Research Journal of Applied Sciences, Engineering and Technology 9(10): 834-840, 2015
DOI:10.19026/rjaset.9.2632
ISSN: 2040-7459; e-ISSN: 2040-7467

Submitted: November 10, 2014 Accepted: January 27, 2015 Published: April 05, 2015

Research Article

A Hybrid Feature Subset Selection using Metrics and Forward Selection

[1]K. Fathima Bibi and [2]M. Nazreen Banu
[1]Department of Computer Science, Bharathiar University, Coimbatore, TN, India
[2]Department of MCA, MAM College of Engineering, Tiruchirappalli, TN, India

Abstract: The aim of this study is to design a Feature Subset Selection Technique that speeds up the Feature Selection (FS) process in high dimensional datasets with reduced computational cost and great efficiency. FS has become the focus of much research on decision support system areas for which data with tremendous number of variables are analyzed. Filters and wrappers are proposed techniques for the feature subset selection process. Filters make use of association based approach but wrappers adopt classification algorithms to identify important features. Filter method lacks the ability of minimization of simplification error while wrapper method burden weighty computational resource. To pull through these difficulties, a hybrid approach is proposed combining both filters and wrappers. Filter approach uses a permutation of ranker search methods and a wrapper which improves the learning accurateness and obtains a lessening in the memory requirements and finishing time. The UCI machine learning repository was chosen to experiment the approach. The classification accuracy resulted from our approach proves to be higher.

Keywords: Algorithm, filter, machine learning, ranker search, repository, wrapper

INTRODUCTION

With sufficient data, it is well to use all the features, together with those unconnected ones. But in practice, there are two problems which may be evolved by the irrelevant features involved in the learning process:

- The irrelevant input features will induce greater computational cost.
- The unrelated features may lead to over fitting. Example: If the identification number of the patient is by fault taken as one input feature, the finale may be that, the sickness is determined by this feature.

Feature selection estimates the primary function between the input and the output; it is reasonable and important to ignore those input features with little effect on the output, so as to keep the size of the approximation model small.

Feature selection is a diminution technique which consists of detecting the related features and discarding the unrelated ones and has been studied for many years. An accurate selection can upgrade the learning speed (Guyon *et al*., 2006). Feature selection has won triumph in many diverse real world cases (Yu and Liu, 2004; Bolon-Canedo *et al*., 2011; Forman *et al*., 2003; Saari *et al*., 2011).

Filters and wrappers are the two estimation strategies. In filters individual features are evaluated autonomously of the knowledge algorithms whereas wrappers use the knowledge algorithm to assess feature subsets. In this study, we bring in a hybrid technique to choose features more accurately by using a combination of ranker search methods and a wrapper algorithm which uses a distributed learning method from multiple subsets of data processed concurrently. Here the learning is parallelized by distributing the subsets of data to multiple processors and then combining the obtained results into a single subset of relevant features. In this way, the cost of computation and the time required will be appreciably reduced. The experiments are made to study the importance of the feature selection process in the lung cancer data set collected from the UCI machine learning repository.

LITERATURE REVIEW

Wrappers for Feature Subset Selection (Kohavi and John, 1997) searches for an optimal feature subset tailored to a particular algorithm and a domain. Significant improvement in accuracy was obtained using decision trees and naïve-bayes.

Rough-Set Based Hybrid Feature Selection Method for Topic-Specific Text Filtering (Li *et al*., 2004) selects features using x^2 statistic, information gain and then by means of rough set. Naïve-bayes was used to assess the method.

Corresponding Author: K. Fathima Bibi, Department of Computer Science, Bharathiar University, Coimbatore, TN, India

Evaluating feature selection methods for learning in data mining applications (Piramuthu, 2004) evaluates several probabilistic distance-based feature selection methods for inducing decision trees using five-real world data sets.

