Journal of Information Technologies, Vol. 10, No. 1, 2017, ISSN: 1337-7469

Data mining applied to Breast Cancer Wisconsin data

Jan Bohacik*[1]*, Department of Informatics at the University of Zilina, Slovakia

Abstract: Breast cancer is belonging to one of the most prevalent cancers diagnosed in women nowadays. In this paper, several data mining methods are applied to classification of female patients with potential breast cancer into cancerous and non-cancerous. Employed Breast Cancer Wisconsin data are described and analyzed. Experimental results are provided showing the classification performance of particular methods with measures such as sensitivity, specificity, positive predictive value, negative predictive value and accuracy.

Key words: classification, data mining, breast cancer

1. Introduction

In the European Union, there are almost half a million new cases of breast cancer diagnosed annually (2) and related medical costs are about fifteen billion euros a year (5). The start of breast cancer is defined as growing cells in the breast out of control (1) and these cells usually form some mass of abnormal tissue. It is interesting to classify breast cancer masses so that benign ones and malignant ones are recognized early on. An automatic way for classification of masses is the use of classification data mining methods. Data mining finds patterns or trends with statistical pattern recognition and math that can predict patterns of patient data that already exist to classify new patient data (4). It is also an important step of the Knowledge Discovery in Databases process (3). Well-known data mining classification methods include Naive Bayes classifiers, decision tree classifiers, nearest neighbor classifiers and neural network classifiers. The methods require available data with known output status so that the data is used for creation of a pattern and so the diagnostic Breast Cancer Wisconsin data is employed.

The organization of the paper is as follows. The breast cancer data employed in the presented study of breast cancer classification is described in Section 2. In Section 3, classification models and experimental results are analyzed. The conclusion of the paper is in Section 4.

[1] Ján Boháčik with all Slovak diacritics.

2. Breast cancer data

As a dataset, the Wisconsin Breast Cancer data from the UCI Repository of Machine Learning Databases is used (6). This data was collected by Dr. William H. Wolberg (1989–1991) at the University of Wisconsin–Madison Hospitals. It has 699 instances (patients in set $\boldsymbol{V}$) classified into benign ones (non-cancerous) and malignant ones (cancerous). The description of the instances and their summary are in Table. 1 and Table. 2. Describing attributes $\boldsymbol{A}$ are defined as $\boldsymbol{A}$ = $\{A_1$; …; A_k; …; $A_9\}$ = {*Clump Thickness*; *Uniformity of Cell Size*; *Uniformity of Cell Shape*; *Marginal Adhesion*; *Single Epithelial Cell Size*; *Bare Nuclei*; *Bland Chromatin*; *Normal Nucleoli*; *Mitoses*}. The value of attribute A_k for some patient $\boldsymbol{p} \in \boldsymbol{V}$ is marked as $A_k(\boldsymbol{p})$. Class attribute C = *Severity* classifies patients into *benign* and *malignant*, which is denoted by C = $\{c_1; c_2\}$ = {*benign*; *malignant*}. The value of class attribute C for some patient $\boldsymbol{p} \in \boldsymbol{V}$ is marked as $C(\boldsymbol{p})$. There are 16 missing values for attribute A_6 and these are replaced with the median value of 1 computed for all patients without missing values.

Table. 1: Data description.

Attribute	**Data Type**	**Value Range**
Clump Thickness (A_1)	Ordinal	1, 2, 3, …, 10
Uniformity of Cell Size (A_2)	Ordinal	1, 2, 3, …, 10
Uniformity of Cell Shape (A_3)	Ordinal	1, 2, 3, …, 10
Marginal Adhesion (A_4)	Ordinal	1, 2, 3, …, 10
Single Epithelial Cell Size (A_5)	Ordinal	1, 2, 3, …, 10
Bare Nuclei (A_6)	Ordinal	1, 2, 3, …, 10
Bland Chromatin (A_7)	Ordinal	1, 2, 3, …, 10
Normal Nucleoli (A_8)	Ordinal	1, 2, 3, …, 10
Mitoses (A_9)	Ordinal	1, 2, 3, …, 10
Severity (C)	Categorical	*benign* (c_1)
		malignant (c_2)

Table. 2: Data analysis.

