IJCSE International Journal of Computer Sciences and Engineering Open Access

Research Paper Volume-6, Issue-3 E-ISSN: 2347-2693

Enhancing Wrapper Based Algorithms for Selecting Optimal Features from Thyroid Disease Dataset

K. Pavya[1*], B. Srinivasan[2]

[1*] Department of Computer Science, Vellalar College for Women, Bharathiar University, Tamilnadu, India
[2]Department of Computer Science, Gobi Arts and Science College, Bharathiar University, Tamilnadu, India

**Corresponding Author: pavyavcw@gmail.com*

Available online at: www.ijcseonline.org

Received: 19/Feb//2018, Revised: 25/Feb2018, Accepted: 17/Mar/2018, Published: 30/Mar/2018

Abstract: Advances in medical information technology have enabled healthcare industries to automatically collect huge amount of data through clinical laboratory examinations. Thyroid disease (TD) is a study of Endocrinology and is considered as one of the most common diseases that is frequently misunderstood and misdiagnosed. Machine learning techniques are increasingly introduced to construct the CAD systems owing to its strong capability of extracting complex relationships in the biomedical data. Feature selection is a technique to choose a subset of variables from the multidimensional data which can improve the classification accuracy in diversity datasets. In addition, the best feature subset selection method can reduce the cost of feature measurement. This work focuses on enhancing the wrapper based algorithms for feature selection.

Keywords: Data Mining, Feature Selection, Wrapper Method, Genetic Algorithm, Ant Colony Optimization

I. INTRODUCTION

Proper analysis of the thyroid data besides clinical examination and complementary investigation is a vital issue in the identification of thyroid disease. Doctors can incorporate numerous factors, including clinical evaluation, blood tests, imaging tests, biopsies, and other tests to diagnose thyroid disease. A common used method is a test, called the thyroid-stimulating hormone (TSH) test, which can identify thyroid disorders even before the onset of symptoms. Usage of CAD systems for diagnosis provides multiple advantages [1]

- Can minimize the operator-dependent nature inherent in medical imaging systems and can make the diagnostic process reproducible.
- Help to improve the accuracy of diagnosis
- Can work with features (like computational features and statistical features) that cannot be obtained through visual analysis or through intuitive examinations.

These features can be extracted and analyzed automatically using data mining techniques. Data mining plays an essential role in medical field for disease diagnosis. It offers lot of classification techniques to predict the disease accuracy [2]. The computer based analysis system indicates the mechanized medical diagnosis system. This mechanized diagnosis system support the medical practitioner to make good decision in treatment and disease [3].Classification maps data into predefined groups or classes. It is frequently referred to as supervised learning because the classes are determined before examining the data [1].The classification performance can be improved using Feature selection. Feature selection is used to eliminate the irrelevant features from the data without much loss of the information. There are two types of feature selection algorithm namely, Filter based and Wrapper based. Filter based method selects the feature without depending upon the type of classifier used. The advantage of this method is that, it is simple and independent of the type of classifier used so feature selection need to be done only once [4].In Reference [5],an enhanced filter based algorithm was proposed to improve the performance of thyroid disease classification.

In this paper two wrapper based algorithms based on Hold-out Recursive feature elimination-Support vector machine (HRS), Ant Colony Optimization (ACO) and Genetic Algorithm (GA) is proposed. The first method enhances RFE method using sequential forward and sequential backward search algorithm and in the second method ACO is combined with Genetic Algorithm to form a hybrid wrapper based feature selection algorithm. To further improve the feature selection process both these enhanced methods are combined. The steps involved are shown in figure1.The rest of the paper is organized as follows: Section II presents the literature review related to the topic and Section III presents the methodology used to design the proposed feature selection algorithm. The experimental results are discussed in Section IV. Finally, Section V concludes the paper.

