International Journal of Innovative
Computing, Information and Control
Volume **9**, Number **2**, February **2013**

ICIC International ©2013 ISSN 1349-4198
pp. **727–739**

A NEW HYBRID FEATURE SELECTION METHOD BASED ON ASSOCIATION RULES AND PCA FOR DETECTION OF BREAST CANCER

Onur Inan[1], Mustafa Serter Uzer[2] and Nihat Yilmaz[2]

[1]Computer Engineering Department
[2]Electrical-Electronics Engineering Department
Faculty of Engineering and Architecture
Selcuk University
Konya, Turkey
{ oinan; msuzer; nyilmaz }@selcuk.edu.tr

Received November 2011; revised March 2012

Abstract. *In this study, a new hybrid feature selection method named as AP has been formed to detect breast cancer, using association rules (Apriori algorithm) and Principal Component Analysis (PCA) together with artificial neural network classifier. Thanks to this hybrid system, both the decrease in the size of data and the successful and fast training of classifiers have been achieved. In order to detect the accuracy of the suggested system, Wisconsin breast cancer data have been used. 10-fold cross-validation has been used on the classification phase. The average classification accuracy of the developed AP + NN system is 98.29%. Among the studies performed through cross-validation method for breast cancer, our study result appears to be very promising. As the results suggest, this system, which is performed through size reduction, is a feasible system for faster and more accurate diagnosis of diseases.*

Keywords: Breast cancer diagnosis, Feature selection, Neural network, Apriori, PCA

1. **Introduction.** By the way electronic and information technologies are used to solve medical problems new diagnosis systems are developed for medical doctor. The developed systems are decreasing diagnosis error and they provided standardization of data acquisition. The most studied medical problem is early diagnosis of cancers.

To solve medical problems in the use of electronic and information technologies developed new disease diagnostic systems for doctors. Developed in the medical field, these systems reduce errors and standardize the data collected. The early diagnosis of cancers is the most studied medical problem.

Cancer is the state where the body cells lose their functions, start multiplying and dividing uncontrollably. Cancer cells accumulate and form tumors (bulks). Tumors can be benign (well-tempered) or malignant (ill-tempered) [1]. The leading cause of death in the world is still cancer based on malignant tumors. It has been recorded that 7.9 million people died of cancer in 2007 [2]. Cancers are named according to the organ they originate from. Therefore, the cell growth on breast tissues is called breast cancer. Breast cancer holds the 2nd spot among the cancer-related causes of death, after lung cancer for women [1]. According to the data of World Health Organization, around 460,000 women died of breast cancer in 2008 [2]. Though very rarely seen in males, breast cancer affects one out of eight females at least one time in their lives. Moreover, a recent research performed in Canada indicates that this rate is 1/3. Although scientists are aware of the effects such as genetic factors, obesity and aging, they have not achieved to develop a treatment that will prevent contracting this disease [3]. In this situation, the most suitable type of

treatment is early diagnosis. Diagnosis of the disease in the early phase saves many lives. The medical analysis results of those who are suspected to carry the disease are used to diagnose the cancer. These analyses can be sorted as social statistics, blood values, prints of medical imaging devices (Roentgen, MR, Doppler and Mammography). As is seen, a great deal of data emerges to be evaluated after the medical analysis results. The obtained large amounts of data must be analyzed efficiently for medical diagnostics. It has become widespread in the recent years to use the fast-growing information technologies and especially data mining techniques to analyse these data. Data mining is a very common technique used for determining, verifying and estimating data. Recently referred to as data classification methods, data mining techniques have more common use in pattern recognition and have been used frequently for identifying the cancer disease [4].

In this study, AP + NN method is suggested to be used in the diagnosis of breast cancer. Our method consists of 2 phases. In the first phase, the feature selection methods, namely Apriori and PCA are used respectively. Therefore, the important features have been selected and feature vector size has been reduced. In the second phase, these reduced data are used for the artificial neural network and classification has been performed.

The rest of the paper is organized as follows. In Section 2, we present the related works of breast cancer classification and feature selection. In Section 3, we describ Wisconsin breast cancer database which is used for the proposed system. Our used algorithm and methods (Apriori, Principal Component Analysis and Artificial Neural Networks) are given in detail in Section 4. The experimental application and the experimental results are given to show the effectiveness of our method, respectively in Section 5 and Section 6. Finally, we conclude this paper in Section 7.

