International Journal of Computer Applications (0975 - 8887)
Volume 62 - No. 1, January 2013

Breast Cancer Diagnosis by using k-Nearest Neighbor with Different Distances and Classification Rules

Seyyid Ahmed Medjahed
Department of Computer Science
University of Science and Technology Oran
USTOMB, Algeria

Tamazouzt Ait Saadi
Department of Computer Science
University of LeHavre
LeHavre, France

Abdelkader Benyettou
Department of Computer Science
University of Science and Technology Oran
USTOMB, Algeria

ABSTRACT

Cancer diagnosis is one of the most studied problems in the medical domain. Several researchers have focused in order to improve performance and achieve to obtain satisfactory results. Breast cancer is one of cancer killer in the world. The diagnosis of this cancer is a big problem in cancer diagnosis researches. In artificial intelligent, machine learning is a discipline which allows to the machine to evolve through a process. Machine learning is widely used in bio informatics and particularly in breast cancer diagnosis. One of the most popular methods is K-nearest neighbors (K-NN) which is a supervised learning method. Using the K-NN in medical diagnosis is very interesting. The quality of the results depends largely on the distance and the value of the parameter "k" which represent the number of the nearest neighbors. In this paper, we study and evaluate the performance of different distances that can be used in the K-NN algorithm. Also, we analyze this distance by using different values of the parameter "k" and by using several rules of classification (the rule used to decide how to classify a sample). Our work will be performed on the WBCD database (Wisconsin Breast Cancer Database) obtained by the university of Wisconsin Hospital.

Keywords:

Classification, Diagnosis, Breast Cancer, K-Nearest Neighbors, Distance, Classification Rule

1. INTRODUCTION

Early detection of cancer is essential for a rapid response and better chances of cure. Unfortunately, early detection of cancer is often difficult because the symptoms of the disease at the beginning are absent. Thus, cancer remains one of the topics of health research, where many researchers have invested with the aim of creating evidence that can improve treatment, preventions and diagnostics.
Research in this area is a quest of knowledge through surveys, studies and experiments conducted with applications in order to discover and interpret new knowledge to prevent and minimize the risk adverse consequences. To understand this problem more precisely, tools are still needed to help oncologists to choose the treatment required for healing or prevention of recurrence by reducing the harmful effects of certain treatments and their costs.
To develop tools for cancer management, machine learning methods and clinical factors, such as : patient age and histopathological variables form the basis for daily decision making are used. Several studies have been developed in this topic by using the gene expressions [17, 7, 9] or using image processing [2, 16].
In machine learning there are two types: the supervised and unsupervised learning. The first admits that the classes used to classify the data are known in advance and the second, the classes are not known. Among the methods, there are: Support Vector Machines, Decision Tree, Neural Network, Bayesian networks, k-nearest neighbors, etc.
The algorithm k-nearest neighbors is widely used in data classification [12, 18, 8]. The k-nn permits the classification of a new element by calculating its distance from all the other elements. The proper functioning of the method depends on the choice of the parameter k which represents the number of neighbors chosen to assign the class to the new element and the choice of the distance.
In this paper, we study and analyze several distances and different values of the nearest neighbors parameter k, by using different classification rules in the k-nearest neighbors algorithm. The performance will be evaluated in term of the classification accuracy rate and classification time and to validate the results obtained by these approaches, we use several tests with different training and tasting sets. This experimentation will be conducted on the Wisconsin Breast Cancer Database (WBCD) which is the publicly available breast cancer database [13, 11, 4, 14, 1] and is a widely studied.
The paper is organized as follows: First, an overview of the K-Nearest Neighbors method is given. In $Section.3$, different distances defined in the literature are illustrated. In $Section.5$, we present the results obtained by these approaches. Finally, we conclude and describe future work in the $Section.6$.

