ORIGINAL ARTICLE

Iranian Quarterly Journal of Breast Disease 2015; 8(1).

Discovery of Hidden Patterns in Breast Cancer Patients Data Using Data Mining to Examine the Data with a Real Data Set

Alireza Atashi: Student Research Committee, Mashhad University of Medical Sciences, Mashhad, Iran. & Breast Cancer Research Center, ACECR, Tehran, Iran
Behzad Kiani: Student Research Committee, Mashhad University of Medical Sciences, Mashhad, Iran
Ebrahim Abbasi: Student Research Committee, Mashhad University of Medical Sciences, Mashhad, Iran. & Breast Cancer Research Center, ACECR, Tehran, Iran
Najmeh Nazeri: Breast Cancer Research Center, ACECR, Tehran, Iran

Corresponding Author: Behzad Kiani, kianib911@mums.ac.ir

Abstract

Introduction: The growing trend of breast cancer in recent years has shown the necessity of reliance on new and safe methods to diagnose and control of this disease. Data mining, as one of the most accurate and reliable techniques, has been frequently used to diagnose or predict various cancers. Data mining is one of this techniques that a well-known applications of is "Pattern Recognition" among large patient data sets. In this paper, the researches recognize and explore the unknown patterns in a real breast cancer data set using data mining algorithms.

Method: Due to excessive missing data in the collection, only data on 665 patients were available. Since the number of fields in the remaining records had null values, as one of the data preprocess and preparation phases, these values were estimated using the EM algorithm in SPSS20 software. Fields have been converted into discrete fields and finally the APRIORI algorithm has been used to analyze and explore the unknown patterns. After the rule extraction, an expert in the field of breast cancer eliminated redundant and meaningless relations.

Results: 100 association rules with a confidence value of more than 0.9 explored by the APRIORI algorithm and after the clinical expert feedback, 8 clinically meaningful relations have been detected and reported.

Conclusion: In this study, a number of less -known and interesting patterns have been extracted using the data mining algorithms. The use of data mining concepts, especially in medical data is very useful with considering the large volume of data and unknown relationships between causes of disease and the demographic characteristics. Due to the high number of risk factors, the use of these techniques is effective for cancer data. After pattern recognition these patterns or models provide the future study hypotheses and the next studies, for example RCTs will confirm or reject the hypotheses.

Keywords: Breast Cancer- Data Mining- Association Rules.