2. **Related Works.**

2.1. **Breast cancer classification.** In the last decades, a lot of researchers have studied on prediction and classification of breast cancer pattern. For breast cancer problem, Goodman et al. used three different methods. The first, optimized-LVQ method performance obtained 96.7% classification accuracy, the second, big-LVQ method reached 96.8% and the last method, AIRS, which he proposed depending on the artificial immune system, obtained 97.2% classification accuracy [5]. Abonyi and Szeifert obtained 95.57% classification accuracy for 10-fold validation experiment with the application of supervised fuzzy clustering [6]. A least square support vector machine (LS-SVM) classifier algorithm was proposed for breast cancer diagnosis in [7] and classification accuracy was obtained 98.53% using 10-fold cross validation. Akay used SVM-based method combined with feature selection for breast cancer diagnosis and it was observed that the proposed method yields the highest classification accuracies without cross-validation (98.53%, 99.02%, and 99.51% for 50-50% of training-test partition, 70-30% of training-test partition, and 80-20% of training-test partition, respectively) for a subset that contained five features [1]. Yeh et al. proposed a new hybrid approach using discrete particle swarm optimization and statistical method for mining breast cancer pattern and it reached accuracy of 98.71% [4]. Karabatak and Ince proposed an AR + NN method to use in breast cancer diagnosis problem. This method consists of two-stages. In the first stage, the input feature vector dimension is reduced by using association rules. This provides elimination of unnecessary data. In the second stage, neural network uses these inputs and classifies the breast cancer data. In test stage, 3-fold cross validation method was applied. The average correct classification rate of proposed system is 95.6% for four inputs and 97.4% for eight inputs [8]. Marcano-Cedeño et al. presented an Artificial Neural Network based on the

biological metaplasticity property for Classification of Breast Cancer. The obtained Artificial metaplasticity Multilayer Perceptron (AMMLP) classification accuracy is 99.26% without cross-validation [9]. A rough set (RS) based supporting vector machine classifier (RS_SVM) is proposed for breast cancer diagnosis by Chen et al. and it was observed the proposed method achieved (the highest classification accuracies 99.41%, 100%, and 100%, average accuracies 96.55% 96.72% 96.87% for 50-50% of training-test partition, 70-30% of training-test partition, and 80-20% of training-test partition, respectively) for a subset that contained five features via 5-fold cross-validation [10].

2.2. **Feature selection.** Feature selection plays a very significant role for the success of the system in fields like pattern recognition and data mining. Feature selection provides a smaller but more distinguishing subset compared with the starting data, selecting the distinguishing features from a set of features and eliminating the irrelevant ones. This results in both reduced processing time and increased classification accuracy.

For feature selection, there are many methods in the literature covering a wide range from filtering to wrapping approaches [11,12]. In the filter approach, the goodness of an attribute or set of attributes is estimated by using only intrinsic properties of the data, while in the wrapper approach, the merit of a given candidate subset is obtained by learning and evaluating a classifier using only the variables included in the proposed subset [13]. Principal component analysis (PCA) and linear discriminant analysis (LDA) are the popular feature selection methods to reduce size [14].

In the recent years, many methods have been used for feature selection; particularly artificial intelligence, feature conversion methods and statistical methods: boosting feature selection for neural network based regression [15], filter model for feature subset selecting based on genetic algorithm [16], application of ant colony algorithm for feature selection [17], feature selection using particle swarm optimization [18], a discrete particle swarm optimization method for feature selection [19], feature selection by Weighted-SNR for cancer microarray data classification [20], Bhattacharyya space for feature selection [21], subspace based feature selection method [12], support vector-based feature selection using fisher's linear discriminant and support vector machine [22], HMM (Hidden Markov Models) based feature space transform for voice pathology detection [23] have been used.

3. **Wisconsin Breast Cancer Database.** We have used the Wisconsin breast cancer database (WBCD taken from the UCI machine learning repository) which was obtained from the University of Wisconsin Hospitals, Madison from Dr. W. H. Wolberg in our experiments. This dataset is commonly used among researchers who use machine learning methods for breast cancer classification, so it provides us to compare the performance of our method with that of others [1]. There are 699 records in this database. Each record in the database has nine attributes, each of which is represented as an integer between 1 and 10. The features are clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli and mitoses. This dataset contains 16 instances with missing attribute values. We substituted the missing data by frequently encountered values of own class. In this database, four hundred and fifty eight samples of the dataset (65.5%) belong to benign class, and two hundred and forty one samples of the dataset (34.5%) are of malignant class. The nine attributes are detailed in Table 1 [1,8,9,24,25].

TABLE 1. Wisconsin breast cancer data description of attributes

Attribute number	Attribute Description	Values of attributes
1	Clump thickness	1-10
2	Uniformity of cell size	1-10
3	Uniformity of cell shape	1-10
4	Marginal adhesion	1-10
5	Single epithelial cell size	1-10
6	Bare nuclei	1-10
7	Bland chromatin	1-10
8	Normal nucleoli	1-10
9	Mitoses	1-10
N = 699 observations, 241 malignant and 458 benign		

4. **Methods.**

4.1. **Association rules.** Data mining is the whole study of extracting meaningful and useful data from the current, meaningless data. Association rules find the relations between the values that features can take in large data sets [26]. For researchers and companies who are processing data, discovering useful association relations inside large amounts of data enables the researchers' studies and companies' activities to become much more efficient. While finding association rules in large databases, the criteria are whether each element is repeated as many as the pre-determined minimum support number and whether the frequently repeated elements constitute strong association rules [27].

