Using Frequent Itemset Mining for Breast Cancer Detection

Syed Zubair Ahmad Shah* and M. Amjad
Jamia Millia Islamia, New Delhi - 110025, Delhi, India; zub12345@gmail.com, mamjad@jmi.ac.in

Abstract

Objectives: This paper introduces an improvement over the use of Artificial Neural Networks (ANN) for breast cancer detection. Methods: It suggests use of frequent pattern mining for minimizing the dimensions of breast cancer database. After reduction step, the database is then input to an ANN for classification. Findings: We have shown through experimentation that the proposed model not only reduces the input database dimensions but also produces better classification results. Application: The proposed model will be highly beneficial in the field of medicine and will help in precise detection of breast cancer.

Keywords: Artificial Neural Network, Association Rule Mining, Breast Cancer Detection, Dimension Reduction, Frequent Itemset Mining

1. Introduction

Data mining has been found applicable to problems of medicine, engineering, genetics, image recognition etc. In the field of medicine - in addition to data mining - artificial intelligence, neural networks and expert systems have also been thoroughly used in the development of medicines, in proposing models for surgeries, in disease diagnosis etc.

Many systems that can diagnose cancer have been created over the last few years. These systems help doctors to make correct diagnostic decisions which have resulted in saving of hundreds of millions of lives and have helped doctors to suggest appropriate cure for the deadly as well as non-deadly diseases alike.

Analyzing historical databases using various data mining techniques – frequent pattern mining, classification and clustering – has been found much useful in disease diagnosis of new patients. The work of creating such databases has already been done by many hospitals and researchers around the world. The relationships between the data within these databases can easily be found by various data mining techniques.

One of the serious problems in the field of medicine has been found to be the detection of cancers. Our work is based on breast cancer detection. Breast cancer develops from breast tissue. Symptoms include a gob in the breast, an alteration in breast shape, dimples on skin, fluid release off the nipple, or reddish patch(es) on skin. Of all cancers, breast cancer is the main type of cancer found in women. It accounts to 25 percent of all cases. In 2012, it resulted in more than half a million deaths.

Many methods have been used to detect breast cancers including Mammography and FNAC. It has been found difficult to interpret the former very precisely and for the latter average identification rate has been found to be only 90%. Later, researchers turned to artificial intelligence for finding ways to detect breast cancers. For predicting and classifying breast cancers, introduced the use of ANN and multi-variate adaptive regression splines. used decision trees and ANN for predicting breast cancer relapse. used isotonic-separation method to detect breast cancer. proposed use of fuzzy artificial immune system and k-nearest neighbors for detection of breast cancer.

In this article, we propose use of frequent pattern miner accompanied by an ANN for breast cancer diag-
Using Frequent Itemset Mining for Breast Cancer Detection

nosis problem. Our method takes as input historical datasets, processes it and constructs rules and finally classifies the data. The frequent pattern/ itemset miner helps in dimension reduction i.e. its main job is to minimize the inputs to the ANN.

The field of frequent pattern mining is concerned about developing data mining algorithms to identify intriguing, unanticipated and useful patterns in datasets. There are many frequent pattern mining algorithms – some matrix based and some graph based. We have used FTMBG (Frequent Termsets Mining using Bipartite Graphs) algorithm.

The concept of ANN developed from the study of biological nervous system. In an ANN, simple fictitious nodes known as neurons or neurodes are linked in order to create a network that imitates a biological neural network. There is no standard definition of ANN, nevertheless a model can be called neural (i) if it contains sets of robust weights, i.e. arithmetic parameters which are regulated by some learning algorithm and(ii) if it is able to proximate nonlinear functions of their inputs.

