Feature Subset Selection and different Algorithms for Feature Selection in Data Mining

Nitin kumar[1]
[1]M. Tech. Student
[1]Computer Science and Engineering
[1]Galgotias University

Abstract---In machine learning and statistics, feature selection plays a very important role as a process for the selection of the relevant features from the pool of features for use in model construction and for the better classification and understanding of the data. There are many irrelevant features or attributes in the data that does not give us any information at all or gives very less information. So it is very much compulsory to have a subset of features that can classify describe the data at its best. In this paper a discussion about Feature subset selection algorithm is done. Discussion about the different methods by which we can extract relevant features or methods for the selection of the best subset of features is done.

I. INTRODUCTION

Feature subset selection is the process of selecting a subset of relevant feature for the construction of a model or better classification and description of the data. The core concept of the feature subset selection technique is that the raw data we are using has many un appropriate, redundant and irrelevant features. Redundant features are those features those provide no information as compared to already selected feature and irrelevant feature can be considered as a feature of no use in context of the information. Feature selection techniques are the subset of the Feature extraction field. Feature extraction makes new features from the attribute set or the original features, whereas feature selection returns a subset of the features.

Feature selection is also employed for the reduction of the dimensions that is we can impose random or some predefined constraints on the number of attributes considered during the construction of a model. We can also take an attribute in consideration as feature as our choice or we can discard an attribute based on the usefulness of that specific attribute for analysis.

Feature selection can be classified as supervised and unsupervised learning techniques. In supervised feature selection technique the main aim is to extract a subset of feature that has a higher accuracy of classification. In unsupervised feature selection technique the main aim is to find a good subset of feature that has high quality of clusters for a given number of clusters.

It is very much worth noticing that which type of feature selection we are applying for the effective analysis, because of the datasets we are taking for analysis gives us far more information then we need to build the model. For example dataset contains 1000 columns that describes the characteristics of a disease a person suffering from, but if some of the attributes are irrelevant or gives us very little information then we can discard them on the other hand more memory and CPU time is needed to complete the model.

The unneeded attributes can degrade the quality of the patterns extracted because of the following reason:

- Some columns are noisy or redundant. This noise makes it more difficult to discover meaningful patterns from the data.
- To discover quality patterns, most data mining algorithms require much larger training data set on high-dimensional data set. But the training data is very small in some data mining applications.

II. HOW THE FEATURE SELECTION WORKS?

In general feature selection finds out a score for each attribute in the data and then selects only that attribute which has the best score. We can also adjust the threshold for the top the top score. There are many methods for calculating these scores, and the method applied depends upon the following factors:

- The algorithm used in your model
- The data type of the attribute
- Any parameters that you may have set on your model

Feature selection is applied to inputs, already known or predicted attributes. Now each attribute is scored for feature selection and as soon as the scoring is done only those attributes are used which are selected by algorithm and are used for building the model and for prediction. If we chose the attribute that does not satisfy the threshold then it can be used for building the model and prediction, but the prediction will be based solely on the global statistics that exist in the model.

A. Definition of Feature Selection Methods for Attribute Scoring

There are many well-established and popular methods for scoring attributes. The method that is applied in any algorithm or data set depends on the data types, and the attribute used.

The *interestingness* score provided by SQL Analysis services is used to sort the attributes in columns that contain non binary data.

Shannon's entropy and *two Bayesian* scores are available for the discrete data. However, if the model contains any continuous columns, the interestingness score will be used to assess all input columns.

The following section describes each method attribute scoring:

- Interestingness score:

Interestingness score describes that how much the feature is interesting that is how much useful information it is giving to us. But the usefulness of that feature depends upon scenario. So data mining industry has developed many methods to calculate *interestingness.*

The interestingness of the attribute is entropy-based, meaning that attributes that are randomly distributed has higher entropy and less information gain, so attributes like this are less interesting.

Central entropy, m means the entropy of the entire feature set, when we subtract the entropy of the target attribute with that of the m, we can deduce out that how much information the attribute provides.

- Shannon’s Entropy:

It measures the uncertainty of a random variable for a particular outcome.

Analysis services use the following formula to calculate Shannon’s entropy:

$H(X) = -\sum p(x_i) \log (p(x_i))$

- Bayesian with k2 prior:

A Bayesian network is a directed or acyclic graph of states and transitions between states that mean that some states have priority more than other states, some states are posterior, and the graph does not repeat. By this discussion we can say that we require prior knowledge in Bayesian networks. However, which prior states to use in calculating probabilities of later states are important for algorithm design.