The mathematical model of association rules was expressed by Agrawal et al. in 1993. In this model, $I = \{i_1, i_2, \ldots, i_m\}$ is defined as the set of objects and D is defined as the set of operations. Each operation within the database D is called T. The sets of objects are adjusted, providing $T \subseteq I$. Let us take A and B as sets of objects. If $A \subseteq T$, T set of operations contains A, which is formed by some of the elements of I. This association rule is expressed as $A \Rightarrow B$. Here, $A \subset I$, $B \subset I$ and $A \cap B = \emptyset$ [28,29]. Mathematical expressions related to Apriori algorithm are defined as in Equations (1)-(3):

$$\text{support}\,(A \Rightarrow B) = P\,(A \cup B) \tag{1}$$

$$\text{confidence}\,(A \Rightarrow B) = P\,(B/A) \text{ or} \tag{2}$$

$$\text{confidence}\,(A \Rightarrow B) = \text{support}\,(A \Rightarrow B)/\text{support}\,(A) \tag{3}$$

where support (A) = support ($A \Rightarrow A$).

In this algorithm, the first thing to do is to define the support value of the rule $A \Rightarrow B$. Support value is the probability of the operation T to contain $A \cup B$. Afterwards, the confidence value of the rule $A \Rightarrow B$ is defined. This probability is the probability for the operation T to contain A along with B. The associations that exceed the defined threshold values are taken into consideration. These are called exceptional patterns. The aim of association rules is to help find $A \Rightarrow B$ rules larger than the minimum support and confidence values provided externally to the algorithm [30].

Apriori algorithm. One of the significant algorithms for the introduction of association rules in the history of data mining, apriori algorithm was developed by Agrawal et al. in 1994. Apriori algorithm is one of the most frequently used algorithms for the extraction of association rules. The apriori algorithm takes its name from the word root "prior", since it acquires the information of frequently encountered (common) objects from the

previous step. This algorithm is based on the rule that "all the subsets of a common set of objects should also be common" [29]. The algorithm's pseudo-code is given in Table 2.

As shown in pseudo-code at Table 2, in order to find frequent data sets with k elements in a priori algorithm, database is scanned for k times. Subsets of values allocated for intensive object sets have to be frequent data sets. Product codes to be endingly then their numbers in the set are counted in order to be checked with support value. When data sets pass to the next stage, numbers are re-counted. Nominee set values are shown with C. Nominee sets of K data set include $c[1]$ elements data set, $c[2]$ elements data set .... $C[k]$ elements data sets. Data sets over support value of C nominee data sets are transmitted into intensive object sets. Intensive object sets are represented with L. Find candidate objects. Support values are below the values considered [21]. C candidate objects are all sub-clusters. Dense clusters of objects; this procedure is carried out non-candidates.

After finding intensive object sets over database, strong rules can be constituted with these object sets. These rules provide both minimum support and reliability values. These rules can be obtained by using Equations (4) and (5) [30].

$$\textit{Confidence}\,(\mathrm{A} \Rightarrow \mathrm{B}) : \mathrm{P}\,(\mathrm{B} \backslash \mathrm{A}) = \textit{Support Value}\,(\mathrm{A} \cup \mathrm{B}) / \textit{Support Value}\,(\mathrm{A}) \tag{4}$$

where (A ∪ B) *Support Value* is set numbers in which A and B objects are together, (A) *Support Value* is objects in which A is alone. Association rules like following can be obtained from this Equation (4). Scan X object and non-empty sub sets of X are found.

For non-empty sets;

$$\text{"s} \Rightarrow (\mathrm{X} - \mathrm{s})\text{" if support (X)/supportdestek (s)} \geq \min\_\mathrm{rel} \tag{5}$$

Equation is implemented. Values over Min_rel (minimum reliability value) are taken.

TABLE 2. Pseudo-code of Apriori Algorithm

```
Input: Database (D), minimum support min_sup.
Output: L, D as well as dense clusters of objects.
L1=find_frequent_1-itemsets(D)
//Single-element clusters are dense objects.
for (k=2;L(k-1)≠0;k++)
{  Ck=apriori_gen(L(k-1), min_sup)
    for each transaction t∈D
    { // Scan count for D
        Ct=subset(Ck,t);
        for each candidate c∈Ct
        c.count++;      }
                        Lk={ c∈Ck | c.count ≥min_sup   }     }
RT dense clusters of objects, objects to obtain the candidate Ck.
  return L=Uk Lk;

procedure apriori_gen (L(k-1);min_sup)
{ for each itemset l1∈L(k-1)
   for each itemset l2∈L(k-1)
    if (l1[1]= l2[1] )∧(l1[2]=l2[2]) ∧…∧ (l1[k-1]=l2[k-1]) then
     { c=l1x l2; // Generation of candidate objects.
       if has_infrequent_subset(c,L(k-1)) then
       delete c; // Deletion of unnecessary candidates.
       else add c to Ck;      }
   return Ck;}

procedure has_infrequent_subset(c,L(k-1));
{   for each (k-1)-subset s of c
   if s∉L(k-1) then
     return TRUE
   return FALSE }
```

