Study on Data Mining Techniques for Cancer Prediction System

Roseline Jecintha I¹, Poonguzhali V²
¹Assistant Professor, ²Research Scholar
PG & Research Department of Computer Applications, St.Joseph’s College of Arts & Science, Cuddalore
Email: roseline_jessy@yahoo.co.in, mpss1806@gmail.com

Abstract - Cancer occurs when changes called mutations obtain situate in genes that control cell growth. The mutations permit the cells to divide and multiply in an uncontrolled, hectic way. The cells are multiplying, producing copies that obtain increasingly more abnormal. In the majority cases, the cell copies finally form a tumor. Cancer is the most vital reason for death in the world. The most of common cancers diagnosed in the world are those of the breast, lung, and blood cancers. The prognosis of different cancer is extremely variable. Several cancers are curable with early detection and treatment. Cancers that are aggressive at a later stage may be more difficult to cure. Knowledge Discovery in the database (KDD), which includes data mining techniques are has been used in healthcare. This study paper we have discussed various data mining techniques that have been utilized for the breast cancer, lung cancer, blood cancer. We focus on present research being carried out using the data mining approach to enhance the breast, lung, blood cancers risk factors are diagnosis and prognosis.

Keywords - Breast cancer, lung cancer, blood cancer, data mining and KDD.

I. INTRODUCTION

One of the popular research tools for medical field is Data Mining. Researchers to identify and relationships among a huge number of variables. Able to predict the result of a disease using the historical cases stored within the dataset. The aim of our study is to use various data mining techniques to classify the risk factors of breast cancer, lung cancer, and the blood cancer. This paper reviews the classification techniques in health domain.

1.1 Knowledge discovery process and data mining

The term Knowledge Discovery in Databases or KDD for short, refers to the wide route of finding knowledge in data, and emphasizes the "high-level" function of particular data mining methods. The combined aim of the KDD process is to dig up knowledge from the large repositories. Data mining methods to pull out the information according to the condition of actions, using a database along with any vital preprocessing, sampling, and transformations of that database.

1.2.1 The list of steps involved in the knowledge discovery process:

- **Data Cleaning** - The noise and inconsistent data is removed.
- **Data Integration** - many data sources are combined.
- **Data Selection** - Data relevant to the analysis task are retrieved from the database.
- **Data Transformation** - Data is transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations.
- **Data Mining** - smart methods are applied in order to pull out data patterns.
- **Pattern Evaluation** - Data patterns are evaluated.
- **Knowledge Presentation** - Knowledge is represented.

1.2.2. How to use KDD and data mining in research

1. Developing an understanding of the application domain, the related prior information, and the goals of the user.
2. Creating a target data set: selecting a data set, or focusing on a separation of variables, or data samples, on which detection is to be performed
3. Data cleaning and preprocessing. Strategies for handling missing data fields. Accounting for time series information and known changes.
4. Data reduction and projection. Using dimensionality decrease or transformation methods to decrease the useful number of variables under reflection or to find invariant representations for the data.
5. Choosing the data mining task. Decide whether the goal of the KDD process is classification, regression, clustering, etc.
6. Choosing the data mining algorithms. Deciding which models and parameters may be appropriate. Matching a particular Data mining method with the overall criteria of the KDD process.
7. Data mining. Searching for patterns of interest in a particular representational form Consolidating discovered knowledge.

1.2.3. Classification and Prediction
Classification is the route of finding a model that describes the data classes. The reason is to be able to use this model to guess the class of objects whose class label is unidentified. This obtained model is based on the analysis of sets of training data. The obtained model can be presented in the following forms:
- Classification (IF-THEN) Rules
- Decision Trees
- Mathematical Formulae
- Neural Networks

Classification - It predicts the class of objects whose class label is unidentified. Its purpose is to find an obtained model that describes and distinguishes data classes or concepts. The Derived Model is based on the study set of training data i.e. the data object whose class label is well known.

Prediction - It is used to predict missing or engaged numerical data values rather than class labels. Regression study is usually used for prediction. Prediction can also be used for recognition of distribution trends based on available data.

II. PRECEDING STUDIES OF CANCER PREDICTION

<table>
<thead>
<tr>
<th>AUTHORS</th>
<th>PAPER</th>
<th>TECHNIQUE USED</th>
<th>RESULTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jaimini Majali, Rishikesh (2015)</td>
<td>Data Mining Techniques For Diagnosis And Prognosis Of Cancer</td>
<td>The Developing system with two approaches of data mining association rule mining and classification techniques, FP-growth algorithm for diagnosis of cancer. Applied Wisconsin dataset to FP–growth algorithm Using Wisconsin data set.</td>
<td>Among the various data mining classifiers and soft computing approaches, the decision tree is found to be the best predictor on Wisconsin data set.</td>
</tr>
<tr>
<td>Dr.T.Christopher, J.Jamera banu(2016)</td>
<td>Study of Classification algorithm for lung cancer prediction</td>
<td>Classification algorithms such as Naïve Bayes, the Bayesian network and 348 algorithms are used to analyze lung cancer prediction.</td>
<td>The naïve Bayes algorithm gives a better performance over the other classification algorithm.</td>
</tr>
<tr>
<td>Alaa M. El-Halees, Asem H. Shurrab (2017)</td>
<td>Blood tumor prediction using data mining techniques</td>
<td>Rule Induction, Association rules, Deep learning.</td>
<td>The experiments gave different accuracy ratio according to the type of blood disease and the type of the classifier. We found that the deep learning classifier has the best ability to detect tumor from blood samples.</td>
</tr>
</tbody>
</table>


### III. BREAST CANCER

Breast cancer is the cancer that has spread from the ducts or glands to other parts of the breast. More than 40,000 women were expected to die from the disease. Breast cancer can also be diagnosed in men.

**Risk factors for breast cancer**
- Age, drinking alcohol, Having dense breast tissue, Gender genes, Giving birth at an older age, never being pregnant these are included in risk factors for breast cancer.

### IV. LUNG CANCER

Lung cancer is the uncontrolled growth of anomalous cells that starts in one or both lungs; regularly in the cells that line the air passages. The abnormal cells do not expand into healthy lung tissue; they divide quickly and form tumors. Two types of lung cancer are SCLC and NSCLC.

SCLC stands for Small Cell Lung Cancers
- NSCLC stands for Non-Small Cell Lung Cancers

**Risk factors for lung cancer**
- Smokings, passive smoking, exposure to asbestos fibers, exposure to radon gas, familial predisposition are included in risk factors for lung cancer.

**Blood Cancer**
- Blood cancers affect the produce and function of blood cells. These cancers start in bone marrow. The most common are leukemia, lymphoma, and myeloma.

**Risk factors for blood cancer**
- Certain blood disorders, genetic syndromes, family history, older age, male gender are some of the risk factors of blood cancer.

### V. CONCLUSION

This paper discusses various data mining techniques used to researchers in their research papers. The data mining is used in the field of medical prediction are discussed. The main focus is on using different algorithms for cancer prediction using data mining. In future the work may be extended and improved for the automation of breast cancer, lung cancer, blood cancer prediction. We will use the other types of data mining techniques to predict. Some politeness restrictions such as contractual relationships between researcher and health care organization are compulsory to overcome the safety issues. There is also a need of standardized approach for constructing the data warehouse. In recent years due to an enhancement of internet facility, huge datasets (text and non-text form) are also available on the website. So, there is also an essential need for effective data mining techniques for analyzing this data to uncover hidden information.
References