- Bayesian Dirichlet Equivalent with Uniform prior:

The Bayesian Dirichlet Equivalent with uniform prior method assumes a special case of Dirichlet distribution, in which a constant is used for construct a fixed or uniform distribution of prior states. The BDE score also assumes likelihood equivalence.

III. TABLES SHOWING THE USE OF ATTRIBUTE SCORING METHODS

ALGORITHM	METHOD OF ANALYSIS
Decision Tree	Interestingness score Shannon’s Entropy Bayesian with K2 Prior Bayesian Drichilet with uniform prior
Neural Network	Interestingness Score Shannon’s Entropy Bayesian Drichlet with uniform prior
Clustering	Interestingness
Linear regression	Interestingness score
Association rules	Not used
Logistic regression	Interestingness Score Shannon’s Entropy Bayesian With K2 Prior Bayesian Dirichlet with uniform prior

Table. 1: Use of Attribute Scoring Methods

IV. FEATURE SUBSET SELECTION USING GEOMETRIC ALGORITHM

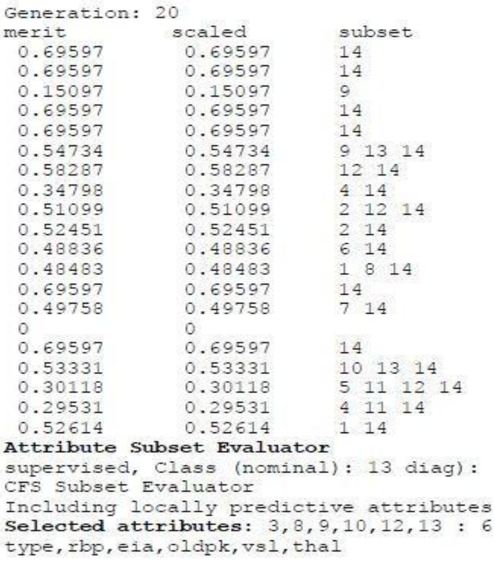
As we know that feature extraction is the process of detecting and eliminating Irrelevant, weakly relevant or redundant attributes or dimensions in a given data set. The goal of feature selection is to find the minimal subset of attributes such that the resulting probability distribution of data classes is close to original distribution obtained using all attributes.

For a dataset of size D, with n attributes, 2^n subsets are possible. So search for an optimal subset would be highly expensive. The problem increases when the value n and the number of classes increase. So sometimes feature selection technique are heuristic methods. These heuristic techniques are greedy in nature that tries to analyze the reduced search space. As we know that the feature selection technique contains two steps for selecting best possible subset of features. First one is called as ranking method or scoring attribute method and the second one is feature subset selection method. In the first step features are given a rank or score by metric like information gain, chi square test etc.

The Feature that does not accumulate the desired or appropriate score is eliminated. In the second step the search is for optimal subset of features that would be equivalent to original subset of features. The Subset of features is evaluated using distance based on the distance measuring techniques like Euclidean distance, Hamming etc. or filter metrics like probabilistic distance. Some of the commonly used feature subset selection is greedy forward selection, backward attribute selection, and genetic algorithms.

Genetic algorithm uses the concept of natural evolution methodology that describes that how population evolved around. The genetic search starts with zero attributes, and an initial population with randomly generated rules. On the basis of the idea of the survival of the fittest, new population is constructed to comply the best fittest rule in the current population. The process of generation continues until a population P is generated where every rule in p satisfies the fitness threshold. Fr more understanding let us take an example of the data related to the heart disease. With initial population of 20 instances, generation continues until the twentieth generation with cross over probability of .6 and mutation probability of .033. The genetic search gives six attributes in result satisfying the fittest rule.

The Attribute subset evaluator uses the following info:



```
Generation: 20
merit            scaled           subset
  0.69597          0.69597         14
  0.69597          0.69597         14
  0.15097          0.15097         9
  0.69597          0.69597         14
  0.69597          0.69597         14
  0.54734          0.54734         9 13 14
  0.58287          0.58287         12 14
  0.34798          0.34798         4 14
  0.51099          0.51099         2 12 14
  0.52451          0.52451         2 14
  0.48836          0.48836         6 14
  0.48483          0.48483         1 8 14
  0.69597          0.69597         14
  0.49758          0.49758         7 14
  0                0
  0.69597          0.69597         14
  0.53331          0.53331         10 13 14
  0.30118          0.30118         5 11 12 14
  0.29531          0.29531         4 11 14
  0.52614          0.52614         1 14
Attribute Subset Evaluator
supervised, Class (nominal): 13 diag):
CFS Subset Evaluator
Including locally predictive attributes
Selected attributes: 3,8,9,10,12,13 : 6
type,rbp,eia,oldpk,vsl,thal
```

