ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Feature Subset Selection Methods for High Dimensional Data

B.Anitha
M.Tech Student,
Department of CSE,
M.V.S.R.Engineering College,
Hyderabad, R.R Dist.

B.Venkataramana
Assistant Professor,
Department of CSE,
M.V.S.R.Engineering College,
Hyderabad, R.R Dist.

Abstract:

Feature selection involves identifying a subset of most useful features that produces compatible results as the original entire set of feature A feature selection algorithm may be evaluated from both the efficiency and effectiveness point of view . While the efficiency concerns the time required to find a subset of features, the effectiveness is related to the quality of the subset of features. The core idea of feature selection process is improve accuracy level of classifier; reduce dimensionality; speedup the clustering task etc, this paper mainly focuses on comparison of various techniques and algorithms for feature selection process.

Keyword:

Feature Selection, Feature Clustering, Filter method.

1.Feature Selection:

A "feature" or "variable" or "attributes" refers to an aspect of the data .Usually before collection data, features are specified or chosen. Feature can be discrete, continuous, or nominal. Generally, feature are characterized as:

1.Relevant:

These are features which have an influence on the output and their role cannot be assumed by the rest .

2. Irrelevant:

Irrelevant features are defined as those feature not having any influence on the output, and whose values are generated at random for each example.

3. Redundant:

A redundancy exists whenever a feature can take the role of another, (simplest way to model redundancy).

Problem of selecting some subset of a learning algorithms input variables upon which it should focus attention , while ignoring the rest. Feature selection is the process of selecting, selecting the best feature among all the feature because all the features are not useful in constructing the clusters: some feature may be redundant or irrelevant thus not contributing to the learning process. Feature selection is a process commonly used in machine learning, wherein a subset of the features available from the data are selected for application of a learning algorithm. The best subset contains the least number of dimensions that most contribute to accuracy; we discard the remaining, unimportant dimensions. This is an important stage of preprocessing and is one of two ways of avoiding the curse of dimensionality (the other is feature extraction). The main aim of feature selection is to determine a minimal feature subset from a problem domain while retaining a suitably high accuracy in representing the original features, learning from data techniques can benefit. To be completely sure of the attribute selection, we would ideally have to test all the enumerations of attribute subsets, which is infeasible in most cases as it will result in 2n subsets of n attributes. Feature selection has been an active research area in pattern recognition, statics, and data mining communities.

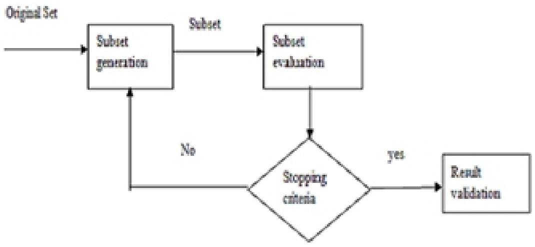


Fig 1.1 : General procedure for feature selection.

1.2 Advantages of Feature selection:

•It dimensionality reduces the Feature space, to limit storage requirements and increase algorithm speed;

•It removes the redundant, irrelevant or noisy data
•The immediate effects for data analysis tasks are speeding up the running time of the learning algorithms.
•Improving the data quality.
•Increasing the accuracy of the resulting model.
• Feature set reduction ,to save resources in the next round of data collection or during utilization;
•Performance improvement, to gain in predictive accuracy;
•Data understanding, to gain knowledge about the process that generated the data or simply visualize the data.

1.3 Algorithms for Feature selection (FSA):

A Feature selection algorithm(FSA) is a computational solution that is motivated by a certain definition of relevance .The purpose of a FSA is to identify relevant feature according to a definition of relevance.

1.3.1 Characterization of FSAs:

There exist in the literature several considerations to characterize feature selection algorithms.

1.3.2 Search Organization:

A search algorithm is responsible for driving the feature selection process using a specific strategy. We consider three types of search: exponential, sequential and random.

1.3.2.1 Generation of Successors:

Mechanism by which possible variants (successor candidates) of the current hypothesis are proposed. Up to five different operators can be considered to generate a successor for each state: Forward, Backward , Compound, Weighting and Random.

1.3.2.2 Evaluation Measure:

Function by which successor candidates are evaluated, allowing to compute different hypothesis to guide the search process. Some of the evaluation measures are probability of error, Divergence, Dependence, interclass distance, information or Uncertainty and consistency.

