules to the entire dataset. Our training phase accuracy was ~94%, while our testing phase accuracy was ~93%. We hope that our research will lead to the creation of more efficient and accurate automated diagnostic tools to aid in the battle against cancer, where prompt treatment may save lives if detected at an early stage.

REFERENCES: -

1. G. Naga Rama Devi, Dr. K. Usha Rani "Importance of Feature Extraction for classification of Breast Cancer Datasets – S Study", International Journal of Scientific and Innovative Mathematics Research(IJSIMR), Vol. 3, Special Issue 2, July-2015, PP 763-768, ISSN 2347-307X.
2. G. Naga Rama Devi, Dr. K. Usha Rani "Evaluation of Classifier Performance using Resampling on Breast Cancer", International Journal of Science and Engineering Research, Vol. 6, Issue 2, February 2015, ISSN 2229-5518.
3. G. I. Salama, M. B. Abdelhalim, and M. A. Zeid, "Breast Cancer Diagnosis on Three Different Datasets Using Multi-Classifiers", Int. J. of Comput. and Inform. Technology, vol. 1, no. 1, pp. 2277 – 0764 , Sept. 2012.
4. M. M. Beg and M. Jain, "An Analysis Of The Methods Employed For Breast Cancer Diagnosis", Int. J. of Research in Comput. Sci., vol. 2, no. 3, 2012, pp. 25-29.
5. Ankit Agrawal, Sanchit Misra, Ramanathan Narayanan, Lalith Polepeddi, Alok Choudhary, "A Lung Cancer Mortality Risk Calculator Based on SEER Data", IEEE 2011.

6. F. Paulin and A. Santhakumaran, "Classification of breast cancer by comparing Back propagation training algorithms", Int. J. on Comput. Sci. and Eng., vol. 3, no. 1, Jan. 2011, pp. 327-332.
7. Farzaneh Keivanfard , Mohammad Teshnehlab , Mahdi Aliyari Shoorehdeli , "Feature Selection and Classification of Breast Cancer on Dynamic Magnetic Resonance Imaging by Using Artificial Neural Networks", Proceedings of the 17th Iranian Conference of Biomedical Engineering (ICBME2010), 3-4 November 2010.
8. W. C. Yeh, W. W. Chang and Y. Y. Chung, A new hybrid approach for mining breast cancer pattern using discrete particle swarm optimization and statistical method, Expert Systems with Applications, vol.36, no.4, pp.8204-8211, 2009.
9. Santi Wulan Purnami, S.P. Rahayu and Abdullah Embong, "Feature selection and classification of breast cancer diagnosis based on support vector machine", IEEE 2008.