IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

EFFECTUAL ILLNESS ANALYSIS USING HURISTIC CLASSIFICATION METHODOLOGIES IN DATAMINING

[1]Dr.Vineetha KR, [2]Muhsina K P
[1]Associate Professor, [2]MCA Scholar
[1]Department of MCA
[1]Nehru College of Engineering and Research Centre, Pambady, India
vpvprakash@gmail.com, muhsinamuthalib9582@gmail.com

Abstract : Several combinations of database and machine learning approaches are used to extract hidden and unknown patterns from massive data sets. Data mining is essential for dealing with large amounts of data. Data mining is concerned with data heterogeneity and correctness. Furthermore, medical data mining is an extremely essential research subject, and substantial efforts have in this area in recent years since accuracy in medical data systems can lead to seriously deceptive medical treatments. Suitable mining methods should be used to examine medical data collections. Data mining techniques have been employed in constructing medical systems for disease prediction using a set of medical data sets to execute related tasks. In this an examination of the survivability rate prediction of breast cancer patients using data mining approaches is presented in this paper. SEER Public-Use Data was used as the source of information. The preprocessed data collection contains 151,886 records, each of which has all 16 fields from the SEER database. We looked into three data mining methods: Nave Bayes, back-propagated neural networks, and C4.5 decision tree algorithms. These algorithms have been used in a number of experiments. The achieved prediction results are comparable to those of other methods. However, we discovered that the C4.5 method performs far better than the other two strategies.

IndexTerms - Data Minining, SEER, WEKA, Breast cancer survivability

I. INTRODUCTION

Mining is a knowledge finding process that involves evaluating data that may be concealed in incredibly large amounts of data. It is a method of extracting data from historical records in order to make significant decisions for future projections. Image mining, opinion mining, web mining, text mining, graph mining, and medical data systems are all examples of data mining applications. It has grown in importance in medical research as a means of uncovering previously unknown patterns in medical data. Medical practitioners can examine diseases based on the predictions provided by the prediction mode. In the United States today, one out of every eight women will acquire breast cancer during their lifetime.

According to the most recent data, the survival rate is 88 percent five years after diagnosis and 80 percent ten years after diagnosis. Extraction of knowledge from data related to an illness allows for the finding of the survival rate or survivability of that disease[1]. SEER (Surveillance Epidemiology and End Results) is one of these data sources, and it is a one-of-a-kind, dependable, and crucial resource for researching various aspects of cancer. The SEER database brings together patient-level data on cancer location, tumor histology, stage, and cause of death[2].The features of a population can be studied to determine the elements that influence a given outcome. Observational research, such as statistical learning and data mining, can establish the relationship between the variables and the outcome, but not always the cause-and-effect relationship. Many scientific fields, such as medicine and biotechnology, are increasingly relying on data-driven statistical research. The current study uses data mining techniques to predict the survival rate of breast cancer patients. The researchers analyzed SEER data and developed a pre-classification method that considers three variables: Survival Time Recode (STR), Vital Status Recode (VSR), and Cause of Death (COD)[3,4].

II. LITERATURE SURVEY

Data Mining Technique for Medical Diagnosis Using a New Smooth Support Vector Machine is a research done by Santi Wulan Purnami , Jasni Mohamad Zain & Abdullah Embong which is deal with idea of data mining technology.The usage of data mining techniques in medical investigations has gradually increased over the previous decade. The purpose of this paper is to describe a recent study on the use of data mining techniques to solve medical diagnosis issues. Multiple Knot Spline Smooth Support Vector Machine is the proposed data mining technique (MKS-SSVM).MKS-SSVM is a new SSVM that approximates the plus function using several knot spline functions rather than the integral sigmoid function as in SSVM[5]. The findings of this investigation revealed that MKS-SSVM was effective in diagnosing medical datasets, particularly diabetes and heart illness, which is a promising outcome when compared to earlier results[6].

Breast cancer prediction using data mining techniques research done by S. Padma Priya1, Assistant Professor & Head Department of Information Technology Sri Adi Chunchanagiri Women's College, Cumbum,(India),P Soumya, Research Scholar, Department of Computer Science ,Sri Adi Chunchanagiri Women's College, Cumbum,(India) deals with breast cancer prediction .Data mining is, in fact, a component of a bigger knowledge discovery process. Breast cancer is the worst disease and the most frequent of all cancers, and it is the top cause of cancer deaths in women globally, according to one of the latest data mining studies. In DM research, categorization is one of the most important strategies. Breast cancer detection and prognosis are two medical applications that present a significant challenge to medical researchers[7,8,9]. This study examines at a number of review and technical papers about breast cancer diagnosis. His research looks at a variety of review and technical articles on breast cancer diagnosis. The primary purpose of this study is to provide an overview of current research employing data mining approaches to improve breast cancer diagnosis. This survey focuses on the utilization of the ID3 and C4.5 classification algorithms in breast cancer analyses.