Euclidean Based Feature Selection for Network Intrusion Detection (Suebsing and Hiransakolwong, 2009) applies Euclidean distance for selecting a subset of robust features using smaller storage space and getting higher intrusion detection performance. Three different test data sets are used to weigh up the management of the proposed technique.

PROPOSED METHODOLOGY

In our proposed methodology, we make use of a hybrid approach. A first step was added in order to rank the features. After obtaining this ranking, the wrapper model will be the focus of our attention.

Ranker search methods:

- Select from the list given below until all the feature selection methods are applied with ranker search on the given dataset
  List of feature selection methods:
  - The Euclidean distance:

$$d(i,j) = \sqrt{(xi1 - xj1)^2 + (xi2 - xj2)^2 + \cdots + (xin - xjn)^2} \quad (1)$$

where, $i = (x_{i1}, x_{i2} \ldots x_{in})$ and $j = (x_{j1}, x_{j2} \ldots x_{jn})$, are two n-dimensional vectors

  - The manhattan (or city block) distance:

$$d(i,j) = |xi1 - xj1| + |xi2 - xj2| + \ldots . + |xin - xjn| \quad (2)$$

  - The Minkowski distance between two objects, using p = 3:

$$d(i,j) = (|xi1 - xj1|\hat{}3 + |xi2 - xj2|\hat{}3 + \ldots . + |xin - xjn|\hat{}3)\hat{}(\tfrac{1}{3}) \quad (3)$$

- To calculate, rank for a feature, compute:
  - Midrange:

(mr) = (largest assessment-smallest assessment) /2 (4)

  - Euclidean distance:

$$d(i,j) = \sqrt{(mr - data1)^2 + (mr - data2)^2 + \cdots + (mr - datan)^2} \quad (5)$$

  - Manhattan distance:

$$d(i,j) = |mr - data1| + |mr - data2| + \ldots . + |mr - datan| \quad (6)$$

  - Minkowski distance:

$$d(i,j) = (|mr - data1|^3 + |mr - data2|^3 + \ldots . + |mr - datan|^3)\hat{}(\tfrac{1}{3}) \quad (7)$$

- The output lists the features in a descending array
- Assign the weights for the features in the array from n to 1
- Add the weights of all the methods and store it in a descending list. Now grade the features from n to 1

Wrapper model: The idea of the wrapper approach is to select a feature subset using a learning algorithm. Searching procedure consists of two basic issues. Forward selection technique begins with an empty set and adds features successively. Backward elimination technique begins with a full set and removes features successively (Das, 2001); but forward selection is far less delayed therefore this issue will be used in our experimental research.
The data is split into groups where each group consists of *k* features obtained successively over the grading.

Wrapper is applied to the displace datasets DS_i. A range SL_i is now returned for each subset. Now combine the result. Calculate the classification accuracy for the first range SL_1 and the features in it are part of the final range. Mark it as support. The remaining ranges SL_j will be added to the final range SL if they pick up the support accuracy. Final classification accuracies are now obtained thru range SL. With this ultimate step the unrelated and unnecessary features are removed.