Fig. 1 : system design of a feature subset selection using GA

V. Feature Subset Selection using Classifiers

Classification is a supervised learning method that extracts models describing important data classes or the future trends. Classification methods are generally used in machine learning, pattern recognition and artificial intelligence. Classification methods have many applications which includes risk analysis, credit card fraud detection, target marketing, manufacturing and medical diagnosis. We can use three classifiers Decision Tree, Naïve Bayes and classification by using clustering to diagnosis for the selection of the features.

VI. Feature Subset Selection using Naïve Bayes Method

It is a statistical classifier that assumes no dependency between the attributes of the data. It works to maximize the posterior probability in determining the class. Theoretically it has minimum rate but that is not always the case. However, in accuracies are due to the class conditional independence and the lack of variable of availability probability data.

VII. Feature Subset Selections using the Decision Tree

Decision tree is a commonly used classifier that is used for classifying the data. It can also be used for the construction of the trees containing attributes as nodes and then counting the branches iteratively for the selection of the best subset of features. In this process we first of select the specific branch from the tree that traces some of the attributes and then the selected attributes are evaluated or scored for their ability to construct the model in a best possible way.

The performance of the decision tree can be increased with the suitable attribute selection. Correct selection of the branch or the corresponding attributes will give us a best feature subset for the construction of the model.

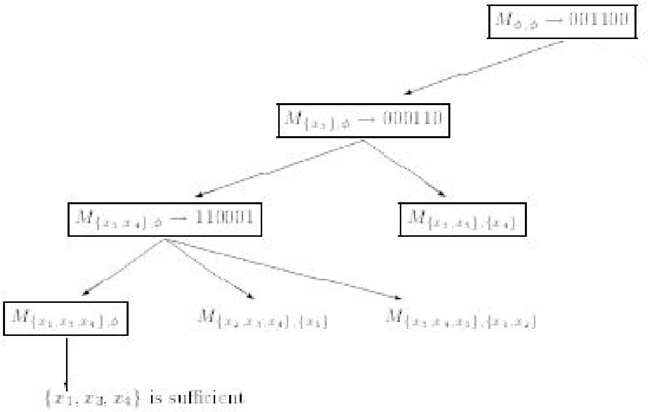

$M_{\phi,\phi} \rightarrow 001100$
$M_{\{x_3\},\phi} \rightarrow 000110$
$M_{\{x_3,x_4\},\phi} \rightarrow 110001$
$M_{\{x_3,x_1\},\{x_4\}}$
$M_{\{x_1,x_3,x_4\},\phi}$
$M_{\{x_2,x_3,x_4\},\{x_1\}}$
$M_{\{x_3,x_4,x_5\},\{x_1,x_2\}}$
$\{x_1, x_3, x_4\}$ is sufficient

Fig. 2: Tracing the branches of a decision tree

VIII. Feature Subset Selection Clustering

Clustering as we know is the process of collecting or grouping the similar type of objects or attribute together in one group. Feature set is first of all clustered by using a appropriate clustering technique and then clustering is repeated until we get a cluster with attribute with maximum similarity within the attributes with in that particular cluster and the minimum similarity between the other different clusters.

IX. Conclusion

Feature subset selection is the process of selecting the most appropriate feature subset by using different types of methods being discussed above the attributes are to be selected with more precision so that we can find a best subset of the features for the construction of the model. For this best selection many different methods have been implemented depending upon the type of data and attributes.

References

[1] Asha Rajkumar and Mrs. G.Sophia Reena (2010): Diagnosis Of Heart Disease Using Datamining Algorithm, GJCST,Vol. 10 Issue 10 Ver. 1.0 Sep2010, pp. 38-43.

[2] http://technet.microsoft.com/enus/library/ms175382.aspx

[3] Bressan, M. and J. Vitria (2003): On the selection and classification of independent features, Pattern Analysis and Machine Intelligence, IEEE Transactions. pp. 1312-1317.

[4] Shantakumar B.Patil and Y.S.Kumaraswamy (2009): Intelligent and Effective Heart Attack Prediction System Using Data Mining and Artificial Neural Network, European Journal of Scientific Research ISSN 1450- 216X Vol.31 No.4, pp. 642-656.

[5] Chen J and Greiner R (1999): Comparing Bayesian Network Classifiers. In Proc. of UAI-99, pp. 101–108.