4.2. **Principal component analysis.** Principal Component Analysis (PCA) is a statistical method frequently used for data analysis. PCA is a conversion technique which makes it possible to reduce the size of data sets which include a large number of interrelated features, so that the current data can be expressed with a fewer number of variables. The converted variables are named as the principal components of the first variables, and the first of the fundamental variables is the highest variance value. Other principal components are ordered with descending variance values [31]. The feature reduction method of PCA can be explained as below:

The M is a t-dimensional data set. The n principal axes $G_1, G_2, \ldots, G_n$ here $1 \leq n \leq t$, are orthonormal axes onto which the retained variance is maximum in the projected space [32]. Commonly $G_1, G_2, \ldots, G_n$ can be given by the n leading eigenvectors of the sample covariance matrix:

$$C = \left(\frac{1}{L}\right) \sum_{k=1}^{L} (x_k - \bar{x})^T (x_k - \bar{x}). \tag{6}$$

Here $x_k \in M$, $\bar{x}$ is the mean of samples, L is the number of samples. According to this:

$$UG_k = v_k G_k, \quad k \in 1, \ldots, n, \tag{7}$$

Here v_k is the kth largest eigenvalue of U. The n principal components of a given observation vector $x_k \in M$ are given as below:

$$Q = [q_1, q_2, q_3, \ldots, q_n] = \left[\left[G_1^T x, G_2^T x, \ldots, G_n^T x\right] = G^T x\right] \tag{8}$$

There, q is the n principal components of x [33].

4.3. **Artificial neural networks.** Artificial Neural Networks (NN) are biologically inspired, intelligent techniques and they have a number of simple and highly interconnected layers of neurons. Multilayered perceptron neural networks (MLPNNs) are the simplest NN architectures, and therefore most commonly used [34,35]. The MLP structure is seen in Figure 1.

An MLPNN has mainly three layers: an input layer, an output layer, and an intermediate or hidden layer. The input layer neurons distribute the input signals x_i to neurons in the hidden layer (s). Each hidden layer neuron j sums up its input signals x_i after weighing them with the strengths of the respective connections w_{ji} from the input layer

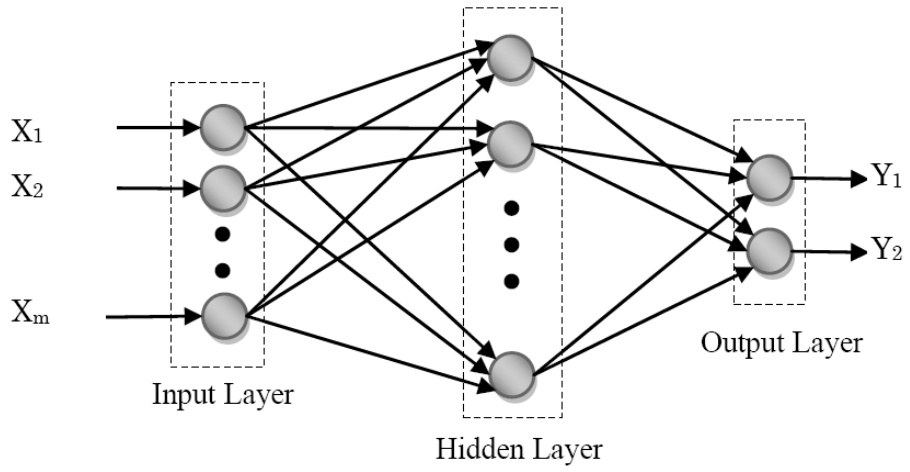

FIGURE 1. Multi-layer perceptron structure

and computes its output y_j as a function f of the sum:

$$y_j = f\left(\sum w_{ji}x_i\right) \tag{9}$$

where f can be a sigmoid or hyperbolic tangent function. The output of neurons in the output layer is computed similarly.

Training a network consists of adjusting weights of the network using a learning algorithm. The Back-Propagation [36] learning algorithm is used in this study. It is a gradient descent with momentum and adaptive learning rate backpropagation that gives the change $\Delta w_{ij}(k)$ in the weight of a connection between neurons i and j as follows:

$$\Delta w_{ji}(k) = \alpha\delta_j x_i + \mu\Delta w_{ji}(k-1) \tag{10}$$

where x_i is the input, α is the learning coefficient, μ is the momentum coefficient, and δ_i is a factor depending on whether neuron j is an output neuron or a hidden neuron. For output neurons,

$$\delta_j = \frac{\partial f}{\partial net_j}\left(y_j^T - y_i\right) \tag{11}$$

where $net_j = \sum x_i w_{ji}$ and y_j^T is the target output for neuron j. For hidden neurons,

$$\delta_j = \frac{\partial f}{\partial net_j}\sum_q w_q\delta_q \tag{12}$$

As there are no target outputs for hidden neurons in Equation (12), the difference between the target and actual output of a hidden neuron j is replaced by the weighted sum of the δ_q terms already obtained for neurons q connected to the output of j. Thus, iteratively beginning with the output layer, the δ term is computed for all neurons in all layers except the input layer and weights were then updated according to Equation (10).