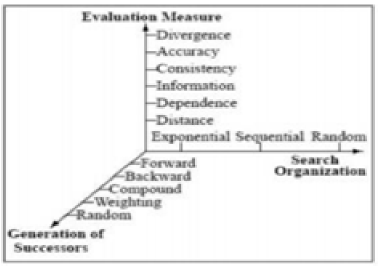


Fig 1.3 Characterization of a FSA

1.4 Approaches:

There are two approaches in Feature selection :

1.Forward selection: Start with no variables and add them one by one, at each step adding the one that decreases the error the most , until any further addition does not significantly decrease the error.

2. Backward selection: Start with all the variables and remove them one by one, at each step removing the one that decreases the error the most(or increases it only slightly),until any further removal increases the error significantly. To reduce overfitting , the error referred to above is the error on a validation set that is distinct from the training set.

1.5 General Schemes For Feature Selection:

Feature selection is similar to data preprocessing technique .it is an approach of identifying subset of features that are mostly related to target model. The main aim is to remove irrelevant and redundant features, it is also known as attribute subset selection. Feature extraction creates new feature from function of the original features. Where as feature selects returns a subset of the feature. Feature selection techniques are often used in domains Where there are many features and comparatively few sample steps in a Feature Selection Method:

•Invention Procedure: Produce candidate subset from original feature set.

•Estimation Function: Estimate the candidate subset.

•Evaluation: Compare with user defined threshold value.

•Verification Method: Test out whether the subset is valid.

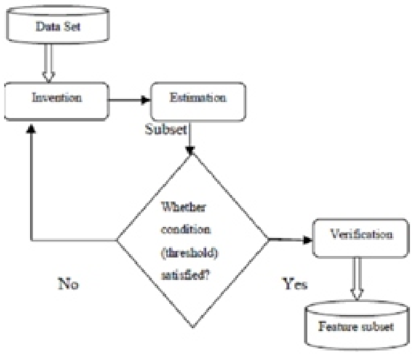


Figure 3.5: Steps involved in feature selection

1.6 Categories of Feature Subset Selection Algorithm:

The feature subset selection algorithm are generally categories into main four categories as Filter, Wrapper, Embedded and Hybrid Method.

Wrapper Method:

Wrapper method use a predictive model to score feature subsets .Each new subset is used to train a model, which is tested on a hold-outset. Counting the number of mistakes ,very computationally intensive.

Filter Method:

Filter method measure is chosen to be fast to compute, and still capturing the usefulness of the feature subset. Common measures include the Mutual Information, Pearson product-moment correlation coefficient, and inter/intra class distance.

Embedded Method :

Embedded method is a catch-all group of techniques which perform feature selection as part of the model construction process. One other popular approach is the Recursive Feature Elimination algorithm, commonly used with Support Vector Machines to repeatedly construct a model and remove feature with low weights .

Hybrid Method:

Hybrid method is combination of filter and wrapper methods. The mainly focus on combination of filter and wrapper methods in accordance to achieve best performance with particular learning algorithm with similar time complexity of the filter methods.

Comparison Of Various Algorithms:

In this section , we present the comparison results in terms of

•Classification accuracy (Accuracy of the selected features)
•Runtime (time to obtain the feature subset)
•Proportion of selected features(ratio of the number of features selected by a feature selection algorithm)

Comparison of Various Algorithm and techniques are discussed as follows:

S.NO	Techniques(or) Algorithms	Advantages	Disadvantages
1	FAST Algorithm	Improve the performance of classifiers	Required more time
2	Consistency Measure	Fast, Remove noisy and irrelevant data	Unable to handle large volume of data
3	Wrapper Approach	Accuracy is high	Computational complexity is large
4	Filter Approach	Suitable for very large features	Accuracy is not guaranteed
5	Agglomerative Linkages Algorithm	Reduce complexity	Decrease the Quality when dimensionality become high
6	INTERCAT Algorithm	Improve Accuracy	Only deal with irrelevant data
7	Distributional clustering	High classification accuracy	Difficult to evaluation
8	Relief Algorithm	Improve efficiency and Reduce Cost	Powerless to detect Redundant features

Table: Comparison of various techniques and algorithms

2. Classification:

The basic classification techniques are Decision tree induction; Bayesian classification, and Rule-based classification task play an important role in clustering process. Classification is performed via following two step process:

1.Model construction:

Describing a set of predetermined classes. Model is represented as classification rules, decision tree and mathematical formulae.

2.Model usage:

It is used to estimate the accuracy of model. Accuracy rate is the percentage of test set sample that are correct classified by the model.

