APPLICATION OF CLASSIFICATION ALGORITHMS IN BREAST CANCER DIAGNOSIS

Jasmina Novakovic
Academy of Business Applied Studies Belgrade, jasmina.novakovic@bpa.edu.rs

Suzana Marković
Academy of Business Applied Studies Belgrade

Abstract

The main objective of this paper is to present the impact on classification accuracy different feature selections techniques on benchmark medical data set for breast cancer. Six feature selection techniques have been used for feature selection, evaluated and compared using different classification algorithms. Accuracy of the classifier is influenced by the choice of feature selection techniques and thresholds. In our experiment, the highest classification accuracy was obtained with the J48 algorithm using the IG method.

Keywords: breast cancer, classification accuracy, feature selection.

1. INTRODUCTION

There is a growing research interest in fields of machine learning and knowledge discovery. Recent development of technologies for collecting and storing data has led to huge data repositories, which are extremely difficult for humans to analyze. That is why, many data mining techniques are being developed in order to support extracting various knowledge representations from such large data bases.

In knowledge discovery, one of the main tasks considered is supervised classification, where learning process is provided with a set of training examples of target classes. Each example corresponds to a single object to be classified and is described by a finite of features. The goal of learning is to discover a rule or a function which maps such descriptions into those classes.

For the diagnosis and treatment of cancer is critically important precise prediction of tumors. In research of cancer, biologists have still used the traditional microscopic technique to assess tumor behavior for breast cancer patients.

In addition to microscopic technique, biologists are increasingly use modern machine learning techniques to obtain proper tumor information from the databases. In cancer diagnosis, supervised learning methods are the most popular among the existing techniques [1].

Particularly, following the authors used the machine learning techniques: Bellaachia and Guven [2], Delen, Walker and Kadam [3], used the above methods to find the most suitable one for predicting survivability rate of breast cancer patients. Soria, Garibaldi, Biganzoli and Ellis used the C4.5 tree classifier, MLP and Naïve Bayes classifier in prediction of breast cancer [4].

An algorithm, which consists of knowledge representation (learned from some training set) and the strategy of its usage, forms a classifier, which can be used to predict classes of new coming objects. Classification accuracy is typical measure used to evaluate classifier's performance. Several algorithms have been proposed over the years for inducing various knowledge representations and various classifiers [5, 6, 7]. For many classification problems those algorithms are very effective, but they don`t always lead to satisfactory classification accuracy in more complex and difficult cases. As it`s shown in theoretical studies and confirmed in empirical comparative studies there is no single best algorithm to be used for all data sets. It means that every algorithm has its own area of

superiority and specialized to solve some classes of learning problems.

The main aim of this paper is to experimentally verify, on benchmark medical data set, the impact on classification accuracy different feature selections techniques. In our paper we use classification algorithms such as: IBk (k-nearest neighbours classifier), Naïve Bayes, SVM (Support Vector Machine), J48 (J48 decision tree) and RBF (Radial Basis Function network). We also use different ranking and feature selection techniques to improve the classification accuracy of the underlying algorithm.

This paper is organized as follows. In the next section we briefly describe classification algorithms. Section 3 contains general issues concerning ranking and feature selection techniques. Section 4 presents experimental evaluation. Final section contains discussion of the obtained results and some closing remarks.

2. CLASSIFICATION ALGORITHMS

This section gives a brief overview of classification algorithms: IBk, Naïve Bayes, SVM, J48 and RBF.

The k-nearest neighbor classification was developed from the need to perform discriminant analysis when reliable parametric estimates of probability densities are unknown or difficult to determine. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors. In the k-nearest neighbor algorithm k is a positive integer, typically small. This algorithm is a type of instance-based learning, where the function is only approximated locally and all computation is deferred until classification.