2. Preliminaries

We used frequent pattern mining in finding association rules. Association rules are statements which assist in discovering relations within seemingly unrelated data of a data repository. Let I={item1,item2,...} be a set of items. Let DB be a transaction-database. Let X be an itemset. The number of transactions that contain X is its support. If support of X is at least equal to minimum support threshold then X will be called a frequent itemset. Confidence of a rule X→Y is defined by

\[
\text{confidence}(X \rightarrow Y) = \frac{\text{support}(X \cup Y)}{\text{support}(X)}
\]

i.e.the ratio of fraction of transactions in which both occur together to the fraction of transactions in which A is present. If a minimum confidence value is passed then the rule is established as an association rule and implies that the presence of A in a transaction almost guarantees presence of Y. In our experimentation we have used such rules on the characteristics (columns) of Wisconsin database with the intention to reduce them in number.

3. The Approach and its Evaluation

In our study, we used Wisconsin breast cancer dataset for evaluation. This dataset is provided by the University of Wisconsin, Madison. Tables 1 and 2 summarize the records and characteristics of this database. Each value in the database has been normalized to keep it between 1 and 10.

### Table 2. Summary of database characteristics

| Characteristic Code | Characteristic Description | Mean | Standard Deviation |
|---------------------|---------------------------|------|--------------------|
| I                   | Clump Thickness           | 4.4  | 2.8                |
| II                  | Uniformity of Cell Size   | 3.1  | 3.1                |
| III                 | Uniformity of Cell Shape  | 3.2  | 3.0                |
| IV                  | Marginal Adhesion         | 2.8  | 2.9                |
| V                   | Single Epithelial Cell Size | 3.2  | 2.2                |
| VI                  | Bare Nuclei               | 3.5  | 3.6                |
| VII                 | Bland Chromatin           | 3.4  | 2.4                |
| VIII                | Normal Nucleoli           | 2.9  | 3.1                |
| IX                  | Mitoses                   | 1.6  | 1.7                |

FTMBG algorithm that we have used in this paper is a frequent pattern mining algorithm that uses bi-graphs to find frequent itemsets. It works iteratively. First it finds size 1 itemsets, then size 2 itemsets and so on till no more itemsets exist in the database.

The results of FTMBG algorithm lead to dimensionality reduction. Then those reduced dimensions act as inputs to the next stage i.e. the Neural Network as shown in Figure 1.

A neural network takes many inputs. It constitutes of neurons that are inter-connected and the interconnections have weights. Non-linear activation functions are used by each neuron to produce output. The neural network uses past experiences to produce a final output
which in turn helps in decision making\cite{ref2}. Since using neural networks for cancer detection has been already researched in great detail we, in this paper, mainly focus on reducing the inputs of the neural network. We suggest two methods to reduce the inputs.

### 3.1 Method I

All rows and columns of the database are used to find relations between the input characteristics. Those association rules that pass the support and confidence thresholds are used to eliminate some input characteristics. For example if $X \rightarrow Y$ is such a rule, then $Y$ can be eliminated. The reduced input set is adequate for the ANN to produce appropriate results.

We kept confidence 100% and found the following rule (at value of each input=1).

**Input:** \{I, III, VIII, IX\} $\rightarrow$ \{II\}

This means if the characteristics I, III, VIII and IX have value 1, then the characteristic II also has value 1, which implies we can eliminate the characteristic II. Using method I we were successful in eliminating only one characteristic.

### 3.2 Method II

Large itemsets for every class are found out. Only those items are used for further consideration that is part of at least one of these large itemsets.

In our input database, there are two classes – benign and malignant. We found large itemsets of each class as follows.

Class Benign (Value=1):

**Input:** \{II, VIII, IX\}

Class Malignant (Value=10):

**Input:** \{VI\}

This implies that the characteristics II, VIII and IX are adequate to describe a benign class and the characteristic VI alone is adequate to describe a malignant class. So, we can use only these inputs for the ANN stage.