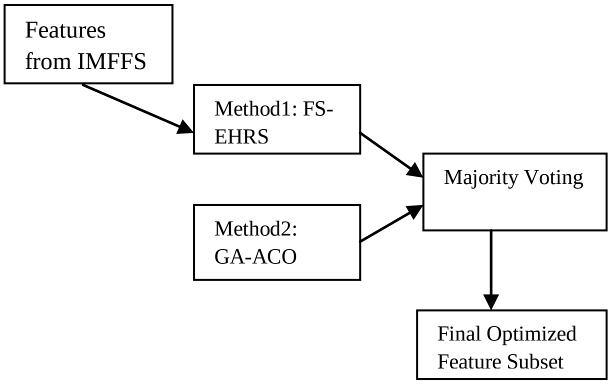


Fig.1 Steps involved in Improved Multiple Wrapper-Based FS (IMWFS) Algorithm

II. LITERATURE REVIEW

The feature selection is to select a subset of variables from the input data which can efficiently describe the input data while reducing effects from noise or relevant variables and still provide good prediction results[2,3]. Wrapper methods perform a search in the space of feature subsets such as classification performances on a cross-validation of the training set which provided better results than filter methods. But wrapper approaches increase the computational cost [6].

Donghai Guan, et al. [7] reviewed and compared two techniques of integrating feature selection and ensemble learning, (1) Feature selection for ensemble learning (ENfs) and (2) Ensemble learning for feature selection (FSen). This approach obtained predictive accuracy superior to conventional feature selection methods for supervised learning. Moreover, its most prominent advantage is the ability to handle stability issue that is usually poor in existing feature selection methods.

S´anchez-Maro˜no, et al. [8] proposed a new wrapper method, called Incremental ANOVA and Functional Networks-Feature Selection (IAFN-FS) for dealing with multiclass problems based in classical algorithms, such as C4.5 and Naïve Bayes. The multiple binary classifiers approach obtained better results in accuracy, although it has the drawback of selecting a higher number of features.

Akin Ozcift and Arif Gulten [9] used a rotation forest ensemble decision tree algorithm wrapped with best first search strategy. The wrapper uses forward selection to choose the optimum subset on the Erythemato-Squamous diseases dataset. The discrimination ability of selected features is evaluated using several machine learning algorithms and the diversity of the training data using the bagging algorithm.

Yvan Saeys, et al. [6] proposed the method of ensemble feature selection techniques for high dimension data which can be used to yield more robust feature selection techniques. As well Sangkyun Lee, et.al [10] presented a method of an extension to Rapid Miner which delivers implementations of algorithms which is well suited for very high-dimensional data. These experiments were conducted on a microRNA-expression dataset.

L. Yu and H. Liu [11] propose a new framework of feature selection which avoids implicitly handling feature redundancy and turns to efficient elimination of redundant features via explicitly handling feature redundancy. Relevance definitions divide features into strongly relevant features, weakly relevant features and irrelevant features; redundancy definition divides weakly relevant features into redundant and non redundant ones. Thus produces the final subset. Its advantage is decoupling relevance and redundancy analysis and allows a both efficient and effective way in finding a subset that approximates an efficient subset. It uses C-correlation for relevance analysis and both C & F correlations for redundancy analysis.

Asha Gowda Karegowda, et al.[12] described the feature subset selection problem using wrapper approach in supervised learning. The experimented wrapper method used Genetic algorithm as random search technique wrapped with different classifiers/ induction algorithm namely decision tree C4.5, NaïveBayes, Bayes networks and Radial basis function as subset evaluating mechanism. Relevant attributes identified by different wrappers were compared using different classifiers in validation step. The results prove that there is no one standard wrapper approach, which is best for different datasets, however experiment results show that employing feature subset selection, surely enhances the classification accuracy.

III. METHODOLOGY

In wrapper method the feature is dependent upon the classifier used, i.e. it uses the result of the classifier to determine the goodness of the given feature or attribute. The advantage of this method is that it removes the drawback of the filter method, i.e. it includes the interaction with the classifier and also takes the feature dependencies and drawback of this method is that it is slower than the filter method because it takes the dependencies also. The quality of the feature is directly measured by the performance of the classifier [4].

