CSE497- PROJECT SPECIFICATION REPORT

BREAST CANCER DETECTION USING DATA MINING APPROACHES WITH GENETIC ALGORITHMS

GROUP MEMBERS

BETÜL ESMER	150110024
FADİME SAYIN	150110006
SEMRA MEMİŞOĞLU	150110052

ADVISOR

Asst. Prof. FATMA CORUT ERGİN

1. PROBLEM STATEMENT

In this project, a data mining application that works on breast cancer cells will be designed. This application is proposed to detect whether the given data is of benign or malignant cancer cells with high accuracy estimates.

2. PROBLEM DESCRIPTION

Breast cancer is one of the most important problems that threaten women's health although rarely seen in men. Breast cancer is counted as the second type of cancer after lung cancer, which causes the largest number of deaths. Breast cancer is seen in one of every 10 women in Western European countries and every 8 women in United States. In our country, 30 thousand women diagnosed with breast cancer every year. Nowadays, increasing the possibility of early diagnosis and many studies on this subject, rate of death due to breast cancer is reduced to some extent. Mammography is one of the most used methods to detect breast cancer [1]. In literature, radiologist show considerable variation in interpreting a mammography [2]. So, it is necessary to develop effective and efficient identification methods to diagnose breast cancer disease as early as possible.

We decided to use data mining algorithms and genetic algorithm in our project, because constantly growing volume of data makes it impossible to analyze and capture the valuable knowledge among large amounts of data using the current statistical methods. Because of the insufficiency of the current analysis tools, new solutions have been found for extracting the valuable but hidden knowledge among huge data. These solutions are data mining. Data mining include descriptive and predictive techniques for meaningful knowledge which is unknown early from data.

Genetic algorithms have been successfully applied to solve search and optimization problems. The basic idea of a genetic algorithm is to search a hypothesis space to find the best hypothesis. A pool of initial hypothesis called a population is randomly generated and each hypothesis is evaluated with a fitness function. Hypotheses with greater fitness have higher probability of being chosen to create the next generation. Some fraction of the best hypotheses may be retrained into the next generation, the rest undergo genetic operations such as crossover and mutation to generate new hypotheses. The most important components in GA consist of representation (definition of individuals), fitness function, population, parent selection mechanism, variation operators (crossover and mutation), and survivor selection mechanism (replacement).

3. AIMS OF THE PROJECT

- Decreasing the death rate

  The most important purpose in our project is to decrease the death rate from breast cancer with early diagnosis. For this disease, early diagnosis plays a vital role, because

it can be cured in its early phases. However, if the patient is late, this disease can be incurable.

- More reliable solution
  Another aim in our project is to achieve more reliable solutions in respect to previous studies on this problem. Previous studies did not obtain hundred percent successes; in our project we will try to catch maximum success.

- Prevent other similar diseases
  If the project is successful, algorithm which is developed for our problem can be used easily for another similar problems.

- Time saving
  Anybody who uses this tool will save time, because it will run faster than clinic tests.

4. RELATED WORK

In [3], association rules (AR) is used reducing the dimension of breast cancer database and Neural networks (NN) is used for intelligent classification. Wisconsin breast cancer database was used to evaluate the proposed system performance. AR finds interesting associations or relationships among large set of data items, so some input may be eliminated. AR1 and AR2 elimination methods were performed. AR1 technique uses all input parameters and their all records to find relations among the input parameters. AR2 is used with classification problems. In AR2, large itemsets found for every class and all items in these itemsets do not have the same importance. If an item of large itemset of any class is used in other classes and it has different value, this item must be used as NN inputs. Only one input parameter of NN was eliminated using AR1 technique.

Wisconsin database used in [3] includes 9 attributes and 699 records. If AR1+NN with 8 attributes were used, its correct classification rate is %97.4. The correct classification rate of NN with 9 attributes is %95.2 and correct classification rate of AR2+NN with elimination is %95.6. So, AR1+NN can be used for best classification performance and AR2+NN for using input parameters at minimum number.

In [4], three different types of classification models were used; artificial neural networks (ANN), decision trees, and logistic regression. ANN architecture used in this study includes multi-layer perceptron (MLP) with back propagation. The MLP is known to be a powerful function approximator for prediction and classification problems. Secondly, decision tree generation is done using C5. Finally, logistic regression was used as the third model. They used the SEERS database in tests. Test result show that decision tree gives the best classification accuracy of 93.6%, the ANN model came out to be the second best with a classification accuracy of 91.2% and the logistic regression model came out to be the worst with a classification accuracy of 89.2%.

In [5], a new hybrid approach of using both integrated statistical method and discrete particle swarm optimization were used. Wisconsin breast cancer database was used for tests. According to test results, the proposed hybrid approach can improve the accuracy to 98.71%, sensitivity to 100% and specificity to 98.21 %. These results are very promising compared to the previously reported classification techniques for mining breast cancer data.

In [6], 400 patients (cancer and non-cancer) are used to collect data. K-means clustering algorithm is used for clustering to identify the relevant and non-relevant data. In this project, to discover the frequent pattern AprioriTid and Decision Tree algorithms are used. In the paper, it is said that they developed a successful lung cancer prediction system with significant pattern prediction tools, however, there are no numerical results given.

In [7], NN is used for prediction and classification tasks. Discriminant analysis and logistic regression are two common data mining methods to construct classification models. NN have reported to have better classification capability than linear discriminant analysis and logistic regression. Purpose of the study is to show the performance of data classification by integrating ANN with the MARS (multivariate adaptive regression splines). The classification accuracy of the networks is improved with obtained variables from MARS. In conclusion, in this project diagnosis system to detect breast cancer based on NN and AR. AR is used to decrease the number of items in the breast cancer database and NN is used for classification as intelligently. In the tests, they used Wisconsin breast cancer database. Again, there are no numerical results in the paper.

