ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

Comparison of Feature Selection Strategies for Classification using Rapid Miner

Dr.J.Arunadevi[1], M.Josephine Nithya[2]

Assistant Professor, PG & Research Department of Computer Science, Raja Doraisingam Govt. Arts College, Sivaganga, Tamilnadu, India[1]

Research Scholar, PG & Research Department of Computer Science, Raja Doraisingam Govt. Arts College, Sivaganga, Tamilnadu, India[2]

ABSTRACT: Feature selection is an important part in any of the data processing algorithms as it reduces the complexity of the processor by the reduction of the feature space. In this paper we discuss three different strategies for the reduction or the selection of the feature of the given dataset. They are backward elimination, forward selection and optimized selection using genetic algorithm. Five dataset were employed to test the three strategies and out of the five four dataset gives better results while using optimized selection and the forward selection for the rest one. The results are discussed in the detailed manner.

KEYWORDS: Feature selection, Optimized selection, Genetic algorithm, Rapid miner

I. Introduction

Feature selection is observed to be an lively and vigorous research area in many fields such as pattern recognition, statistics, machine learning and data crunching applications [1,2]. The core objective of the feature selection is to choose a subset of input variables from the available set of features. Feature selection has proven in both theory and practice to be effective in enhancing learning efficiency, increasing predictive accuracy and reducing complexity of learned results [3,4].

It is being realized that the feature selection process is inevitable because of the enormous increase of data in terms of volume, velocity, variety and veracity. Data crunching applications require the data to be processed in such a manner that the value and the quality of the data should not be affected during the process of feature selection.

This paper concentrates on the feature selection strategies that have been studied using the rapid miner tool [5]. Three types of feature selection strategies have been analyzed using the test dataset. They are backward elimination, forward selection and optimized selection using genetic algorithm.

The section 2 deals with the necessary background study of the paper, section 3 describes the other methodologies involved for the feature selection. Section 4 talks about the problem formulation and the section 5 discuss the environment of the experiment. Section 6 discusses the results and the section 7 gives the conclusion.

II. Related work

A typical feature selection process contains two phases: feature selection, and model fitting and performance evaluation [6]. The feature selection phase contains three steps:
(1) Generating a candidate set containing a subset of the original features via certain research strategies;
(2) Evaluating the candidate set and estimating the utility of the features in the candidate set. Based on the evaluation, some features in the candidate set may be discarded or added to the selected feature set according to their relevance; and
(3) Determining whether the current set of selected features are good enough using certain stopping criterion. If it is, a feature selection algorithm will return the set of selected features, otherwise, it iterates until the stopping criterion is met.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

The focal point of feature selection is to decide on a subset of variables from the input which can efficiently describe the input data while dipping effects from noise or inappropriate variables and still provide good prediction results [7]. Feature selection is different from dimensionality reduction. Both methods seek to reduce the number of attributes in the dataset, but a dimensionality reduction method do so by creating new combinations of attributes, where as feature selection methods include and exclude attributes present in the data without changing them [8].

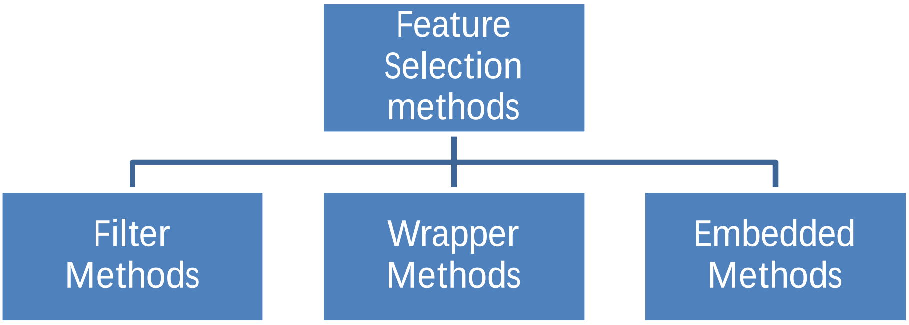

Fig1 : Classification of feature selection methods

Feature selection could be classified into three methods. We can discuss them one by one

(1) **Filter Methods**: Filter feature selection methods apply a statistical measure to assign a scoring to each feature. The features are ranked by the score and either selected to be kept or removed from the dataset. The methods are often univariate and consider the feature independently, or with regard to the dependent variable. It is based on the intrinsic characteristics of the data.

