AWERProcedia
Information Technology
& Computer Science

Vol 04 (2013) 401-407

3rd World Conference on Innovation and Computer Sciences 2013

# Breast Cancer Diagnosis Based on Naïve Bayes Machine Learning Classifier with KNN Missing Data Imputation

**Ceren Güzel *,** Gazi University, Faculty of Engineering, Computer Engineering Department, 06570, Ankara, Turkey.

**Mahmut Kaya,** Gazi University, Faculty of Engineering, Computer Engineering Department, 06570, Ankara, Turkey.

**Oktay Yıldız,** Gazi University, Faculty of Engineering, Computer Engineering Department, 06570, Ankara, Turkey.

**Hasan Şakir Bilge,** Gazi University, Faculty of Engineering, Computer Engineering Department, 06570, Ankara, Turkey.

**Suggested Citation:**

Güzel C., Kaya M. & Yıldız O. Hasan Şakir Bilge. Breast Cancer Diagnosis Based on Naïve Bayes Machine Learning Classifier with KNN Missing Data Imputation. *AWERProcedia Information Technology & Computer Science*. [Online]. 2013, 04, pp 401-407. Available from: www.awer-center.org/pitcs

Received December 25, 2012; revised January 18, 2013; accepted March 16, 2013.
Selection and peer review under responsibility of Prof. Dr. Fahrettin Sadıkoglu, Near East University.

## Abstract

Cancer is one of the most mortal diseases in the world. Breast cancer is the second leading cause of death in women. Mammography is a method which is used to detect breast cancer in the initial stage. It helps physicians about in their decisions whether biopsy is necessary or not with respect to tissue shape, border and density. According to researches, 70% of biopsies have done without a need. Because of its cost and complications, it is essential to decide whether biopsy is necessary or not. To achieve this aim, lots of machine learning algorithms are developed to help medical diagnosis in literature, but data sets include some missing values in many real world tasks. These missing values adversely affect classifier performance. Our approach is to impute missing values with k Nearest Neighbor algorithm (kNN) and Naïve Bayes. Then, the performance of the system is evaluated by kNN and Naïve Bayes classifiers to detect breast cancer. Our proposal is measured by performance criteria such as accuracy, sensitivity, specificity and ROC analysis. With this approach, 95 out of 131 missing data which is 9.89% of all data are filled. The experimental results on Mammographic Mass database demonstrate the effectiveness of our proposal with 82.49% accuracy while 81.69% accuracy is obtained without any imputation using same

* ADDRESS FOR CORRESPONDENCE: **Ceren Güzel**, Gazi University, Faculty of Engineering, Computer Engineering Department, 06570, Ankara, Turkey, *E-mail Address*: cerenguzel@gazi.edu.tr / Tel.: +90-312-582-3119

training and test sets. It can be concluded that this approach can also be used for other medical diagnosis problems with high accuracy.

Keywords: Breast cancer, missing data imputation, kNN, Naïve Bayes, Mammographic Mass;

## 1. Introduction

Cancer is one of the leading causes of death, exceeded only by diseases of the circulatory system. According to data taken from 40 European countries, new 3.45 billion cancer cases and 1.75 billion deaths are predicted in 2012 [1]. Data taken from 2008 show that there are about 452,5 cancer cases in every 100.000 people [2]. According to statistics from WHO, 7.6 billion people were dead because of cancer in 2008. Moreover, breast cancer mortality rate was 458.000 equivalently 13 percent of all deaths [3]. Because of this fact, early and true diagnosis is an important issue and plays a key role to cure this disease.

Mammography is at the heart of diagnosis of breast cancer [4]. But, investigations show us these surgical interventions are done while there is no need in the range of 65% and 80% patients. In addition, these interventions cause side effects in patients and specialists. Computer based diagnostic systems are composed to minimize the number of biopsy and to contribute specialist evolutions about disease. These systems control mammography findings and make a decision about biopsy is necessary or not [5].

Economical and social importance of early and true diagnosis of cancer bring into open a question about machine learning methods whether they can be used for diagnosis of cancer or not. According to researches, machine-learning approaches have 91.10% success in diagnostic detection while experienced specialists have 79.97% correct diagnosis detection [6]. The purpose of using machine learning techniques is to minimize mistakes in diagnosis, which is done by specialists.