2. K-NEAREST NEIGHBORS METHOD

The k-nearest neighbors algorithm is one of the most used algorithms in machine learning [19, 15, 6]. It is a learning method bases on instances that does not required a learning phase.
The training sample, associated with a distance function and the choice function of the class based on the classes of nearest neighbors is the model developed. Before classifying a new element, we must compare it to other elements using a similarity measure. Its k-nearest neighbors are then considered, the class that appears most among the neighbors is assigned to the element to

be classified. The neighbors are weighted by the distance that separate it to the new elements to classify.

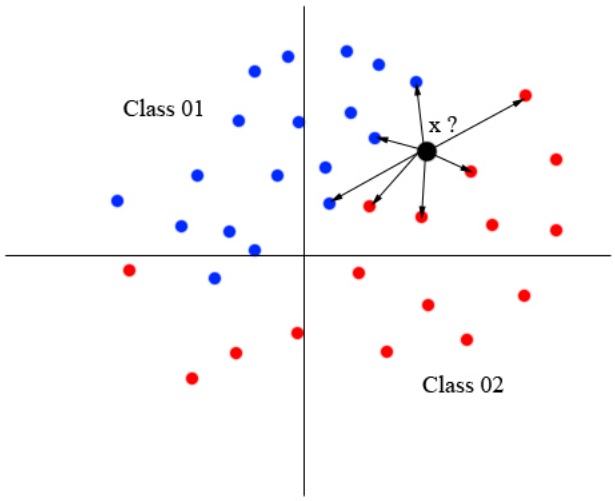

Fig. 1. The K-nearest neighbors method.

The proper functioning of the method depends on the choice of some number of parameter such as the parameter k which represents the number of neighbors chosen to assign the class to the new element, and the distance used.

2.1 The K-Nearest Neighbors Algorithm

Choose a value for the parameter k.
Input : Give a sample of N examples and their classes.
The classe of an sample x is $c(x)$.
Give a new sample y.
Determine the k-nearest neighbors of y by calculating the distances.
Combine classes of these y examples in one class c
Output : The class of y is $c(Y) = c$

3. THE DISTANCES

In mathematics, a distance is an application that formalizes the idea of the distance which is the length between two points. The distance will allow us to group the individuals which are similar and separate those that do not resemble.
A distance $d(x_i, x_j)$ in a space E is an application $E \times E$ in $\mathbb{R}$ satisfying the following axioms:

—Non-negativity : $d(x_i, x_j) \geq 0$
—Symmetry : $d(x_i, x_j) = d(x_j, x_i)$
—Reflexivity : $d(x_i, x_j) \Leftrightarrow x_i = x_j$
—Triangle inequality : $d(x_i, x_j) \leq d(x_i, x_k) + d(x_k, x_j)$

In the case of a vector space, the distances are often differents norms.
We can defined in many ways the distances between two points, although it is generally given by the Euclidean distance (or 2-distance). Given two points x_{ir} and x_{ij} from E,we define the different distances as follows:

3.1 Cityblock distance (1-distance)

Called also Manhattan distance, the cityblock distance is associated to the $1 - norm$, for two vectors x_{ir}, x_{jr} the Manhattan distance is defined by:

$$d(x_i, x_j) = \sum_{r=1}^{n} |x_{ir} - x_{jr}|$$

It represents the sum of absolute differences.

3.2 Euclidean distance (2-distance)

Euclidean distance is the most universal, between two vectors x_{ir} and x_{jr}, the euclidean distance is defined as :

$$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n} (x_{ir} - x_{jr})^2}$$

Noting that the Euclidean distance is a special case of the Minkowski metric when $p = 2$