(2) Wrapper Methods: Wrapper methods consider the selection of a set of features as a search problem, where different combinations are prepared, evaluated and compared to other combinations. A predictive model us used to evaluate a combination of features and assign a score based on model accuracy.

(3) Embedded Methods: Embedded methods learn which features best contribute to the accuracy of the model while the model is being created. The most common type of embedded feature selection methods are regularization methods.

The filter, wrapper, and embedded models are the major models used in feature selection for algorithm design. In [9], an interesting hybrid approach is proposed to combine the wrapper with the filter model through a so-called greedy randomized adaptive search procedure (GRASP). The advantage of the method is that it can inherit the strength of both models to improve the performance of feature selection.

The intension of feature selection is to decide a subset of features for enhancing the prediction accuracy or minimizing the size of the structure without drastically reducing prediction accuracy of the classifier built using only the selected features [10]. The filter approach operates independently of any learning algorithm. These methods rank the features by some criteria and omit all features that do not achieve a sufficient score. Due to its computational efficiency, the filter methods are very popular to high-dimension data [11]. Some popular filter methods are F-score criterion [12], mutual information [13], information gain [14] and correlation [15]. The wrapper approach involves with the predetermined learning model, selects features on measuring the learning performance of the particular learning model [15-16]. Although wrappers may produce better results, they are expensive to run and can break down with very large numbers of features. This is due to the use of learning algorithms in the evaluation of feature subsets every time [17]. Filter and wrapper are two complementary approaches, then the hybrid approach attempts to take advantage of the filter and wrapper approaches by exploiting their complementary strengths [18-20].

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

III. PROBLEM FORMULATION

Due to the immense need for the feature selection strategies, this paper tries to address the problem as the comparison of the various feature selection methods such as backward elimination, Forward selection and Optimized selection using evolutionary approach based in the Genetic algorithm [21].
Forward selection : This strategy initially uses only attribute subsets which exactly one attribute. Then additional attributes are added heuristically, until there is no more performance gain by adding an attribute.
Backward elimination: In contrast to the forward selection strategy, the backward elimination strategy starts with the complete attribute set as initial subset and iteratively (and also heuristically) removes attributes from that subset, until no performance gain can be achieved by removing another attribute.
Evolutionary strategy: An optimal attribute subset might also be found by an evoluationary strategy. Therefore, every attribute subset is considered as an individual. An evolutionary algorithm works on a population of such individuals which may be selected to mutate or experience a cross over. For feature selection, a mutation might switch features on and off, a cross over might interchange features between individuals. An evolutionary feature selection strategy is implemented using the Genetic Algorithm.

IV. EXPERIMENTAL ENVIRONMENT

The experiment is carried out with the Rapid miner tool. RapidMiner is a software platform developed by the company of the same name that provides an integrated environment for machine learning, data mining, text mining, predictive analytics and business analytics. It is used for business and commercial applications as well as for research, education, training, rapid prototyping, and application development and supports all steps of the data mining process including data preparation, results visualization, validation and optimization. RapidMiner is developed on an open core model, with the RapidMiner Basic Edition available for download under the AGPL license [22].

4.1 Backward Elimination : The Backward Elimination starts with the full set of attributes and, in each round, it removes each remaining attribute of the given ExampleSet. For each removed attribute, the performance is estimated using the inner operators, e.g. a cross-validation. Only the attribute giving the least decrease of performance is finally removed from the selection. Then a new round is started with the modified selection. This implementation avoids any additional memory consumption besides the memory used originally for storing the data and the memory which might be needed for applying the inner operators. The stopping behavior parameter specifies when the iteration should be aborted. There are three different options:
 with decrease: The iteration runs as long as there is any increase in performance.
 with decrease of more than: The iteration runs as long as the decrease is less than the specified threshold, either relative or absolute. The maximal relative decrease parameter is used for specifying the maximal relative decrease if the use relative decrease parameter is set to true. Otherwise, the maximal absolute decrease parameter is used for specifying the maximal absolute decrease.
 with significant decrease: The iteration stops as soon as the decrease is significant to the level specified by the alpha parameter.
The speculative rounds parameter defines how many rounds will be performed in a row, after the first time the stopping criterion is fulfilled. If the performance increases again during the speculative rounds, the elimination will be continued. Otherwise all additionally eliminated attributes will be restored, as if no speculative rounds had executed. This might help avoiding getting stuck in local optima.

