Medical Technologies Journal, Volume: 1, Issue: 4, October-December 2017, Pages:76-136. Doi : https://doi.org/10.26415/2572-004X-vol1iss4

CREATING A SEMI-SUPERVISED DATA MINING MODEL FOR PREDICTING BREAST CANCER RELAPSE

Type of article: conference abstract

Mahboobeh Zaremoayedi1, Zahra Mohammadi2, Mahdi Nasiri3*, Alireza Atashi4
1: MSc Student of Medical Informatics, School of Management and Medical Informatics, Shiraz University of Medical Sciences, Shiraz, Iran.
2: MSc Student of Medical Informatics, School of Management and Medical Informatics, Shiraz University of Medical Sciences, Shiraz, Iran.
3: PhD in Software Engineering, School of Management and Medical Informatics, Shiraz University of Medical Sciences, Shiraz, Iran.
4: PhD in Medical Informatics, Cancer Informatics Department, Breast Cancer Research Center, ACECR, Tehran, Iran.
*Tel: +98. 9126946632, E-mail: mn.nasiri@gmail.com

ABSTRACT
Introduction: Breast cancer is one of the most common types of cancer and malignancy in Iranian women,that recently has been a growing increase. there is always a possibility of recurrence In persons afflicted by thisdisease.In regarding to the complexity of analysis, data mining is among the best solutions that is used to detect or predict cancers.
Methods: In this retrospective study, data of 809 patients with breast cancer from center of breast cancer research of Tehran's Academic Center for Education, Culture and Research (ACECR) and 26 features from each patients were used. In regarding to high number of missing data in this collection, only information of 655 patients and 14 features of each patient were usable. Many features in records have null values, thus as one of data pre-processing and preparing phases, via Auto-Clustering algorithm, the data was divided into 10 clusters and according to dominant values in each cluster for each feature, these null features has been valuing. Data was divided to recurrence and non- recurrence classes. Semi-supervised method has been used in this study. the modeling was performed using labeled data and then a hybrid model for giving label to nonlabled data has been created. For this, data with recurrence lable divided proportional 30 percent for testing and 70 percent for training,and got to decision tree algorithms C5.0, Chaid, Quest, CRT, Autoclassifier as inputs. Then, the model was formed by mixing classifier algorithms with Confidence-Weighting-Voting method for predicting cancer recurrence, and used K-Fold (K=10) method for evaluating created model.
Results: The sensitivity of developed hybrid model was 82.93% and its specificity was 93.93%. The precision value of the model is 89.47% and its accuracy is 89.72%. this model mistakenly labelled only 10% of recurrence in patients of breast cancer as non-recurrence.
Conclusion: Creating predictive models with an appropriate sensitivity and specificity isimportant since, if the possibility of recurrence is high, could perform special preventive proceedings before its spreading. The false negative percentage is also important in medical prediction models, as it can have very dangerous consequences. In the prediction model presented in this study, the value of this parameter was 10%, and in this regard, this model can be considered acceptable.
KEYWORDS: Breast cancer, Data mining, Semi-Supervised Model, Cancer prediction

1. Declaration of conflicts

This abstract is selected from the First International Congress of Diseases and Health Outcomes Registry and First National Congress of Medical Informatics, 14-17 February 2017, Mashhad, Iran

2. Authors' biography

No biography.

3. References

No references.