In [8], fine needle aspirate (FNA) is used with a data mining & statistical method to get an easy way to achieve the best result. They have combined some statistical methods such as principle component analysis (PCA), partial least squares (PLS) linear regression analysis with data mining methods such as select attribute, decision trees and association rules to find the unsuspected relationships. This approach follows a seven step process to extract useful information from data: goal identification, get a target data set, data preprocessing, data transformation, data mining & statistical analysis, interpretation & evaluation, and writing report. In step data mining & statistical analysis they have followed another five step process: select attribute (eliminate irrelevant or unhelpful attributes), PCA analysis (projection method for qualitative analysis), PLS1 analysis, decision tree analysis (expression of the rule for classification), and association rules analysis (find correlation relationship among all the attributes). In this study, as in [3], the Wisconsin breast cancer database is used. Similarly, there are no test results in the paper.

In[9], this article includes providing a comparison among the capabilities of various neural networks, such as Multiplayer Perceptron (MLP), Self Organizing Map (SOM), Radial Basis Function (RBF) and Probabilistic Neural Network (PNN). Wisconsin breast cancer database and Shiraz Namazi hospital breast cancer data were used in the tests. In this study, RBF and PNN were proved to be the best classifiers in the traininig set. No numerical results are given in the paper.

5. SCOPE OF THE PROJECT

Constraints

- We can access only limited data. We will use a benchmark dataset, since we cannot find real data for the given problem.
- Group members are not expert about the medical terms in this problem.
- We can just use the data given in the dataset, we do not have the chance to communicate with the patients.

Compare

In previous studies, many algorithms are proposed to solve the problem. Some of them give numerical results, and some of them even do not have any numerical results. The best solution approach proposed in the literature gives a 98.21% success rate. However, since the problem considers human life, it is very important to get more accurate results. In our project, we will try to find a better approach that gives higher accuracy. Moreover, we are planning to work on different databases.

6. SUCCESS FACTORS AND BENEFITS

We can conclude that our project is successful if we can detect whether the given data is for the benign or malignant cancer cell.

At the end of the testing phase, if we can find high values for the sensitivity and accuracy in a reasonable amount of time, we can conclude that we proposed a successful method.

This project aims to detect breast cancer as early as possible. Since early diagnosis is important for cancer, patient will have the chance to survive.

Our Project can be applied faster and minimum amount of medical tests will be needed. In these medical tests, the patient may be exposed to radiation. So, using a computerized approach will be safer.

The project will provide academical benefit to us.

7.METHODOLOGY AND TECHNICAL APPROACH

In our project ,we need some resources such as software ,hardware and specific dataset.

We will use Netbeans as programming IDE for java language , Microsoft Excel to show statistical results of test As hardware resource ,We need a high quality computer to run the tests.

We will use specific dataset which was found from this web site [9] http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/

8.MANAGEMENT PLAN

Literature survey ; We researched methods of data mining ,breast cancer on data mining from articles, books and web and this phase will continue.

Studying methods of data mining ; We study methods of data mining to use appropriate methods in the project.

Searching for dataset ; Data set will be needed for the project so we researched dataset to find an appropriate dataset which was used in the articles.The dataset will be divided as training set and test set.

Preparation of specification report ; Report was prepared containing such as problem description,aim and scope of the project, related work etc.

Studying of genetic algorithm; We study genetic algorithm which we will use to generate a decision tree .

Implementation ; We will implement our solution approach with genetic algorithm and decision tree generation.

TASKS/MONTHS	Sep	Oct	Nov	Dec	Oct	Feb	Mar	Apr	May	Jun
Literature survey										
Studying methods of data mining										
Searching for dataset										
Preparation of specification report										
Studying of genetic algorithm										
CSE497-REPORT										
Implementation										
Test phase										
Cse498-REPORT										

Test phase ; We will run a set of tests first using the specified dataset [9], and then, we will test with other datasets given in the literature.

Division of responsibilities and duties among team members

Betul ESMER: Litarature survey , generating the initial decision tree to be used as the initial population in genetic algorithms, test phase

Semra MEMİŞOĞLU: Litarature survey, implementing the genetic algorithm designed for the problem, test phase

Fadime SAYIN: Literature survey, implementing the genetic algorithm designed for the problem, test phase

References

1. Choua,S.-M Leeb , T.-S., Shaoc,Y.E., &Chenb,I.-F.(2004).Mining the breast cancer pattern using artifical neural Networks and multivariate adaptive regression splines. Expert System with Applications,27,133-142

2. Elmore,J.,Wells,M.,Carol,M.,Lee,H.,Howard,D.,&Feinstein, A.(1994).Variability in radiologists interpretation of mammograms. New England Journal of Medicine ,331(22),1493-1499.

3. Murat Karabatak, "An expert system for detection of breast cancer based on association rules and neural network",Expert System with Applications , 2009

4. Dursun Delen,Glenn Walker,Amit Kadam, "Predicting breast cancer survivability: Comparison of three data mining methods",Artifical Intelligence in Medicine,2004

5. Wei-Chang Yeh, Wei-Wen Chang , Yuk Ying Chung " A new hybrid approach for mining breast cancer pattern using discrete particle swarm optimization and statistical method", Expert systems with Applications, 2009

6. Mukti, Md Zamilur Rahman, and Farzana Ahmed. "Early detection of lung cancer risk using data mining." Asian Pacific Journal of Cancer Prevention 14.1 (2013): 595-598.