4.4. **Performance evaluation.** Four methods for performance evaluation of breast cancer diagnosis are used. These methods are classification accuracy, confusion matrix, analysis of sensitivity and specificity, and k-fold cross validation.

4.4.1. *Classification accuracy.* In this study, the classification accuracies for the datasets are measured using the equation:

$$\text{accuracy } (T) = \frac{\sum_{i=1}^{N}\text{assess } (t_i)}{N}, \quad t_i \in T \tag{13}$$

$$\text{assess } (t_i) = \begin{cases} 1, & \text{if classify } (t_i) \equiv \text{correctclassication}, \\ 0, & \text{otherwise}, \end{cases} \tag{14}$$

where T is the set of data items to be classified (the test set). N is the number of testing samples of the dataset. We will also show the accuracy of our performed k-fold cross validation (CV) experiment.

4.4.2. *Confusion matrix.* The confusion matrix contains four classification performance indices: true positive, false positive, false negative, and true negative as shown in Table 3. These four indices are also usually used to evaluate the performance the two-class classification problem [37].

TABLE 3. The four classification performance indices included in the confusion matrix

Actual class	**Predicted class**	
	Positive	Negative
Positive	True positive (TP)	False negative (FN)
Negative	False positive (FP)	True negative (TN)

4.4.3. *Analysis of sensitivity and specificity.* For sensitivity, specificity, positive predictive value and negative predictive value, we use the following expressions [7].

$$\text{Sensitivity } (\%) = \frac{TP}{TP + FN} \times 100 \tag{15}$$

$$\text{Specificity } (\%) = \frac{TN}{TN + FP} \times 100 \tag{16}$$

$$\text{Positive predictive value } (\%) = \frac{TP}{TP + FP} \times 100 \tag{17}$$

$$\text{Negative predictive value } (\%) = \frac{TN}{TN + FN} \times 100 \tag{18}$$

4.4.4. *k-fold cross-validation.* k-fold cross-validation is used for the test result to be more valuable [38]. In k-fold cross-validation, the original sample is partitioned into k subsamples randomly. Of the k sub-samples, a single sub-sample is retained as the validation data for testing the model, and the remaining $k - 1$ sub-samples are used as training data. The cross-validation process is then repeated k times (the folds), with each of the k sub-samples used exactly once as the validation data. The average of k results from the folds gives the test accuracy of the algorithm [39].

5. **Experimental Application.** In this study, AP + NN method is proposed to be used in the problem of breast cancer. Our method consists of 2 phases, as shown in Figure 2. In the first phase, the feature selection methods, namely Apriori and PCA are used respectively. Therefore, the important features have been selected and feature vector size has been reduced. In the second phase, these reduced data are used for the artificial neural network and classification has been performed.

5.1. **Apriori + PCA (AP).** Apriori + PCA (AP) is a hybrid feature selection method combining Apriori, which is used to find the relations between features in large databases, and PCA, which is a conversion technique which makes it possible to reduce the size of data sets which contain inter-related variables. The advantages of this hybrid system are that it enables selecting of significant inputs by eliminating unnecessary ones, makes it possible to deal with less size of data, increases the classification competence of the

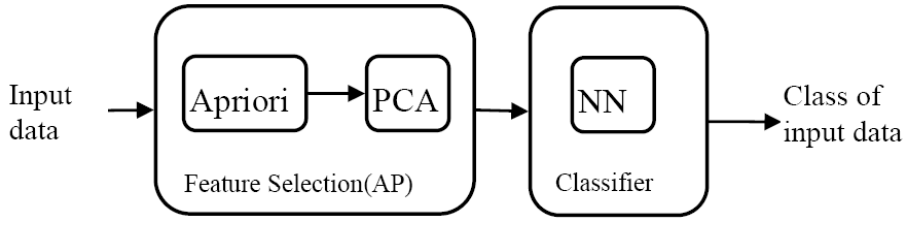

FIGURE 2. Block diagram of the proposed system

system, and decreases the requirement for memory and capacity thanks to more efficient work in smaller-size spaces.

On the first phase of the AP technique, the association rules between input parameters have been detected. While detecting these relations, rules that possess a sufficient support value and confidence value have been extracted. Two rules that have turned out to hold the highest level of security and that can be used in feature selection are provided on Table 4.

Table 4. Rules taken into consideration

Input	Value	confidence
$1, 2, 3, 7 => 9$	$1, 1, 1, 3 => 1$	100%
$1, 3, 8, 9 => 2$	$1, 1, 1, 1 => 1$	100%

According to the first rule, if the 1, 2, 3, 7 input parameters have the same value, 9nd input parameter should also have the same value. Therefore, we can say that the ninth input value depends on the other inputs. That is why, the 9nd input parameter is not used while inputting for NN. The other inputs have been converted into PCA space and one more input parameter have been eliminated from this space. The parameters reduced in size have been provided into NN as input parameters.

According to the second rule, if the 1, 3, 8, 9 input parameters have the same value, 2nd input parameter should also have the same value. Therefore, we can say that the second input value depends on the other inputs. That is why, the 2nd input parameter is not used while inputting for NN. The other inputs have been converted into PCA space and two input parameters have been eliminated from this space. The parameters reduced in size have been provided into NN as input parameters.

