International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611

Survey on Methodologies and Techniques Involved in Feature Selection

Pallavi Malji[1], Dr. Sachin Sakhare[2]

[1]Savitribai Phule, Pune University, Pune , Maharashtra, India

[2]Professor, Savitribai Phule, Pune University, Pune , Maharashtra, India

Abstract: ***Feature selection is an important data processing step where irrelevant and redundant attributes are removed for shorter learning time, better accuracy and better comprehensibility. It involves identifying an optimal subset of the most useful features from the original set of features. The efficiency of feature selection algorithm concerns the time required to find a subset of features and the effectiveness is related to the quality of the subset of feature. This paper focuses on the study of different feature selection models, techniques and evaluation measures used for feature selection and their merits and limitations. Then we talk about the different frameworks involved in feature selection and find the gap between them which needs to be bridged so as to improve the feature selection efficiency and performance.***

Keywords: Evaluation measures, frameworks, feature selection models, irrelevant, redundancy.

1. Introduction

Earlier the data used to be stored in a single Relation but in today's technological world the data to be stored is increasing day by day. Consider classification as a mining task, as there are various tables they lead to huge number of Features and all the Features does not add value to the classification task . Usually it is assumed that more the data, it is good, but more the data there are more chances for a classifier to get confused and it might give incorrect results. Therefore, only relevant non redundant data is needed. Thus to improve the efficiency and accuracy of classifiers, we need to select optimal features. Feature selection solves this problem and hence is required for the classifier. The complete set of features contains many attributes but only a few of them are sufficient and required for performing the data mining tasks successfully. Thus we select only those which are important with the mechanism of feature selection.

Feature Selection Models:

There are mainly three models of Feature selection:
1. Filter model
2. Wrapper model
3. Hybrid model

1. **Filter Model:** In filter model the characteristics of training data are used to perform attributes selection.eg: Entropy, distance. It does not use any learning algorithm. Filter model can be implemented in two ways: Feature weighting approach and Subset search method.

In the first approach weights are assigned to individual features using some measures such as Information gain, Symmetric uncertainty, Distance, Consistency and Classifier error rate. Then threshold value is calculated and all the attributes which are above certain threshold value are selected as relevant attributes but this method does not handle redundancy between the attributes, as redundant attributes mostly have same weight or rank.

In second approach, optimal subset of features is found out using different search strategies. Time complexity of subset search approach is high. Therefore, it is less suitable for high dimensional data sets.

Advantages of filter model are:
It is computationally cheaper than Wrapper model.
It is useful for large data sets.

Disadvantage of Filter model is:
It does not consider the effect of the selected features on the performance of classifier

2. Wrapper Model: In Wrapper model, subset search method is used. The subset can be obtained from the above mentioned different strategies and then this subset is given as an input to the Learning algorithm and if for a particular subset the performance of the algorithm improves then it is selected otherwise it is not selected.

Advantage of this model is: It gives accurate results than Filter model.

Disadvantage of this model is: It is costlier in terms of computation and hence, not suitable for large data sets.

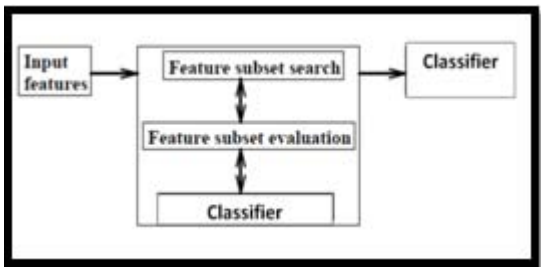


Figure 1: Wrapper Model

3. Hybrid Model*:* A Hybrid model can be implemented by combining the Filter and Wrapper together. In this approach the Filter model provides a smaller subset compared to the entire subset and on this subset the Wrapper is applied.

Advantages of Hybrid Model are:
It is faster than directly applying the Wrapper model on the complete set. It will give more accurate result than Filter model.

Disadvantages of Hybrid Model are:
It will take more time than Filter model.
The Wrapper model selects the exact set of Features but in Hybrid model it is possible that some important features are already lost by the Filter model those will not be considered in the Final subset. Therefore, it will give less accurate results than Wrapper model.