Naïve Bayes classifier is based on the elementary Bayes' theorem. Naïve Bayes can achieve relatively good performance on classification tasks. This classifier greatly simplifies learning by assuming that features are independent given the class variable. In simple terms, a Naïve Bayes classifier assumes that the presence (or absence) of a particular feature of a class is unrelated to the presence (or absence) of any other feature. Naïve Bayes classifiers have worked quite well in many complex real-world situations in spite of their naive design and apparently over-simplified assumptions. An advantage of the Naïve Bayes classifier is that it requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Since independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix.

SVM is a discriminative classifier formally defined by a separating hyperplane. In other words, given labeled training data (supervised learning), the algorithm outputs an optimal hyperplane which categorizes new examples. Many hyperplanes might classify the data; the best hyperplane is the one that represents the largest separation, or margin, between the two classes. In general, the larger the margin it is the lower the generalization error of the classifier. We choose, the maximum–margin hyperplane, such the hyperplane in which the distance from it to the nearest data point on each side is maximized. It happens that in a finite dimensional space the sets to be discriminated are not linearly separable. It was proposed that the original finite-dimensional space be mapped into a much higher-dimensional space to make the separation easier. In that way, mapping into a larger space, cross products may be computed easily in terms of the variables in the original space, making the computational load reasonable.

J48 decision tree works by recursively partitioning the training data set according to tests on the potential of feature values in separating the classes. The decision tree is learned from a set of training examples through an iterative process, of choosing a feature and splitting the given example set according to the values of that feature. For this algorithm, the most important question is which of the features is the most influential in determining the classification and hence should be chosen first. Entropy measures or equivalently, information gains are used to select the most influential, which is intuitively deemed to be the feature of the lowest entropy (or of the highest information gain). J48 decision tree works by: a) computing the entropy measure for each feature, b) partitioning the set of examples according to the possible values of the feature that has the

lowest entropy, and c) for each are used to estimate probabilities, in a way exactly the same as with the Naïve Bayes approach. Although feature tests are chosen one at a time in a greedy manner, they are dependent on results of previous tests.

The classification of neural networks proved to be very good just for serious classification problems, problems where is difficult or impossible to use the classical technique. Besides, neural networks are well suited to work in conditions of noise in the data. RBF is an artificial neural network that uses radial basis functions as activation functions. The output of the network is a linear combination of radial basis functions of the inputs and neuron parameters. RBF have many uses, including function approximation, time series prediction, classification, and system control.

3. FEATURE RANKING AND SELECTION

Different feature ranking and feature selection techniques have been proposed in the machine learning literature. The purpose of these techniques is to discard irrelevant or redundant features from a given feature vector.

We consider evaluation of the practical usefulness of the following ranking methods:

- Information Gain (IG) attribute evaluation,
- Gain Ratio (GR) attribute evaluation,
- Symmetrical Uncertainty (SU) attribute evaluation,
- Relief-F (RF) attribute evaluation,
- One-R (OR) attribute evaluation,
- Chi-Squared (CS) attribute evaluation.

Entropy is a commonly used in the information theory measure, which characterizes the purity of an arbitrary collection of examples. It is in the foundation of the IG, GR, and SU attribute ranking methods. The entropy measure is considered as a measure of system's unpredictability. The entropy of Y is

$$H(Y) = -\sum_{y \in Y} p(y) \log_2(p(y)) \quad (1)$$

where $p(y)$ is the marginal probability density function for the random variable Y. If the observed values of Y in the training data set S are partitioned according to the values of a second feature X, and the entropy of Y with respect to the partitions induced by X is less than the entropy of Y prior to partitioning, then there is a relationship between features Y and X. Then the entropy of Y after observing X is:

$$H(Y/X) = -\sum_{x \in X} p(x) \sum_{y \in Y} p(y/x) \log_2(p(y/x)) \quad (2)$$

where $p(y|x)$ is the conditional probability of y given x.

3.1. INFORMATION GAIN

Given the entropy as a criterion of impurity in a training set S, we can define a measure reflecting additional information about Y provided by X that represents the amount by which the entropy of Y decreases. This measure is known as IG. It is given by

$$IG = H(Y) - H(Y/X) = H(X) - H(X/Y) \quad (3)$$

IG is a symmetrical measure (refer to equation (3)). The information gained about Y after observing X is equal to the information gained about X after observing Y. A weakness of the IG criterion is that it is biased in favor of features with more values even when they are not more informative.