Table 1: Ranking of features using the different metrics

Euclidean			Manhattan			Minkowski		
Weight	Attribute	Rank	Weight	Attribute	Rank	Weight	Attribute	Rank
56	X9	3.162277660	56	X9	10	56	X3	2.223980091
55	X19	3.162277660	55	X19	10	55	X9	2.154434690
54	X20	3	54	X20	9	54	X19	2.154434690
53	X33	3	53	X33	9	53	X20	2.080083823
52	X3	2.915475947	52	X3	8	52	X33	2.080083823
51	X11	2.828427125	51	X11	8	51	X11	2
50	X15	2.828427125	50	X15	8	50	X15	2
49	X24	2.828427125	49	X24	8	49	X24	2
48	X31	2.828427125	48	X31	8	48	X31	2
47	X32	2.828427125	47	X32	8	47	X32	2
46	X26	2.645751311	46	X16	7	46	X26	1.912931183
45	X34	2.645751311	45	X26	7	45	X34	1.912931183
44	X35	2.645751311	44	X34	7	44	X35	1.912931183
43	X8	2.449489743	43	X35	7	43	X8	1.817120593
42	X12	2.449489743	42	X8	6	42	X12	1.817120593
41	X14	2.449489743	41	X12	6	41	X14	1.817120593
40	X37	2.449489743	40	X14	6	40	X37	1.817120593
39	X41	2.449489743	39	X37	6	39	X41	1.817120593
38	X16	2.345207880	38	X41	6	38	X6	1.709975947
37	X4	2.236067977	37	X2	5	37	X7	1.709975947
36	X6	2.236067977	36	X4	5	36	X13	1.709975947
35	X7	2.236067977	35	X5	5	35	X25	1.709975947
34	X13	2.236067977	34	X6	5	34	X29	1.709975947
33	X25	2.236067977	33	X7	5	33	X38	1.587401052
32	X29	2.236067977	32	X10	5	32	X42	1.587401052
31	X36	2.236067977	31	X13	5	31	X40	1.442249570
30	X38	2	30	X17	5	30	X4	1.401019665
29	X42	2	29	X21	5	29	X16	1.401019665
28	X40	1.732050808	28	X22	5	28	X36	1.401019665
27	X2	1.581138830	27	X23	5	27	X28	1.259921050
26	X5	1.581138830	26	X25	5	26	X39	1.259921050
25	X10	1.581138830	25	X27	5	25	X43	1.259921050
24	X17	1.581138830	24	X29	5	24	X44	1.259921050
23	X21	1.581138830	23	X30	5	23	X45	1.259921050
22	X22	1.581138830	22	X36	5	22	X2	1.077217345
21	X23	1.581138830	21	X51	5	21	X5	1.077217345
20	X27	1.581138830	20	X54	5	20	X10	1.077217345
19	X30	1.581138830	19	X55	5	19	X17	1.077217345
18	X46	1.581138830	18	X56	5	18	X21	1.077217345
17	X49	1.581138830	17	X38	4	17	X22	1.077217345
16	X50	1.581138830	16	X42	4	16	X23	1.077217345
15	X51	1.581138830	15	X40	3	15	X27	1.077217345
14	X54	1.581138830	14	X46	3	14	X30	1.077217345
13	X55	1.581138830	13	X49	3	13	X46	1.077217345
12	X56	1.581138830	12	X50	3	12	X49	1.077217345
11	X28	1.414213562	11	X28	2	11	X50	1.077217345
10	X39	1.414213562	10	X39	2	10	X51	1.077217345
9	X43	1.414213562	9	X43	2	9	X54	1.077217345
8	X44	1.414213562	8	X44	2	8	X55	1.077217345
7	X45	1.414213562	7	X45	2	7	X56	1.077217345
6	X18	0	6	X18	0	6	X18	0
5	X47	0	5	X47	0	5	X47	0
4	X48	0	4	X48	0	4	X48	0
3	X52	0	3	X52	0	3	X52	0
2	X53	0	2	X53	0	2	X53	0
1	X1	0	1	X1	0	1	X1	0