The artificial neural network results of these two rules are 98.14% and 98.29%, respectively. Since the second rule has a better result according to these ANN results, the second rule has been selected to compare with the others.

5.2. **NN layer.** Values obtained from AP algorithm, are used as input of multi-layer perceptron neural network classifier. MLP architecture and training parameters used in the study are given in Table 5. The length of the whole data was 699 lines. The inputs were clump thickness, uniformity of cell size, uniformity of cell shape, marginal adhesion, single epithelial cell size, bare nuclei, bland chromatin, normal nucleoli and mitoses. Malignant and benign were the NN output. Firstly, these inputs analyze apriori algorithm for feature selection. According to Apriori algorithm results, one of inputs named as uniformity of cell size and mitoses can be eliminated. So, input numbers can be reduced to eight.

Table 5. MLP architecture and training parameters

Architecture	
The number of layers	3
The number of neuron on the layers	Input:6, Hidden:18, Output:1
The initial weights and biases	Random
Activation Functions	Tangent-sigmoid (for hidden layer) Log-sigmoid (for output layer)
Training parameters	
Learning rule	A gradient descent with momentum and adaptive learning rate backpropagation
Goal	0.000001

Obtained data set converted PCA data space and then most significant six data columns are used NN input. Consequently, the number of nodes for input NN and output NN were formed as 6 and 1, respectively. For best NN structure, the number of hidden layer nodes obtained 18. The developed NN structure included one hidden layer. MLP feed forward back-propagation was used as an NN structure. In the training of neural network, A gradient descent with momentum and adaptive learning rate backpropagation algorithm is used. We used this method as a 10-fold cross-validation in our applications.

6. **Experimental Results and Discussion.** Using AP hybrid feature selection method on the first phase, related features have been selected and feature vector size has been reduced. Using 10-fold cross-validation method in the second phase, NN classification has been performed. In the cross-validation, whole data are divided into equal ten parts named as CV-1, CV-2, ..., CV-10. A column has been eliminated in the apriori result, as well as another column in PCA space. The result on Table 6 has been achieved by eliminating the 9th and 2nd columns respectively from the apriori and reducing a column from the PCA space. According to these results, it has been seen that the result achieved by eliminating the 2nd column and reducing a column from the PCA space is better, thus selected for comparison.

The NN performance results (sensitivity, specificity, positive predictive value, negative predictive value, correct classification rate) acquired by means of implementing the 10-fold cross-validation are indicated in Table 7.

TABLE 6. Correct classification rate for breast cancer detection using AP + NN

The classifier	**Epochs**	**The neuron number of the hidden layer**	**Average of Correct classification rate (%)**
AP + NN (Except for 2) $(6, 18, 1)$	325	18	98.29
AP + NN (Except for 9) $(6, 3, 1)$	350	3	98.14

TABLE 7. Performance results for breast cancer detection using AP + NN (except for 2)

The classifier	**Cross validation partitions**	**Sensitivity (%)**	**Specificity (%)**	**Positive predictive value (%)**	**Negative predictive value (%)**	**Correct classification rate (%)**
AP + NN (Except for 2) $(6, 18, 1)$	CV-1	100	100	100	100	100
	CV-2	100	100	100	100	100
	CV-3	100	92.59	95.56	100	97.14
	CV-4	95.45	96.15	97.67	92.59	95.71
	CV-5	97.78	100	100	96.15	98.57
	CV-6	97.92	100	100	95.65	98.57
	CV-7	97.50	100	100	96.77	98.57
	CV-8	97.73	100	100	96.30	98.57
	CV-9	97. 73	100	100	96.30	98.57
	CV-10	95.92	100	100	91.30	97.14
	Average	98.03	98.87	99.32	96.51	**98.29**

TABLE 8. Confusion matrices obtained using our method on cross-validation partitions

Cross validation partitions	**Actual class**	**Number of predicted "benign"**	**Number of predicted "malignant"**	**Cross validation partitions**	**Actual class**	**Number of predicted "benign"**	**Number of predicted "malignant"**
CV-1	Benign	49	0	CV-6	Benign	47	1
	Malignant	0	20		Malignant	0	22
CV-2	Benign	52	0	CV-7	Benign	39	1
	Malignant	0	18		Malignant	0	30
CV-3	Benign	43	0	CV-8	Benign	43	1
	Malignant	2	25		Malignant	0	26
CV-4	Benign	42	2	CV-9	Benign	43	1
	Malignant	1	25		Malignant	0	26
CV-5	Benign	44	1	CV-10	Benign	47	2
	Malignant	0	25		Malignant	0	21

TABLE 9. Classification accuracies obtained with our proposed system and other classifiers from literature

