HYBRID FEATURE SELECTION TECHNIQUE USING FILTER (FCBF) & WRAPPER METHOD (SFFS) FOR DISCRETE CLASS DATA MINING

[1]ABHILASHA SHARMA, [2]ARTI DESHPANDE

[1]M.E Student, [2]Assistant Professor, Department of Computer Engineering, Thadomal Shahani Engineering College, Mumbai, India

Abstract - Feature Selection can extensively increase the scope of classifiers to work on high dimensional data by reducing the variables or features used for classification. A dataset may represent a number of attributes which have little relevance to classification problem associated with it. By choosing a subset of the most useful attributes, classification is best aided to produce most accurate classification results. This paper uses a hybrid feature selection technique using supervised feature selection techniques. The filter method is used to sort the features in order of their 'usefulness' and the wrapper method is used for picking out the best subset from these features to simplify classification. The last step is to create a comparison in performance between attributes acquired as a result of the hybrid model and the total set by passing both via the classifier to calculate error and accuracy.

Keywords - data mining, feature selection, fast correlation based filter, sequential forward floating selection, hybrid model, classification, naïve bayes

I. INTRODUCTION – FEATURE SELECTION

Feature (or variable, or attribute) subset selection (FSS) is the process of identifying the input variables which are relevant to a particular learning (or data mining) problem Feature selection can play a good role in reducing the complexity and enhancing the performance of the recognition system by selecting salient features and discard irrelevant or redundant features. In other words, selecting a prominent subset of features is the aim of the feature selection technique which not only decreases the dimensionality, but also increases the accuracy rate of the classifier.

The proposed system is a hybrid model which aims at combining two known techniques of feature selection. These primary models are combined with the hope that they will eliminate each other's disadvantages to produce much stronger results than what would have been possible individually. Using classification as a final step, using the features identified by the hybrid model, the proposed system will be tested for accuracy using the subset of features against the entire set.

II. FEATURE SELECTION METHODS

A. Fast Correlation Based Filter - FCBF

FCBF is an efficient feature selection algorithm based on relevance among features and redundancy values. It is a multivariate feature selection method starting with a full set of features, using symmetrical uncertainty (SU) to calculate the dependences of features [5]. The algorithm consists of two stages: the first one is a relevance analysis, aimed at ordering the input variables depending on a relevance score, which is computed as SU with respect to the target output. This stage is also used to discard irrelevant variables, which are those whose ranking score is below a predefined threshold. The second stage is a redundancy analysis, aimed at selecting predominant features from the relevant set obtained in the first stage. However, in the proposed model only the first stage, that is, relevance analysis will be implemented. The output from filtering will be a list of features in ranked order based on SU.

B. Sequential Forward Floating Selection - SFFS

SFS is the simplest greedy search algorithm – It starts from an empty set gradually adding feature x^+ after every iteration. It adds x^+ to a set Yktill ($Yk + x^+$) is a subset of a full set of features. SFS performs best when the optimal subset is small. A criteria (objective function) is chosen in order to determine the most significant attribute in the subset. The creation of a subset is complete when the repetitive iterations result in the same subset.The main disadvantage of SFS is that it is unable to remove features that become obsolete after the addition of other features. [13][10]

SBS on the other hand starts with a full set and iteratively removes a feature x^- that reduces the value of the criteria function. This results in a subset of the features which will result in the maximum value of the objective function. The main disadvantage of SBS is that a feature once discarded from the full set cannot be added again, that is, the usefulness of a feature cannot be revaluated once removed.

For this purpose SFFS has been introduced. Steps performed in SFFS are [13]:

- **Step 1:** Inclusion. Use the basic SFS method to select the most significant feature with respect to X and include it in X. Stop if d features have been selected, otherwise go to step 2.
- **Step 2:** Conditional exclusion. Find the least significant feature k in X. If it is the feature

just added, then keep it and return to step 1. Otherwise, exclude the feature k. Note that X is now better than it was before step 1. Continue to step 3.