Feature selection i.e. the question for the most relevant features for classification or regression problems, is one of the main data mining tasks. A wide range of search methods have been integrated into RapidMiner including evolutionary algorithms. For all search methods we need a performance measurement which indicates how well a search point (a feature subset) will probably perform on the given data set [23].

4.2 Forward Selection : The Forward Selection operator starts with an empty selection of attributes and, in each round, it adds each unused attribute of the given ExampleSet. For each added attribute, the performance is estimated using the inner operators, e.g. a cross-validation. Only the attribute giving the highest increase of performance is added to the selection. Then a new round is started with the modified selection. This implementation avoids any additional memory

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

consumption besides the memory used originally for storing the data and the memory which might be needed for applying the inner operators. The *stopping behavior* parameter specifies when the iteration should be aborted. There are three different options:

- without increase : The iteration runs as long as there is any increase in performance.
- without increase of at least: The iteration runs as long as the increase is at least as high as specified, either relative or absolute. The minimal relative increase parameter is used for specifying the minimal relative increase if the use relative increase parameter is set to true. Otherwise, the minimal absolute increase parameter is used for specifying the minimal absolute increase.
- without significant increase: The iteration stops as soon as the increase is not significant to the level specified by the alpha parameter.

The *speculative rounds* parameter defines how many rounds will be performed in a row, after the first time the stopping criterion is fulfilled. If the performance increases again during the speculative rounds, the selection will be continued. Otherwise all additionally selected attributes will be removed, as if no speculative rounds had executed. This might help avoiding getting stuck in local optima [24].

4.3 Optimize Selection (Evolutionary) : A genetic algorithm (GA) is a search heuristic that mimics the process of natural evolution. This heuristic is routinely used to generate useful solutions to optimization and search problems. Genetic algorithms belong to the larger class of evolutionary algorithms (EA), which generate solutions to optimization problems using techniques inspired by natural evolution, such as inheritance, mutation, selection, and crossover.

In genetic algorithm for feature selection 'mutation' means switching features on and off and 'crossover' means interchanging used features. Selection is done by the specified selection scheme which is selected by the selection scheme parameter. A genetic algorithm works as follows:

Generate an initial population consisting of p individuals. Each attribute is switched on with probability p_i. The numbers p and p_i can be adjusted by the population size and p initialize parameters respectively.

For all individuals in the population Perform mutation, i.e. set used attributes to unused with probability p_m and vice versa. The probability p_m can be adjusted by the p mutation parameter. Choose two individuals from the population and perform crossover with probability p_c. The probability p_c can be adjusted by the p crossover parameter. The type of crossover can be selected by the crossover type parameter. Perform selection, map all individuals according to their fitness and draw p individuals at random according to their probability where p is the population size which can be adjusted by the population size parameter. As long as the fitness improves, go to step number 2.

4.4 Dataset used

There are five dataset has been used for this experiment which are available in Rapidminer.

Table 1: Dataset description

Name of the Dataset	No. of attributes	No. of examples
Deals	4	1000
Golf	5	14
Labor Negotiations	17	40
Sonar	61	208
Weighting	7	500

4.5 Experiment test bed

The test bed has been created in the Rapidminer miner tool. The dataset is loaded and the noise has been added to the database and then the feature selection operators are loaded. The validation operator is used for validating the learner.

4.5.1 X- Validataion Operator

The X-Validation operator is a nested operator. It has two subprocesses: a training subprocess and a testing subprocess. The training subprocess is used for training a model. The trained model is then applied in the testing subprocess. The performance of the model is also measured during the testing phase. The input ExampleSet is partitioned into k subsets of equal size. Of the k subsets, a single subset is retained as the testing data set (i.e. input of the testing subprocess), and

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

the remaining $k-1$ subsets are used as training data set (i.e. input of the training subprocess). The cross-validation process is then repeated *k* times, with each of the *k* subsets used exactly once as the testing data. The *k* results from the *k* iterations then can be averaged (or otherwise combined) to produce a single estimation. The value *k* can be adjusted using the *number of validations* parameter.