Table 2: Combined weights of the different metrics

Attributes	Euclidean	Manhattan	Minkowski	Combined weight
X9	56	56	55	167
X19	55	55	54	164
X20	54	54	53	161
X3	52	52	56	160
X33	53	53	52	158
X11	51	51	51	153
X15	50	50	50	150
X24	49	49	49	147
X31	48	48	48	144
X32	47	47	47	141
X26	46	45	46	137
X34	45	44	45	134
X35	44	43	44	131
X8	43	42	43	128
X12	42	41	42	125
X14	41	40	41	122
X37	40	39	40	119
X41	39	38	39	116
X16	38	46	29	113
X6	36	34	38	108
X7	35	33	37	105
X4	37	36	30	103
X13	34	31	36	101
X25	33	26	35	94
X29	32	24	34	90
X2	27	37	22	86
X5	26	35	21	82
X36	31	22	28	81
X38	30	17	33	80
X10	25	32	20	77
X42	29	16	32	77
X40	28	15	31	74
X17	24	30	19	73
X21	23	29	18	70
X22	22	28	17	67
X23	21	27	16	64
X28	11	25	27	63
X27	20	25	15	60
X30	19	23	14	56
X39	10	10	26	46
X51	15	21	10	46
X46	18	14	13	45
X43	9	9	25	43
X54	14	20	9	43
X49	17	13	12	42
X44	8	8	24	40
X55	13	19	8	40
X50	16	12	11	39
X45	7	7	23	37
X56	12	18	7	37
X18	6	6	6	18
X47	5	5	5	15
X48	4	4	4	12
X52	3	3	3	9
X53	2	2	2	6
X1	1	1	1	3

Table 3: Classification results for implementations of wrapper

Classifier	Accuracy				Reduced features		Time HH:MM:SS	
	Data set taken as a whole		Data set divide into subsets		Data set taken as a whole	Data set divide into subsets	Data set taken as a whole	Data set divide into subsets
	Train	Test	Train	Test				
C4.5	94.56	82.63	98.75	86.50	49	38	06:30:51	00:01:08
Naïve Bayes	95.66	92.04	97.65	91.65	49	37	06:45:08	00:00:29

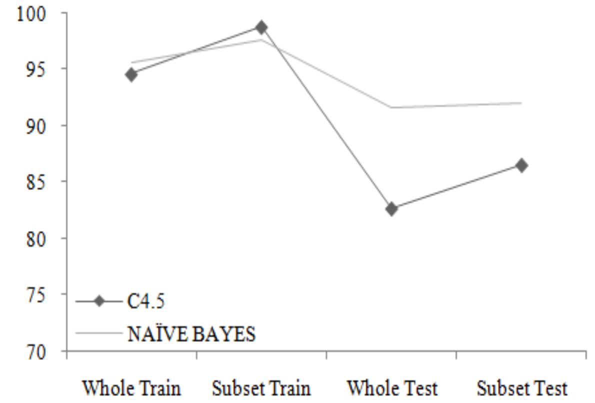


Fig. 1: Increase in accuracy when the dataset was split into subsets

Pseudo-code for the proposed algorithm:

$DS_{(p\ x\ r)}$ = Training dataset with *p* samples and *r* features

q_r = Number of subset of *k* features

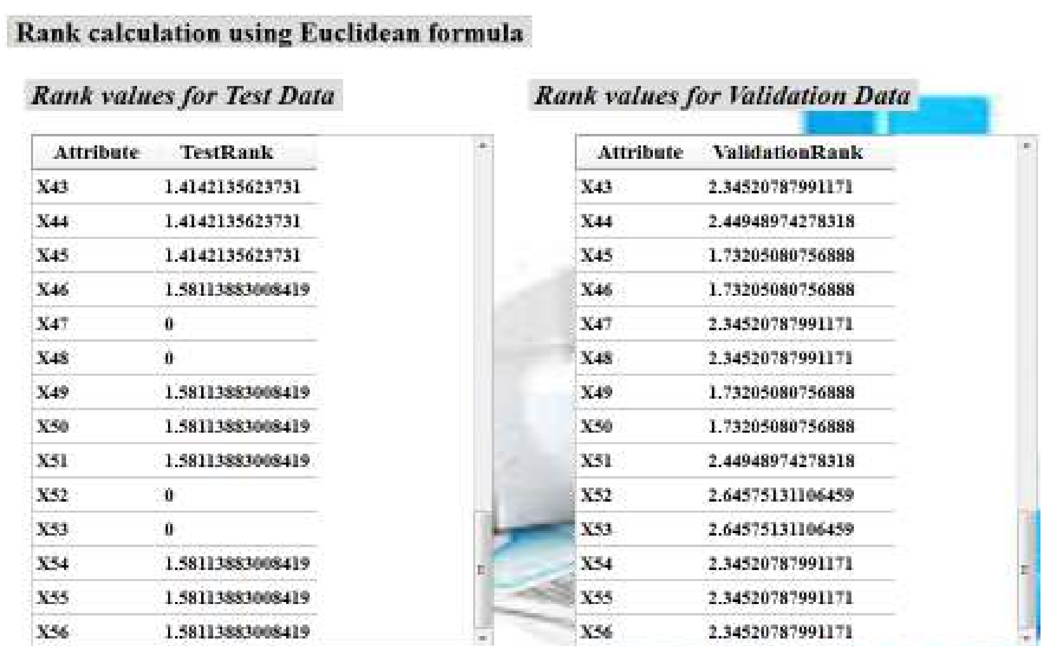
1. Apply ranker search methods over DS and obtain a ranking R1 of the features
2. for i = 1 to q_r
   (a) $R1_i$ = first k features in R1
   (b) R1 = R1 \ R_i
   (c) $DS_i = DS_{(p\ x\ Ri)}$
3. for i = 1 to q_r
   (a) SL = subset of features obtained after applying wrapper over DS_i.
4. SL = SL_1
5. support = accuracy classifying subset DS (p x r_i) with classifer CL.
6. for i = 2 to q_r
   (a) $DS_{(backup)}$ = DS U DS_i
   (b) accuracy = classifying subset DS (p x $DS_{(backup)}$) with classifer CL
   (c) if accuracy > support
      i. DS = $DS_{(backup)}$
      ii. support = accuracy
7. Build classifier CL with $DS_{(p\ x\ DS)}$
8. Obtain prediction P

RESULTS AND DISCUSSION

Lung cancer data set is used for the experiments. 50% of the data set is taken as the training dataset and the remaining are taken as the test data set. The univariate filters Euclidean, Manhattan and Minkowski provides an ordered ranking of all the features. Table 1 shows the ranking of features and Table 2 shows the combined ranking of features.

To show the adequacy of the proposed wrapper, it will be compared with the performance when applying the wrapper over the whole set of features directly. Table 3 shows the classification accuracy, the number of features and the execution time required on the dataset. Figure 1 show the increase in accuracy when the data set was split into subsets.

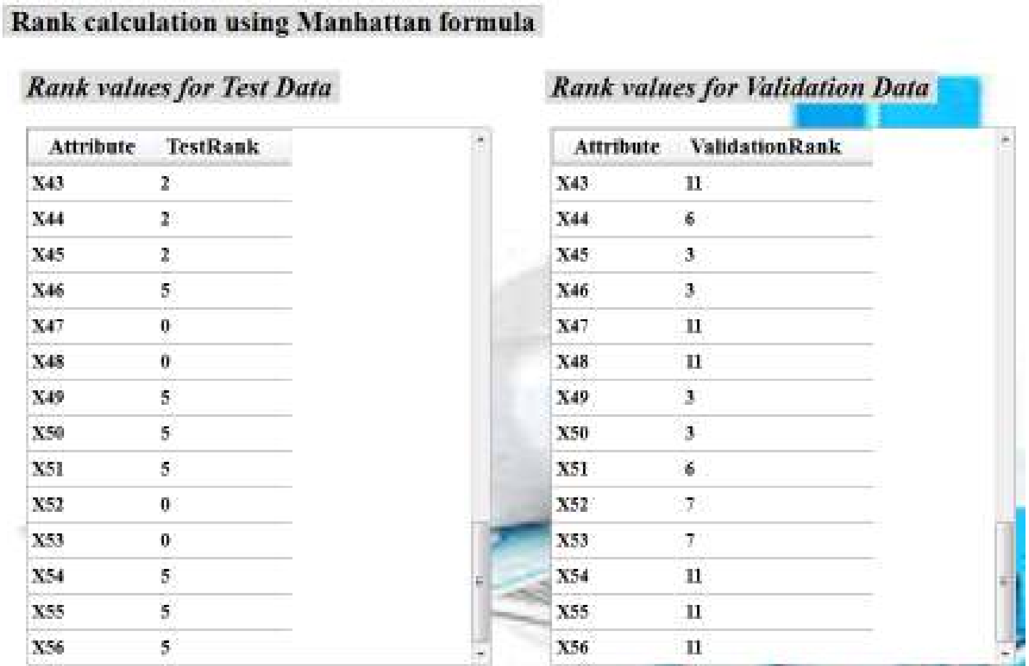
Screen shots for ranking of features using the different metrics: Experiments are carried out using dot net technology for ranking of features using the different metrics such as Euclidean (Fig. 2), Manhattan (Fig. 3) and Minkowski (Fig. 4) on test and validation data.