3.3 Minkowski distance (p-distance)

The most frequently distances are the Minkowski distance which is defined as follows :

$$d(x_i, x_j) = \sqrt[p]{\sum_{r=1}^{n} |x_{ir} - x_{jr}|^p}$$

where $p = 1, 2, ..., \infty$

3.4 Tchebychev distance (∞-distance)

Tchebychev distance is also called Maximum value distance. It is the maximum of the absolute rank:

$$d(x_i, x_j) = \lim_{p \to \infty} \sqrt[p]{\sum_{r=1}^{n} |x_{ir} - x_{jr}|^p} = \sup_{1 \leq r \leq n} |x_{ir} - x_{rj}|$$

3.5 Cosine distance

Giving the angular distance between the cosines, the cosine distance is written as follows:

$$d(x_i, x_j) = \frac{\sum_{r=1}^{n} x_{ir} \times x_{ij}}{\sqrt{\sum_{r=1}^{n} (x_{ir})^2} \times \sqrt{\sum_{r=1}^{n} (x_{jr})^2}}$$

3.6 Correlation distance

The distance correlation of two random variables is obtained by dividing their distance covariance by the product of their distance standard deviations. The distance correlation is :

$$d(x_i, x_j) = 1 - c_{xi,xj}$$

where c is the coefficient of Pearson correlation.
The euclidean, cityblock, cosinus and correlation types distances are the most used in the k-nn algorithms. Therefore, we used this algorithm for classification based on these different distances to find the class of a new element.

4. THE CHOICE OF THE PARAMETER K (THE NUMBER OF NEAREST NEIGHBORS)

The choice of the parameter k ($k \in N$) is determined by the user, this choice depends on the data. The effect of noise on the classification is reduced when the values chosen for k is greater, but this makes the boundaries between classes less distinct. A good choice of the value of k can be selected by different heuristic

techniques such as cross-validation. In this study we choose the value of k that minimizes the classification error.
In the case of a binary classification, it is more inviting to choose an odd value for k, it avoids the equal votes. In case of equality, we can increase the value of k from 1 to decide [5].

5. EXPERIMENTATION

We have used the k-nearest neighbors that is experimented with several variants of distances, different values of k and different classification rules for the choice of nearest neighbors. The k-nearest neighbors does not require learning phase.
To analyze the relevance of these different distances, we have defined a protocol of comparison which includes the classification accuracy rate and time classification by using different values of the parameter k and classification rules. This experiments are conducted on the database WBCD (Wisconsin breast cancer database) [10, 3] obtained by the university of Wisconsin. This database contains information about the breast cancer which are taken by the Fine Needle Aspirate (FNA) of human breast tissue. These data correspond to 699 clinical cases where $458\ (65,50\%)$ are a benign cases and $241\ (34,50\%)$ are malignant cases. The WBCD database contains missing data for 16 observations, which limited this experimentation to 683 clinical cases.
The evaluation of performance of the learning methods need the separation of the database in two parts: The training set that represents the initial basis for which the classes of different clinical cases are known and the testing set. In this study we have used the **holdout** method which is a kind of cross validation to separate randomly the database, and we have obtained: $455\ (65,10\%)$ clinical cases for the training phase and $244\ (34,90\%)$ for the testing phase.
The performance evaluations of each types distances and classification rules are chosen in function of : classification accuracy rate and time classification for each value of the nearest neighbors parameter. Several tests have been conducted to validate our results.
For the classification rules we have used:
The **nearest rule** recommends that the new element will be assigned to the majority class among the nearest neighbors. If k is an even number, the class of the new element will be the class of the nearest neighbor.
The **random rule** recommends that the new element will be assigned to the majority class among the nearest neighbors. If k is an even number, the class of the new element will be the class of on the nearest neighbors and it will be done randomly.
For the **consensus rule**, the new element will only be affected if all the nearest neighbors are of the same class.
Figure 2 illustrates the classification accuracy rate in function of the value of k (the number of nearest neighbors) based on the use of **the nearest rule** to classify a new element.
The high classification accuracy rate, $98,70\%$ is recorded by the algorithm that uses the Euclidean distance with a value of $k = 1$. The same algorithm used with Manhattan distance and with $k = 1$ gives rather promising result $(98,48\%)$.
Figure 2 shows that when k increases over the classification accuracy rate decreases and stabilizes at a value close to 50, with a classification accuracy rate close to $94,40\%$ (Table 5). However, the best result is obtained with the Euclidean distance $(94,45\%)$, this corroborates with what was presented in the literature.
The minimum classification time is recorded for both cosine and correlation distances. In contrast, Euclidean and Manhattan distances are time-consuming classification, these results are illustrated in Figure 3.
Figure 4 illustrates the classification accuracy rate in function of the value of k (the number of nearest neighbors) based on the use of **the random rule** for classifying a new element.