In this paper we present the feature selection classical and new frameworks. In addition we also determine research gaps in feature selection technologies on the basis of study of various research papers in the feature selection technique of data mining domain.

2. Feature Selection Frameworks

There are mainly two types of Framework for feature selection:
1. Classical Framework
2. New Framework

2.1 Classical Framework

It is four step process, which consists of generation of subset, evaluation measures, stopping criteria and validation.

a) **Generation of subset** can be done in following ways described below
1. Exhaustive search
2. Heuristic search
3. Random search

Exhaustive search: The strategy explores all the possible subsets to find the best subset. Accuracy of this strategy is good but it is time consuming.
Heuristic search: It can be further classified as:-
Forward subset selection: It starts from an empty set. The attribute is added in the set based on some evaluation measure.
Backward elimination: It starts with a complete set. The attribute is removed from the set if it results into better subset.
Random search: In this search strategy the attributes are selected in random fashion.

b) Evaluation measures involved in feature selection are mentioned below:
1. Information
2. Distance
3. Consistency
4. Classifier error rate
5. Dependency

c) Stopping criteria: The subset search algorithm stops when one of the following criteria is met
1. Desired number of features are selected
2. Desired number of iterations are completed
3. Optimal subset has obtained
4. No further change in result after addition or removal of features

d) Validation: Accuracy of algorithm for known data can be found by comparing the optimal subset and selected subset and for real (unknown) data we need to use the predictive accuracy as we are not aware of optimal subset. Thus, in classical framework described above we verify results using classifier. This framework can be used for subset search method.

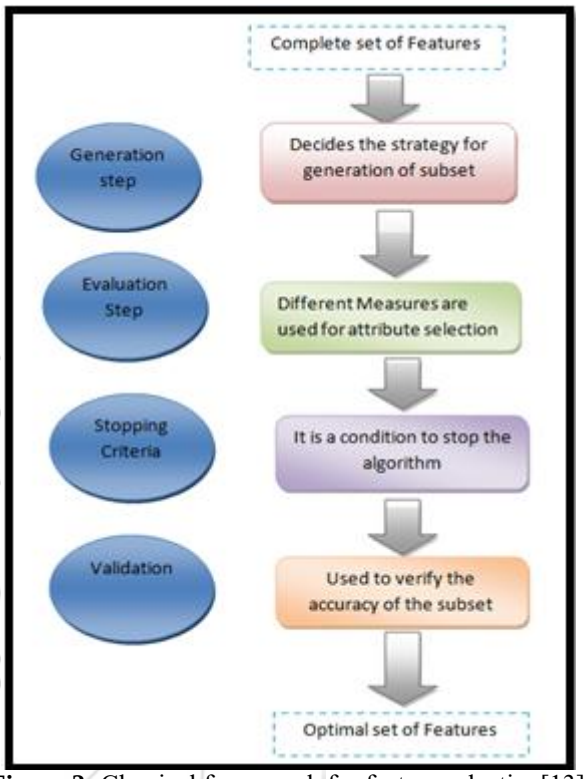


Figure 2: Classical framework for feature selection[13]

2.2 New Framework

The classical framework is a four step process involving generation, evaluation, stopping criteria and validation but the new framework that is widely used now a days is a two step process which involves relevance analysis and redundancy analysis.
Relevance analysis: Attributes which are relevant to our task are selected using different measures as below:
1. Information gain
2. Distance
3. Consistency
4. Classifier error rate
5. Dependency or correlation

Definition 2.1 Distance measure:
Distances between the instances are found out using Euclidean distance, Manhattan distance.
Euclidean distance between the variables i and j can be computed as below:

$$d(i,j)=\sqrt{(x_{i1}-x_{j1})^2+(x_{i2}-x_{j2})^2+..+(x_{in}-x_{jn})^2}$$

Example: - Relief algorithm

Definition 2.2 Consistency measure: If two tuples contain similar data but they belong to different classes then they leads to inconsistency. If the two tuples are exactly similar and they belong to same class then they leads to consistency.

Example:- Las Vegas Filter algorithm

Definition 2.3 Classifier error rate: The classifiers use the error rate measure to rank the feature. It gives accurate results but it is time consuming. It is used in Wrapper model. Consider, if we select Information gain as our measure then features which are above some threshold value are considered as relevant features. Information gain and Symmetric uncertainty are useful for finding relevant features. The Information Gain and correlation based measures are explained in detail in the next section.

