Annals of University of Craiova, Math. Comp. Sci. Ser.
Volume 36(2), 2009, Pages 92–96
ISSN: 1223-6934

Prediction of recurrent events in breast cancer using the Naive Bayesian classification

Diana Dumitru

Abstract. Breast cancer is considered to be the second leading cause of cancer deaths in women today. One of the main problems is to predict recurrent and non-recurrent events, probably more important than the first breast cancer diagnosis. The goal of this paper is to investigate the potential contribution of the Naive Bayesian classification methodology as a reliable support in computer-aided diagnosis of such events, using the well-known Wisconsin Prognostic Breast Cancer dataset. The results showed that the Naive Bayes classifier provides performances equivalent to other machine learning techniques with low computational effort and high speed.

2000 Mathematics Subject Classification. Primary 62C10.
Key words and phrases. Breast Cancer, Recurrent Events, Naive Bayesian Classifier.

1. Introduction

Breast cancer is considered to be the second leading cause of cancer deaths in women today, after the lung cancer, and is the most common cancer among women excluding non-melanoma skin cancers.

Occasionally, breast cancer can return after primary treatment. Consequently, the main problem under these circumstances is to predict such a recurrent event, because only expertise from experience is not enough.

An automatic reliable prediction of these recurrent events in breast cancer is an important real-world medical problem to solve. Our study tries to give a Machine Learning-based solution to this problem, using the Naive Bayes classifier.

Naive Bayes classifier is a probabilistic classifier based on the Bayes' theorem, considering a strong (Naive) independence assumption. Thus, a Naive Bayes classifier considers that all attributes (features) independently contribute to the probability of a certain decision. Taking into account the nature of the underlying probability model, the Naive Bayes classifier can be trained very efficiently in a supervised learning setting, working much better in many complex real-world situations, especially in the computer-aided diagnosis than one might expect [2], [4].

The Naive Bayes classifier has been applied to the Wisconsin Prognostic Breast Cancer (WPBC) dataset, concerning a number of 198 patients and a binary decision class: non-recurrent-events totalling 151 instances and recurrent-events totalling 47 instances. The testing diagnosing accuracy, that is the main performance measure of the classifier, was about 74.24%, in accordance with the performance of other well-known Machine Learning techniques.

2. Naive Bayes classifier

The essence of the Bayesian approach is to provide a mathematical rule explaining how you should change your existing beliefs in the light of new evidence. In other words, it allows scientists to combine new data with their existing knowledge or expertise.

Less implemented in data exploration software, Naive Bayes technique is a classification method which owes its name to the British reverend Thomas Bayes (1702 - 1761). Despite its name (also known as "Idiot's Bayes") it is one of the most efficient and effective inductive learning algorithms for machine learning and data mining.

Using a training dataset, the Bayesian classifiers use methods based on the Bayes' Theorem in order to determine the probability of associating certain classes at certain instances given the values of the predictor variables.

A Naive Bayes classifier can classify a certain instance supposing that its attributes are conditional independent given the class. Even if this assumption of attributes independence does not reflect the reality in many areas, this classification method has been founded efficient, effective and robust to noise. Its efficiency applies even in domains where the attributes are not independent: document classification is a domain for which the Naive Bayesian classifiers are often used with success, even if it is a certain dependency between the attributes (between the words in the document).

This classifier learns from training data the conditional probability of each variable X_k given the class label C. Classification is done by applying Bayes rule to compute the probability of C given the particular instance X_1, ..., X_n, by the formula:

$$P\left(C = c | X_1 = x_1, ..., X_n = x_n\right) \tag{1}$$

Because the classifier is founded on the assumption that variables are conditionally independent, the posterior probability of the class is formulated as follows:

$$P\left(C = c | X_1 = x_1, ..., X_n = x_n\right) = P\left(C = c\right) * \prod_{X_k} P\left(X_k = x_k | C = c\right) \tag{2}$$

The result of the classification is the class with the highest posterior probability:

$$max_C \prod_{X_k} P\left(X_k = x_k | C = c\right) \tag{3}$$

The Naive Bayes technique is particularly suited when the dimension of the feature space (of the inputs) is high. Despite its simplicity, the Naive Bayes classifier can often outperform more sophisticated classification methods in several specific domains.

A classification procedure consists in the following: consider a simple setting of classification learning, in which the goal is to predict the class $c \in C = \{c_1, ..., c_m\}$ of a query input $x = \{a_1, ..., a_n\}$, given a training set of pre-classified examples. Instances are characterized in terms of an attribute-value representation, and a_i is the value of the i^{th} attribute.

Naive Bayes classifier algorithm

(1) The class C will be the class with maximum posterior probability:

$$max_{c_i \in C} P\left(c_i | x\right) \tag{4}$$

(2) To identify this class, the posterior probabilities $P(c_i|x)$ must be estimated. Using Bayes formula:

$$P(c_i|x) = \frac{P(x|c_i)\,P(c_i)}{P(x)} \tag{5}$$

it will be sufficient to estimate the probability of an input given a class.

(3) Because Naive Bayes classifier assumes the probabilities of attributes to be conditionally independent given the class, the probability $P(x|c_i)$ can be estimated using the formula:

$$P(x|c_i) = \prod_{j=1}^{n} P(a_j|c_i) \tag{6}$$

(4) The final step is to estimate the probabilities for $P(c_i)$ and $P(a_j|c_i)$ using the training set.