III. METHODOLOGY

Breast cancer prediction

Breast cancer is now a very frequent cancer in women. The traditional method for detecting breast cancer is mammography. However, radiologists differ greatly in their interpretations of mammograms. Furthermore, Elmore stated that 90 percent of radiologists recognized less than 3% of malignancies, while 10% recognized roughly 25% of cases. Fine needle aspiration cytology is another method for diagnosing breast cancer that provides a more precise prediction accuracy. The average correct recognition rate, on the other hand, is around 90%[1]. The goal of all linked studies is to discriminate between patients with malignant breast cancer and individuals without breast cancer in the benign group. Cancer prognosis is divided into three categories:

i)cancer susceptibility prediction (risk assessment),

ii) cancer recurrence prediction

iii) cancer survivability prediction The American Joint Commission on Cancer is the recognized predictive factor for breast cancer (AJCC).

It is a stage method based on the TNM system (T, tumor; N, node; M, metastasis), and survival is defined as any case of breast cancer in which the patient is still alive six months after diagnosis. C4.5 is a well-known classification strategy in decision tree induction that has been employed by Abdelghani Bellaachia and Erhan Gauven in conjunction with two techniques: Nave Bayes and Back-Propagated Neural Network. They apply the aforesaid data mining techniques to conduct an analysis of the forecast of breast cancer patients' survivability rates, which is incorporated in the updated version of the SEER Breast Cancer Data. An examination of the survivability rate prediction of breast cancer patients using data mining approaches is presented in this paper. SEER Public-Use Data was used as the source of information. We looked into three data mining methods: Nave Bayes, back-propagated neural networks, and C4.5 decision tree algorithms. These algorithms have been used in a number of experiments. The achieved prediction results are comparable to those of other These algorithms were used to forecast the survivorship rate of the SEER breast cancer data set. These three categorization algorithms were used to discover the best appropriate one for forecasting cancer survivor rates[10]. The

Nave Bayes strategy is based on the well-known Bayesian approach, which uses a simple, clear, and rapid classifier. It is known as 'Naive' because it presupposes mutually independent qualities. In fact, this is almost never the case, although it is possible to remove the dependent categories by preprocessing the data. This approach has been used to encode, use, and develop probabilistic information in a variety of fields, with notable achievements in machine learning methods. Artificial neural networks are used in the second technique. A multi-layer network with back-propagation (also known as a multi-layer perceptron) is employed in this research. The C4.5 decision-tree generation algorithm is the third technique. The ID3 algorithm is used to create C4.5[11]. It has been demonstrated that the last two strategies perform better. These three data mining methods were tested using the Weka toolbox. Weka is a software suite that includes tools for data classification, regression, clustering, association rules, and visualization. The toolkit is written in Java and is free and open source software licensed under the GNU General Public License. A collection of tools was created to extract and clean the raw SEER data. A simple study reveals that the SEER data contains missing data. The SEER data lacks missing information in the fields of Extent of Disease (EOD) and Site Specific Surgery (SSS) for about half of the cases, according to a basic study.

The majority of the missing information can be found in documents acquired prior to 1988. We eliminated these entries from the test data set because we wanted to use all of the accessible fields in the SEER database. The EOD Coding System code for these records is '4'. After 1998, the SSS field's usage altered. The information is split into five different fields instead of the standard field. To complete the empty SSS fields, a mapping technique from new SSS to old SSS is devised. The records with missing information are then eliminated after this stage. The EOD field consists of five fields, one of which is the EOD code[12]. These parameters (tumor size, number of positive nodes, number of nodes, and number of primary) have missing information coded as '999, 99, or 9' to reflect 'unknown' data. Please note that fields with 'unknown' values are not included in the data in Table 1. The fields used in our study are also listed in the table.

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 noted in the previous section, this study took a unique approach to the pre-classification process. Unlike here, three fields were included: STR, VSR, and COD. In the SEER database, the STR column contains values ranging from 0 to 180 months. The following is a breakdown of the pre-classification procedure[10].

// 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

The records that are omitted in the above approach are those of patients who have a STR of less than 60 months and are still alive, or those of patients who have a STR of less than 60 months but died from a reason other than breast cancer. Tables 2 and 3 illustrate the pre-classification process classes and the approach utilized in, 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

Following the preprocessing step, assessing the effect of the attributes on the prediction, or attribute selection, is a typical approach. The information gain metric was chosen to rank the qualities because it is a common method and the C4.5 decision tree technique uses it. When an attribute gives additional information about a class, information gain (IG) is measured as the difference in entropy (H).The information gain and entropy before and after observing the attribute Xi for the class C are as follows:

H(C)
$= -\sum p(c)\log p(c)$, $c \in C$
$H(C|Xi) = -\sum p(x) \sum p(c|x)\log p(c|x)$, $x \in Xi$, $c \in C$
$IGi = H(C) - H(C|Xi)$

The rated survival properties of data as determined by the Weka toolbox are shown in Figure 1. It is apparent that Tumor Extension has a better ranking than Tumor Size.