- Wrapper-based algorithms require one predetermined mining algorithm and uses its performance as the evaluation criterion.
- It searches for features better suited to the mining algorithm aiming to improve mining performance

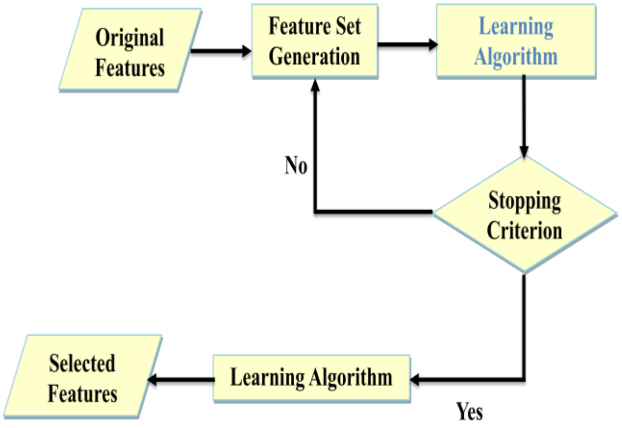


Fig 2.Wrapper based method

Table1.characteristics of wrapper method

Items	Wrappers
Processing Speed	Slow
Classification Accuracy	High
Depend on Learning Methods	Yes
Computational Cost	High

3.1 Selection using Enhanced Hold out –Recursive Feature Elimination-Support Vector Machine [FS-EHRS]

In holdout method, the given data are randomly partitioned into two independent sets, a training set and a test set. Typically, two-thirds of the data are allocated to the training set, and the remaining one-third is allocated to the test set. The training set is used to derive the model, whose accuracy is estimated with the test set (Fig 3).The estimate is pessimistic because only a portion of the initial data is used to derive the model [13].

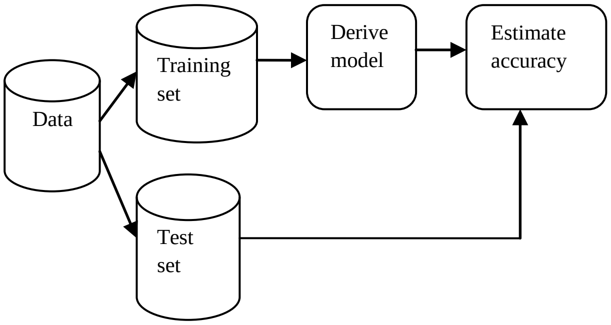


Fig 3. Estimating accuracy with the holdout method.

3.1.1 Method 1: FS-EHRS:

The first method enhances the conventional Recursive Feature Elimination (RFE) method.RFE method uses SVM during classification and it differs from other techniques in two ways.

- Method used to select the features during each iteration
- Stopping criterion

The algorithm applies a hold-out technique during feature selection that considers the whole dataset instead of just training data and uses a stopping criterion that considers the classification error of both training and testing data, instead of just training data (as the other methods does). The consideration of whole dataset improves the performance of feature selection. The stopping criterion helps to reduce computational costs required to determine when feature elimination degrades classification performance. This algorithm is called Feature Selection using Hold-Out RFE-SVM (FS-HRS)

3.1.2 FS-HRS Algorithm:

- Initialization
- Repeat
    - Random split of the training data
    - SVM Training
    - for each feature, determine the number of classification errors (CE) when the current feature is removed
    - remove feature j with the smallest CE using FS-SBSF algorithm
- Until the smallest CE is greater than number of CEs in test set using all features.

3.1.3 Sequential Backward and Sequential Forward Search (FS-SBSF) Algorithm:

A sequential forward selection (SFS) is the simplest greedy search algorithm [14]. SFS starts with an empty selection of attributes and, in each round, it adds each unused attribute of the given example set. For each added attribute, the performance is estimated using the cross validation. Only the attribute giving the highest performance is added to the selection for the object function. Then a new round is started with the modified selection. Therefore, the SFS algorithm adds features which give a high value to the object function [15].

Sequential backward selection (SBS) works in the opposite direction to SFS [14]. SBS starts with the full set of attributes and, in each round, it removes each remaining attribute of the given example set. For each iteration or attribute removal, the performance is estimated using the inner operators, such as a cross validation. Only the attribute giving the least decreasing performance is finally removed from the selection. Then a new round is started with the modified selection [15]. This elimination process has two advantages: first, it can discard several features and second, it allows for backtracking, and so, when a subset of features worsens the results obtained by the previous one, some previously eliminated features can be included in the new subset for re-evaluation [8].