Given that all the attributes of the breast cancer dataset are continuous, the Gaussian distribution will be used to estimate the probabilities:

$$P(a_j|c_i) = \frac{1}{\sqrt{2\pi}\sigma_{ji}} e^{\left(-\frac{(a_j-\mu_{ji})^2}{2\sigma_{ji}^2}\right)} \tag{7}$$

Even through the Naive Bayes assumption is usually violated in practice and the probability estimates for $P(c_i|x)$ are often not very accurate, the Naive Bayes classifier achieves surprisingly high classification rate. This is because, for a correct classification, it is only important that the true class receives the highest probability. More details about Naive Bayes classification are to be found in [3]-[8].

3. Dataset

The Naive Bayes classifier has been applied to the Wisconsin Prognostic Breast Cancer (WPBC) dataset (UCI Machine Learning repository: `http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/`). The input features contain 12 relevant attributes describing the characteristics of the cell nuclei present in the digitised image of a fine needle aspirate (FNA) of breast mass. These numerical attributes are displayed in Table 1.

4. Results

In order to evaluate the classification efficiency, two main metrics have been computed for the Naive Bayes classifier performance in terms of correct classification rate (%) in both the training and the testing phase. Moreover, the sensitivity, that is:

$$sensitivity = \frac{number\ of\ True\ Positives}{number\ of\ True\ Positives + number\ of\ False\ Negatives} \tag{8}$$

and specificity, that is:

$$specificity = \frac{number\ of\ True\ Negatives}{number\ of\ True\ Negatives + number\ of\ False\ Positives} \tag{9}$$

of this classification have been computed in addition, to give a deeper insight of the automatic diagnosis [1].

The results of the classification performance for both training and testing procedures are displayed in Table 2.

TABLE 1. The main characteristics of the Wisconsin breast cancer dataset

Attribute	**Type**	**Range**
Radius (mean of distances from the centre to all points on the perimeter)	Numerical	(10.95, 27.22)
Texture (standard deviation of gray-scale values)	Numerical	(10.38, 39.28)
Perimeter	Numerical	(71.90,182.10)
Area	Numerical	(361.60, 2250)
Smoothness (local variation in radius lengths)	Numerical	(0.075, 0.145)
Compactness ($perimeter^2/area - 1.0$)	Numerical	(0.046, 0.311)
Concavity (severity of concave portions of the contour)	Numerical	(0.024, 0.427)
Concave points (number of concave portions of the contour)	Numerical	(0.020, 0.201)
Symmetry	Numerical	(0.131, 0.304)
Fractal dimension ("coastline approximation" - 1)	Numerical	(0.050, 0.097)
Tumour size - diameter of the excised tumour in centimetres	Numerical	(0.400, 10.00)
Lymph node status - number of positive auxiliary lymph nodes observed at time of surgery	Numerical	(0, 27)

TABLE 2. Naive Bayes classification accuracy

Training accuracy (%)	**Testing accuracy (%)**
76.52	74.24

Classification (training)	Estimated (Bayesian)		
		Class = R	Class = N
Effective classes	Class = R	8	21
	Class = N	10	93

Classification (testing)	Estimated (Bayesian)		
		Class = R	Class = N
Effective classes	Class = R	5	13
	Class = N	4	44

As we can see from the above table, the testing accuracy is about 74%, that is consistent with the best results obtained by other machine learning techniques applied to such datasets.

The misclassification matrices, corresponding to the training/testing performances are displayed below.

Finally, the specificity and sensitivity parameters, displayed below in Table 3, are important in analysing the classification performance.

TABLE 3. Sensitivity and specificity of the Bayesian classification

	Training (%)	**Testing (%)**
Sensitivity	27.59	27.78
Specificity	90.30	91.67

A sensitivity of 100% means that the test recognizes all sick people as such. Thus in a high sensitivity test, a negative result is used to rule out the disease.

A specificity of 100% means that the test recognizes all healthy people as healthy. Thus a positive result in a high specificity test is used to confirm the disease.

5. Conclusion

The methodology we have developed enables exploration and analysis, by automatic means, of large quantities of data related to breast cancer characteristics, in order to obtain an optimal prediction of recurrent events. The purpose of this paper is to demonstrate the suitability and ability of the Naive Bayes methodology in classification/prediction problems in breast cancer. The classification results were consistent with some of the highest results obtained from other classifiers published in the literature.

References

[1] Altman, D.G., Bland, J.M.: Diagnostic tests. 1: Sensitivity and specificity. BMJ 308 (6943): 1552, (1994)

[2] Belciug, S: Bayesian classification vs. k-nearest neighbor classification for the non-invasive hepatic cancer detection, Proc. 8th International conference on Artificial Intelligence and Digital Communications, Craiova, September 2008 (Research Notes in Artificial Intelligence and Digital Communications), 31-35, (2008)

[3] Duda, R., Hart, P., Storck, D.: Pattern classification (2nd Edition), Wiley Interscience, New York, (2000)

[4] Gorunescu, F.: Data Mining: Concepts, models and techniques, Blue Publishing House, Cluj-Napoca, (2006)

[5] Rish, I.: An empirical study of the Naive Bayes classifier, IJCAI 2001 Workshop on Empirical Methods in Artificial Intelligence - http://www.research.ibm.com/people/r/rish/ papers/RC22230.pdf, (2001)

[6] Robles , V., Larrañaga , P., Peña , J.M., Menasalvas, E., Pérez M.S.: Interval Estimation Naive Bayes (2003)

[7] Sulzmann, J.N., Furnkranz, J., Hullermeier, E.: On Pairwise Naive Bayes Classifiers (2007)

[8] Zhang, H., Ling, C. X., Zhao, Z.: Data Mining: Concepts, models and techniques, The Learnability of Naive Bayes (2005)

(Diana Dumitru) CS ROMÂNIA SA, CRAIOVA, ROMANIA
E-mail address: d_i_diana@yahoo.com