Extraction of Best Attribute Subset using Kruskal's Algorithm

Sonam R. Yadav
Department of Computer Engineering
Pune

Ravi P. Patki
Department of IT Engineering
Pune

ABSTRACT
Data mining is the technique by which one can extract efficient and effective data from huge amount of raw data. There are various techniques for extracting useful data. Attribute Selection is one of the effective methods for this operation. To obtain useful data from enormous amount of data is not a simple task. It contains several phases like pre-processing, classification, analysis etc.

Attribute Selection is an important topic in Data Mining, as it is the effective way to reduce dimensionality, to remove irrelevant data, to remove redundant data, & to increase accuracy of the data. It is the process of identifying a subset of the most useful attributes that produces attuned results as the original entire set of attribute.

In last few years, there were different techniques for attribute selection. These techniques were judged on two measures i.e. efficiency is time to find the clusters and effectiveness is quality of subset attributes. Some of the techniques are Wrapper approach, Filter Approach, Relief Algorithm, Distributional clustering etc. But each of one having some drawbacks like unable to handle large volumes of data, computational complexity, accuracy is not guaranteed, difficult to evaluate and redundancy detection etc.

To overcome some of these problems in attribute selection method this paper proposes technique that aims to provide an effective clustering based attribute selection method for high dimensional data. Initially this technique removes irrelevant attributes depending on some threshold value. Afterwards, using Kruskal's algorithm minimum spanning tree is constructed from these attributes. From that tree some representative attributes are selected by partitioning. This is nothing but the final set of attributes.

General Terms
Data Mining, Clustering.

Keywords
Attribute Selection, Clustering, Data Mining, Graph-Based Clustering, Minimum Spanning Tree.

1. INTRODUCTION
Attribute selection is the process of identifying a subset of the most useful features that produces compatible results as the original entire set of features. With the point of picking a subset of good attributes as for the target ideas, attribute subset selection is a powerful path for lessening dimensionality, evacuating insignificant information, expanding learning accuracy, furthermore, enhancing result comprehensibility [1], [4]. Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same cluster are more similar to each other than to those in other clusters. It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics Many attribute subset selection methods have been proposed. They can be isolated into four general categories: the Embedded, Wrapper, Filter, and Hybrid methodologies. The embedded methods consolidate attribute selection as a piece of the training process and are typically particular to given learning algorithms, and consequently might be more proficient than the other three categories [5].Traditional machine learning algorithms like decision trees or artificial neural networks are cases of embedded approaches[2]. The wrapper systems utilize the prescient exactness of predefined learning algorithms to focus the integrity of the chose subsets; the precision of the learning algorithms is generally high. However, the consensus of the selected attributes is restricted and the computational many-sided quality is huge. The filter methods are free of learning algorithms with great consensus. Their computational many-sided quality is low, yet the exactness of the learning algorithms is not ensured [6], [7], [8]. Concerning the filter attribute selection methods, the use of clusters examination has been shown tobe more viable than traditional attribute selection algorithms. Pereira et al. [9], Baker et al. [4], and Dhillon et al. [10] utilized the distributional grouping of cluster to decrease the dimensionality of content text data. In cluster analysis, graph-theoretic systems have been decently mulled over and utilized as a part of numerous applications. Their outcomes sometimes have the best concurrence with human performance [11]. The general graph-theoretic clustering is basic: Compute an area graph of instances, at that point delete of any edge in the diagram that is much longer/shorter than its neighbours. The result is a backwoods and every tree forest represents a cluster. In our study, we apply graph theoretic clustering methods to attributes. Specifically, we embrace the minimum spanning tree(MST) based grouping algorithms.

An attribute selection algorithm proposed in this paper can be seen as the combination of a search technique for proposing new feature subsets along with an efficiency and effectiveness. The simplest step is to test each possible subset of attributes finding the one which increases the accuracy.

The proposed algorithm consists of three parts:

(i) removing irrelevant attributes
(ii) constructing a MST from relative one
(iii) Partitioning the MST and selecting representative attributes.

