Filter Based Feature Selection Methods for Prediction of Risks in Hepatitis Disease

Pinar Yildirim

Abstract—**Recently, large amount of data is widely available in information systems and data mining has attracted a big attention to researchers to turn such data into useful knowledge. This implies the existence of low quality, unreliable, redundant and noisy data which negatively affect the process of observing knowledge and useful pattern. Therefore, researchers need relevant data from huge records using feature selection methods. Feature selection is the process of identifying the most relevant attributes and removing the redundant and irrelevant attributes. In this study, a comparison between filter based feature selection methods based on a well-known dataset (i.e., hepatitis dataset) was carried out and four classification algorithms were used to evaluate the performance of the algorithms. Among the algorithms, Naïve Bayes and Decision Table classifiers have higher accuracy rates on the hepatitis dataset than the others after the application of feature selection methods. The study revealed that feature selection methods are capable to improve the performance of learning algorithms. However, no single filter based feature selection method is the best. Overall, Consistency Subset, Info Gain Attribute Eval, One-R Attribute Eval and Relief Attribute Eval methods performed better results than the others.**

Index Terms—**Feature selection, hepatitis, J48, naïve bayes, IBK, decision table.**

I. Introduction

Recently, thanks to innovations of computer and information technologies, huge amounts of data can be obtained and stored in both scientific and business transactions. This amount of data implies low quality, unreliable, redundant and noisy data to observe useful patterns [1]. Therefore, researchers need relevant and high-quality data from huge records using feature selection methods.

Feature selection methods reduce the dimensionality of feature space, remove redundant, irrelevant or noisy data. It brings the immediate effects for application: speeding up a data mining algorithm, improving the data quality and the performance of data mining and increasing the comprehensibility of the mining results [2].

In this study, the great interest of hepatitis disease was considered which is a serious health problem in the world and a comparative analysis of several filter based selection algorithms was carried out based on the performance of four classification algorithms for the prediction of disease risks [3]. The main aim of this study is to make contributions in the prediction of hepatitis disease for medical research and introduce a detailed and comprehensive comparison of popular filter based feature selection methods.

II. Feature Selection Methods

Several feature selection methods have been introduced in the machine learning domain. The main aim of these techniques is to remove irrelevant or redundant features from the dataset. Feature selection methods have two categories: wrapper and filter. The wrapper evaluates and selects attributes based on accuracy estimates by the target learning algorithm. Using a certain learning algorithm, wrapper basically searches the feature space by omitting some features and testing the impact of feature omission on the prediction metrics. The feature that make significant difference in learning process implies it does matter and should be considered as a high quality feature. On the other hand, filter uses the general characteristics of data itself and work separately from the learning algorithm. Precisely, filter uses the statistical correlation between a set of features and the target feature. The amount of correlation between features and the target variable determine the importance of target variable [1], [4]. Filter based approaches are not dependent on classifiers and usually faster and more scalable than wrapper based methods. In addition, they have low computational complexity.

A. Information Gain

Information gain (relative entropy, or Kullback-Leibler divergence), in probability theory and information theory, is a measure of the difference between two probability distributions. It evaluates a feature X by measuring the amount of information gained with respect to the class (or group) variable Y, defined as follows:

$$\mathrm{I}(X) = \mathrm{H}\ (\mathrm{P}(Y)\text{-}\mathrm{H}\ (\mathrm{P}(Y/X)) \tag{1}$$

Specifically, it measures the difference the marginal distribution of observable Y assuming that it is independent of feature $X(\mathrm{P}(Y))$ and the conditional distribution of Y assuming that is dependent of X $(\mathrm{P}(Y/X))$. If X is not differentially expressed, Y will be independent of X, thus X will have small information gain value, and vice versa [5].

B. Relief

Relief-F is an instance-based feature selection method which evaluates a feature by how well its value distinguishes samples that are from different groups but are similar to each other. For each feature X, Relief-F selects a random sample and k of its nearest neighbors from the same class and each of different classes. Then X is scored as the sum of weighted differences in different classes and the same class. If X is differentially expressed, it will show greater differences for

Manuscript received November 5, 2014; revised February 20, 2015.