Data sets are as important as learning algorithms in machine learning. Data sets include missing values because of wrong measurements and empty features which is generated during collecting data. Removing missing values is the most common and frequently used way in machine learning. However, we need as much as possible training data to generate and examine the classification model correctly. While approach of removing missing data can be used in big scale data set, which include few missing data, it cannot be used correctly because of all data have critical importance to evaluation in small scale data set [7]. Due to this reason, missing data imputation concept is proposed. More valid analysis can be done using this concept.

A considerable amount of literature has been published on detection of breast cancer using machine learning techniques. Albrecht et al. proposed a hybrid method, which combines perceptron, and logarithmic simulated annealing algorithms [8]. Abbass examined memetic pareto artificial neural network which is based on pareto differential evaluation algorithms [9]. Abonyi and Szeifert developed a classifier based on fuzzy aggregation [10]. Şahan et al. proposed a hybrid method which is combination of fuzzy artificial immune system and kNN algorithms. Training data is weighted by fuzzy system and extracted by artificial immune system (AIS) and then classified with kNN algorithm [11]. Polat and Güneş analysed least square support vector machine classifier in Wisconsin data set [12]. Guijarro-Berdias et al. presented a new method based on linear least squares [13]. Akay identified best features with F-score then generated 9 models [14]. Subashini et al. normalized all features between -1 and +1 in database. They found hidden layer weights using k-means supervised algorithm and predicted output layer weights in RBFNN [15]. Peng et al. conducted feature characterization, feature pre-selection using filter method, wrapper feature selection and feature research using sequential forward floating search [16]. Conforti and Guido optimized kernel function using semi-definite programming and combined SVM and this approach in classification [17]. Fan et al. developed a hybrid method combining data aggregation based on rules and fuzzy decision tree [18]. Marcano-

Cedeno et al. inspired neurons biologic meta-plasticity feature and Shannon's information theory and generated a new ANN classifier [19].

In this study, missing data are imputed apart from literatures, which eliminate them. After imputation training and testing dataset are identified then classification is occurred as seen in Fig. 1.

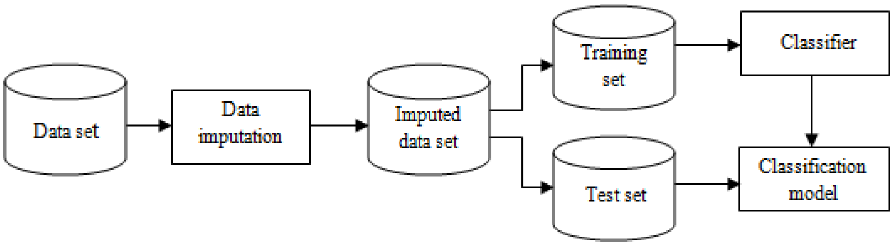

Figure1. Flow chart of proposed approach

## 2. Material and Method

### *2.1. Mammographic Mass Data Set*

Mammographic Mass data set, based on BI-RADS (Breast Imaging Reporting and Data System) features, was used to classify mammographic mass to be benign or malign [20]. Each sample in database has BIRADS features, age, shape, margin, density and severity features as seen in Table 1.

Table1. Features of Mammographic Mass data set

| Feature | Feature Description | Value Range | Mean | Standard Deviation | Missing Data (%) |
|---|---|---|---|---|---|
| 1 | BI-RADS | 0 - 5 | 4.35 | 1.78 | 0.21 |
| 2 | Age | 18 - 96 | 55.49 | 0 | 0.52 |
| 3 | Shape | 1 - 4 | 2.72 | 1.24 | 3.23 |
| 4 | Margin | 1 - 5 | 2.80 | 1.57 | 4.99 |
| 5 | Density | 1 - 4 | 2.91 | 0.38 | 7.91 |
| 6 | Severity | 0 - 1 | 0.46 | 0.50 | 0 |

BI-RADS is developed to standardize terminology between radiologist and doctors. BI-RADS assessment is done based on age, shape, margin density. BI-RADS assessment contains 6 categories as seen in Table 2. Severity feature is class label 0 (benign) or 1 (malignant). 510 of 961 samples are classified being benign. Rests of 961 are classified being malignancy among samples. In mammographic database, 162 data include missing values.