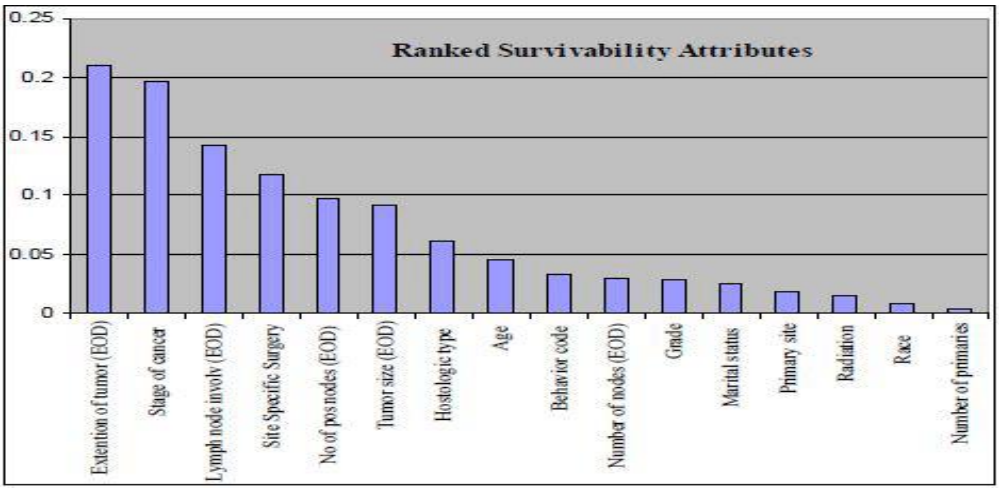

Figure 1: Survivability Attributes in Order

The performance parameters of accuracy, precision, and recall are used to compare the three strategies. In order to have an accurate assessment of the classifier's performance; A cross-validation with ten folds was used. Cross-validation is the process of splitting data into k subgroups in its most basic form[3]. The estimated error rate is the average error rate from these k subgroups, and each subgroup is projected using the classification rule created from the remaining (k-1) subgroups. The error rate can then be calculated in an unbiased manner. The estimated error rate is the average error rate from these k subgroups, and each subgroup is projected using the classification rule created from the remaining (k-1) subgroups. The error rate can then be calculated in an unbiased manner. The final classifier rule is derived from the complete collection of data. We acquire the metrics of precision, recall, accuracy A I and Cross Validation Accuracy (CVA) to indicate a classifier performance after running the classifier 10 times with 10 folds:
CVA = (1/10)∑Ai i = 1, 2, ..., 10 Ai = # records correctly classified / total # records
After running a specified k-fold cross-validation, the Weka toolbox can determine all of these performance measures.

IV. RESULT ANALYSIS

The accuracy of three data mining approaches is compared in this study. Along with high precision and recall metrics, the goal is to have high accuracy. Although these measures are more commonly employed in the field of information retrieval, we have included them because they are related to other existing metrics such as specificity and sensitivity. Results are presented in table4.

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 seen in Table 4, neural networks and decision trees have similar performance. Table 5 illustrates the experimental findings utilizing the same dataset as our approach and the pre-classification approach used in. The results clearly reveal that the categorization rate (81%) is significantly lower than our approach's (87%) classification rate. On an AMD Athlon 64 4000+ system, the computation times of the methods Nave Bayes, neural net, and C4.5 were in the range of 1 minute, 12 hours, and 1 hour, respectively. The results obtained in this study differ from those of Delen et al. Due to the fact that we used a fresher database (2000 vs. 2002), as well as a different pre-classification (109,659 and 93,273 vs. 35,148 and 116,738), and different toolkits (industrial grade tools vs. Weka).

V. CONCLUSION

The difficulties, algorithms, and techniques for the problem of breast cancer survivability prediction in the SEER database were defined, addressed, and resolved in this study. Aside from the Survival Time Recode (STR), the Vital Status Recode (VSR), and the Cause of Death are taken into account in our approach (COD). The results of the experiments reveal that our strategy outperforms the other. This study clearly demonstrates that preliminary results for the use of data mining approaches to the survivability prediction problem in medical databases are promising . Our study excludes records with incomplete data; future work will incorporate missing data in the EOD field from previous to 1988 EOD fields. Because the size of the data set will grow significantly, this may improve performance.

VI. 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.a