The learning processes usually optimize the model to make it fit the training data as well as possible. If we test this model on some independent set of data, mostly this model does not perform that well on testing data as it performed on the data that was used to generate it. This is called 'over-fitting'. The Cross-Validation operator predicts the fit of a model to a hypothetical testing data. This can be especially useful when separate testing data is not present [26].

4.6 Performance Metrics used

The performance metrics used for the experiment is given below

4.6.1 Accuracy:

Accuracy is how close a measured value is to the true value. It expresses the correctness of a measurement and determined by absolute and comparative way.

$$Accuracy = \frac{Sum\ of\ true\ positives + Sum\ of\ true\ negatives}{Total\ population}$$

4.6.2 Precision:

It refers to the closeness of the set of values obtained from identical measurements of a quantity. It represents the reproducibility of a measurement.

$$Precision = \frac{Sum\ of\ true\ positives}{Sum\ of\ true\ positives + Sum\ of\ false\ positives}$$

4.6.3 Recall

It measures the proportion of positives that are correctly identified as such. It is also called as the sensitivity of the classifier.

$$Recall = \frac{Sum\ of\ true\ positives}{Sum\ of\ true\ positives + Sum\ of\ false\ negatives}$$

V. RESULTS AND DISCUSSIONS

The five dataset have been employed with the three types of feature selection strategies. The results are presented below. Table 2 exhibits the results obtained from the backward elimination when it is applied for the five dataset described above.

Table 2. Results for the Backward Elimination strategy

	Accuracy	Precision	Recall
Deals	81.8	94.65	68.71
Golf	60	71.43	55.56
labor negotiations	87.5	93.33	90
Sonar	65.07	65.65	74.39
Weighting	82	81.05	85.06

Table 3 exhibits the results obtained from the Forward selection when it is applied for the five dataset described above. The values for the three metrics are displayed.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

Table 3. Results for the Forward selection strategy

	Accuracy	Precision	Recall
Deals	83.2	82.14	87.05
Golf	65	65	100
labor negotiations	90	91.67	96.67
Sonar	77.88	78.44	83.03
Weighting	83.2	82.14	87.05

Table 4 exhibits the results obtained from the optimized selection strategy using the genetic algorithm when it is applied for the five dataset described above. The values for the three metrics are displayed.

Table 4. Results for the Optimized selection (Using genetic Algorithm) strategy

	Accuracy	Precision	Recall
Deals	94.5	94.08	95.33
Golf	85	87.5	77.78
labor negotiations	92.5	91.67	95
Sonar	70.69	71.69	78.41
Weighting	84.6	83.5	87.8

Figure 2 displays the comparison of the three selection strategies used based on the performance metric accuracy. Based on the comparison of the values obtained optimized selection strategy performs better than other two for the four dataset except for the sonar dataset. For the sonar dataset forward selection aids the good result.

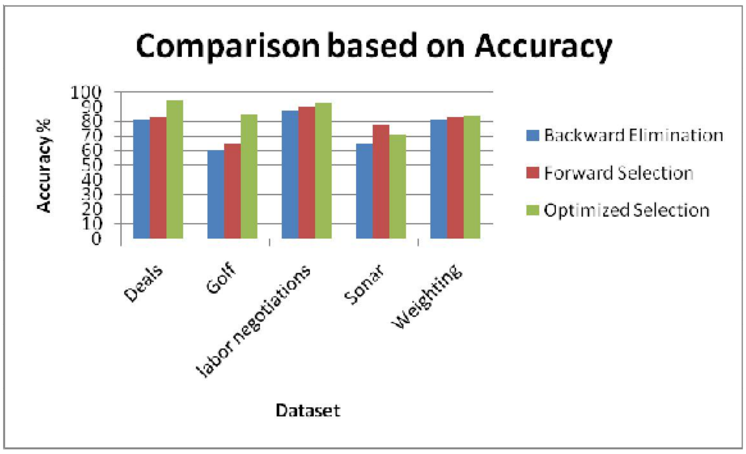

Fig 2: Comparison based on the accuracy measure for the classifier

Figure 3 displays the comparison of the three selection strategies used based on the performance metric precision. Based on the comparison of the values obtained optimized selection strategy performs better than other two for the four dataset except for the sonar dataset. For the sonar dataset forward selection gives the good result.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