Table2. BI-RADS assessment

| BI-RADS | Description | Risk of Malignancy | Recommendation |
|---|---|---|---|
| 0 | Incomplete assessment | N/A | Workup |
| 1 | Negative | N/A | Mammogram in 1 year |
| 2 | Benign | N/A | Mammogram in 1 year |
| 3 | Malignancy | < 2% | Mammogram in 6 mounts |
| 4 | Suspicious | 20% | Biopsy |
| 5 | Probably malignancy | 90% | Medical intervention |
| 6 | Malignancy | N/A | Medical intervention |

### *2.2. Machine Learning Techniques*

In breast cancer detection, various machine learning algorithm are used as computer based diagnostic systems. K nearest neighbor and Naïve Bayes are two of the most common machine learning techniques.

#### *2.2.1. K nearest neighbor*

K nearest neighbor is a basic and wellknown classification method [21]. In k nearest neighbor method, distance between testing data point $x$ and training data points $x_i$, $i = 1,\ldots,n$ are calculated.

$$d_E(x, x_i) = \sum_{i=1}^{n} |x - x_i| \tag{1}$$

$$x : d_E(x, x_i) < d_E(x, x_j), i \neq j \tag{2}$$

Nearest k points are determined. Testing data points are classified with respect to specified k nearest neighbors.

#### *2.2.2. Naïve Bayes*

Naïve Bayes is a statistical classification algorithm based on Bayes theorem. In training stage, using class probabilities and conditional probabilities, class label of testing data point is estimated. For two classed data set, data point is classified with respect to which class probability is higher [22].

In Bayes theorem, posterior probability of samples in c class are calculated using (3). Since features are independent, Naïve Bayes classifier are defined as seen in (4). $p(x_1,\ldots,x_n)$ can be ignored from (4) because it is same for all samples.

$$p(c \mid x_1,\ldots,x_n) = \frac{p(x_1,\ldots,x_n \mid c)\, p(c)}{p(x_1,\ldots,x_n)} \tag{3}$$

$$p(c \mid x_1,\ldots,x_n) = \frac{p(c) \prod_{i=1}^{n} p(x_i \mid c)}{p(x_1,\ldots,x_n)} \tag{4}$$

### *2.3. Missing Data Imputation*

Missing value imputation methods are firstly developed by using statistical algorithms. These methods have a range from basic imputation techniques like mean imputation to complex methods like likelihood parameter estimation. Nowadays, machine-learning algorithms begin to be used to impute missing values [7, 23]. Apart from statistical approaches, missing values are imputed using classification. Naïve Bayes and kNN are widely used machine learning methods in missing value imputation.

In Naïve Bayes missing value imputation, each feature is trained as a class feature including missing values. Firstly, conditional and prior probabilities are calculated during training phase. Then, each feature having missing values is imputed using $\arg\max_c\, p(c \mid z_1,\ldots,z_n) = \arg\max_c\, p(c) \prod_{i=1}^{n} p(z_1,\ldots,z_n \mid c)$ equations.

In kNN missing value imputation, missing values are imputed using similar cases in missing values. Assuming $x$ is a missing value in feature $j$. For finding x missing value, nearest k neighbours, $v = \{v_k\}_{k=1}^{K}$ not having missing value are found. Using $v$, feature $j$ including nearest neighbours, missing $j$ feature value is found. In categorical data, common technique is $\{v_{kj}\}_{k=1}^{K}$. Another approach is as an $\lambda_k$ in $v_k$ identifying a decision scheme.

$$\lambda_k = \begin{cases} \dfrac{d(v_K,x)-d(v_k,x)}{d(v_K,x)-d(v_1,x)}, \text{ if } d(v_k,x) \neq d(v_1,x) \\ 1, \text{ else.} \end{cases} \qquad (5)$$

Using this method $\bar{x}_j$ is imputed with respect to weights having biggest sum among k nearest neighbors.

## 3. Experimental Results

95 of 131 missing values in Mammographic Mass data set are imputed by kNN and Naïve Bayes approaches. To examine and validate Naïve Bayes and kNN classifiers on imputed data, 10 fold cross validation is used.