Classification methods:
Bayesian classification:

Bayesian classifiers are statistical classifiers used to predict class membership probabilities. It is also known as naïve Bayesian classifiers based on Bayes theorem. compare to other classifiers it have the minimum error rate.

Decision tree induction:

Decision tree are constructed in a top-down recursive divide-and-conquer method . It consist of three algorithms such as ID3 (Iterative Dichotomiser), C4.5 (successor of ID3),CART(classification and Regerssion tree). The procedure employ an attribute selection measure such as giniindex,information gain and gain ratio. Attribute seiectionmeasure[1] is used to separates the original data set(D)into individual classes.

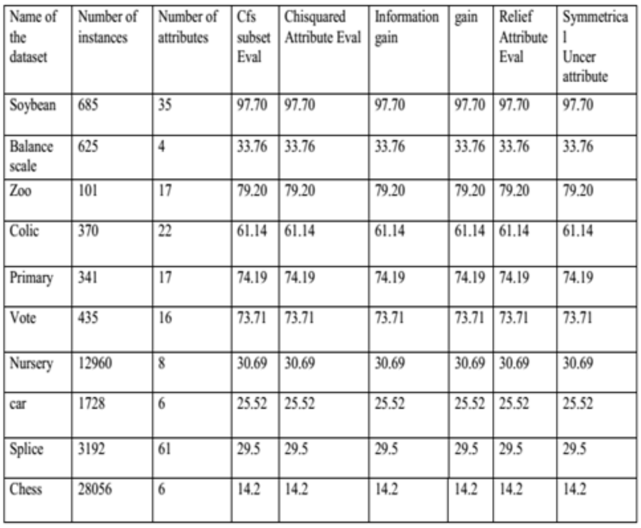
Rule Based Classification:

A rule-based classifiers used a set of rules for classification task. This method effectively produces the subset of features using different heuristic techniques.

3..Result And Analysis:

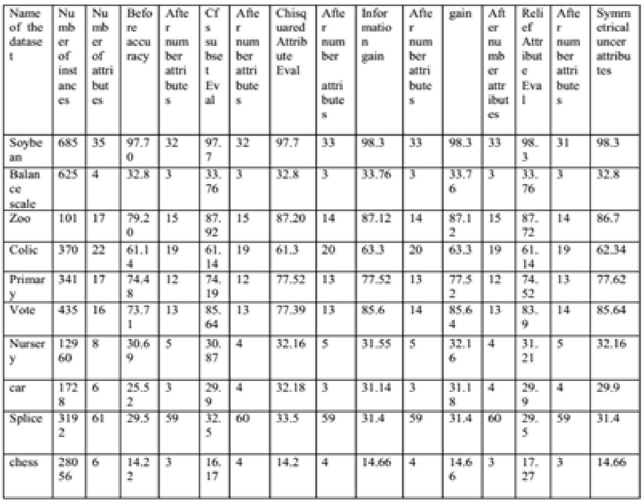
In this section ,we present the experimental results in terms of the proportion of selected features, the time to obtain the feature subset, the classification accuracy. attribute evaluator is basically used for ranking all the features according to some metric, various attribute evaluators are available in WEKA. We used (weka 3.7.8) a learning machine tool this work which include Feature selection is normally done by searching the space of attribute subset evaluating each one . This is achieved by combining attribute subset evaluator with a search method. In the present investigation, an evaluation of six filter feature subset methods with rank search or Greedy search method was performed.the correctly\incorrectly classified instances are defined the case where the instances are used as test data .

Before Accuracy feature selection (J-48 decision tree)



Name of the dataset	Number of instances	Number of attributes	Cfs subset Eval	Chisquared Attribute Eval	Information gain	gain	Relief Attribute Eval	Symmetrical Uncer attribute
Soybean	685	35	97.70	97.70	97.70	97.70	97.70	97.70
Balance scale	625	4	33.76	33.76	33.76	33.76	33.76	33.76
Zoo	101	17	79.20	79.20	79.20	79.20	79.20	79.20
Colic	370	22	61.14	61.14	61.14	61.14	61.14	61.14
Primary	341	17	74.19	74.19	74.19	74.19	74.19	74.19
Vote	435	16	73.71	73.71	73.71	73.71	73.71	73.71
Nursery	12960	8	30.69	30.69	30.69	30.69	30.69	30.69
car	1728	6	25.52	25.52	25.52	25.52	25.52	25.52
Splice	3192	61	29.5	29.5	29.5	29.5	29.5	29.5
Chess	28056	6	14.2	14.2	14.2	14.2	14.2	14.2