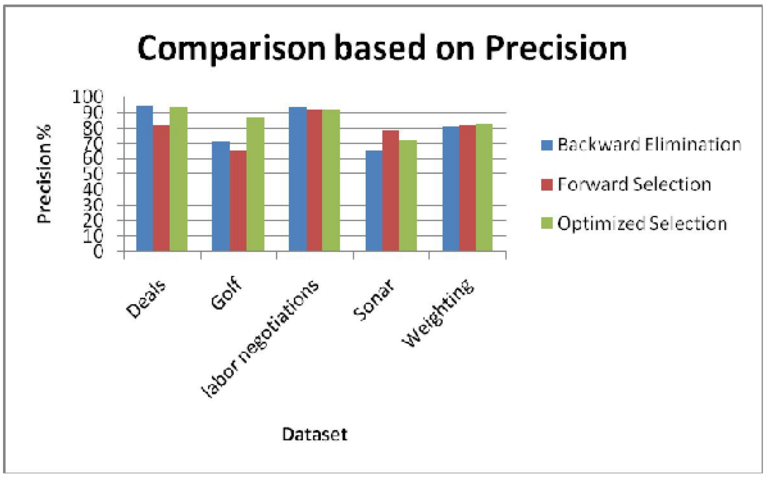

Fig 3: Comparison based on the precision measure for the classifier

Figure 4 displays the comparison of the three selection strategies used based on the performance metric recall. Based on the comparison of the values obtained optimized selection strategy performs better than other two for the three dataset except for the golf and sonar dataset. For the golf and sonar dataset forward selection gives the good result

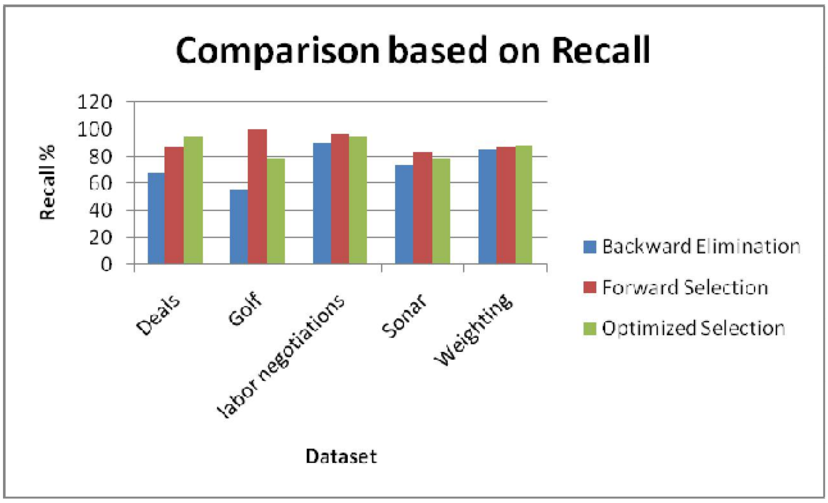

Fig 4: Comparison based on the recall measure for the classifier

Based on the observation of the above results we come to a conclusion that the optimized selection could be used for the feature selction.

IV. CONCLUSION

This paper discusses the feature selection strategies in detail. The various types of feature selection and its importance are neatly sketched. The experiment is carried out through the Rapidminer tool and with the three main selection methods namely backward elimination, forward selection and the evolutionary approach. The evolutionary approach seems to return good results based on the performance metric used for the four dataset out of five dataset. The sonar dataset gives better result for the forward selection strategy. The future study could be concentrated on employing various evolutionary algorithms for the feature selection.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 7, July 2016

