BREAST CANCER CLASSIFICATION USING K-NEAREST NEIGHBORS ALGORITHM

Can EYUPOGLU

Istanbul Commerce University, Department of Computer Engineering, Istanbul-Turkey

ceyupoglu@ticaret.edu.tr

Abstract: Breast cancer is a very common disease around the world and the second leading cause of cancer death in women. Expert systems are developed using data mining methods in order for disease diagnosis and significant tools for assisting medical doctors in their clinical decision. In this study, *k*-Nearest Neighbors algorithm (*k*-NN) was used in order to classify breast cancer disease. Besides, *k*-NN was implemented for different *k* values and the obtained classification accuracies were compared with each other. According to the study results, it was seen that breast cancer disease was successfully classified using *k*-NN.

Keywords: Breast Cancer, Classification, Data Mining, Expert Systems, *k*-Nearest Neighbors Algorithm

Introduction

According to the recent studies, breast cancer is the most common cancer in young women (Jelen et al., 2016). Among all cancers, it causes the second highest number of deaths around the world (Mittal et al., 2017; DeSantis et al., 2014; Misek and Kim, 2011). Nearly 1.7 million women suffer from breast cancer every year (Tang et al., 2016). In 2012, breast cancer was the reason for 18.3% of all cancer cases in Egypt (Abdel-Zaher and Eldeib, 2016; Salama et al., 2012). In 2015, approximately 481 women were diagnosed with breast cancer every week in Canada according to the statistics of the Breast Cancer Society of Canada (2015).

Disease diagnosis is a very sophisticated process in medicine. In order to diagnose a disease accurately, several tests are required. Computer-aided diagnostic tools, expert systems, assist physicians to make primary decision for early diagnosis. Thanks to early detection of a disease, treatment process can be minimized and patients' lives may be saved. Especially in breast cancer, medical doctors desire to know the condition of the patient early who has benign or malignant case. Computer-aided diagnostic tools classify benign or malignant cases successfully in order for detection of breast cancer disease (Jelen et al., 2016). In this study, *k*-Nearest Neighbors algorithm (*k*-NN) is utilized to classify breast cancer disease as benign or malignant.

The rest of the paper is organized as follows. Materials and Methods Section introduces the classification of breast cancer disease using *k*-NN and the dataset used. In Results and Discussion Section, the experiments including classification accuracies and error values of the methods used in this study are demonstrated. Finally, conclusions being under study are summarized in Conclusion Section.

Materials and Methods

In this section, the dataset of breast cancer diagnosis used in this study is described. Then, the algorithm used for classification of breast cancer disease is explained.

Dataset Description

In order to classify breast cancer disease, Wisconsin Breast Cancer Database was used. The dataset was created by Dr. William H. Wolberg from the University of Wisconsin (Wolberg, 1991). It is available online in Machine Learning Repository at the University of California - Irvine (Lichman, 2013). The dataset contains 699 instances. There are 11 attributes which are sample code number, clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli, mitoses and class (benign or malignant). 458 (65.5%) instances are benign and 241 (34.5%) instances are malignant. There are 16 instances containing a single missing attribute value. The description of the attributes is shown in Table 1.

Table 1. Attributes of the Wisconsin Breast Cancer Database.

Attribute #	Attribute name	Domain	Mean	Standard deviation
1	Sample code number	Id number	-	-
2	Clump thickness	1 - 10	4.418	2.816
3	Uniformity of cell size	1 - 10	3.134	3.051
4	Uniformity of cell shape	1 - 10	3.207	2.972
5	Marginal adhesion	1 - 10	2.807	2.855
6	Single epithelial cell size	1 - 10	3.216	2.214
7	Bare nuclei	1 - 10	3.545	3.644
8	Bland chromatin	1 - 10	3.438	2.438
9	Normal nucleoli	1 - 10	2.867	3.054
10	Mitoses	1 - 10	1.589	1.715
11	Class	2 for benign 4 for malignant	-	-

Classification with *k*-Nearest Neighbors Algorithm

k-NN is a widely used pattern classification technique because of its simplicity and efficiency (Eyupoglu, 2016; Zhang et al., 2012; Wang et al., 2007). Furthermore, *k*-NN, a versatile multivariate statistical method, utilizes standard Euclidean distance and evaluates the distinguishing features. It makes no assumption with regard to the statistical structure of the data (Niwas et al., 2013; Shakhnarovish et al., 2005).