### Table 3. Multi-layer perceptron architecture

| No. of Layers | III |
|---------------|-----|
| No. of Neurons on the layers | Input: IV, VIII, IX |
|                | Hidden: XI |
|                | Output: I |
| Initial Weights/Biases | Randomized |
| Activation Functions | Tangent Sigmoid |
|                      | Tangent Sigmoid |
|                      | Linear |

### Table 4. Multi-layer perceptron training parameters

| Learning Rule                  | LevenbergMarquardt |
|-------------------------------|--------------------|
| Sum of Squared Error          | .01                |

### Table 5. Performance comparison

| Classification Technique Used                          | Epochs | Correctly Classified | Misclassified | Correct Classification Rate |
|--------------------------------------------------------|--------|----------------------|---------------|-----------------------------|
| Neural Network (9, 11, 1)                              | 56     | 202                  | 25            | 89.0                        |
| Neural Network (Input Controlled by Method I) (8, 11, 1)| 52     | 215                  | 12            | 94.7                        |
| Neural Network (Input Controlled by Method II) (4, 11, 1)| 39     | 209                  | 18            | 92.1                        |
In the next phase, a Multi-Layer Perceptron (MLP) is used which takes as input the characteristics obtained in the first phase i.e. the frequent pattern mining phase. MLP’s architecture and training parameters are shown in Tables 3 and 4.

4. Results

Table 5 shows the results of our experimentation on the Wisconsin database. The best classification was produced by the neural network when Method I inputs were given to it. The table shows that if our goal predominantly is to minimize the input parameters of the ANN then Method II should be preferred and if our main goal is to obtain the most appropriate results then Method I should be used. Another point worthy of consideration is that whether we use Method I or Method II the results are always better than when no frequent pattern mining is used.

5. Conclusion

In this paper, a method was proposed for diagnosis of breast cancer. This method uses the output of a frequent pattern miner as input to a neural network. The suggested method is focused on the reduction of characteristics in the input dataset. The experimental results show that not only were the inputs to neural network minimized but also that the classification was a better one. We proposed two methods for dimensionality reduction. One reduced the number of dimensions of our input database from 9 to 8 and another from 9 to 4.

6. References

1. World cancer report 2014. World Health Organization; 2014.

2. Chou SM, Lee TS, Shao YE, Chenb IF. Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines. Expert Systems with Applications. 2004; 27(1):133–42. Crossref.

3. Dillon WR, Goldstein M. Multivariate analysis methods and applications. Wiley: New York; 1984.

4. Hand DJ. Discrimination and classification. Wiley: New York; 1981.

5. Aragones JMJ, Ruiz JAG, Jimenez GR, Perez JM, Conejo EA. A combined neural network and decision trees model for prognosis of breast cancer relapse. Artificial Intelligence in Medicine. 2003; 27(1):45–63. Crossref.

6. Ryu YU, Chandrasekaran R, Jacob VS. Breast cancer prediction using the isotonic separation technique. European Journal of Operational Research. 2007; 181(2):842–54. Crossref.

7. Sahan S, Polat K, Kodaz H, Gunes S. A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer diagnosis. Computers in Biology and Medicine. 2007; 37(3):415–23. Crossref.

8. Krishnamurthy M, Manivannan K, Chilambuchelvan A, Rajalakshmi E, Kannan A. Enhanced Candidate Generation for Frequent Item Set Generation. Indian Journal of Science and Technology. 2015; 8(13):1–7. Crossref.

9. Tynchenko VS, Tynchenko VV, Bukhtoyarov VV, Tynchenko SV, Petrovskyi EA. The multi-objective optimization of complex objects neural network models. Indian Journal of Science and Technology. 2016; 9(29):1–11.

10. Bennett KP, Mangasarian OL. Robust linear programming discrimination of two linearly inseparable sets. Optimization Methods and Software. 1992; 1(1):23–34. Crossref.

11. Bishop CM. Neural networks for pattern recognition. Ist edn, Clarendon Press: Oxford; 2005.

12. Hanbay D, Turkoglu I, Demir Y. An expert system based on wavelet decomposition and neural network for modeling Chua’s circuit. Expert Systems with Applications. 2007; 34(4):2278–83. Crossref.