Pinar Yildirim is with the Okan University, Istanbul, Turkey (e-mail: pinar.yildirim@ okan.edu.tr).

DOI: 10.7763/IJMLC.2015.V5.517

samples from different classes, thus it will receive higher score (or vice versa) [5].

C. One-R

One-R is a simple algorithm proposed by Holte [6]. It builds one rule for each attribute in the training data and then selects the rule with the smallest error. 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 [2].

D. Principal Component Analysis (PCA)

The aim of PCA is to reduce the dimensionality of dataset that contains a large number of correlated attributes by transforming the original attributes space to a new space in which attributes are uncorrelated. The algorithm then ranks the variation between the original dataset and the new one. Transformed attributes with most variations are saved; meanwhile discard the rest of attributes. It’s also important to mention that PCA is valid for unsupervised data sets because it doesn’t take into account the class label [1], [7].

E. Correlation Based Feature Selection (CFS)

CFS is a simple filter algorithm that ranks feature subsets and discovers the merit of feature or subset of features according to a correlation based heuristic evaluation function. The purpose of CFS is to find subsets that contain features that are highly correlated with the class and uncorrelated with each other. The rest of features should be ignored. Redundant features should be excluded as they will be highly correlated with one or more of the remaining features. The acceptance of a feature will depend on the extent to which it predicts classes in areas of the instance space not already predicted by other features. CFS’s feature subset evaluation function is shown as follows [8]:

$$Merit_s = \frac{k r_{cf}}{\sqrt{k + (k+1) r_{ff}}} \quad (2)$$

where $Merit_s$ is the heuristic “merit” of a feature subset S containing k features, r_{cf} is the mean feature-class correlation ($f \in s$), and r_{ff} is the average feature-feature intercorrelation. This equation is, in fact, Pearson’s correlation, where all variables have been standardized. The numerator can be thought of as giving an indication of how predictive of the class a group of features are; the denominator of how much redundancy there is among them. The heuristic handles irrelevant features as they will be poor predictors of the class. Redundant attributes are discriminated against as they will be highly correlated with one or more of the other features [9].

F. Consistency Based Subset Evaluation (CS)

CS adopts the class consistency rate as the evaluation measure. The idea is to obtain a set of attributes that divide the original dataset into subsets that contain one class majority [8]. One of well known consistency based feature selection is consistency metric proposed by Liu and Setiono [10].

$$Consistency_s = 1 - \frac{\sum_{j=0}^{k} |D_j| - |M_j|}{N} \quad (3)$$

where s is feature subset, k is the number of features in s, $|D_j|$ is the number of occurrences of the jth attributes value combination, $|M_j|$ is the cardinality of the majority class for the jth attribute’s value, and N is the number of features in the original dataset [10].

III. Classification Algorithms

A wide range of classification algorithms is available, each with its strengths and weaknesses. There is no single learning algorithm that works best on all supervised learning problems. This section gives a brief overview of four supervised learning algorithms used in this study, namely, J48, Naïve Bayes, IBK and Decision Table [2].

A. J48

J48 is the Weka implementation of the C4.5 algorithm, based on the ID3 algorithm. The main idea is to create the tree by using the information entropy. For each node the most effectively split criteria is calculated and then subsets are generated. To get the split criteria the algorithm looks for the attribute with highest normalized information gain.

The last step is called pruning, the algorithm starts at the bottom of the tree and removes unnecessary nodes, so the height of the tree can be reduced by deleting double information.