2. BASIC CONCEPTS
In this section, a brief introduction about the basic concepts of the Data Mining, Clustering, Minimum Spanning tree, Prim's Algorithm and Kruskal's Algorithm is provided.

2.1 Data Mining:

Data mining is the way of discovering the interesting knowledge from large amounts of information sources or data warehouses. When there is huge amount of data and certain information is to be found out from that data, then different stages of data mining is applied to it to gain information from it. Data mining tasks classified into two forms: 1. Descriptive mining tasks: Represent the general properties of the data. 2. Predictive mining tasks: Perform the implication on the current data.

Different Data mining Functionalities are: Characterization and Discrimination, Mining Frequent Patterns, Association and Correlations, Classification and Prediction, Cluster Analysis, Outlier Analysis, Evolution Analysis. Out of these functionalities this paper focuses on the cluster Analysis.

2.2 Cluster Analysis:

Clustering is the grouping similar objects into one class. A cluster is an association of data objects that are similar to one another within the same cluster and are dissimilar to the objects in different clusters. Document clustering (Text clustering) is closely related to the concept of data clustering. Document clustering is a more exact technique for unsupervised document organization, automatic topic extraction and fast information retrieval or filtering. Clustering helps to reduce the dimension and it simplifies the task as number of dataset is minimized to form a cluster.

2.3 Minimum Spanning Tree:

A minimum spanning tree (MST) is an undirected, connected, acyclic weighted graph with minimum weight. 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 attribute subset clustering but it take time to construct the cluster.

Various Applications:

- Design of computer networks and Telecommunications networks
- Transportation networks, water supply networks, and electrical grids.
- Cluster analysis
- Constructing trees for broadcasting in computer networks
- Image registration and segmentation

2.4 Prim's Algorithm

In computer science, Prim's algorithm is a greedy algorithm that finds a minimum spanning tree for a weighted undirected graph. This means it finds a subset of the edges that forms a tree that includes every vertex, where the total weight of all the edges in the tree is minimized. The algorithm operates by building this tree one vertex at a time, from an arbitrary starting vertex, at each step adding the cheapest possible connection from the tree to another vertex.

2.5 Kruskal's Algorithm

Kruskal's algorithm is a minimum-spanning-tree algorithm where the algorithm finds an edge of the least possible weight that connects any two trees in the forest. It is a greedy algorithm in graph theory as it finds a minimum spanning tree for a connected weighted graph at each step. This means it finds a subset of the edges that forms a tree that includes every vertex, where the total weight of all the edges in the tree is minimized. If the graph is not connected, then it finds a minimum spanning forest (a minimum spanning tree for each connected component).

3. BACKGROUND AND COMPARATIVE ANALYSIS

Many algorithms & techniques have been proposed up till now for clustering based feature/ attribute selection. Some of them have focused on minimizing redundant data set and to improve the accuracy whereas some other features subset selection algorithm focuses on searching for relevant features.

In paper [1], Attribute subset selection includes recognizing a subset of the most helpful attributes that delivers perfect results as the first whole arrangement of attributes. An attribute selection algorithm calculation may be assessed from both the efficiency and effectiveness perspectives. While the efficiency concerns the time needed to discover a subset of attributes, the adequacy is identified with the nature of the subset of attributes. Current existing algorithms for attributes subset choice works just in view of directing factual test like Pearson test or symmetric vulnerability test to discover the connection between the highlights and apply edge to channel repetitive and superfluous attributes (Quick calculation employments symmetric instability test for attributes subset determination). In this work, the FAST algorithm works on the Shared data and maximal data coefficient to enhance the efficiency and effectiveness of the attributes subset choice.