REFERENCES

[1] P. Mitra, C. A. Murthy and S. K. Pal. "Unsupervised feature selection using feature similarity," IEEE Transactions on Pattern Analysis and Machine Intelligence , vol.24, no.3, pp. 301–312, 2002.
[2] Miller, "Subset Selection in Regression," Chapman & Hall/CRC (2nd Ed.), 2002.
[3] H. Almuallim and T. G. Dietterich. "Learning boolean concepts in the presence of many irrelevant features", Artificial Intelligence, vol. 69, no.1-2, pp. 279–305, 1994.
[4] D. Koller and M. Sahami, "Toward optimal feature selection," In Proceedings of the Thirteenth International Conference on Machine Learning , pp. 284–292, 1996.
[5] https://rapidminer.com
[6] Huan Liu, Hiroshi Motoda, Rudy Setiono, Zheng Zhao." Feature Selection: An Ever Evolving Frontier in Data Mining", Journal of Machine Learning Research - JMLR , vol.10, pp. 4-13, 2010.
[7]I. Guyon , A. Elisseeff . "An introduction to variable and feature selection", J Mach Learn Res. vol.3, pp.1157–1182, 2003.
[8] http://machinelearningmastery.com/an-introduction-to-feature-selection
[9] M. A. Esseghir. "Effective wrapper- lter hybridization through grasp schemata". In The 4th Workshop on Feature Selection in Data Mining , 2010.
[10] D. Koller, and M. Sahami, "Toward optimal feature selection". In:Proceedings of International Conference on Machine Learning, 1996.
[11] Shima Kashef, Hossein Nezamabadi-pour, "An advanced ACO algorithm for feature subset selection", Neurocomputing vol.147, pp.271–279 , 2015.
[12] S. Ding. " Feature selection based F-score and ACO algorithm in support vector machine", 2nd International Symposium on Knowledge Acquisition and Modeling, 2009.
[13] L.T. Vinh, S. Lee, Y.-T. Park and B.J. d'Auriol, "A novel - feature selection method based on normalized mutual information", Appl Intell vol.37, pp.100-120, 2010.
[14] M.H. Aghdam, N. Ghasem-Aghaee and M.E. Basiri, "Text feature selection using ant colony optimization, Expert Systems with Applications", vol.36, pp.6843-6853, 2009.
[15] M. Kabir, Shahjahan, K. Murase, "A new local search based hybrid genetic algorithm for feature selection", Neurocomputing, vol.74, pp.2914–2928, 2011.
[16] R. Kohavi, G. John, "Wrappers for feature selection", Artificial Intelligence, vol.97, no. 1-2, pp. 273-324, 1997.
[17] M.E. Basiri, N. Ghasem-Aghaee, M.H. Aghdam, "Using ant colony optimization-based selected features for predicting post-synaptic activity in proteins", EvoBIO in: Lecture Notes in Computer Science, vol.4973, pp.12-23, 2008.
[18] P. Bermejo, J.A. Gámez and J.M. Puerta, "A GRASP algorithm for fast hybrid (filter-wrapper) feature subset selection in highdimensional datasets", Pattern Recognition Letters, vol.32, pp. 701–711, 2001.
[19] J. Huang, Y. Cai and X. Xu, "A hybrid genetic algorithm for feature selection wrapper based on mutual information," Pattern Recognition Letters Vol. 28, pp.1825-1844, 2007.
[20] R.K. Sivagaminathan, S. Ramakrishnan, "A hybrid approach for feature subset selection using neural networks and ant colony optimization", Expert Systems with Applications, vol.33, pp. 49-60, 2007.
[21] https://rapid-i.com/wiki/index.php?title=Feature_selection
[22] https://en.wikipedia.org/wiki/RapidMiner
[23]http://docs.rapidminer.com/studio/operators/modeling/optimization/feature_selection/optimize_selection_backward.html
[24]http://docs.rapidminer.com/studio/operators/modeling/optimization/feature_selection/optimize_selection_forward.html
[25]http://docs.rapidminer.com/studio/operators/modeling/optimization/feature_selection/optimize_selection_evolutionary.html
[26] http://docs.rapidminer.com/studio/operators/validation/x_validation.html

BIOGRAPHY

Dr.J.Arunadevi, Assistant Professor in PG & Research Department. of Computer Science, Raja Doraisingam Govt.Arts College, Sivaganga, Tamilnadu, India. She received her MCA from Manonmanium Sundaranar University, Trinelvelli in 2004, she is the University Gold Medallist and obtained Ph.D from Mother Teresa Women's university in 2012. She has published nearly 20 research papers in the national and international journals. She is currently pursuing research in the area of Data mining, Machine learning and Evolutionary Computing.

M. Josephine Nithya is the M.Phil research scholar in PG & Research Department. of Computer Science, Raja Doraisingam Govt.Arts College, Sivaganga, Tamilnadu, India.