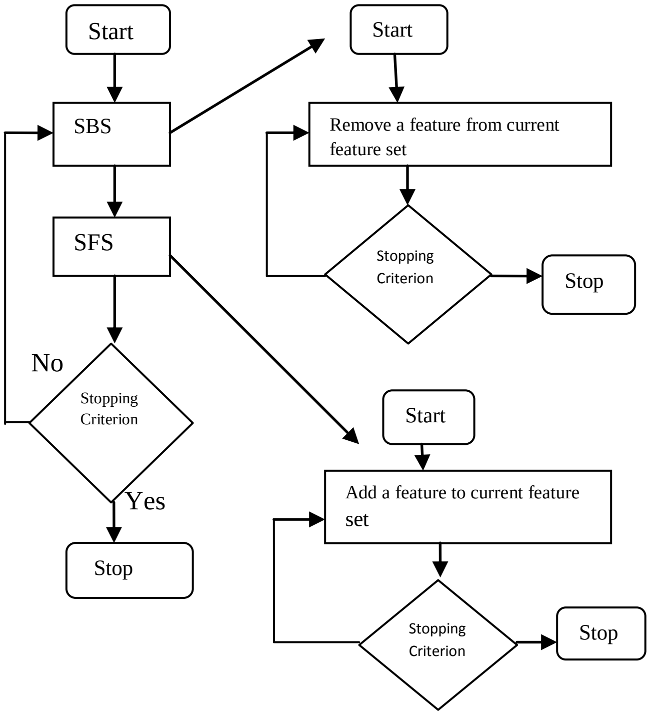


Fig 4. SBS – Sequential Backward Search and SFS – Sequential Forward Search

3.1.4 Steps in Proposed FS-EHRS Algorithm:

- Perform Initialization
- Repeat
    - Split data from IMFFS using Random split method
    - SVM Training
    - for each feature, determine the number of classification errors (CE) when the current feature is removed
    - remove feature j with the smallest CE using Sequential Backward and Sequential Forward Search (FS-SBSF) Algorithm
- Until the smallest CE is greater than number of CEs in test set using all features.

3.2 Ant Colony Optimization-Genetic Algorithm [ACO-GA]

Genetic algorithm attempts to incorporate ideas of natural evolution. In general, genetic learning starts as follows. An initial population is created consisting of randomly generated rules. Each rule can be represented by a string of bits. As a simple example, suppose that samples in a given training set are described by two Boolean attributes,A_1 and A_2, and that there are two classes,C_1 and C_2.The rule “IF A_1 AND NOT A_2 THEN C_2” can be encoded as the bit string “100” where the two leftmost bits represent attributes A_1 and A_2,respectively,and the rightmost bit represents the class. Similarly, the rule “IF A_1 AND NOT A_2 THEN C_1” can be encoded as “001”. If an attribute has k values, where k>2, then k bits may be used to encode the attribute’s values. Classes can be encoded in a similar fashion [13].

Ant colony optimization is a meta-heuristic method that uses artificial ants to get solutions to combinatorial optimization problems. ACO is based on the behavior of real ants and possesses enhanced abilities such as memory of past actions and knowledge about the distance to other locations. In character, an individual ant is unable to communicate or successfully hunt for food, but as a group, ants hold the ability to solve complex problems and effectively find and collect food for their colony. Ants communicate using a chemical substance called pheromone. As an ant travels, it deposits a constant amount of pheromone that other ants can follow. Each ant travels in a somewhat random manner, but when an ant encounters a pheromone trail, it must make a decision whether to go after it. If it follows the trail, the ant’s own pheromone reinforces the existing trail, and the increase in pheromone increases the probability of the next ant selecting the path. Therefore, the more ants that travel on a path, the more attractive the path becomes for subsequent ants. Additionally, an ant using a short route to a food source will return to the nest sooner and therefore, mark its path twice, before other ants return. This directly influences the selection probability for the next ant leaving the nest [16].

In common the ACO allocate two ants such as forward and backward ant. The forward ant is used for searching the food while the backward ant is used to get back to their host. For transferring the information, only the forward ant can be used. There is no use of backward ant in the transmitting process, but it can be used for the acknowledgement purpose. Here the current node is assign as a forward ant during the transformation and it can also act as backward ant during the acknowledgement.