k-NN estimates class attribute depending the *k* nearest training examples in the feature space. When a dataset is given, it chooses the *k* nearest samples from the classified training data and determines the class considering the most representative samples. Euclidean distance similarity metric is used to select the neighborhoods and calculated using Eq. (1) as follows (Shen et al., 2016):

$$\text{Euclidean distance} = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2} \quad (1)$$

where x_i and y_i are two points in Euclidean *n*-space. After all test samples are classified by *k*-NN, the classification accuracy is calculated with dividing the number of correctly classified samples by the total number of samples. Mean absolute error (MAE) is calculated according to the following Eq. (2) as follows (Willmott and Matsuura, 2005):

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i - x_i| \quad (2)$$

where y_i is the prediction value and x_i is the real value. Root mean square error (RMSE) is calculated using Eq. (3) as follows (Willmott and Matsuura, 2005):

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n} |y_i - x_i|^2} \quad (3)$$

Results and Discussion

In this paper, *k*-NN was used to classify breast cancer disease and implemented for different k-fold cross-validation and *k* values. Then, the obtained classification accuracies were compared with each other. The application used for classification was implemented using Weka 3.8 software. Weka (Waikato Environment for Knowledge Analysis) was developed by Machine Learning Group at the University of Waikato in New Zealand. It is a collection of machine learning algorithms and data preprocessing tools for data mining tasks (Witten et al., 2011). The tests were performed for 2-fold, 5-fold and 10-fold cross-validation and *k* values were ranging from 1 to 20. The classification accuracies of *k*-NN for different k-fold cross-validation and *k* values were presented in Table 2.

Table 2. Classification accuracies of *k*-NN for different k-fold cross-validation and *k* values.

k	2-fold cross validation	5-fold cross validation	10-fold cross validation
	Classification accuracy (%)	Classification accuracy (%)	Classification accuracy (%)
1	95.8512	95.4220	95.1359
2	94.5637	94.5637	94.4206
3	96.8526	96.7096	96.8526
4	95.9943	96.4235	96.7096
5	96.1373	96.5665	96.9957
6	96.1373	96.2804	96.7096
7	96.4235	96.2804	96.7096
8	96.2804	96.1373	96.1373
9	96.4235	96.2804	96.2804
10	96.2804	95.9943	96.4235
11	96.1373	96.2804	96.2804
12	96.1373	96.1373	96.2804
13	96.1373	96.2804	96.1373
14	96.1373	96.2804	96.1373
15	96.1373	96.4235	96.4235
16	95.9943	96.4235	96.7096
17	96.1373	96.4235	96.5665
18	95.9943	96.4235	96.4235
19	96.1373	96.4235	96.5665
20	95.8512	96.4235	96.4235

As seen from Table 2, the classification accuracies of *k*-NN range from 94.4206% to 96.9957%. For 2-fold, 5-fold and 10-fold cross validation techniques, the classification accuracies vary between 94.5637% - 96.8526%, 94.5637% - 96.7096% and 94.4206% - 96.9957%, respectively. The best classification accuracies of 2-fold, 5-fold and 10-fold cross validation techniques are obtained for k=3, k=3 and k=5, and these rates are 96.8526%, 96.7096% and 96.9957%, respectively. As a result, the classification accuracy of nearly 97% was achieved using *k*-NN. Moreover, the error values of *k*-NN for different k-fold cross-validation and *k* values were shown in Table 3.

Table 3. Error values of *k*-NN for different k-fold cross-validation and *k* values.