- **Step 3:** Continuation of conditional exclusion. Again find the least significant feature in X. If its removal will (a) leave X with at least 2 features, and (b) the value of J(X) is greater than the criterion value of the best feature subset of that size found so far, then remove it and repeat step 3. When these two conditions cease to be satisfied, return to step 1.

III. IMPLEMENTATION

The proposed hybrid model will filter out relevant features using Symmetrical Uncertainty (SU) for FCBF. The result will be a ranked list of features in order of their SU. This will allow for narrowing down of features using a threshold. The formula of SU is as follows:

$$SU(X,Y) = 2\left[\frac{IG(X|Y)}{H(X) + H(Y)}\right]$$

The algorithm for ranking features in order of their SU is described in the diagram below:

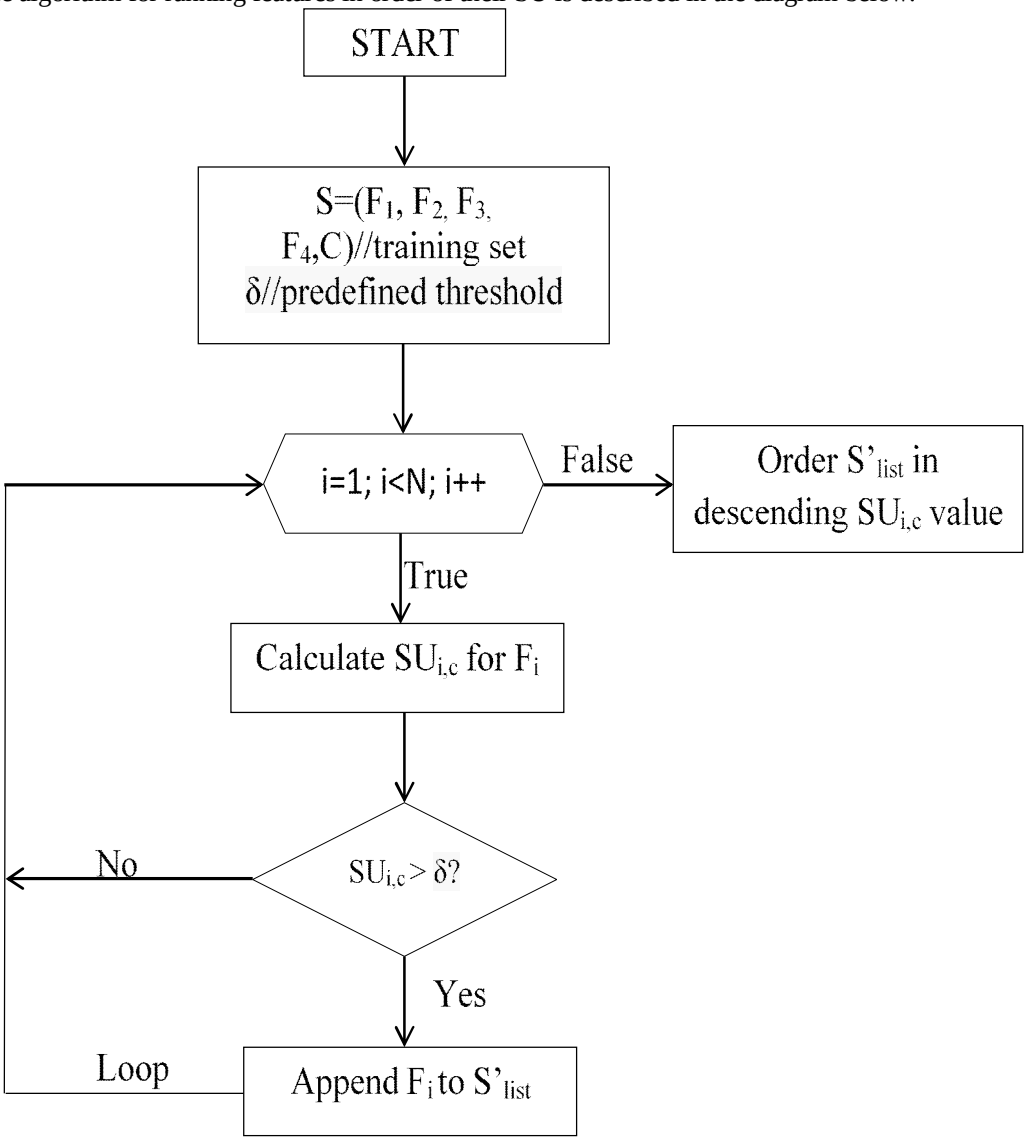

Fig.1: Relevance Analysis in Filter Based Correlation Filter

Using a threshold to select an optimal subset of the features in a dataset, the next step is to further optimize this set. Using SFFS, the feature set received as a result of SU ranking will be optimized. The resulting set will be a smaller set but will be 'best' amongst all features.

The wrapper method is described as follows:

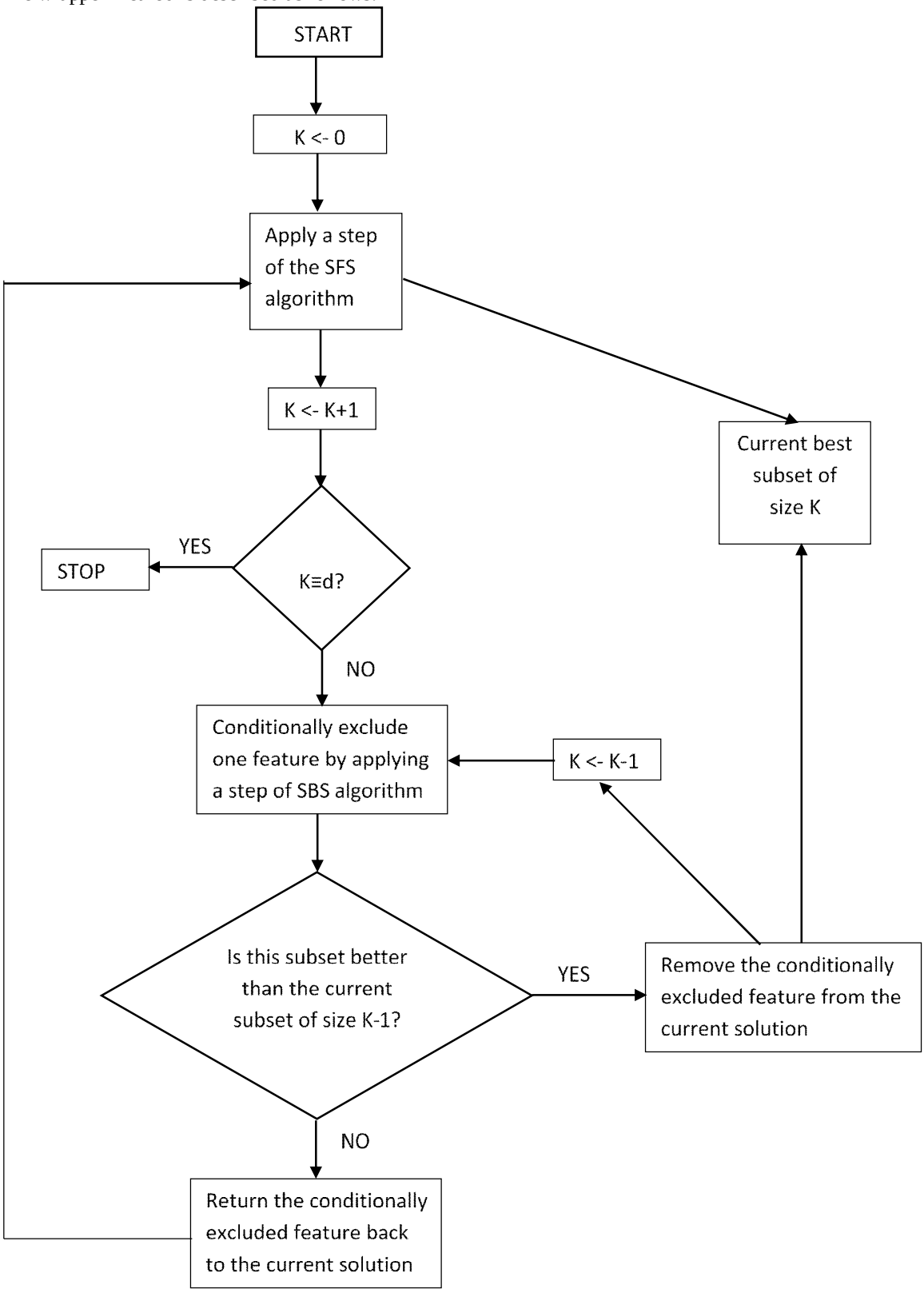

Fig.2: Algorithm of Sequential Forward Filter Selection (Wrapper Feature Selection Technique)

As described in the above diagram, SFS is Sequential Forward selection and SBS is Sequential Backward Selection.

The final step in the hybrid model is to use a classifier to classify the data set using the subset of features as a result of the filter method & wrapper method. The hybrid model is shown as follows:

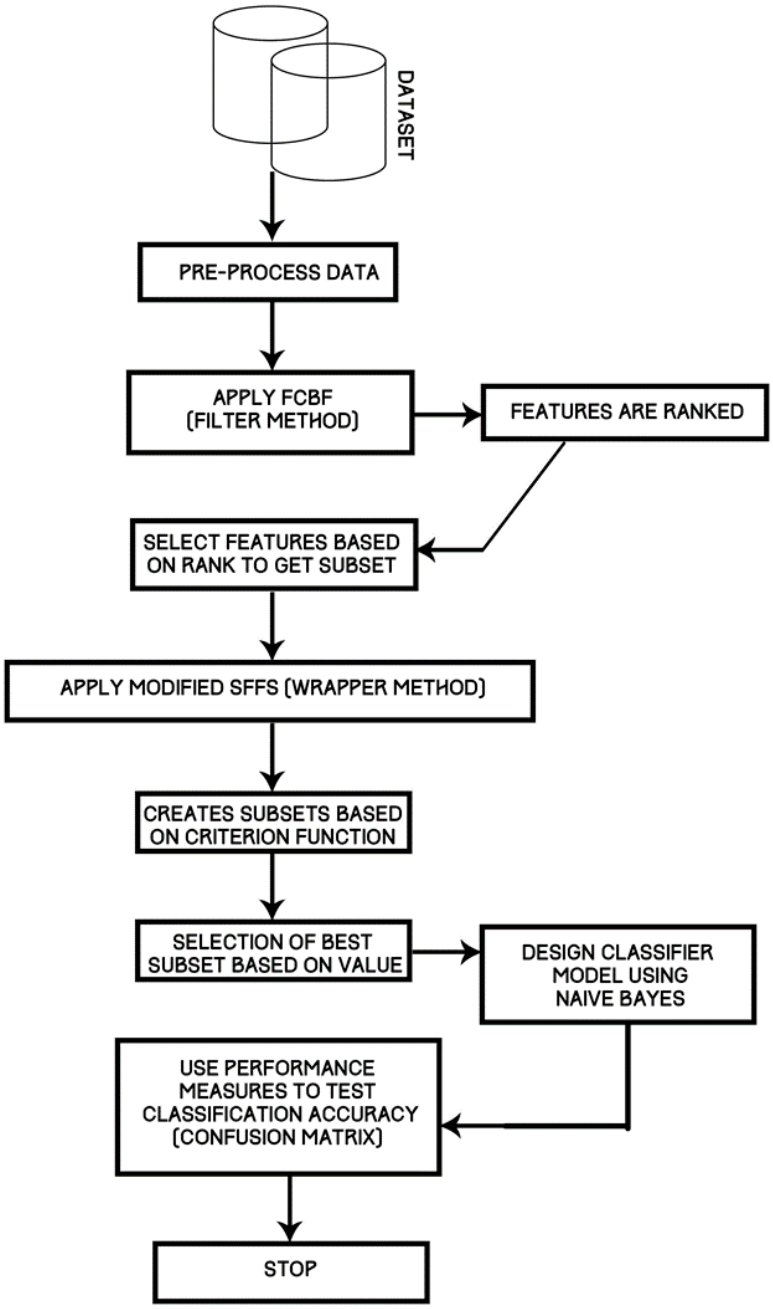

Fig.3: Algorithm of the Hybrid Feature Selection Technique using FCBF and SFFS

The Naïve Bayes classifier will run 3 times. One, on the feature subset acquired after Filter Method. The second, using the feature subset acquired after the Hybrid Model and lastly, using the entire feature set. In order to test accuracy of the hybrid model, confusion matrix will be used on the classifier.A confusion matrix is used to test the performance of a classifier. Confusion matrices are popular in use for supervised classification problems. The confusion matrix is as follows: [14]

	PREDICTED NO	PREDICTED YES
ACTUAL NO	TRUE NEGATIVE	FALSE POSITIVE
ACTUAL YES	FALSE NEGATIVE	TRUE POSITIVE

TABLE I
Confusion Matrix

Formulae using confusion matrix:

$$\text{Accuracy} = \frac{TP + TN}{TP + FN + FP + FN}$$

$$\text{Error} = \frac{FN + FP}{TP + FN + FP + FN}$$

IV. DATASETS

Datasets used have been picked out from UCI's Machine Learning Repository. Three datasets have been picked out for the purpose of this implementation.[15][16][17]

	Chess Data Set	Dermatology Data Set	Zoo Name Data Set
Data Set Characteristics	Multivariate	Multivariate	Multivariate
Attribute Characteristics	Categorical	Categorical, Integer	Categorical, Integer
Number of Instances	3196	366	101
Number of Attributes	36	34	17

TABLE II
Datasets

V. RESULTS

A. Threshold

Since the filter method (FCBF) uses a threshold for filtering out relevant features, the following thresholds were used for the purpose of this execution. The thresholds used were decided after testing the data set over a multitude of thresholds.

Dataset	Threshold
Chess Data Set	0.7
Dermatology Data Set	0.25
Zoo Name Data Set	0.18

TABLE IIII
Thresholds considered for datasets

B. Results

In order to properly compare how classification One of the obvious observations during the implementation is that while Wrappers are heavy on resource consumption, they contribute to better classification results.

Parameters	Filter	Wrapper	Naïve Bayes
No of Attributes	17	12	36
Time	673	2128	853
Accuracy	87.804	59.523	86.757
Memory	735.66 KB	66844.49 KB	45537.56 KB

TABLE IV
Results (Chess Dataset)

As shown in Table IV, the chess dataset has a total of 36 features. Of the 36, 17 features were obtained as a subset when the filter method was implemented. These 17 features were input to the wrapper method in order to further optimize the feature set obtained. When passed through the wrapper method, the result was a mere feature set of 9 features. Lastly, the classifier was run on 3 different feature sets to determine if the classification accuracy obtained for one method was greater than the other. While the hybrid model performed the best, resources consumed for the same were considerable. The filter method on the other hand was able to produce nearly accurate results with the least memory consumption and time. For high – dimensional data, this is important as this implies that filter methods can handle highly dimensional data and still produce better classification results than simply using the entire feature set as input to the classifier.