3.2.1 Method 2: ACO-GA Algorithm:

- ➢ This algorithm combines the advantages of GA and ACO algorithm for fast and better optimal feature search capability.
- ➢ ACO advantage : Can perform local searching

- GA advantage: Considers a global perspective by operating on the complete population from the very beginning.
- Thus, combining ACO and GA can take advantage of each other and perform in a better manner.
- ACO and GA are used to explore the space of all subsets of given feature set.
- The performance of selected feature subsets is measured by invoking an evaluation function with the corresponding reduced feature space and measuring the specified classification result.
- The best feature subset found is then output as the recommended set of features to be used in the actual design of the classification system.

3.3 Majority voting

The majority voting is an algorithm for finding the majority of a sequence of elements using linear time and constant space. The algorithm finds a majority element, if there is one: that is, an element that occurs repeatedly for more than half of the elements of the input. However, if there is no majority, the algorithm will not detect that fact, and will still output one of the elements. A version of the algorithm that makes a second pass through the data can be used to verify that the element found in the first pass really is a majority.

IV. EXPERIMENTAL ANALYSIS

The datasets are taken from UCI Thyroid dataset with the number of Instances 7200 and 21 Attributes. Performance metrics taken are Accuracy, Sensitivity, Speed and Specificity. The performance of the feature selection algorithms are evaluated using three classifiers namely BPPN (Back Propagation Neural Network), KNN (K-Nearest Neighbor) and SVM (Support Vector Machine).

Table 2 . Shows the coding scheme used in experimental analysis

Code	**Description**
IMWFS	Improved Multiple Wrapper-Based Feature Selection
FS-EHRS	Feature Selection using Enhanced HO-RFE-SVM
FS-ACO	Feature Selection using Ant Colony Optimization
FS-GA	Feature Selection using Genetic Algorithm
FS-GA-ACO	Feature Selection using GA and ACO

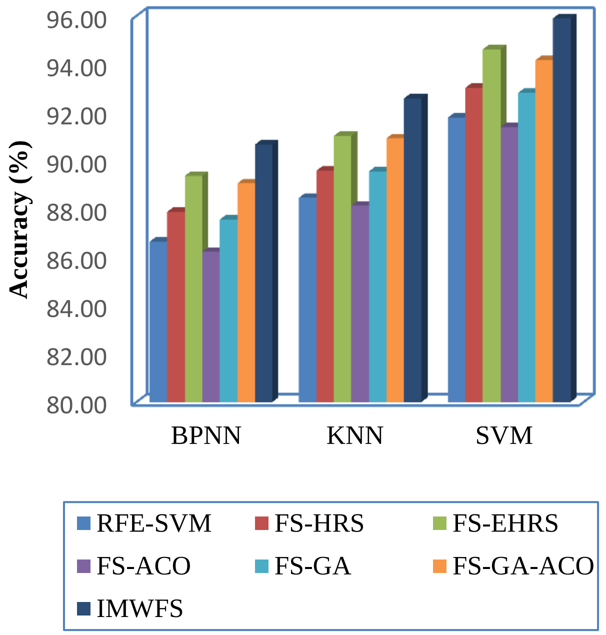


Fig5. Shows the accuracy of the proposed algorithm and compares the results with existing algorithms