Table3. Performance of kNN and Naïve Bayes classifiers without any imputation

| Classifier | Specificity | Sensitivity | Accuracy (%) |
|---|---|---|---|
| Naïve Bayes | 0.79 | 0.84 | **81.69** |
| kNN (k=1) | 0.78 | 0.74 | 75.90 |
| kNN (k=3) | 0.80 | 0.78 | 79.40 |
| kNN (k=5) | 0.80 | 0.80 | 80.24 |
| kNN (k=7) | 0.80 | 0.82 | 80.96 |
| kNN (k=9) | 0.78 | 0.81 | 79.76 |

Table4. Performance of kNN and Naïve Bayes classifiers with imputation methods

| Value Imputation | Classifier | Specificity | Sensitivity | Accuracy (%) |
|---|---|---|---|---|
| Naïve Bayes | Naïve Bayes | 0.80 | 0.84 | 81.73 |
| Naïve Bayes | kNN (k=1) | 0.78 | 0.74 | 76.22 |
| Naïve Bayes | kNN (k=3) | 0.81 | 0.77 | 78.92 |
| Naïve Bayes | kNN (k=5) | 0.81 | 0.80 | 80.11 |
| Naïve Bayes | kNN (k=7) | 0.79 | 0.82 | 80.43 |
| Naïve Bayes | kNN (k=9) | 0.78 | 0.80 | 79.14 |
| kNN (k=3) | Naïve Bayes | 0.81 | 0.84 | **82.49** |
| kNN (k=3) | kNN (k=1) | 0.78 | 0.71 | 74.70 |
| kNN (k=3) | kNN (k=3) | 0.81 | 0.78 | 79.14 |
| kNN (k=3) | kNN (k=5) | 0.81 | 0.78 | 79.57 |
| kNN (k=3) | kNN (k=7) | 0.82 | 0.79 | 80.43 |
| kNN (k=3) | kNN (k=9) | 0.81 | 0.79 | 79.68 |

To analyze proposed approach, accuracy, specificity, sensitivity and ROC analysis were conducted. Without any imputation (extracted missing value) obtained accuracy, specificity and sensitivity scores are shown in Table 3. According to classification results obtained after imputation in Table 4, best classification (82.49%) is obtained using kNN (k=3) missing value imputation approach and Naïve Bayes classifier. Using proposed approach, classification accuracy increased from 81.69% to 82.49%.

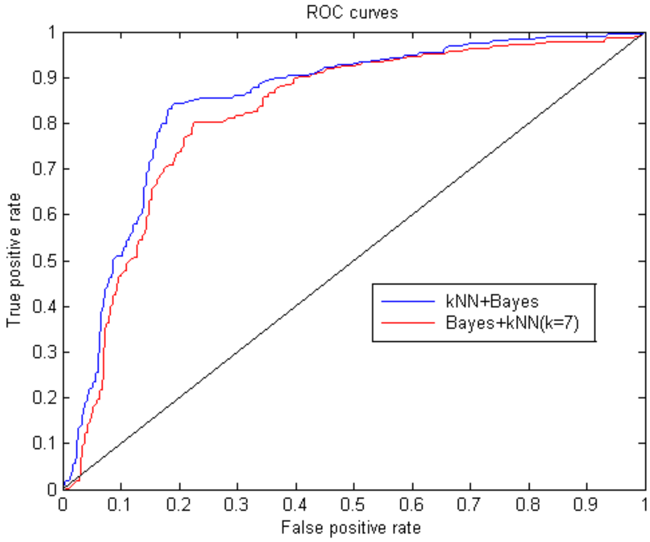

Figure2. ROC Curves with kNN and Naïve Bayes missing value imputation

According to ROC analysis, the highest area under ROC curve is 0.8260 with kNN imputation and Naïve Bayes classification, while area under curve is 0.7772 using Naïve Bayes imputation and kNN (k=7) classification as seen in Figure 2. The results of this study indicate that successful classification is achieved using kNN missing value imputation and Naïve Bayes classification.

## 4. Conclusion

In this study, computer detection system based on Naïve Bayes and kNN approaches is represented for breast cancer. Since Mammographic Mass data set includes missing values and has not enough samples for correct diagnosis, missing value imputation techniques were used instead of removing these values. The following conclusions can be drawn from the present study that classification accuracy increase after imputation. The findings of this study suggest that making more correct diagnosis can be possible using different missing data imputation approaches.