Attribute	TestRank
X43	1.4142135623731
X44	1.4142135623731
X45	1.4142135623731
X46	1.58113883008419
X47	0
X48	0
X49	1.58113883008419
X50	1.58113883008419
X51	1.58113883008419
X52	0
X53	0
X54	1.58113883008419
X55	1.58113883008419
X56	1.58113883008419

Attribute	ValidationRank
X43	2.34520787991171
X44	2.44948974278318
X45	1.73205080756888
X46	1.73205080756888
X47	2.34520787991171
X48	2.34520787991171
X49	1.73205080756888
X50	1.73205080756888
X51	2.44948974278318
X52	2.64575131106459
X53	2.64575131106459
X54	2.34520787991171
X55	2.34520787991171
X56	2.34520787991171

Fig. 2: Rank calculation using Euclidean metric on test and validation data



Rank calculation using Manhattan formula

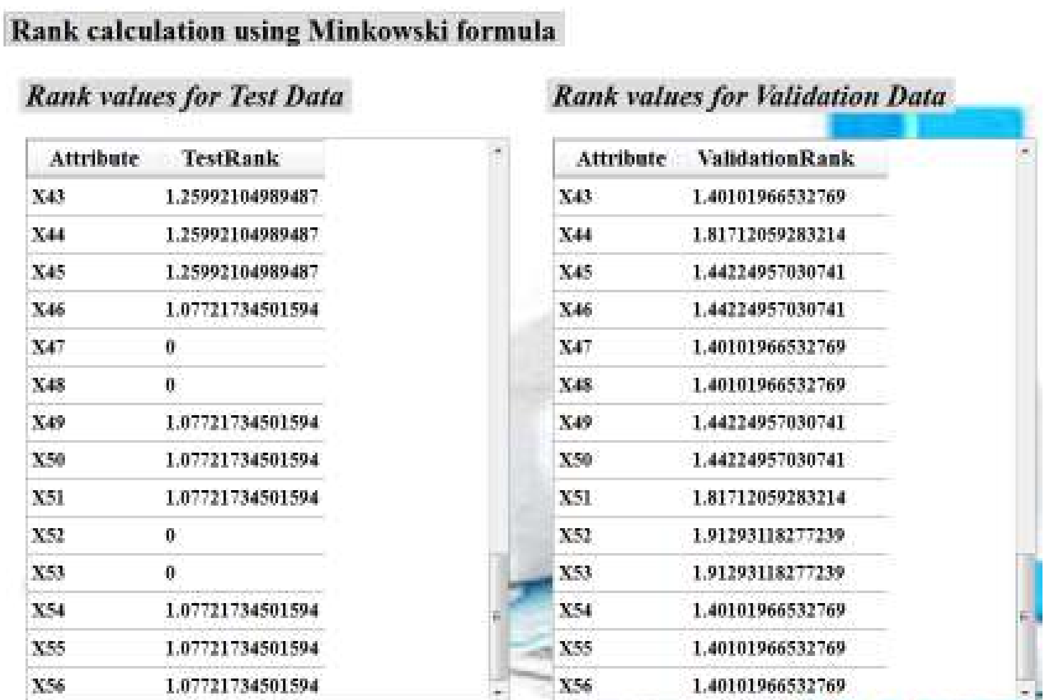
Rank values for Test Data

Attribute	TestRank
X43	2
X44	2
X45	2
X46	5
X47	0
X48	0
X49	5
X50	5
X51	5
X52	0
X53	0
X54	5
X55	5
X56	5