3.2. GAIN RATIO

The Gain Ratio is the non-symmetrical measure that is introduced to compensate for the bias of the IG. GR is given by

$$GR = \frac{IG}{H(X)} \quad (4)$$

As equation (4) presents, when the variable Y has to be predicted we normalize the IG by dividing by the entropy of X and vice-versa. Due to this normalization, the GR values fall always in the range [0, 1]. A value of GR = 1 indicates that the knowledge of X completely predicts Y, and GR = 0 means that there is no relation between Y and X. In opposite to IG, the GR favors variables with fewer values.

3.3. SYMMETRICAL UNCERTAINTY

The Symmetrical Uncertainty criterion compensates for the inherent bias of IG by dividing it by the sum of the entropies of X and Y. It is given by

$$SU = 2\frac{IG}{H(Y) + H(X)} \quad (5)$$

SU takes values, which are normalized to the range [0, 1] because of the correction factor 2. A value of SU = 1 means that the knowledge of one feature completely predicts and the other SU = 0 indicates that *X* and *Y* are uncorrelated. Similarly, to GR, the SU is biased toward features with fewer values.

3.4. CHI-SQUARED

Chi-squared attribute evaluation evaluates the worth of a feature by computing the value of the chi-squared statistic with respect to the class. The initial hypothesis H0 is the assumption that the two features are unrelated, and it is tested by chi-squared formula:

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \quad (6)$$

where *Oij* is an observed frequency and *Eij* is an expected (theoretical) frequency, asserted by the null hypothesis. The greater the value of $\chi 2$, the greater the evidence against the hypothesis H0.

3.5. ONE-R

This attribute evaluation evaluates features individually by using the OneR classifier. OneR classifier ranks features according to error rate (on the training set). It treats all numerically valued features as continuous and uses a straightforward method to divide the range of values into several disjoint intervals. It handles missing values by treating "missing" as a legitimate value.

This is one of the most primitive schemes. It produces simple rules based on one feature only. Although it is a minimal form of classifier, it can be useful for determining a baseline performance as a benchmark for other learning schemes.

3.6. RELIEF-F

Relief-F attribute evaluation evaluates the worth of a feature by repeatedly sampling an instance and considering the value of the given feature for the nearest instance of the same and different class. This attribute evaluation assigns a weight to each feature based on the ability of the feature to distinguish among the classes, and then selects those features whose weights exceed a user-defined threshold as relevant features. The weight computation is based on the probability of the nearest neighbors from two different classes having different values for a feature and the probability of two nearest neighbors of the same class having the same value of the feature. The higher the difference between these two probabilities, the more significant is the feature. Inherently, the measure is defined for a two-class problem, which can be extended to handle multiple classes, by splitting the problem into a series of two-class problems.

4. EXPERIMENT AND RESULTS

In this section, the performance of different classifiers examined and compared using benchmark real-world classification mammographic mass data set. An experiment was set up to demonstrate the effectiveness of different classifiers. In order to achieve good experimental results, we use different feature ranking and feature selection techniques.

Mammographic mass data set [8], taken from UCI repository of machine learning databases, was used for discrimination of benign and malignant mammographic masses based on BI-RADS attributes and the patient's age. Mammographic mass data set contains a BI-RADS assessment, the patient's age and three BI-RADS attributes together with the ground truth (the severity field) for 516 benign and 445 malignant masses that have been identified on full field digital mammograms collected at the Institute of Radiology of the University Erlangen-Nuremberg between 2003 and 2006.

In this data set, each instance has associated BI-RADS assessment ranging

from 1(definitely benign) to 5 (highly suggestive of malignancy) assigned in a double-review process by physicians. Assuming that all cases with BI-RADS assessments greater or equal a given value (varying from 1 to 5), are malignant and the other cases benign, sensitivities and associated specificities can be calculated. These can be an indication of how well machine learning techniques perform compared to the radiologists. Figure 1. shows a way to extract contours based on the shadows of the mammographic mass.