## References

International Agency for Research on Cancer. 2013 Feb. (Last Modified), Cited 2013 April 10, from: http://www.iarc.fr/en/mediacentre/iarcnews/pdf/EUCAN EJC Feb2013.pdf

Türkiye Halk Sağlığı Kurumu Kanser Daire Başkanlığı. 2013 Jan. 23 (Last Modified), Cited 2013 April 10, from http://kanser.gov.tr/index.php/daire-faaliyetleri/kanser-istatistikleri.html

World Health Organization website. 2013 Jan. (Last Modified), Cited March 13, from http://www.who.int/mediacentre/factsheets/fs297/en/index.html

Türkiye Halk Sağlığı Kurumu Kanser Daire Başkanlığı. 2013 Feb. 6 (Last Modified), Cited April 10, form http://kanser.thsk.gov.tr/Dosya/Bilgi-Dokumanlari/raporlar/mamografi.pdf

Elter M., Schulz-Wendtland R., Wittenberg T. (2007). The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process, Medical Physics, Volume 34, 41-64.

Brause R. W. (2001). Medical Analysis and Diagnosis by Neural Networks, ISMDA '01 Proceedings of the Second International Symposium on Medical Data Analysis, 1-13

Farhangfar A., Kurgan L., Dy. J. (2008). Impact of imputation of missing values on classification error for discrete data, Pattern Recognition Letters, Volume 41, 3692-3705.

Albrecht A. A., Lappas G., Vinterbo S. A., Wong C. K., Ohno-Machado L. (2002). Two applications of the lsa machine, Proceeding of the 9th International Conference on Neural Information Processing (ICONIP'02), Volume 1, 184-189.

Abbass H. A. (2002). An evolutionary artificial neural networks approach for breast cancer diagnosis, Artificial Intelligence in Medicine, Volume 25, 265-281.

Abonyi J., Szeifert F. (2003). Supervised fuzzy clustering for the identification of fuzzy classifiers, Pattern Recognition Letters, Volume 14, 2195-2207.

Şahan S., Polat K., Kodaz H., Güneş S. (2007). A new hybrid method based on fuzzy-artificial immune system and k-nn algorithm for breast cancer Diagnosis, Computers in Biology and Medicine, Volume 37, 415-423.

Polat K., Güneş S. (2007). Breast cancer diagnosis using least square support vector machine, Digital Signal Processing, Volume 17, 694-701.

Guijarro-Berdinas B., Fontenla-Romero O., Perez-Sanchez B., Fraguela P. (2007). A Linear Learning Method for Multilayer Perceptrons Using Least-Squares, Lecture Notes in Computer Science, 365-374.

Akay M. F. (2009). Support vector machines combined with feature selection for breast cancer diagnosis, Expert Systems with Applications, Volume 36, 3240-3247.

Subashini T. S., Ramalingam V., Palanivel S. (2009). Breast mass classification based on cytological patterns using RBFNN and SVM, Expert Systems with Applications, Volume 36, 5284-5290.

Peng Y., Wu Z., Jiang J. (2010). A novel feature selection approach for biomedical data classification, Journal of Biomedical Informatics, Volume 43, 15-23.

Conforti D., Guido R. (2010). Kernel based support vector machine via semidefinite programming: Application to medical diagnosis, Computers & Operations Research, Volume 37, 1389-1394.

Fan C., Chang P., Lin J., Hsieh J. C. (2011). A hybrid model combining casebased reasoning and fuzzy decision tree for medical data classification, Applied Soft Computing, Volume 11, 632-644.

Marcano-Cedeno A., Quintanilla-Dominguez J., Andina D. (2011). WBCD breast cancer database classification applying artificial metaplasticity neural network, Expert Systems with Applications, Volume 38, 9573-9579.

UCI Machine Learning Repository. 2007 April 29 (Last Modified), Cited March 13, from http://archive.ics.uci.edu/ml/datasets/Mammographic+Mass

Bishop C. M. (1995). Neural Networks for Pattern Recognition, Oxford University Press, 1995.

Pearl J. (1998). Probabilistic reasoning in intelligent systems, Morgan Kaufmann.

Lakshminarayan K., Harp S. A., Goldman R., Samad T. (1996). Imputation of missing data using machine learning techniques, Knowledge Discovery and Data Mining KDD-96, 140-145.