Classification Accuracy of feature selection (J-48 decision tree)



Name of the dataset	Number of instances	Number of attributes	Before accuracy	After number attributes	Cfs subset Eval	After number attributes	Chisquared Attribute Eval	After number attributes	Information gain	After number attributes	gain	After number attributes	Relief Attribute Eval	After number attributes	Symmetrical uncer attributes
Soybean	685	35	97.70	32	97.7	32	97.7	33	98.3	33	98.3	33	98.3	31	98.3
Balance scale	625	4	32.8	3	33.76	3	32.8	3	33.76	3	33.76	3	33.76	3	32.8
Zoo	101	17	79.20	15	87.92	15	87.20	14	87.12	14	87.12	15	87.72	14	86.7
Colic	370	22	61.14	19	61.14	19	61.3	20	63.3	20	63.3	19	61.14	19	62.34
Primary	341	17	74.48	12	74.19	12	77.52	13	77.52	13	77.52	12	74.52	13	77.62
Vote	435	16	73.71	13	85.64	13	77.39	13	85.6	14	85.64	13	83.9	14	85.64
Nursery	12960	8	30.69	5	30.87	4	32.16	5	31.55	5	32.16	4	31.21	5	32.16
car	1728	6	25.52	3	29.9	4	32.18	3	31.14	3	31.18	4	29.9	4	29.9
Splice	3192	61	29.5	59	32.5	60	33.5	59	31.4	59	31.4	60	29.5	59	31.4
chess	28056	6	14.22	3	16.17	4	14.2	4	14.66	4	14.66	3	17.27	3	14.66

Minimum Spanning tree:

A minimum spanning tree (MST) is an undirected connected, weighted graph is a spanning tree of minimum weight. A tree is an acyclic graph. the idea is to start with an empty graph and try to add edges one at a time ,the resulting graph is a subset of some minimum spanning tree .Each graph has several spanning trees. this method is mainly used to make the appropriate feature subset clustering.

Modules Description:

1. Removal of Irrelevant features:

feature subset selection methods have been proposed for machine learning applications. The generality of the selected feature is limited and the computational complexity is large.

2. F-Correlation calculation :

F-Correlation is the measure of correlation between the every two attribute of the dataset. High F-Correlation value implies the attribute are highly correlated.

3. MST construction:

A minimum spanning tree is constructed using either Prims or Kruskal's algorithm. After applying any of these algorithms the resulting tree will have n nodes and n-1 edges. In this we adopt Kruskal's algorithm

Conclusion:

In this project different UCI dataset downloaded in weka using classification accuracy finding six types of attribute method applied and filter method used. they are CFS subsetEval(CFS),Chi-SquareAttributeevaluation(CH), Gain-ratio Attribute evaluation(GR), Information-Gain-Attribute evaluation,Relief Attribute evaluation(RF)and Symmetrical Uncertainty Attribute evaluation(SU).The algorithm involes(i)removing irrelevant feature (ii)construcing a minimum spanning tree from relative ones,and (iii) partitioning the MST and selecting representative feature.in the proposed algorithm,a cluster consists of features.Each cluster is treated as a single feature and thus dimensionality is drastically reduced.

References:

[1] Almuallim H. and Dietterich T. G. ,Algorithm 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 concept in the presence of many irrelevant feature, Artificaial Intelligence,69(1-2),pp 279-305, 1994.

[3] Arauzo-Azofra A., Benitez J.M.Castro J.L.,A feature set measure based on relief, in proceeding of the fifth international conference on Recent Advances in soft Computing, pp 104-109,2004.

[4] Baker L.D. and McCallum A.K., Distibutional clustering of wortds for text classification, In proceedings of the 21st Annual international ACM SIGIR conference on Research and Development in information Retrieval,pp-103,1998.

[5] Battiti R., using mutual information for selecting feature in supervised neural net learning, IEEE Transaction 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 175-1905,2000

[7] BiesiadaJ.andDuchW.,m Feature election for high- dimensional data pearson redundancy based filter, Advances in Soft computring,45, 242C249,2008.

[8] Butterworth R., piatetsky-shapiroG.andsimovici D.A., on Feature Selection through Clustering,. In proceedings of the Fifth IEEE international conference on Data Mining, pp 581-584,2005.