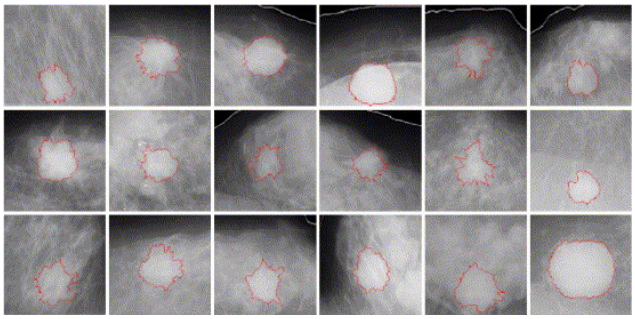

Fig. 1. *Contour extraction based on shadow mammographic mass [9]*

Table 1 presents a description of this data set. Table 2 presents number of attributes in the original dataset and number of attributes selected using filtering methods. Searching for a set of all possible solutions for each method has found the optimal number of attributes.

The table shows the original size of the set, in order to compare the effects of the reduction of the dimensionality of data. Using filter methods, four methods reduce the number of attributes more than half compare with the original data set.

Table 1. *Description of data set*

Attributes			Number of classes	Size for training
in total	categorically	numerical		
5	0	5	2	961

Table 2. *The number of attributes obtained by filter methods*

Without filtering	IG	GR	SU	RF	OR	CS
5	3	2	2	4	2	2

We present classification accuracy of different classifiers for predicting the outcomes of breast biopsies from BI-RADS findings, which have the potential to reduce the number of unnecessary breast biopsies in clinical practice.

In the table of classification accuracy "+" indicates a significantly higher value for classification accuracy, while "−" indicates a significantly lower value for classification accuracy. In the table of training time "+" indicates a significantly smaller value for training time, while "−" indicates a significantly higher value for training time. Comparison is such that the second algorithm is an algorithm in which was performed pre-selection attributes, and the first algorithm is a standard algorithm without pre-selection of attributes.

Further experimental research, the optimal number of selected attributes for each data set and filtering method, checked the accuracy of the classification algorithms.

Table 3. *Classification accuracy with filter methods*

	Without filtering	IG	GR	SU	RF	OR	CS
IBk	75.60	82.27 +	83.49 +	83.38 +	75.18	82.75 +	83.36 +
Naïve Bayes	82.64	81.62	81.58	81.26	79.59 -	80.29 -	81.25
SVM	80.29	82.68 +	83.15 +	83.06 +	79.95	82.46	83.03 +
J48	82.19	83.57 +	83.29	83.19	80.76	82.60	83.16
RBF	77.31	77.66	79.67	79.24	77.07	77.51	79.16

The highest classification accuracy without filtering for a given data set was obtained with the Naïve Bayes

algorithm (Table 3). However, the filter methods in this algorithm failed to increase the classification accuracy. The highest classification accuracy was obtained with the J48 algorithm using the IG method. Using different classifiers, we can conclude that the IG method of filtering in most cases led to better results in the classification accuracy.

Table 4. *Standard deviation with filter methods*

	Without filtering	IG	GR	SU	RF	OR	CS
IBk	3.90	3.37	3.13	3.10	3.64	3.08	3.09
Naïve Bayes	3.11	3.59	3.05	3.33	3.68	3.54	3.33
SVM	3.41	3.18	3.14	3.12	3.61	3.20	3.10
J48	3.21	3.14	3.13	3.09	3.62	3.11	3.07
RBF	3.31	3.67	4.14	4.51	3.83	4.35	4.50

Table 4 shows the standard deviation for the classification accuracy with original and reduced data set using filter methods. From the table it can be seen that the standard deviations generally not much different from standard algorithm and algorithms that use a reduced set of data.

When we show the results for the time required for training data, they were expressed in units of CPU seconds. The experiment was performed on the AMD Phenom (tm) 9650 Quard-Core Pocessor 2.31 GHz with 4GB of RAM.