B. Naïve Bayes

The Naïve Bayes algorithm is a simple probabilistic classifier that calculates a set of probabilities by counting the frequency and combinations of values in a given data set. The algorithm uses Bayes theorem and assumes all attributes to be independent given the value of the class variable. This conditional independence assumption rarely holds true in real world applications, hence the characterization as Naïve yet the algorithm tends to perform well and learn rapidly in various supervised classification problems [11], [12]

C. IBK

IBK is an instance-based learning approach like the K-nearest neighbour method. The basic principle of this algorithm is that each unseen instance is always compared with existing ones using a distance metric; most commonly Euclidean distance and the closest existing instance are used to assign the class for the test sample [13].

D. Decision Table

Decision Table summarizes the dataset with a ‘decision table’, a decision table contains the same number of attributes as the original dataset, and a new data item is assigned a category by finding the line in the decision table that matches the non-class values of the data item. This implementation employs the wrapper method to find a good subset of attributes for inclusion in the table. By eliminating attributes that contribute little or nothing to a model of the dataset, the algorithm reduces the likelihood of over-fitting and creates a smaller, more condensed decision table [14], [15].

IV. Data Description

Hepatitis dataset is available at UCI machine learning data repository contains 19 fields with one class attribute. The dataset includes both numeric and nominal attributes. The class shows whether patients with hepatitis are alive or dead. The intention of the dataset is to forecast the presence or absence of hepatitis virus given the results of various medical tests carried out on a patient (Table I). The hepatitis dataset contains 155 samples belonging to two different target classes. There are 19 features, 13 binary and 6 features with 6-8 discrete values. Out of total 155 cases, the class variable contains 32 cases that died due to hepatitis [3], [16].

TABLE I: Hepatitis Dataset

No	Variable	Values
1	Age	10,20,30,40,50,60,70,80
2	Sex	Male,Female
3	Steroid	No,Yes
4	Antivirals	No,Yes
5	Fatique	No,Yes
6	Malaise	No,Yes
7	Anorexia	No,Yes
8	Liver Big	No,Yes
9	Liver Firm	No,Yes
10	Pleen Palpable	No,Yes
11	Spiders	No,Yes
12	Ascites	No,Yes
13	Varices	No,Yes
14	Biliburin	0.39,0.80,1.20,2.0,3.0,4.0
15	Alk Phosphate	33,80,120,160,200,250
16	Sgot	13,100,200,300,400,500
17	Albumin	2.1,3.0,3.8,4.5,5.0,6.0
18	Protime	10,20,…,90
19	Histology	No,Yes
20	Class	Die,Alive

V. Literature Review

There are several studies based on data mining of biomedical datasets in the literature. Sathyadevi *et al.*, used CART, C4.5 and ID3 algorithms to diagnose hepatitis disease effectively. According their results, CART algorithm performed best results to identify to disease [17].

Roslina *et al.* utilized Support Vector Machines to predict hepatitis and used wrapper based feature selection method to identify relevant features before classification. Combining wrapper based methods and Support vector machines produced good classification results [18]. Sartakhti *et al.* also presented a novel machine learning method using hybridized Support Vector machine and simulated annealing to predict hepatitis. They obtained high classification accuracy rates [19].

Harb *et al.* proposed the filter and wrapper approaches with Particle Swarm Optimization (PSO) as a feature selection method for medical data. They applied different classifiers to the datasets and compared the performance of the proposed methods with another feature selection algorithm based on genetic approach. Their results illustrated that the proposed model shows the best classification accuracy among the others [20].

Huang *et al.* applied a filter-based feature selection method using inconsistency rate measure and discretization, to a medical claims database to predict the adequacy of duration of antidepressant medication utilization. They used logistic regression and decision tree algorithms. Their results suggest it may be feasible and efficient to apply the filter-based feature selection method to reduce the dimensionality of healthcare databases [21].

Inza *et al.* investigated the crucial task of accurate gene selection in class prediction problems over DNA microarray datasets. They used two well-known datasets involved in the diagnosis of cancer such as Colon and Leukemia. The results highlighted that filter and wrapper based gene selection approaches lead to considerably better accuracy results in comparison to the non-gene selection procedure, coupled with interesting and notable dimensionality reductions [22].