In paper [2], Clustering which tries to gathering an arrangement of points into cluster such that points in the same group are more comparable to one another than points in distinctive cluster, under a specific likeness metric. In the generative clustering model, a parametric type of information era is accepted, and the objective in the most extreme probability definition is to discover the parameters that expand the likelihood of generation of the data. In the most general definition, the number of group's k is additionally thought to be an obscure parameter. Such a clustering definition is known as a "model selection" framework, since it needs to pick the best estimation of k under which the grouping model fits the information. In grouping procedure, semi-supervised learning is a class of machine learning systems that make utilization of both marked and unlabelled information for preparing – commonly a little measure of marked information with a lot of unlabelled information. Semi-supervised learning falls between unsupervised learning (with no marked preparing information) and regulated learning. While the proficiency concerns the time needed to discover a subset of attributes, the viability is identified with the quality of the subset of attributes. Traditional approaches for clustering information are in view of metric similarities ,i.e., non-negative, symmetric, and satisfying the triangle inequality measures using graph-based algorithm to supplant this process a later approaches, in the same way as Affinity Propagation (AP) algorithms can be chosen furthermore take enter as general non metric likenesses.

In paper [3], Clustering is the progression of grouping similar objects into one class. It is the movement of collection comparable articles into one class. A cluster is a gathering of information protests that are like each other inside the in distinguishable group and are unlike the articles in different groups. Archive grouping (Text cluster) is nearly identified with the idea of information grouping. Archive bunching is a more particular system for unsupervised record association, programmed theme extraction and quick data recovery or filtering. Data pre-processing is used to improve the efficiency and ease of the mining process. At whatever point

we need to concentrate some information from the information distribution centre that information may be deficient, conflicting or contain boisterous in light of the fact that information stockroom gather and store the information from different outside assets.

In paper [5], Attribute Selection through Clustering introduces an algorithm for attribute selection that clusters attributes using a special metric. Progressive algorithms create groups that are set in a clusters tree, which is generally known as a dendrogram. Clustering are gotten by separating those clusters that are arranged at a given tallness in this tree. It utilize a few information sets from the UCI dataset archive and, because of space confinements we examine just the outcomes got with the votes and zoo data sets, Bayes algorithms of the WEKA bundle were utilized for developing classifiers on information sets got by anticipating the introductory information sets on the arrangements of agent traits. Way to deal with quality choice is the likelihood of the supervision of the procedure permitting the client to select between semi comparable properties It confront arrangement issues that include a huge number of attribute and moderately couple of illustrations went to the fore. We expect to apply our techniques to this kind of data.

Attribute subset selection can be seen as the procedure of distinguishing and evacuating the same number of irrelevant and redundant attributes as could be expected under the circumstances. Selection of attribute subset is a solid route for dimensionality diminishment, elimination of inappropriate data, rising learning exactness, and recouping result un - ambiguousness. Attribute subset selection can be dissected as the methodology of perceiving and wiping out the same number of unseemly and excess highlights as encouraging since: improper attribute don't put into the predictive accurateness and redundant characteristics don't redound to getting an improved indicator for that they make accessible primarily data which is by presently exhibit in past highlight. We develop a novel calculation that can competently and effectively manage both improper and repetitive qualities, and get hold of a predominant attribute subnet.

Table 1. Comparison of Previous Techniques

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

4. PROPOSED SYSTEM

This section describes the used methodology to identify the attribute subset which is main aim of attribute selection method. This methodology works in the same way as that of previous attribute selection method. But the replaced part is in this methodology minimum spanning tree is generated using Kruskal's Algorithm instead of Prim's Algorithm. Chosen this algorithm on the basis that as the number of nodes increases, Kruskal's Algorithm performs better than Prim's Algorithm. Used methodology works in different phases organised in pipelined fashion as follows.

1. Irrelevant Attributes Removal:

Irrelevant attributes, along with redundant attributes, severely affect the accuracy of the learning machines. Thus, attribute subset selection should be able to identify and remove as much of the irrelevant and redundant information as possible.

a. Information Gain Computation

Relevant attributes have strong correlation with target concept so are always necessary for a best subset, while redundant attributes are not because their values are completely correlated with each other. Thus, notions of attribute redundancy and attribute relevance are normally in terms of attribute correlation and attribute-target concept correlation.

To find the relevance of each attribute with the class label, Information gain is computed in this module. This is also said to be Mutual Information measure. Mutual information measures how much the distribution of the attribute values and target classes differ from statistical independence. This is a nonlinear estimation of correlation between attribute values or attribute values and target classes.