Attribute	**Median**	**Mode**	**Missing Values**
Clump Thickness (A_1)	4	1	0 (0%)
Uniformity of Cell Size (A_2)	1	1	0 (0%)
Uniformity of Cell Shape (A_3)	1	1	0 (0%)
Marginal Adhesion (A_4)	1	1	0 (0%)
Single Epithelial Cell Size (A_5)	2	2	0 (0%)
Bare Nuclei (A_6)	1	1	16 (2.29%)
Bland Chromatin (A_7)	3	2	0 (0%)
Normal Nucleoli (A_8)	1	1	0 (0%)
Mitoses (A_9)	1	1	0 (0%)
Severity (C)	N/A	*benign*	0 (0%)

3. Experiments

Four different classification methods are compared with 10 fold-cross validation on the breast cancer data described in Section II. In the experiments, sensitivity, specificity, positive predictive value, negative predictive value and accuracy are computed. Sensitivity is defined as tp/(tp + fn), specificity as tn/(tn + fp), positive predictive value is tp/(tp + fp), negative predictive value is tn/(tn + fn) and accuracy is (tp + fn)/(tp + fp + fn + tn). In the formulas, tp means true positive, fp denotes false positive, fn is false negative and tn is true negative. Malignant patients are considered to be positive ones and benign patients are labeled as negative ones. Sensitivity measures the ratio of malignant patients which are correctly identified as malignant ones. Specificity measures the ratio of benign patients which are correctly identified as benign ones. Positive predictive value is the probability that patients identified as malignant ones truly have the disease. Negative predictive value is the probability that patients identified as benign ones truly do not have the disease. Measure accuracy is defined as the ratio of correct identifications to all identifications made with the classification model.

Table. 3: Results.

Method	**SEN (%)**	**SPEC (%)**	**PPV (%)**	**NPV (%)**	**ACC (%)**
Bayes	97.5104	95.1965	91.4397	98.6425	95.9943
C4.5	91.7012	94.9782	90.5734	95.6044	93.8484
NN	91.7012	97.1616	94.4444	95.6989	95.2790
MLP	97.5104	95.1965	91.4397	98.6425	95.9943

Achieved results are presented in Table. 3 where columns represent computed measures and rows contain particular classification methods. SEN is sensitivity, SPEC is specificity in percentages, PPV is positive predictive value in percentages, NPV is negative predictive value in percentages and ACC is accuracy in percentages. Bayes denotes a Naive Bayes classifier implemented in Weka (7) as NaiveBayes, C4.5 is a decision tree classifier implemented in Weka as class J48, NN is a nearest neighbor classifier implemented in Weka as IBk and MLP is a neural network classifier implemented in Weka as class MultilayerPerceptron. According to the values in Table. 3, methods Bayes and MLP have the best results among the methods.

4. Conclusions

Four essentially different classification methods were employed on breast cancer data. Specifically, the following methods were considered: Naive Bayes classifier, decision tree, nearest neighbor classifier and neural network classifier. All of the methods achieved good results while the Naive Bayes classifier and neural network classifier had the best results with sensitivity 97.5104%, specificity 95.1965%, positive predictive value 91.4397%, negative predictive value 98.6425% and accuracy 95.9943%. Future investigations could include modifications or combinations of these methods and employment of fuzzy logic.

5. Bibliography

(1) American Cancer Society: Breast Cancer (Published by: American Cancer Society), Year: 2016.

(2) Bernard SW, Wild CP: World Cancer Report 2014 (Published by: International Agency for Research on Cancer, World Health Organization), Year: 2014.

(3) Fayyad, U., Piatetsky-Shapiro, G., Smyth, P.: From data mining to knowledge discovery in databases, AI Magazine (Volume: 17, Number: 3), Pages: 37-54, Year: 1996.

(4) Kurniawati YE, Permanasari AE, Fauziati S: Comparative study on data mining classification methods for cervical cancer prediction using pap smear results, Proceedings of the 1st International Conference on Biomedical Engineering, Pages: 1-5, Year: 2016.

(5) Luengo-Fernandez R, Leal J, Gray A, Sullivan R: Economic burden of cancer across the European Union: A population-based cost analysis, The Lancet Oncology (Volume: 14, Number: 12), Pages: 1165-1174, Year: 2013.

(6) Mangasarian OL, Street WN, Wolberg WH: Breast cancer diagnosis and prognosis via linear programming, Operations Research (Volume: 43, Number: 4), Pages: 570-577, Year: 1995.

(7) Witten IH, Frank E, Hall MA: Practical machine learning tools and techniques (3rd edition). (Published by: Morgan Kaufman Publishers, Published in: Burlington, MA, USA), Pages: 664, Year: 2011.

6. Address of the author:

doc. Ing. Jan Bohacik, PhD.
Department of Informatics
Faculty of Management Science and Informatics
University of Zilina
Univerzitna 8215/1
010 26 Zilina
SLOVAKIA
Jan.Bohacik@fri.uniza.sk

The article was accepted for publication in August 2017 by the publisher of the Journal of Information Technologies (Vol. 10, No. 1, ISSN: 1337-7467), i.e. by the Faculty of Mass Media Communication, University of Ss. Cyril and Methodius in Trnava, Slovakia.