Rank values for Validation Data

Attribute	ValidationRank
X43	11
X44	6
X45	3
X46	3
X47	11
X48	11
X49	3
X50	3
X51	6
X52	7
X53	7
X54	11
X55	11
X56	11

Fig. 3: Rank calculation using Manhattan metric on test and validation data



Rank calculation using Minkowski formula

Rank values for Test Data

Attribute	TestRank
X43	1.25992104989487
X44	1.25992104989487
X45	1.25992104989487
X46	1.07721734501594
X47	0
X48	0
X49	1.07721734501594
X50	1.07721734501594
X51	1.07721734501594
X52	0
X53	0
X54	1.07721734501594
X55	1.07721734501594
X56	1.07721734501594

Rank values for Validation Data

Attribute	ValidationRank
X43	1.40101966532769
X44	1.81712059283214
X45	1.44224957030741
X46	1.44224957030741
X47	1.40101966532769
X48	1.40101966532769
X49	1.44224957030741
X50	1.44224957030741
X51	1.81712059283214
X52	1.91293118277239
X53	1.91293118277239
X54	1.40101966532769
X55	1.40101966532769
X56	1.40101966532769

Fig. 4: Rank calculation using Minkowski metric on test and validation data

CONCLUSION

To sum up this study, obtain an ordered ranking of all features using a combination of ranker search methods. Divide each dataset DS into several small disjoint datasets DS_i vertically by features. The wrapper algorithm is applied to each one of these subsets and a selection SL_i is generated for each subset of data. After all the small datasets DS_i were used the combination method constructs the final selection SL as the result of the feature selection process. The experiments showed that our method led to a reduction in the running time as well as in the storage requirements while accuracy did not drop.

REFERENCES

Bolon-Canedo, V., N. Sanchez-Marono and A. Alonso-Betanzos, 2011. Feature Selection and classification in multiple class datasets: An application to kdd cup 99 dataset. Exp. Syst. Appl., 38(5): 5947-5957.

Das, S., 2001. Filters, wrappers and a boosting-based hybrid for feature selection. Proceedings of the 18th International Conference on Machine Learning, Morgan Kaufmann Publishers Inc., pp: 74-81.

Forman, G., 2003. An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res., 3: 1289-1305.

Guyon, I., S. Gunn, M. Nikravesh and L.A. Zadeh, 2006. Feature extraction: Foundations and applications. Springer, 207: 301-317.

Kohavi, R. and G.H. John, 1997. Wrappers for feature subset selection. Artificial Intell., 97: 273-324.

Li, Q., J.H. Li, G.S. Liu and S.H. Li, 2004. A rough set-based hybrid feature selection method for topic-specific text filtering. Proceedings of the 3rd International Conference on Machine Learning and Cybernetics, Shanghai, pp: 1464-1468.

Piramuthu, S., 2004. Evaluating feature selection methods for learning in data mining applications. Eur. J. Operat. Res., 156: 483-494.

Saari, P., T. Eerola and O. Lartillot, 2011. Generalizability and simplicity as criteria in feature selection: Application to mood classification in music. IEEE T. Audio, Speech Lang. Process., 19(6): 1802-1812.

Suebsing, A. and N. Hiransakolwong, 2009. Euclidean based feature selection for network intrusion detection. Proceeding of International Conference on Machine Learning and Computing (IPCSIT), Singapore, 3: 222-229.

Yu, L. and H. Liu, 2004. Redundancy based feature selection for microarray data. Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp: 737-742.