The symmetric uncertainty (SU) is derived from the mutual information by normalizing it to the entropies of attributes values and target classes, and has been used to evaluate the goodness of attributes for classification.

The symmetric uncertainty is defined as follows:

$$Gain\ X\ Y = H(X) - H\ (X|Y)$$
$$= H(Y) - H(Y \mid X)$$

To calculate gain, we need to find the entropy and conditional entropy values. The equations for that are given below:

$$H\ (X) = -\sum_{x \in X} p\ (x)\ log_2(x)$$

Where p(x) is the probability density function and p (x|y) is the conditional probability density function.

b. T-Relevance Calculation: The relevance between the attribute Fi € F and the target concept C is referred to as the T-Relevance of Fi and C, and denoted by SU(Fi,C). If SU(Fi,C) is greater than a predetermined threshold, we say that Fi is a strong T-Relevance attribute.

$$SU\ (X,) = 2 \times Gain\ (X|Y)/(X)+H(Y)$$

After finding the relevance value, the redundant attributes will be removed with respect to the threshold value

2. F-Correlation Calculation: The correlation between any pair of attributes Fi and Fj (Fi,Fj € ^ F ^ i ≠ j) is called the F-Correlation of Fi and Fj, and denoted by SU(Fi, Fj). The equation symmetric uncertainty which is used for finding the relevance between the attribute and the class is again applied to find the similarity between two attributes with respect to each label.

3. MST Construction: With the F-Correlation value computed above, the graph is constructed. For that, we use Kruskal's algorithm which form MST effectively.

Description:

1. Create a forest F (a set of trees), where each vertex in the graph is a separate tree.

2. Create a set S containing all the edges in the graph

3. While S is nonempty and F is not yet spanning

Remove an edge with minimum weight from S, If that edge connects two different trees, then add it to the forest, combining two trees into a single tree, Otherwise discard that edge. At the termination of the algorithm, the forest forms a minimum spanning forest of the graph. If the graph is connected, the forest has a single component and forms a minimum spanning tree.

Algorithm
inputs: D(A1,A2, ..., Am, C) - the given data set θ- the T-Relevance threshold.
output: S - selected attribute subset

//==== Part 1 : Irrelevant Attribute Removal ==== //
for element od given data seti = 1 to m **do**
find the T-relevance
if T-Relevance > θ
do add it to pair of attributes set
//==== Part 2: Minimum Spanning Tree Construction ====//
G = NULL; //G is a complete graph
for each pair of attributes {A′i, A′j} ⊂ S **do**
F-Correlation = SU (A′i, A′j) **a**dd the edge to the tree //as per the weight of corresponding tree
minSpanTree = KRUSKALS(G); //KRUSKALS Algorithm to generate the minimum spanning tree

//==== Part 3: Tree Partition and Representative Attribute Selection ==== //
Forest = minSpanTree
for each edge ∈Forest **do**
if SU(A′i, A′j) <SU(A′i, C) ∧SU(A′i, A′j) <SU(A′j,C) **then**
Forest = Forest − Eij **//remove the edge**
S = ϕ
for each tree ∈Forest **do**
find the strongest attribute set
S = S ∪ {AjR};
Return S

5. SYSTEM WORKFLOW

Take the input as a dataset consisting of n number of data. Each data is having n number of attributes. Firstly, information gain is to be compute which is then forwarded to T-Relevance calculation. This whole part is nothing but Irrelevant Attribute Removal. Then from the generated result graph is built up and F-Correlation is calculated for each node with all the connected nodes. This is helpful for finding minimum spanning tree generation. This all part covers under MST construction. Once MST is generated, from that attributes are selected depending on the threshold value. And those representative attributes are selected as a final set of Attributes.