VI. Experimental Results

Hepatitis dataset was used to compare different filter based feature selection methods for the prediction of disease risks. Four classification algorithms reviewed above were considered to evaluate classification accuracy.

The feature selection methods are,

- Cfs Subset Eval
- Principal Components
- Consistency Subset Eval
- Info Gain Attribute Eval
- One-R Attribute Eval
- Relief Attribute Eval

At first, feature selection methods were used to find relevant features in the hepatitis dataset and then, classification algorithms were applied to the selected features to evaluate the algorithms. Respectively, 10, 12, 16 and 19 features were selected by the feature selection algorithms. Same experiment was repeated for four classifiers. WEKA 3.6.8 software was used. WEKA is a collection of machine learning algorithms for data mining tasks and is an open source software. The software contains tools for data pre-processing, feature selection, classification, clustering, association rules and visualization [7], [23].

Some performance measures were used for the evaluation of the classification results, where TP/TN is the number of True Positives/Negatives instances, FP/FN is the number of False Positives/Negatives instances.

Precision is a proportion of predicted positives which are actual positive:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4}$$

Recall is a proportion of actual positives which are predicted positive:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{5}$$

Precision and recall measures are utilized to find the best method, but it is not easy to make decision. Thus, F-measure was used to get a single measure to evaluate results. The F-measure is the harmonic mean of precision and recall. The equation of the F-measure is as follows:

$$\text{F-Measure} = \frac{2TP}{2TP + FN + FP} \tag{6}$$

Table II shows the performance metrics of the classification algorithms with 10-fold cross-validation and the number of features which was selected by the feature selection methods. According to table, the highest precision values were obtained for the hepatitis dataset with Na ïve Bayes and Decision Table classifiers with Info Gain Attribute Eval, One-R Attribute Eval and Relief Attribute Eval. For example, the precision of Na ïve Bayes and Decision Table with these feature selection methods is 0.853 which is the highest value in the Table II. In addition, the precision of Na ïve Bayes on Consistency Subseteval is also 0.853. Similarly, Na ïve Bayes classifier with One-R Attribute Eval and Relief Attribute Eval feature selection approaches have the highest recall values.

TABLE II: EVALUATION OF FEATURE SELECTION METHODS FOR HEPATITIS DATASET

Algorithm	Feature selection method	No of features	Precision	Recall
J48	CfsSubsetEval	10	0.616	0.619
	PrincipalComponents	16	0.802	0.819
	ConsistencySubsetEval	12	0.818	0.832
	InfoGainAttributeEval	19	0.825	0.839
	OneRAttributeEval	19	0.825	0.825
	ReliefAttributeEval	19	0.825	0.825
Na ïve Bayes	CfsSubsetEval	10	0.744	0.735
	PrincipalComponents	16	0.849	0.845
	ConsistencySubsetEval	12	0.853	0.845
	InfoGainAttributeEval	19	0.853	0.845
	OneRAttributeEval	19	0.853	0.853
	ReliefAttributeEval	19	0.853	0.853
IBK	CfsSubsetEval	10	0.663	0.665
	PrincipalComponents	16	0.777	0.774
	InfoGainAttributeEval	19	0.794	0.845
	OneRAttributeEval	19	0.794	0.845
	ReliefAttributeEval	19	0.794	0.845
	ConsistencySubsetEval	12	0.815	0.819
Decision Table	CfsSubsetEval	10	0.618	0.619
	ConsistencySubsetEval	12	0.706	0.735
	PrincipalComponents	16	0.78	0.794
	InfoGainAttributeEval	19	0.853	0.845
	OneRAttributeEval	19	0.853	0.845
	ReliefAttributeEval	19	0.853	0.845

The comparison analysis by root mean squared error was also performed and described in Table III. Root Mean Squared Error (RMSE) can be written as follows:

$$\text{RMSE} = \sqrt{\frac{\sum_{m=1}^{n} (y_{p,m} - t_{m,m})^{2}}{n}} \tag{7}$$

where n is the number of data patterns, $y_{p,m}$ indicates the predicted, $t_{m,m}$ is the measured value of one data point m and $\bar{t}_{m,m}$ is the mean value of all measure data points [24].