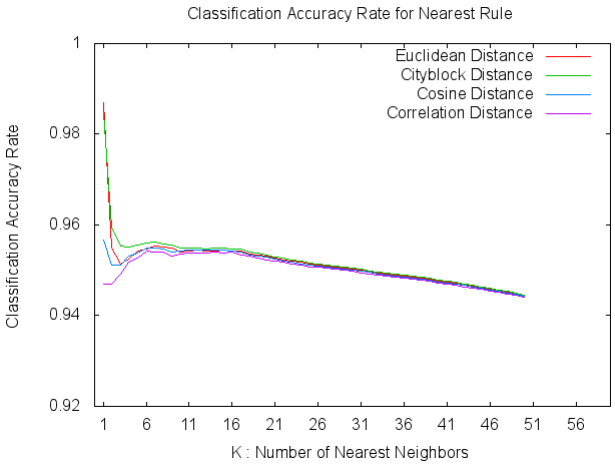

Fig. 2. Representation of classification accuracy rate for each value of parameter k based on nearest rule.

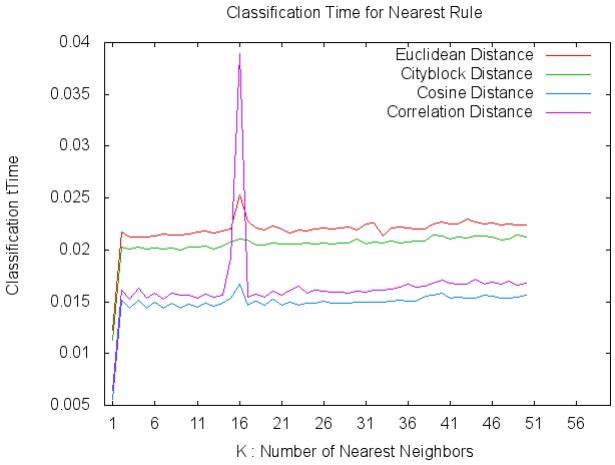

Fig. 3. Representation of classification time for each value of parameter k based on random rule.

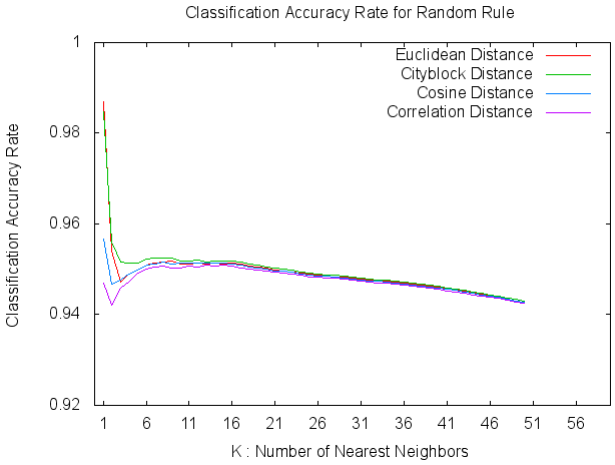

Fig. 4. Representation of classification accuracy rate for each value of parameter k based on random rule.