Author (Year)	Method	Classification accuracy (%)
Goodman et al. (2002)	Optimized-LVQ (10 × CV)	96.70
	Big-LVQ (10 × CV)	96.80
	AIRS (10 × CV)	97.20
Abonyi and Szeifert (2003)	Supervised fuzzy clustering (10 × CV)	95.57
Polat and Gunes (2007)	LS-SVM (10 × CV)	98.53
Karabatak and Ince (2009)	AR1 + NN (3 × CV)	97.4
	AR2 + NN (3 × CV)	95.6
Yeh et al. (2009)	DPSO (without CV)	98.71
Marcano-Cedeño et al. (2011)	AMMLP (without CV)	99.26
Chen et al. (2011)	RS_SVM (5 × CV) (train: 50%-test: 50%)	96.55
	RS_SVM (5 × CV) (train: 70%-test: 30%)	96.72
	RS_SVM (5 × CV) (train: 80%-test: 20%)	96.87
Our study	AP + NN (10 × CV)	98.29

Classification results of the network were displayed by using a confusion matrix. The confusion matrix contains four classification performance indices: true positive, false positive, false negative, and true negative. Confusion matrices obtained using our method on cross-validation partitions are given in Table 8.

The comparison of our suggested system with the other systems has been given in Table 9. The result acquired reveals that the average correctness rate of the studies performed so far on cancer data by employing the method of k-fold cross-validation is a very promising result.

7. **Conclusions.** In this study, a new feature selection method (AP) which combines the Apriori and PCA techniques for the detection of breast cancer has been used. Firstly, all inputs analyze Apriori algorithm for feature selection. According to Apriori algorithm results, one of inputs named as uniformity of cell size and mitoses can be eliminated. So, input numbers can be reduced to eight. Obtained data set converted PCA data space and then most significant six data columns are selected. The output of this newly developed hybrid preprocessing algorithm has been applied to multi-layered feed-forward back-propagation neural network, which is a conventional classifier. For the training and testing phases to be more reliable on a scientific basis, 10-fold cross-validation method has been utilized. Consequently, the number of nodes for input NN and output NN were

formed as 6 and 1, respectively. For best NN structure, the number of hidden layer nodes obtained 18. In order to evaluate the performance of this system, Winconsin breast cancer database has been used. The detailed accuracy values of the suggested AP + NN system are given in Table 7 and the average accurate classification rate of AP + NN system is 98.29%. This average value has been observed to have better results compared with other systems performed on cancer data using cross-validation. As indicated by this research, the success of the artificial neural network increases even further by selecting of related data, converting the data to another space, and eliminating useless and distortive data. We believe that this study will contribute to the development of faster and more reliable automatic diagnostic systems in the area of fight against cancer, where early diagnosis saves lives.

Acknowledgment. The authors are grateful to Selcuk University Scientific Research Projects Coordinatorship for press support of the manuscript. The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] M. F. Akay, Support vector machines combined with feature selection for breast cancer diagnosis, *Expert Systems with Applications*, vol.36, no.2, pp.3240-3247, 2009.
[2] World Health Organization, *Cancer*, http://www.who.int/cancer/en/, 2011.
[3] T. S. Lee, S. M. Chou, Y. E. Shao et al., Mining the breast cancer pattern using artificial neural networks and multivariate adaptive regression splines, *Expert Systems with Applications*, vol.27, no.1, pp.133-142, 2004.
[4] W. C. Yeh, W. W. Chang and Y. Y. Chung, A new hybrid approach for mining breast cancer pattern using discrete particle swarm optimization and statistical method, *Expert Systems with Applications*, vol.36, no.4, pp.8204-8211, 2009.
[5] D. E. Goodman, L. Boggess and A. Watkins, Artificial immune system classification of multiple-class problems, *Proc. of the Artificial Neural Networks in Engineering ANNIE*, pp.179-183, 2002.
[6] J. Abonyi and F. Szeifert, Supervised fuzzy clustering for the identification of fuzzy classifiers, *Pattern Recognition Letters*, vol.24, no.14, pp.2195-2207, 2003.
[7] K. Polat and S. Gunes, Breast cancer diagnosis using least square support vector machine, *Digital Signal Processing*, vol.17, no.4, pp.694-701, 2007.
[8] M. Karabatak and M. C. Ince, An expert system for detection of breast cancer based on association rules and neural network, *Expert Systems with Applications*, vol.36, no.2, pp.3465-3469, 2009.
[9] A. Marcano-Cedeno, J. Quintanilla-Dominguez and D. Andina, WBCD breast cancer database classification applying artificial metaplasticity neural network, *Expert Systems with Applications*, vol.38, no.8, pp.9573-9579, 2011.
[10] H. L. Chen, B. Yang, J. Liu et al., A support vector machine classifier with rough set-based feature selection for breast cancer diagnosis, *Expert Systems with Applications*, vol.38, no.7, pp.9014-9022, 2011.
[11] R. Kohavi and G. H. John, Wrappers for feature subset selection, *Artificial Intelligence*, vol.97, no.1-2, pp.273-324, 1997.
[12] S. Gunal and R. Edizkan, Subspace based feature selection for pattern recognition, *Information Sciences*, vol.178, no.19, pp.3716-3726, 2008.
[13] P. Bermejo, J. A. Gamez and J. M. Puerta, A GRASP algorithm for fast hybrid (filter-wrapper) feature subset selection in high-dimensional datasets, *Pattern Recognition Letters*, vol.32, no.5, pp.701-711, 2011.
[14] M. Karabatak and M. C. Ince, A new feature selection method based on association rules for diagnosis of erythemato-squamous diseases, *Expert Systems with Applications*, vol.36, no.10, pp.12500-12505, 2009.
[15] K. Bailly and M. Milgram, Boosting feature selection for neural network based regression, *Neural Networks*, vol.22, no.5-6, pp.748-756, 2009.
[16] M. E. Elalami, A filter model for feature subset selection based on genetic algorithm, *Knowledge-Based Systems*, vol.22, no.5, pp.356-362, 2009.

