ORIGINAL ARTICLE

Iranian Quarterly Journal of Breast Disease 2016; 9(2).

Imbalanced Data Classification for Primary Diagnosis of Breast Diseases by AdaBoost.M1, K-Nearest Neighbor and Probabilistic Neural Network

Darzi M: Advance Information System Research Group, Research Institute for ICT, ACECR, Tehran, Iran
Olfatbakhsh A: Breast Diseases Department, Breast Cancer Research Center, ACECR, Tehran, Iran
Gorgin S: Department of Electrical and Computer Engineering, Iranian Research Organization for Science and Technology (IROST), Tehran, Iran
Oveisi F: Advance Information System Research Group, Research Institute for ICT, ACECR, Tehran, Iran
Hashemi E: Breast Diseases Department, Breast Cancer Research Center, ACECR, Tehran, Iran
Alavi N: Breast Diseases Department, Breast Cancer Research Center, ACECR, Tehran, Iran

Corresponding Author: Mohammad Darzi, modarzi@yahoo.com

Abstract

Introduction: Breast Cancer is one of the common cancers in Iran. Each Prediagnosis of that can survive women from different risks. The aim of this research is classifying imbalanced dataset for detecting normal vs. abnormal women who came to ACECR Breast Cancer Clinic. Imbalanced datasets are one of the main challenges for designing medical decision support system. So, in this article, imbalanced data classification was addressed via data level solutions.

Methods: In this research for classifying of 918 women' breast situation, the "AdaBoost.M1", "K-nearest neighbor", and "probabilistic neural network" as triple algorithms were used. Because of facing with imbalanced dataset, for solving that, "random over sampling", "Random under sampling", and "Synthetic Minority Over-sampling Technique" were used as 3 re-sampling methods. So, Mat lab and R as software tools were used for implementing of methods and algorithms. Also, the values of 60 features that extracted from women's historical and physical exam forms were used as input data in triple algorithms. Finally, "precision" and "F-Measure" as two criteria were used for evaluating in test state of triple algorithms.

Results: Based on "precision" and "F-Measure" as two useful criteria, the best performance of this research's classification algorithms were through dataset that generated by Synthetic Minority Over-sampling Technique. So, the performance of "AdaBoost.M1", "K-nearest neighbor", and "probabilistic neural network" for classification of that dataset based on "precision" and "F-Measure" were "93.5,93.6", "79.5,87.7",and "86,91.9"respectively.

Conclusion: There are different methods for solving imbalanced datasets problem through classification of that. Re-Sampling is one of the popular data level methods. Through 3 re-sampling methods, the best classification algorithm performance belongs datasets that generated by "Synthetic Minority Over-sampling Technique", So among triple algorithms and four datasets that were used in this research and the based on "precision" and "F-Measure", AdaBoost.M1 had the best performance in classification.

Keywords: Imbalanced Dataset, Classification, Breast Diseases, AdaBoost.M1, K-NN, PNN, SMOTE.

[DOR: 20.1001.1.17359406.1395.9.2.1.5]