The results obtained for the classification of a new element with the $k-$nn algorithm using the random rule are identical with

those provided by the nearest rule and this for all types of distances used. The difference is in the case where k is even, the probability to assign the new element to a class is equiprobable. It remains to emphasize that the best classification accuracry rate, $98,70\%$ is achieved by the algorithm that uses the Euclidean distance with a value of $k = 1$ and the classification accuracy rate decreases when k increases, it begins to stabilize around the value 50 (Figure 4). We record a classification accuracy rate close to $94,25\%$ (Table 5).

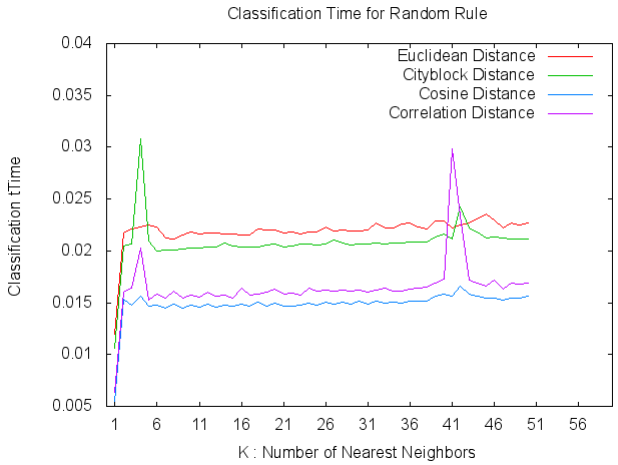

Fig. 5. Representation of classification time for each value of parameter k based on random rule.

Similarly, the minimum classification time is recorded for both cosine and correlation distances. In contrast Euclidean and Manhattan distances are time-consuming classification. . Nevertheless, there is a slight increase in the time of classification for cosine distance, correlation and Manhattan when k takes the values $4, 41$ and 42. Therefore, time classification time of the Euclidean distance shows no disturbance, these results are illustrated in Figure 5.
The figure 6 shows the classification accuracry rate based on the value of k (the number of nearest neighbors) by using the consensus rule for classifying a new element.

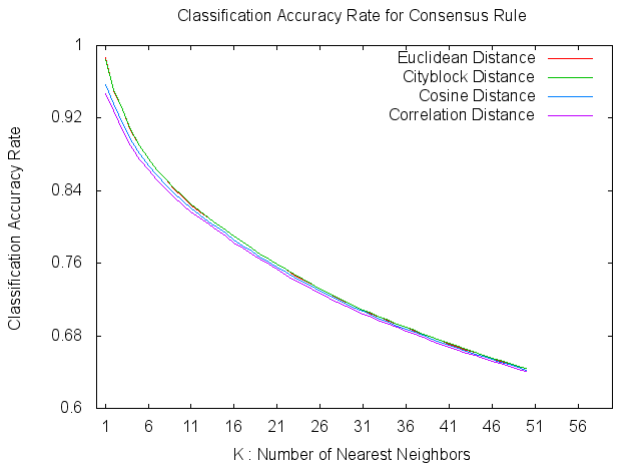

Fig. 6. Representation of classification accuracy rate for each value of parameter k based on consensus rule.

The high classification accuracy rate, $98,70\%$ is reached by the algorithm that uses the Euclidean distance with a value of $k = 1$.

The same algorithm used with distance from Manhattan and with $k = 1$ gives a rather promising result $(98,48\%)$. Also, a difference was observed when using the cosine distance $(95,67\%)$ and correlation $(94,69\%)$. Figure 6 shows that by increasing the value of k, the classification accuracy rate decreases considerably and it is around $64,30\%$ when $k = 50$ (Table 5).