According to Table III results, the root mean square error of Na ïve Bayes with Consistency Subset Eval is 0.3446 which is the lowest error rate among the algorithms. Considering the results of two tables, it is clearly seen that Na ïve Bayes classifer is predominantly better than others.

TABLE III: ROOT MEAN SQUARED ERROR

Feature selection method	J48	Na ïve Bayes	IBK	Decision Table
CfsSubsetEval	0.5117	0.4704	0.5751	0.4891
ConsistencySubsetEval	0.366	0.3446	0.4221	0.4139
PrincipalComponents	0.3853	0.3626	0.4719	0.3715
InfoGainattributeEval OneRAttributeEval ReliefAttributeEval	0.363	0.3638	0.4369	0.3893

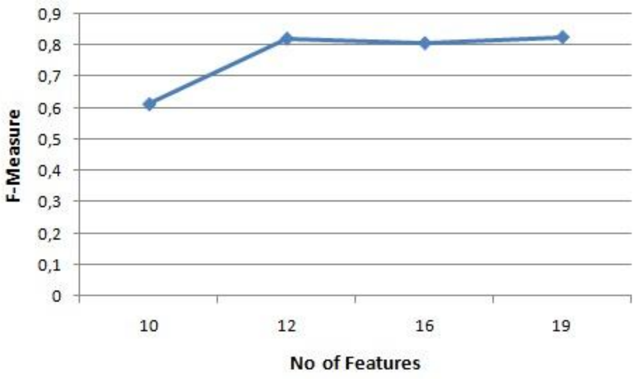

Fig. 1. F-measure using J48 classifier.

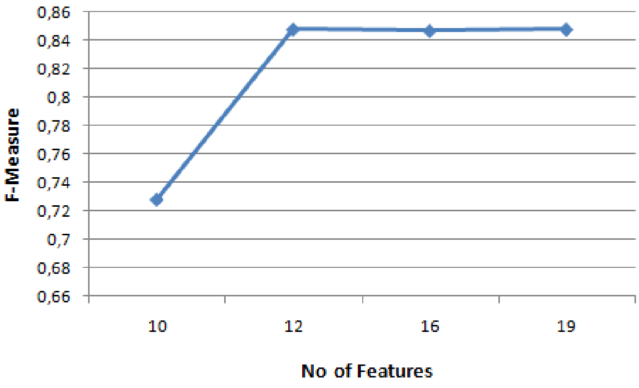

Fig. 2. F-measure using na ïve Bayes classifier.

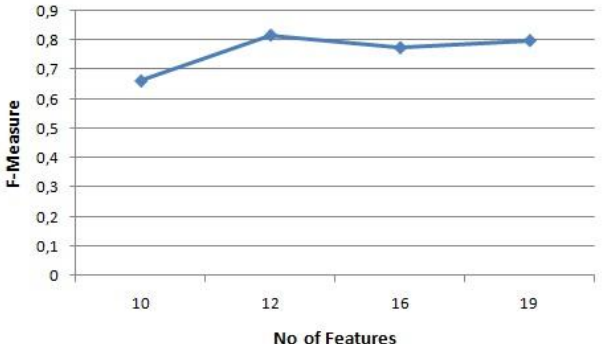

Fig. 3. F-measure using IBK classifier.

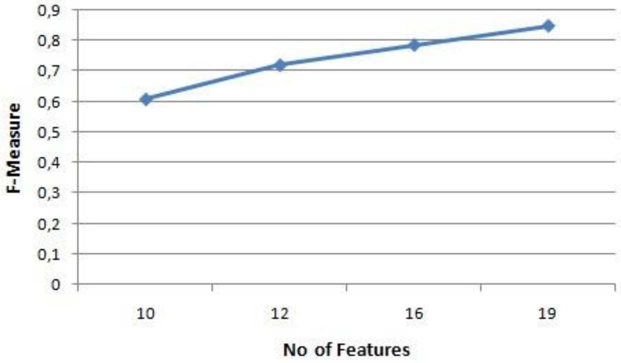

Fig. 4. F-measure using decision table classifier.