Redundancy Analysis: Correlation based measures can be used to remove redundant features. E.g. of these measures are Pearson coefficient (classical linear correlation)Based on Information theory. Thus, this new framework is suitable for Information weighting method.

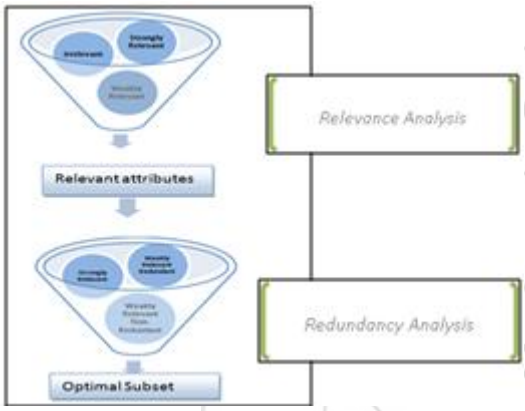


Figure 3: New framework for feature selection [13]

3. Correlation based measures

The correlation based measures can be used to select relevant features or to remove redundant features. Two types of correlation measures are defined below:
1. C-correlation
2. F-correlation

Definition 3.1 C-correlation: It is also known as the *feature - class* correlation. It is used to find the relevant features. The higher the C- correlation value more it is considered to be relevant.

Definition 3.2 F-correlation: It is also known as the *feature - feature* correlation. It is used to find the redundant features. The higher the F- correlation value more it is considered to be redundant. If the F-correlation value of a certain feature is comparative more than its C-correlation value then the feature is considered to be redundant.

Correlation measures: The correlation measures can be furtherclassified into classical linear correlation and information theory based correlation.

Classical linear correlation: This approach has some drawbacks. The linear correlation cannot always be used for feature selection as the features might not be linearly correlated every time. It does not work for nominal attributes. So, we generally use Info theory based correlation measure.

Information theory based correlation: The information theory concept explains that to find the relevant attributes Information gain can be used. Information Gain can be defined, using concept of entropy. Consider a variable X, which takes N values
{Si}i=1 to N

P(Sk) be the probability when X=Sk. Then the information obtained when X= Sk is:

$$I(X) = \log(1/P(X)) = -\log(P(X)) \quad (1)$$

The entropy is a measure of unpredictability. It is the expectation of information required to determine the class label of the tuple. It is average of the information provided by the different values that it might take. The entropy of X can be calculated as below:-

$$E(X) = -\sum_{i=1}^{N} P(x_i) \log_2 P(x_i) = \sum_{i=1}^{N} P(x_i)\, I(x_i) \quad (2)$$

The Entropy of variable X after observing the value of Y can be calculated as below:

$$E(X|Y) = -\sum P(y_i) \sum P(x_i|y_i) \log_2 (P(x_i|y_i)) \quad (3)$$

$$\text{Where } P(x_i|y_j) = P(x_i, y_j)/P(y_j) \quad (4)$$

If the observed values of X in the training data set S are partitioned according to the values of a second feature Y, and the entropy of X with respect to the partitions induced by Y is less than the entropy of X prior to partitioning, then there is a relationship between features X and Y. Given the entropy is a criterion of impurity in a training set S, we can define a measure reflecting additional information about X provided by Y that represents the amount by which the entropy of X decreases .This measure is known as Information Gain or Mutual Information. Below is the equation for calculating the Information gain.

$$IG(X,Y) = E(X) - E(X|Y) = E(X) + E(Y) - E(X, Y) \quad (5)$$

For calculation Symmetric Uncertainty

$$SU(X,Y) = 2\left[\frac{IG(X,Y)}{E(X)+E(Y)}\right] \quad (6)$$

Information gain is biased towards attributes with more number of distinct values. So, it can be normalized using Symmetric uncertainty. The attributes should be more correlated to class and less correlated to each other. It normalizes the values in the range [0, 1]. A value of SU = 1 means one feature completely predicts the other, and SU = 0 indicates, that X and Y are independent

Entropy Correlation Coefficient:

A new measure of dependence called entropy correlation coefficient is denoted by ρ(H) and is given by:

$$\rho(H) = \sqrt{I(X,Y)/\left(\left(\frac{1}{2}\right) * (E(X) + E(Y)\right)}$$

$$= \sqrt{2 * (1 - ((E(X,Y)/(E(X) + E(Y))))}$$

where,

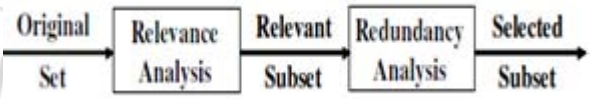
I(X,Y) is the mean dependence information and is denoted by,
I(X,Y)=E(X)+E(Y)-E(X,Y)
E(X)= Entropy value calculated for X variable
E(Y)= Entropy value calculated for Y variable

A value 1 of ρ(H) indicates that knowledge of the value of either one completely predicts the value of the other and the value 0 reveals that X and Y are independent. Thus, this measure can be used to find the inter dependence between the variables.

4. Related Work

By studying different research papers, journals related to this topic provided lot of useful information. Here are some papers that we have referred for literature survey.

There are 6 ranking methods for feature selection which are broadly classified in to two categories:-statistical and entropy-based. There are certain classification algorithms which has the ability to focus on relevant features and ignore irrelevant ones. E.g.;-Decision trees. K-nearest neighbor algorithm uses Feature selection methods to remove noisy features. Entropy-based measures(measure of the system's unpredictability) are Information Gain (IG) attribute evaluation, Gain Ratio (GR) attribute evaluation, Symmetrical Uncertainty (SU) attribute evaluation and Statistical measures are Relief-F (RF) attribute evaluation, One-R (OR) attribute evaluation, Chi-Squared (CS) attribute evaluation.[14] The accuracy of the classifier is influenced by the choice of feature selection techniques and different ranking methods applied on different datasets. [15] The new wrapper method is computationally expensive and does not show significant improvement but found a reduced set of relevant features[16].If feature selection is not performed then the learning algorithm can get confused and also it might become slower. Feature Relevance is classified into 4 categories as strongly relevant features, weakly relevant redundant features, weakly relevant non- redundant features and irrelevant features. So, the Optimal subset=All strongly relevant features+ subset of weakly relevant features (No irrelevant features) .As we don't know which features to be selected from the weakly relevant features redundancy analysis is required. The traditional approach of Feature selection has 4 steps but for efficient Feature Selection they have proposed a new framework as below [17]



Different feature selection algorithms were found during the literature survey. Below table represents a comparison between different algorithms.

Table 1: Existing algorithms feature selection

Algorithm	*Details*
Relief algorithm	It selects relevant Features but it supports only Binary classes, it does not handle redundancy, and user might find difficulty in identifying the proper value for number of samples.
Relief-F algorithm	It selects relevant Features, it solves the problem of supporting only Binary classes, it does not handle redundancy.[11]
FCBF	It handles relevance and redundancy but has been implemented for single table[18]
MR2 algorithm	It has been implemented for Multi Relational Data but uses Pearson's coefficient as a correlation measure which cannot handle Nominal values.
FARS algorithm	It does not perform redundancy analysis. It performs the unnecessary computations for all the attributes like sorting and calculating F-correlation between attributes thus consumes more processing time.
FAST algorithm	It uses the F- correlation measure and computes the independence of features using Symmetric uncertainty. For calculation of symmetric uncertainty we need to calculate information gain which increases the calculation steps. Also the algorithm uses Prim's algorithm to find the minimum spanning tree.[2]

5. Conclusion and Future Scope

To improve the efficiency and accuracy of classifiers, we need to select optimal features from the large features set. Feature selection solves this problem and hence is required for the classifier. In this paper, we have studied different models, frameworks, evaluating measures involved in feature selection and realized that information gain is biased towards attributes with more number of distinct values, symmetric uncertainty is biased towards the attributes with less number of distinct values, and there is a need of measures that can handle all types of values. In Hybrid model it might be possible that the features which are relevant are already removed in the Filter approach. So, even if we go for wrapper those useful features cannot be added. As a part of future scope information gain can be normalized using symmetric uncertainty. Symmetric uncertainty can be normalized using Entropy correlation coefficient. For redundancy analysis we can use functional dependency checking. Functional dependency can be applied on any type of data (continuous or discrete) and it is not biased. New and efficient methods using different generation procedure and different evaluation function can be designed.