Similarly, results obtained for the other two datasets are described below:

Parameters	Filter	Wrapper	Naïve Bayes
No of Attributes	7	6	17
Time	16	116	16
Accuracy	80.198	87.128	86.29
Memory	2240.64 KB	35.71 KB	2491.96 KB

TABLE V
Results (Zoo Names Dataset)

Parameters	Filter	Wrapper	Naïve Bayes
No of Attributes	14	6	34
Time	75	473	182
Accuracy	92.076	94.262	85.715
Memory	10756.49 KB	46335.07 KB	27738.03 KB

TABLE V
Results (Dermatology Dataset)

The following conclusions can be drawn from the results found above:

- The filter method (Symmetrical uncertainty ranking) is significantly better at producing results (accuracy) when the dataset is not highly - dimensional.
- The hybrid model results produced far better accurate results than implementing Naïve Bayes it self
- The wrapper method consumes more time and memory than filter method but overall is better at providing a subset which can classify data faster than the classifier itself
- The subset as a result of the hybrid model can be used on large datasets to filter out relevant features from the entire set of features
- Classification can be aided by the use of feature selection to produce more concrete conclusions whether the entire feature set is essential for correct classification of data

FUTURE WORK

- The proposed model works best on data which has less noisy data
- On increasing the set of missing values in the training data, the filter method is unable to calculate SU

- The future scope will encompass strengthening the filter method to handle nosier data
- The proposed model is able to handle integer values but works best on a mix of categorical data and integer values
- The future scope encompasses broadening the scope of data the hybrid model can handle

REFERENCES

[1] Samuel H. Huang. "Supervised feature selection: A tutorial", Artificial Intelligence Research, Vol. 4, No. 2, 2015.
[2] GirishChandrashekar, FeratSahin. "A survey on feature selection methods", Computers and Electrical Engineering, 2014.
[3] Pablo Bermejo, Luis de la Ossa, José A. Gámez, José M. Puerta, "Fast wrapper feature subset selection in high-dimensional datasets by means of filter re-ranking", Knowledge-Based Systems, Volume 25, Issue 1, Pages 35–44, Special Issue on New Trends in Data Mining, February 2012
[4] SunitaBeniwal, Jitender Arora, "Classification and Feature Selection Techniques in Data Mining", International Journal of Engineering Research & Technology (IJERT), Vol. 1 Issue 6, August 2012.
[5] Lei Yu, Huan Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution", Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003.
[6] Gan, J.Q., AwwadShiekhHasan, B. &Tsui, C.S.L. Int. J. Mach. Learn. & Cyber. (2014) 5: 413. doi:10.1007/s13042-012-0139-z
[7] http://www.saedsayad.com/model_evaluation_c.htm
[8] P. Pudil, J. NovoviEova, J. Kittler, "Floating search methods in feature selection", Pattern Recognition Letters 15 (1994). Doi: 10.1016/0167-8655(94)90127-9
[9] Ted W. Way, BerkmanSahiner, Lubomir M. Hadjiiski, Heang-Ping Chan, "Effect of finite sample size on feature selection and classification: A simulation study", Med Phys. 2010 Feb; 37(2): 907–920. Doi: 10.1118/1.3284974
[10] Pablo Bermejo, Jos´e A. G´amez, Jos´e M. Puerta, Speeding Up Incremental Wrapper Feature Subset Selection with Naive Bayes Classifier, Knowledge-Based Systems, 2014
[11] K.Ramachandra Murthy, Dimension Reduction, www.isical.ac.in/~k.ramachandra/slides/Feature%20Selection.pptx
[12] Kai Ming Ting, Confusion Matrix, Encyclopedia of Machine Learning, doi: 10.1007/978-0-387-30164-8_157
[13] https://archive.ics.uci.edu/ml/datasets/Chess+(King-Rook+vs.+King-Pawn)
[14] http://archive.ics.uci.edu/ml/datasets/dermatology
[15] http://archive.ics.uci.edu/ml/datasets/zoo

★★★