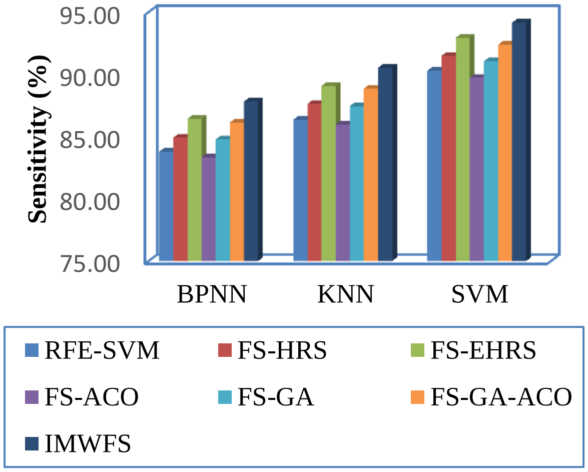
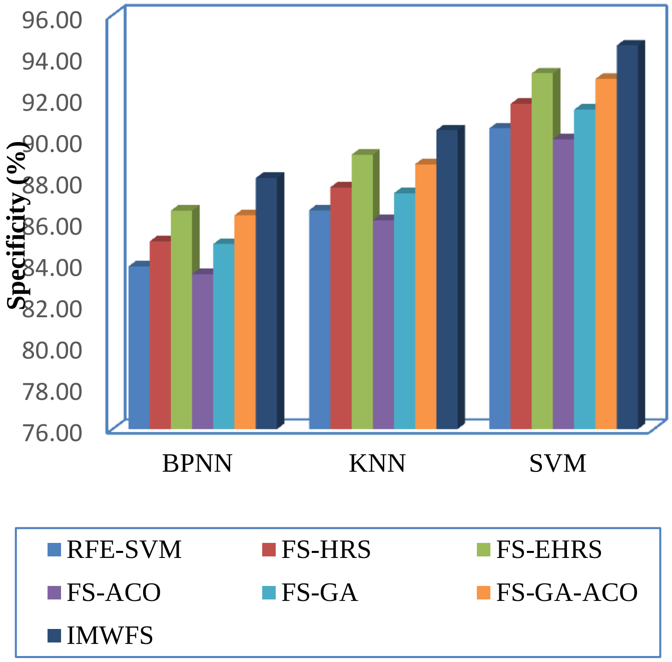


Fig6. Shows the sensitivity of the proposed algorithm and compares the results with existing algorithms

Fig7. Shows the specificity of the proposed algorithm and compares the results with existing algorithms



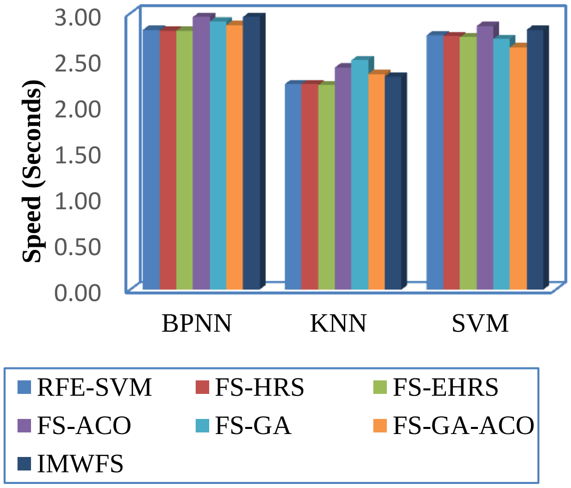


Fig8. Shows the speed of the proposed algorithm and compares the results with existing algorithms

In the experiments discussed, the analysis of the feature selection algorithm performance was done using three frequently used classifiers, namely, BPNN (Back Propagation Neural Network), KNN (K Nearest Neighbor) and SVM (Support Vector Machine) classifier. From the results, it is clear that the SVM produces high accuracy. The FS-HRS algorithm uses backward feature elimination method, which is computationally expensive when the number of input features is large. This is solved by using the multiple filter-based algorithms for selecting only relevant and non-redundant features before using the FS-EHRS method. The output of our previous paper on Enhanced Multiple Filter based feature selection [10] is taken as an input for this Enhanced Multiple Wrapper based feature selection and analyzed with the same classifiers namely, BPPN, KNN and SVM. To further reduce the time complexity, instead of using a backward search algorithm, a Sequential Backward and Sequential Forward Search (FS-SBSF) Algorithms are used and this algorithm is termed as Feature Selection using Enhanced HO-RFE-SVM (FS-EHRS).

V. CONCLUSION

It is of great of importance to remove the noisy and irrelevant features and data samples embedded in data sets before applying some data mining techniques to analyze the data sets. This paper describes a novel idea to identify the noisy and irrelevant features embedded in data sets and detect the quality of the structure of data sets. This experiment revealed that the efficiency of the proposed IMWFS algorithm is better in terms of all the selected performance metrics, when compared to the conventional algorithm. Hence the SVM classifier produces the higher accuracy than the other two classifiers. This indicates that the algorithm is able to remove maximum redundant data preserving the relevant (or important) data.

REFERENCES

[1] B. Srinivasan and K. Pavya, "*A study on data mining prediction techniques in healthcare sector*", International Research Journal of Engineering and Technology, Vol.3, Issue.3, pp. 552-556, 2016.