References

[1] Almuallim H. and Dietterich T.G., Algorithms for Identifying Relevant Features, In Proceedings of the 9th Canadian Conference on AI, pp 38-45,1992.
[2] Almuallim H. and Dietterich T.G., Learning boolean concepts in the presence of many irrelevant features, Artificial Intelligence, 69(1-2), pp 279-305, 1994.
[3] Arauzo-Azofra A., Benitez J.M. and Castro J.L., A feature set measure based on relief, In Proceedings of the fifth international conference on Recent Advances in Soft Computing, pp 104-109, 2004.

[4] Baker L.D. and McCallum A.K., Distributional clustering of words for text classification, In Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in information Retrieval, pp 96-103, 1998.
[5] Battiti R., Using mutual information for selecting features in supervised
neural net learning, IEEE Transactions on Neural Networks, 5(4), pp 537-550, 1994.
[6] Bell D.A. and Wang, H., A formalism for relevance and its application in feature subset selection, Machine Learning, 41(2), pp 175-195, 2000.
[7] Biesiada J. and Duch W., Features election for high-dimensionaldatała Pearson redundancy based filter, AdvancesinSoftComputing, 45, pp 242-249,2008.
[8] Butterworth R., Piatetsky-Shapiro G. and Simovici D.A., On Feature Selection through Clustering, In Proceedings of the Fifth IEEE international Conference on Data Mining, pp 581-584, 2005.
[9] Cardie, C., Using decision trees to improve case-based learning, In Proceedings of Tenth International Conference on Machine Learning, pp 25-32, 1993.
[10] Chanda P., Cho Y., Zhang A. and Ramanathan M., Mining of Attribute Interactions Using Information Theoretic Metrics, In Proceedings of IEEE international Conference on Data Mining Workshops, pp 350-355, 2009.
[11] Chikhi S. and Benhammada S., ReliefMSS: a variation on a feature ranking ReliefF algorithm. Int. J. Bus. Intell. Data Min. 4(3/4), pp 375-390, 2009.
[12] Cohen W., Fast Effective Rule Induction, In Proc. 12th international Conf. Machine Learning (ICML'95), pp 115-123, 1995.
[13] Vimalkumar B.Vaghela , Tejasvi D. Bhaskarwar,"The Impact of feature selection on performance improvement of classifiers in data mining", IJSTM Volume No.04, Special Issue No.01, February 2015
[14] Novaković, Jasmina, Perica ŠTRBAC, and Dušan Bulatović. "Toward optimal feature selection using ranking methods and classification algorithms." Yugoslav Journal of Operations Research ISSN: 0354-0243 EISSN: 2334-6043 21.1 (2011).
[15] Novakovic, Jasmina. "The Impact of Feature Selection on the Accuracy of Naïve Bayes Classifier." 18th Telecommunications forum TELFOR. 2010.
[16] John, George H., Ron Kohavi, and Karl Pfleger. "Irrelevant Features and the Subset Selection Problem." ICML. Vol. 94. 1994.
[17] Yu, Lei, and Huan Liu. "Efficient feature selection via analysis of relevance and redundancy." The Journal of Machine Learning Research 5 (2004): 1205-1224.
[18] Senliol, Baris, et al. "Fast Correlation Based Filter (FCBF) with a different search strategy." International Symposium on Computer and Information Sciences (ISCIS 2008). 2008.
[19] Verma, Vijay Kumar, and Pradeep Sharma. "Data Dependencies Mining In Database by Removing Equivalent Attributes." International Journal of Computer Science and Engineering 1.1 (2013): 13-16.
[20] Dash, Manoranjan, and Huan Liu. "Consistency-based search in feature selection." Artificial intelligence 151.1 (2003): 155-176.
[21] Yu, Lei, and Huan Liu. "Feature selection for high-dimensional data: A fast correlation-based filter solution." ICML. Vol. 3. 2003.

Author Profile

Pallavi H. Malji, Student of Vishwakarma Institute of Information Technology Pune.

Dr. Sachin S. Sakhare, Head of Computer Department of Vishwakarma Instutute of Information technology Pune.