Fig. 1-Fig. 4 shows F-Measure values by number of features. The x-axis is the number of features. F-measures are plotted in the *y*-axis. The results were introduced separately for each classifier. Comparing the algorithms, the highest F-Measure values were obtained with 12 and 19 features and Naïve Bayes and Decision Tree algorithms achieved the best performance among the other classifiers. For example, when the number of features is 19, Naïve Bayes perfomed the highest F-Measure with 0.848. Generally, the performance was varied based on the different features, therefore, we can conclude that feature selection algorithms significantly affect on the accuracy of classifiers.

VII. Conclusion

Feature selection is an important data processing step in data mining studies and many machine learning algorithms can hardly cope with large amounts of irrelevant features. Thus, feature selection approaches became a necessity for many studies.

In this study, a comparative analysis was carried out on the basis of filter based feature selection algorithms to predict the risks of hepatitis disease. Six feature selection algorithms were used to analyze the dataset and their performance was evaluated by using J48, Naïve Bayes, IBK and Decision Table classifiers. The evaluation of results was performed based on four accuracy metrics: precision, recall, root mean squared error and F-Measure. Among the algorithms, Naïve Bayes and Decision Table classifiers have higher accuracy rates on the hepatitis dataset than the others after the application of feature selection methods. In addition, Naïve Bayes has the lowest root mean square error in the others. This study asserted that feature selection methods are capable to improve the performance of learning algorithms. However, no single filter based feature selection method is the best. Overall, Naïve Bayes with ConsistencySubsetEval, InfoGainAttributeEval, OneRAttributeEval and ReliefAttributeEval methods performed better results than the others.

The results of this study can make contributions in the prediction of hepatitis disease in medical research and provide a deep comparison of popular filter based feature selection methods for machine learning studies.

As a future work, a study will be planned to investigate the effects of both continuous and discrete attributes of medical datasets in the performance of feature selection methods and classification accuracy.

References

[1] M. Ashraf, G. Chetty, and D. Tran, "Feature selection techniques on thyroid,hepatitis, and breast cancer datasets," *International Journal on Data Mining and Intelligent Information Technology Applications(IJMIA)*,vol. 3, no. 1, pp. 1-8, 2013.

[2] J. Novakovic, P. Strbac, and D. Bulatovic, "Toward optimal feature selection using ranking methods and classification algorithms," *Yugoslav Journal of Operations Research*, vol. 21, no. 1, pp. 119-135, 2011.

[3] H. Yasin, T. A. Jilani, and M. Danish, "Hepatitis-C classification using data mining techniques," *International Journal of Computer Applications*, vol. 24, no. 3, pp. 1-6, 2011.

[4] M. Leach, "Parallelising feature selection algorithms," University of Manchester, Manchester, 2012.

[5] I. Lee, G. H. Lushington, and M. Visvanathan, "A filter-based feature selection approach for identifying potential biomarkers for lung cancer," *Journal of clinical Bioinformatics*, vol. 1, no. 11, pp. 1-8, 2011.

[6] R. C. Holte, "Very simple classification rules perform well on most commonly used datasets," *Machine Learning*, vol. 11, pp. 63-91, 1993.

[7] I. T. Jolliffe. (2002). Principal Component Analysis. [Online]. Available: http://books.google.com.au

[8] M. A. Hall, "Correlation-based feature selection for machine learning," PhD, Department of Computer Science, The University of Waikato, Hamilton, 1999.

[9] M. A. Hall and L. A. Smith, "Feature selection for machine learning:comparing a correlation-based filter approach to the wrapper," *Proceedings of the Twelfth International Florida Artificial Intelligence Research Society Conference*, pp. 235-239, 1999.