k	2-fold cross validation		5-fold cross validation		10-fold cross validation	
	MAE	RMSE	MAE	RMSE	MAE	RMSE
1	0.0441	0.2031	0.0474	0.2136	0.0501	0.2202
2	0.0435	0.1722	0.0452	0.1813	0.0444	0.1763
3	0.0447	0.1634	0.0463	0.1719	0.0444	0.1672
4	0.0493	0.1648	0.0469	0.1659	0.0454	0.1659
5	0.0523	0.1703	0.0467	0.1594	0.0458	0.1596
6	0.0522	0.1681	0.0480	0.1600	0.0460	0.1559
7	0.0525	0.1668	0.0491	0.1605	0.0472	0.1566
8	0.0541	0.1692	0.0506	0.1630	0.0481	0.1584
9	0.0535	0.1675	0.0514	0.1630	0.0491	0.1595
10	0.0536	0.1684	0.0526	0.1660	0.0503	0.1585
11	0.0545	0.1692	0.0532	0.1665	0.0508	0.1603
12	0.0551	0.1713	0.0540	0.1653	0.0510	0.1612
13	0.0554	0.1715	0.0542	0.1648	0.0522	0.1629
14	0.0557	0.1719	0.0539	0.1641	0.0526	0.1622
15	0.0556	0.1723	0.0543	0.1650	0.0526	0.1617
16	0.0552	0.1719	0.0545	0.1657	0.0524	0.1617
17	0.0559	0.1734	0.0553	0.1672	0.0530	0.1628
18	0.0565	0.1748	0.0549	0.1670	0.0531	0.1625
19	0.0568	0.1756	0.0554	0.1672	0.0537	0.1643
20	0.0563	0.1754	0.0555	0.1678	0.0541	0.1653

According to Table 3, MAE and RMSE values range between 0.0435 - 0.0568 and 0.1559 - 0.2202. For 2-fold, 5-fold and 10-fold cross validation techniques, MAE and RMSE values vary between 0.0435 - 0.0568, 0.0452 - 0.0555 and 0.0444 - 0.0541; 0.1634 - 0.2031, 0.1594 - 0.2136 and 0.1559 - 0.2202, respectively. The minimum MAE values of 2-fold, 5-fold and 10-fold cross validation techniques are obtained for k=2. For RMSE, the minimum values are attained for k=3, k=5 and k=6, respectively. Consequently, the error values of 0.0435 and 0.1559 for MAE and RMSE were procured by *k*-NN. The change of error values with the increase of *k* values can be easily observed from Figure 1.

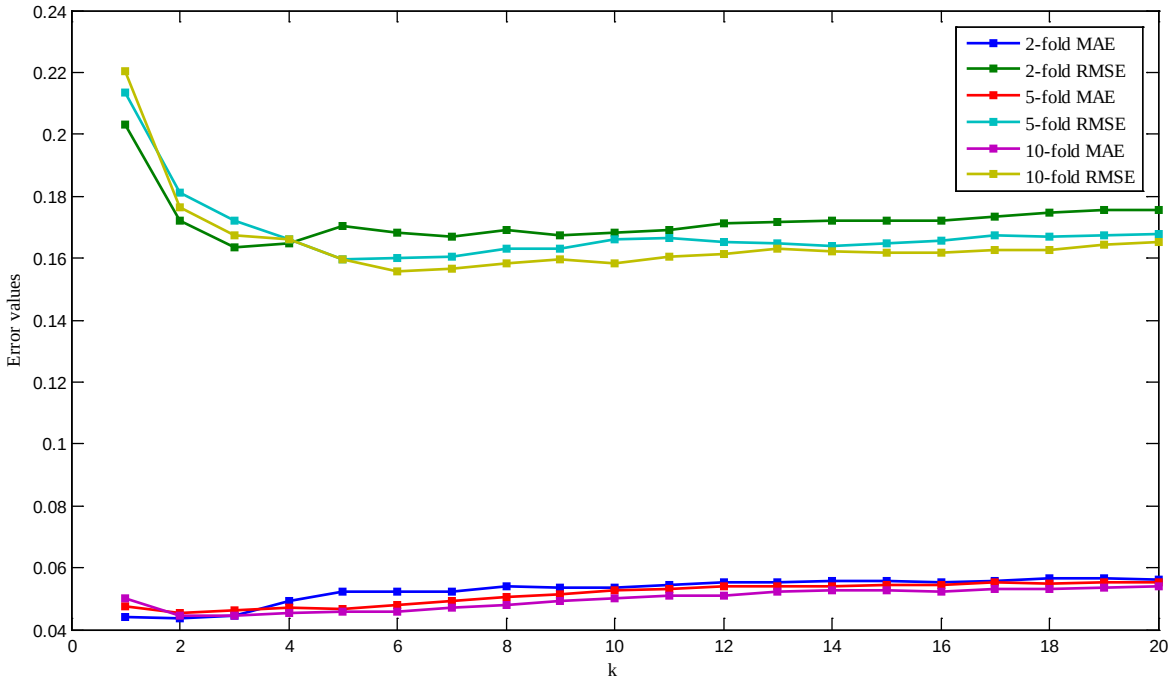

Figure 1. Change of error values with the increase of *k* values.