Table 5. *Training time (in seconds) with filter methods*

	Without filtering	IG	GR	SU	RF	OR	CS
IBk	0.00	0.00	0.00	0.00	0.17 -	0.02 -	0.00
Naïve Bayes	0.00	0.00	0.00	0.00	0.17 -	0.02 -	0.00
SVM	0.13	0.06 +	0.05 +	0.05 +	0.29 -	0.07+	0.05 +
J48	0.01	0.01	0.00	0.00	0.18 -	0.02 -	0.00
RBF	0.02	0.02	0.02	0.02	0.19 -	0.03 -	0.02

Table 5 shows the time required for training algorithms (in seconds) with the original and the reduced data set using the filter methods. Required time to train the data for basic classifiers is small, except SVM algorithm. The required training time is higher with the SVM algorithm compare with other algorithms, and it can be reduced by using the appropriate filter method.

Table 6. *Standard deviation during training (in seconds) with filter methods*

	Without filtering	IG	GR	SU	RF	OR	CS
IBk	0.00	0.01	0.01	0.01	0.01	0.01	0.01
Naïve Bayes	0.00	0.01	0.01	0.01	0.01	0.01	0.01
SVM	0.01	0.01	0.01	0.01	0.01	0.01	0.01
J48	0.01	0.01	0.01	0.01	0.01	0.01	0.01
RBF	0.01	0.01	0.01	0.01	0.01	0.01	0.01

Table 6 shows the standard deviation of the time required for training. Standard deviation for required time to train the data for basic classifiers and classifiers with filter methods is small.

5. CONCLUSION

Problem of discrimination of benign and malignant mammographic masses based on supervised and unsupervised learning methods to help physicians in their decision to perform a breast biopsy on a suspicious lesion seen in a mammogram is our task. According to the obtained results, we can conclude that it is possible to improve the system performance of classification algorithms in the problem of breast cancer, using the filter methods for reducing the dimensionality of the data. To prove the hypothesis, have been implemented and empirically tested filter methods for reducing the dimensionality of the data.

In further research it would be interesting to apply other techniques to solve the problem of dimensionality reduction of data, such as wrapper method and extraction of attributes and analyze and compare the effects of their implementation.

REFERENCE

[1] Nahar J, Chen Y.P, Ali S. Kernel-based naive bayes classifier for breast cancer prediction. Journal of Biological System, 15(1):17–25, 2007.

[2] Bellaachia A, Guven E. Predicting breast cancer survivability using data mining techniques. Scientific Data Mining Workshop in Conjunction with the 2006 SIAM Conference on Data Mining, 2006.

[3] Delen D, Walker G, Kadam A. Predicting breast cancer survivability: a comparison of three data mining methods. Artificial Intelligence in Medicine, 34(2):113–127, 2005.
[4] Soria D, Garibaldi J. M, Biganzoli E, Ellis I. O. A comparison of three different methods for classification of breast cancer data. Seventh International Conference on Machine Learning and Applications, IEEE Computer Society, 2008.
[5] Klosgen W, Zytkow J.M. (eds.). Handbook of Data Mining and Knowledge Discovery. Oxford Press, 2002.
[6] Stefanowski J. Algorithms of rule induction for knowledge discovery. (In Polish), Habilitation Thesis published as Series Rozprawy no. 361, Poznan Univeristy of Technology Press, Poznan 2001.
[7] Kuncheva L. Combining Pattern Classifiers: Methods and Algorithms. Wiley, 2004.
[8] Elter M, Schulz-Wendtland R, Wittenberg T. The prediction of breast cancer biopsy outcomes using two CAD approaches that both emphasize an intelligible decision process. Medical Physics 34(11), pp. 4164-4172, 2007.
[9] Nakagawa T, Harab T, Fujitab H, Iwasec T, Endod T, Horitae K. Automated contour extraction of mammographic mass shadow using an improved active contour model. Elsevier, International Congress Series, Volume 1268, June 2004, pp 882–885.