[10] H. Liu and R. Setiono, "CHI2: Feature selection and discretization of numeric attributes," in *Proc. the 7th IEEE International Conference on Tools with Artificial Intelligence*, 1995.

[11] T. R. Patil and S. S. Sherekar, "Performance analysis of naïve bayes and j48 classification algorithm for data classification," *International Journal of Computer Science And Applications*, vol. 6, no. 2, pp. 256-261, 2013.

[12] G. Dimitoglou, J. A. Adams, and C. M. Jim, "Comparison of the C4.5 and a naïve bayes classifier for the prediction of lung cancer survivability," *Journal of Computing*, vol. 4, issue 8, 2012.

[13] I. H. Witten, E. Frank, *Data Mining: Practical Machine Learning Tool and Technique with Java Implementation*, Morgan Kaufmann, San Francisco, 2000.

[14] R. Kohavi and G. H. John, "Wrappers for feature subset selection," *Artif Intell*, vol. 97, no. 1, pp. 273-324, 1997.

[15] A. Tsymbal and S. Puuronen, "Local feature selection with dynamic integration of classifiers," *Foundations of Intelligent Systems*, pp. 363-375, 2010.

[16] C. L. Blake and C. J. Merz. (1996). UCI repository of machine learning databases. [Online]. Available: http://www.ics.uci.edu/~mlearn/MLRepository.html

[17] G. Sathyadevi, "Application of cart algorithm in hepatitis disease diagnosis," in *Proc. the Recent Trends in Information Technology*, Chennai, Tamil Nadu, pp. 1283-1287, 2011.

[18] A. H. Roslina and A. Noraziah, "Prediction of hepatitis prognosis using support vector machine and wrapper method," in *Proc. the Fuzzy Systems and Knowledge Discovery*, Yantai Shandong, pp. 2209-2211, 2010.

[19] J. S. Sartakhti, "Hepatitis disease diagnosis using a novel hybrid method," *Computer Methods and Programs in Biomedicine*, vol. 108, issue 2, pp. 570-579, 2011.

[20] H. Harb and A. S. Desuky, "Feature selection on classification of medical datasets based on particle swarm optimization," *International Journal of Computer Applications*, vol. 104, no. 5, pp. 14-17, 2014.

[21] S. Huang, L. R. Wulsin, H. Li, and J. Guo, "Dimensionality reduction for knowledge discovery in medical claims database: Application to antidepressant medication utilization study," *Computer Methods and Programs in Biomedicine*, vol. 93, pp. 115-123, 2009.

[22] I. Inza, P. Larranaga, R. Blanco, A. J. Cerrolaza, "Filter versus wrapper gene selection approaches in DNA microarray domain," *Artificial Intelligence in Medicine*, vol. 31, pp. 91-103, 2004.

[23] WEKA: Weka 3: Data Mining Software in Java. [Online]. Available: http://www.cs.waikato.ac.nz/ml/weka

[24] E. U. Küçüksille, R. Selbaş, A. Şencan, "Prediction of thermodynamic properties of refrigerants using data mining," *Energy Conversion and Management*, vol. 52, pp. 836-848, 2011.

Pinar Yildirim is an assistant professor at the Department of Computer Engineering of Okan University in Istanbul. She received her B.S. degree in electronics and communications engineering from Yıldız Technical University, Istanbul, M.S. degree in medical informatics from Akdeniz University, Antalya and PhD degree in the Department of Health Informatics of the Informatics Institute at Middle East Technical University, Ankara, Turkey, in 1989, 1995 and 2011 respectively. She is a member of the Turkish Medical Informatics Association. She was supported by TUBITAK (The Scientific and Technological Research Council in Turkey) and she visited the European Bioinformatics Institute in Cambridge (UK) as an academic visitor between 2008 and 2009. Her research areas include biomedical data mining, machine learning, classification,clustering, missing data and feature selection.