[17] M. H. Aghdam, N. Ghasem-Aghaee and M. E. Basiri, Application of ant colony optimization for feature selection in text categorization, *IEEE Congress on Evolutionary Computation*, pp.2867-2873, 2008.
[18] C.-C. Lai, C.-H. Wu and M.-C. Tsai, Feature selection using particle swarm optimization with application in Spam filtering, *International Journal of Innovative Computing Information and Control*, vol.5, no.2, pp.423-432, 2009.
[19] A. Unler and A. Murat, A discrete particle swarm optimization method for feature selection in binary classification problems, *European Journal of Operational Research*, vol.206, no.3, pp.528-539, 2010.
[20] S. Hengpraprohm and P. Chongstitvatana, Feature selection by weighted-snr for cancer microarray data classification, *International Journal of Innovative Computing Information and Control*, vol.5, no.12(A), pp.4627-4635, 2009.
[21] C. C. Reyes-Aldasoro and A. Bhalerao, The Bhattacharyya space for feature selection and its application to texture segmentation, *Pattern Recognition*, vol.39, no.5, pp.812-826, 2006.
[22] E. Youn, L. Koenig, M. K. Jeong et al., Support vector-based feature selection using fisher's linear discriminant and support vector machine, *Expert Systems with Applications*, vol.37, no.9, pp.6148-6156, 2010.
[23] J. I. Godino-Llorente, J. D. Arias-Londono, N. Saenz-Lechon et al., An improved method for voice pathology detection by means of a HMM-based feature space transformation, *Pattern Recognition*, vol.43, no.9, pp.3100-3112, 2010.
[24] E. D. Ubeyli, Implementing automated diagnostic systems for breast cancer detection, *Expert Systems with Applications*, vol.33, no.4, pp.1054-1062, 2007.
[25] W. H. Wolberg, *Wisconsin Breast Cancer Database 2011*, http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.names.
[26] J. W. Han and W. J. Fu, Mining multiple-level association rules in large databases, *IEEE Transactions on Knowledge and Data Engineering*, vol.11, no.5, pp.798-805, 1999.
[27] M. J. Zaki, Parallel and distributed association mining: A survey, *IEEE Concurrency*, vol.7, no.4, pp.14-25, 1999.
[28] R. Agrawal, T. Imielinski and A. Swami, Mining association rules between sets of items in large databases, *ACM SIGMOD Conference on Management of Data*, Washington, 1993.
[29] R. Agrawal and R. Srikant, Fast algorithms for mining association rules, *Proc. of the 20th VLDB Conference*, Santiago, Chile, 1994.
[30] J. Han and M. Kamber, *Data Mining: Concepts and Techniques*, Morgan Kaufmann Publishers, Burnaby, 2000.
[31] I. T. Jolliffe, *Principal Component Analysis*, 2nd Edition, Springer, New York, 2002.
[32] P. N. Belhumeur, J. P. Hespanha and D. J. Kriegman, Eigenfaces vs. fisherfaces: Recognition using class specific linear projection, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol.19, no.7, pp.711-720, 1997.
[33] D. Calisir and E. Dogantekin, A new intelligent hepatitis diagnosis system: PCA-LSSVM, *Expert Systems with Applications*, vol.38, no.8, pp.10705-10708, 2011.
[34] A. Maren, C. Harston and R. Pap, *Handbook of Neural Computing Applications*, Academic Press, London, 1990.
[35] S. Haykin, *Neural Networks: A Comprehensive Foundation*, Macmillan College Publishing Company, New York, 1994.
[36] D. E. Rumelhart and J. L. McClelland, *Parallel Distributed Processing*, The MIT Press, Cambridge, 1986.
[37] Y. Xu, Z. Qi and J. Wang, Breast cancer diagnosis based on a kernel orthogonal transform, *Neural Comput. & Applic.*, 2011.
[38] D. Francois, F. Rossi, V. Wertz et al., Resampling methods for parameter-free and robust feature selection with mutual information, *Neurocomputing*, vol.70, pp.1276-1288, 2007.
[39] N. A. Diamantidis, D. Karlis and E. A. Giakoumakis, Unsupervised stratification of cross-validation for accuracy estimation, *Artificial Intelligence*, vol.116, pp.1-16, 2000.