[2] N. Sanchez-Marono, A. Alonso-Betanzos, and R.M. Calvo-Estevez, "A Wrapper Method for Feature Selection in Multiple Classes Datasets", J. Cabestany et al. (Eds.): IWANN, pp. 456–463, 2009.

[3] B. Srinivasan and K. Pavya, "*Diagnosis of Thyroid Disease Using Data Mining Techniques: A Study*", International Research Journal of Engineering and Technology, Vol.3, Issue.11, pp. 1191-1194, 2016.

[4] R. G. Osuna, "*Pattern Analysis for Machine Olfaction: A Review*", IEEE SENSORS JOURNAL; pp.189-202, 2002.

[5] K. Pavya and B. Srinivasan, "*Feature Selection Techniques in Data Mining: A Study*", International Journal of Scientific Development and Research (IJSDR), Vol.2, Issue.6, pp. 594-598,2017.

[6] K. Pavya, and B. Srinivasan " *Enhancing Filter Based Algorithms for Selecting Optimal Features from Thyroid Disease Dataset*", *International Journal of Advanced Research in Computer Science*, Vol.8, Issue.9, pp. 184-188,2017.

[7] G. Chandrashekar , F. Sahin, "*A survey on feature selection methods*", Computer and Electrical Engineering, pp.16-28,2014.

[8] Y. Saeys, T. Abeel and Y.V. Peer, "*Robust Feature Selection Using Ensemble Feature Selection Techniques*", W. Daelemans et al.(Eds.): ECML PKDD, pp. 313–325, 2008.

[9] S. Lee , B. Schowe and V. Sivakumar, "*Feature Selection for High-Dimensional Data with RapidMiner*", Technical Report of TU Dortmund University; 2011.

[10] J. Han and M. Kamber, "*Data Mining Concepts and Techniques*", Morgan Kaufmann publishers,Second Edition, 2008.

[11] L.Yu and H. Liu , "Efficient Feature Selection via Analysis of Relevance and Redundancy" , *J.Machine Learning Research*, Vol.10, Issue.5, pp. 1205-1224, 2004.

[12] S. Aravind , G. Michel, "*Hybrid of Ant Colony optimization and Genetic Algorithm for Shortest Path in Wireless Mesh Networks", Journal of Global Research in Computer Science*,Vol.3, Issue.1, pp. 31-34, 2012.

[13] D. Guana, W. Yuana , Y.K. Leea , K. Najeebullaha, M.K. Rasela, "*A Review of Ensemble Learning Based Feature Selection*", IETE Technical Review; 2014.

[14] A. Ozcift and A. Gulten "*A Robust Multi-Class Feature Selection Strategy Based on Rotation Forest Ensemble Algorithm for Diagnosis*", J Med Syst; pp.941–949, 2012.

[15] Asha Gowda, Karegowda, M.A.Jayaram and A.S.Manujunath, "*Feature Subset Selection Problem using Wrapper Approach in Supervised Learning*", International Journal of Computer Applications,Vol.1, Issue.7, pp. 13-17, 2010.

[16] B. Srinivasan and K. Pavya, " *A Comparative Study on Classification Algorithms in Data Mining*", International Journal of Innovative Science, Engineering & Technology, Vol. 3, Issue.3, pp. 415-418, 2016.

AUTHORS PROFILE

Ms. K.Pavya has completed her Master of Computer Applications and Master of Philosophy in Computer Science under Bharathiar University. She is currently working as an Assistant Professor in the Department of Computer Science, Vellalar College for Women under Bharathiar University and also a Part- time Ph.D., Research Scholar, Department of Computer Science, Gobi Arts & Science College under Bharathiar University, Coimbatore, India.
Her research interest is Data Mining.

Dr.B.Srinivasan did his M.C.A in Gobi Arts & Science College, Gobichettipalayam and Ph.D in Computer Science at Vinayaka Missions University, Salem. He is working as an Associate Professor in Computer Science, Gobi Arts & Science College, Gobichettipalayam. He has 25 years of experience in teaching and 18 years of research experience. He has published 70 research papers in the national and international journals. He has guided 65 M.Phils and 5 Ph.Ds and his area of interest includes Network Security and Automata Theory.