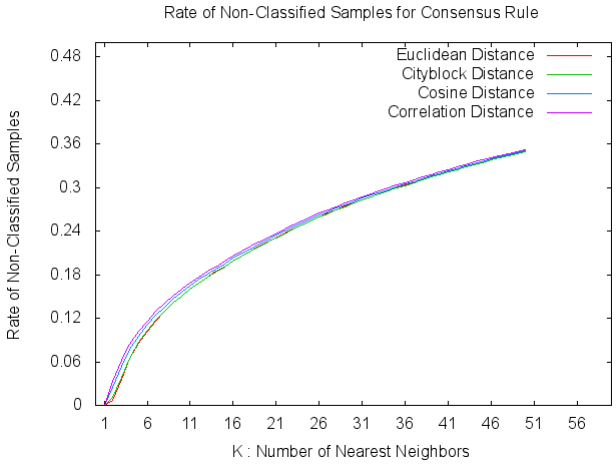

Fig. 7. Representation of the rate of non-classified elements for each value of parameter k based on consensus rule.

Using this rule, some data may not be classified, The k-nn algorithm that uses the rule consensus can not assign a class to a new element if all of these neighbors do not belong to the same class. figure 7 shows the rates of elements that were not classified by the algorithm in function of the parameter k.

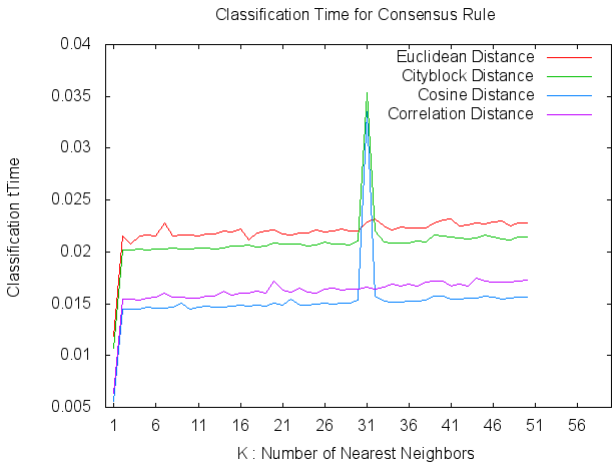

Fig. 8. Representation of classification time for each value of parameter k based on consensus rule.

The minimum time of classification is recorded for both cosine and correlationdistances. By cons, it remains expensive for Manhattan and Euclidean distances. We also observe a slight increase in the time of classification for the cosine and Manhattan distances when $k = 31$. However the time of classification with Euclidean distance and correlation remains stable for all values assigned to the parameter k, these results are illustrated in Figure 8.
Table 5 lists all results presented for the different rules used for calculating distances.

Nearest Rule				
Value of k	Euclidean	Cityblock	Cosine	Correlation
$k = 1$	$98,70$	$98,48$	$95,67$	$94,69$
$k = 10$	$95,41$	$95,48$	$95,41$	$95,35$
$k = 25$	$95,13$	$95,16$	$95,11$	$95,07$
$k = 50$	$94,15$	$94,46$	$94,43$	$94,40$
Consensus Rule				
Value of k	Euclidean	Cityblock	Cosine	Correlation
$k = 1$	$98,70$	$98,48$	$95,67$	$94,69$
$k = 10$	$83,35$	$83,37$	$95,12$	$95,03$
$k = 25$	$73,64$	$73,67$	$94,86$	$94,83$
$k = 50$	$64,36$	$64,39$	$94,27$	$94,24$
Random Rule				
Value of k	Euclidean	Cityblock	Cosine	Correlation
$k = 1$	$98,70$	$98,48$	$95,67$	$94,69$
$k = 10$	$95,12$	$95,19$	$95,12$	$95,03$
$k = 25$	$94,90$	$94,91$	$94,86$	$94,83$
$k = 50$	$94,29$	$94,30$	$94,27$	$94,24$

Table 1. The classification results.

6. CONCLUSION

In this paper, we have highlighted the algorithm K-nearest neighbors for classification. We used this algorithm with several different types of distances and classification rules (majority, consensus and random) in function of the parameter k that we varied in the interval $[1, 50]$. This algorithm was used in the medical diagnosis that is in the diagnosis and classification of cancer. This experiments were conducted on the database WBCD (Wisconsin Breast Cancer Database) obtained by the University Hospital of Wisconsin.