Conclusion

In this paper, in order to classify breast cancer disease as benign or malignant, *k*-NN was utilized. It was implemented for 2-fold, 5-fold and 10-fold cross validation and different *k* values. The classification accuracies and error values were attained to assess the success of *k*-NN. According to the test results, the achieved classification accuracy of *k*-NN is approximately 97%. Besides, the study results show that *k*-NN is an effective classifier in order for classifying breast cancer disease.

References

Abdel-Zaher, A.M., & Eldeib, A.M. (2016). Breast cancer classification using deep belief networks. *Expert Systems With Applications, 46*, pp. 139-144.

Breast cancer society of Canada incidence statistics for 2015. (2015). <http://www.cbcf.org/central/AboutBreastCancerMain/FactsStats/Pages/Breast-Cancer-Canada.aspx> Accessed 15.10.15.

DeSantis, C., Ma, J., Bryan, L., & Jemal, A. (2014). Breast cancer statistics, 2013. *CA: A Cancer Journal for Clinicians, 64*(1), pp. 52-62.

Eyupoglu, C. (2016). Implementation of color face recognition using PCA and k-NN classifier. *2016 IEEE NW Russia Young Researchers in Electrical and Electronic Engineering Conference (EIConRusNW)*, pp. 199-202, St. Petersburg, Russia.

Jeleń, Ł., Krzyżak, A., Fevens, T., & Jeleń, M. (2016). Influence of feature set reduction on breast cancer malignancy classification of fine needle aspiration biopsies. *Computers in Biology and Medicine, 79*, pp. 80-91.

Lichman, M. (2013). UCI Machine Learning Repository, University of California, School of Information and Computer Science, Irvine, California, USA.

Misek, D.E., & Kim, E.H. (2011). Protein Biomarkers for the Early Detection of Breast Cancer. *International Journal of Proteomics, 2011*, pp. 1-9.

Mittal, S., Kaur, H., Gautam, N., & Mantha, A.K. (2017). Biosensors for breast cancer diagnosis: A review of bioreceptors, biotransducers and signal amplification strategies. *Biosensors and Bioelectronics, 88*, pp. 217-231.

Niwas, S.I., Palanisamy, P., Sujathan, K., & Bengtsson, E. (2013). Analysis of nuclei textures of fine needle aspirated cytology images for breast cancer diagnosis using Complex Daubechies wavelets. *Signal Processing, 93*, pp. 2828-2837.

Salama, G. I., Abdelhalim, M.B., & Zeid, M.A. (2012). Breast Cancer diagnosis on three different datasets using multi-classifiers. *International Journal of Computer and Information Technology, 1*, pp. 36-43.

Shakhnarovish, G., Darrell, T., & Indyk, P. (2005). *Nearest-Neighbor Methods in Learning and Vision*, MIT Press.

Shen, L., Cao, D., Xu, Q., Huang, X., Xiao, N., & Liang, Y. (2016). A novel local manifold-ranking based K-NN

for modeling the regression between bioactivity and molecular descriptors. *Chemometrics and Intelligent Laboratory Systems*, *151*, pp. 71-77.

Tang, Y., Wang, Y., Kiani, M.F., & Wang, B. (2016). Classification, Treatment Strategy, and Associated Drug Resistance in Breast Cancer. *Clinical Breast Cancer*, *16*(5), pp. 335-343.

Wang, J., Neskovic, P., & Cooper, L.N. (2007). Improving nearest neighbor rule with a simple adaptive distance measure. *Pattern Recognition Letters*, *28*, pp. 207-213.

Willmott, C.J., & Matsuura, K. (2005). Advantages of the mean absolute error (MAE) over the root mean square error (RMSE) in assessing average model performance. *Climate Research*, *30*, pp. 79-82.

Witten, I.H., Frank, E., & Hall, M.A. (2011). *Data Mining: Practical Machine Learning Tools and Techniques (Third Edition). Chapter 10 – Introduction to Weka*, pp. 403-406, Elsevier, Morgan Kaufmann Publishers.

Wolberg, W.H. (1991). Wisconsin Breast Cancer Database, University of Wisconsin Hospitals, Madison, Wisconsin, USA.

Zhang, N., Yang, J., & Qian, J.J. (2012). Component-based global k-NN classifier for small sample size problems. *Pattern Recognition Letters*, *33*(13), pp. 1689-1694.