The results advocate the use of the k-nn algorithm with both types of Euclidean distance and Manhattan. These distances are effective in terms of classification and performance but are consuming much time. Nevertheless, they remain two types of distance that give the best results ($98,70\%$ for Euclidean distance and $98,48\%$ for Manhattan whith $k = 1$), these values are not significantly affected even when $k = 1$ is increased to 50.

7. REFERENCES

[1] M. F. Akay. Support vector machines combined with feature selection for breast cancer diagnosis. *Expert Systems with Applications*, 2(36), 2009.

[2] B. Alexander, Y. Ran, I. Eran K. Ron, M. Ron, and P. Dori. Breast cancer diagnosis from biopsy images using generic features and svms. *Technical Report - Israel Institute of Technology*, Sep 2006.

[3] K. P. Bennett and O. L. Mangasarian. Robust linear programming discrimination of two linearly inseparable sets. *Optimization Methods and Software 1*, 1992.

[4] E. D. beyli. Implementing automated diagnostic systems for breast cancer detection. *Expert Systems with Applications*, 4(33), 2007.

[5] D. Bremner, E. Demaine, J. Erickson, J. Iacono, S. Langerman, P. M., and Godfried. Output-sensitive algorithms for computing nearest-neighbour decision boundaries. *Discrete and Computational Geometry*, 33(4), 2005.

[6] D. Coomans and D.L. Massart. Alternative k-nearest neighbour rules in supervised pattern recognition. *Analytica Chimica Acta*, 136, 1982.

[7] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3), 2002.

[8] L. Li and C. Weinberg. Gene selection and sample classification using a genetic algorithm and k -nearest neighbor method. *A Practical Approach to Microarray Data Analysis*, 2003.

[9] R. Mallika and V. Saravanan. An svm based classification method for cancer data using minimum microarray gene expressions. *World Academy of Science, Engineering and Technology*, 62, 2010.

[10] O. L. Mangasarian and W. H. Wolberg. Cancer diagnosis via linear programming. *SIAM News*, 5(23), Sep 1990.

[11] A. Marcano-Cedeno, J. Quintanilla-Domnguez, and D. Andina. Wbcd breast cancer database classification applying artificial metaplasticity neural network. *Expert Systems with Applications*, (38), 2011.

[12] M. Martn-Merino and J. De Las Rivas. Improving k-nn for human cancer classification using the gene expression profiles. *Computer Science Advaces in Intelligent Data Analysis VIII*, 5772/2009, 2009.

[13] A. Mert, N. Kilic, and A. Akan. Breast cancer classification by using support vector machines with reduced dimension. *ELMAR Proceedings*, 2011.

[14] K. Polat and S. Gnes. Breast cancer diagnosis using least square support vector machine. *Digital Signal Processing*, 4(17), 2007.

[15] M. Raniszewski. Sequential reduction algorithm for nearest neighbor rule. *Computer Vision and Graphics*, 6375, 2010.

[16] Y. Ireaneus Anna Rejani and S. Thamarai Selvi. Early detection of breast cancer using svm classifier technique. *International Journal on Computer Science and Engineering*, 1(3), 2009.

[17] S. Shah and A. Kusiak. Cancer gene search with data-mining and genetic algorithms. *Computers in Biology and Medicine*, 37, 2002.

[18] P. Shi, S. Ray, Q. Zhu, and M. A Kon. Top scoring pairs for feature selection in machine learning and applications to cancer outcome prediction. *BMC Bioinformatics*, 12, 2011.

[19] J. S. Snchez, R. A. Mollineda, and J. M. Sotoca. An analysis of how training data complexity affects the nearest neighbor classifiers. *Pattern Analysis and Applications*, 10(3), 2007.