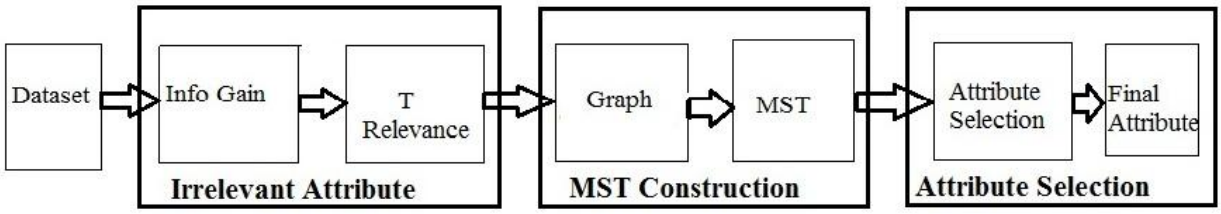
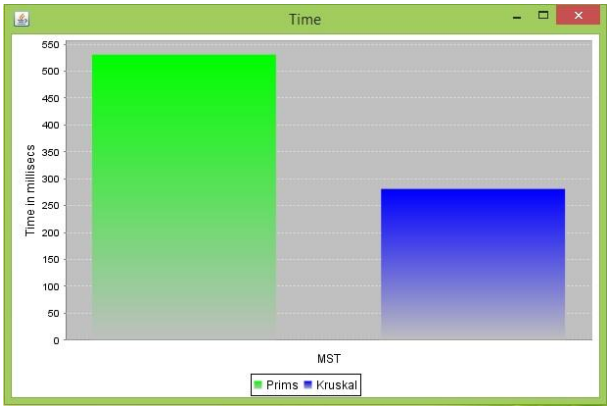


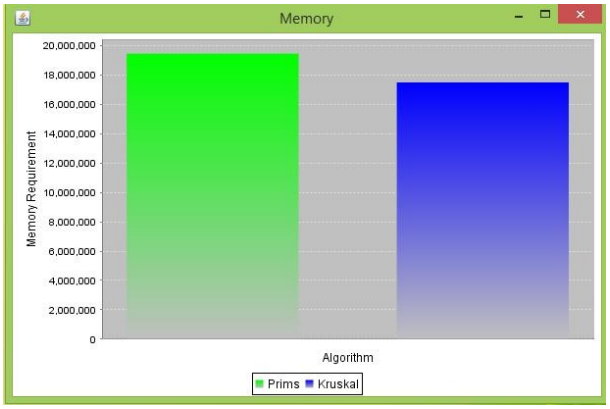
Fig 1. Workflow of System

6. RESULTS

The proposed algorithm is tested against different type and size of data. Comparison between Prim's Algorithm and Kruskal's Algorithm is shown in graph for KDD dataset.



The above graph shows the execution time difference between Prim's and Kruskal's algorithm.



Memory requirement graph is shown above.

From both the graph it is clear that using Kruskal's Algorithm execution time shortens and memory requirement is less.

7. CONCLUSION

In this paper, a hybrid three-phased attribute selection method is proposed. This method takes advantages of mixing FAST algorithm and Kruskal's Algorithm. The first phase analyses relevant attributes to take the best result in the second phase. For the second phase, minimum spanning tree select reliable attributes with a lower cost and higher accuracy. Different performance evaluation parameters are defined and calculated i.e time and memory. The results show that our proposed method outperforms other attribute selection methods on different datasets with different sizes. Furthermore, the proposed method improves the time and memory parameter of FAST algorithm. As FAST is better algorithm of all other. The improvement in FAST i.e. proposed algorithm is also giving the best result.

For future work, one can plan to explore different types of correlation measures to find out the result; as correlation between the attributes is the major thing in this method.

8. REFERENCES

[1] Liu H., Motoda H. and Yu L., Selective sampling approach to active attribute selection, Artif. Intell, 159(1-2), pp 49 -74 (2004)

[2] Modrzejewski M., Attribute selection using rough sets theory, In Proceedings of the European Conference on Machine Learning, pp 213-226, 1993.

[3] Molina L.C., Belanche L. and Nebot A., Attribute selection algorithms: A survey and experimental evaluation, in Proc. IEEE Int. Conf. Data Mining, pp 306-313, 2002.

[4] Guyon I. and Elisseeff A., An introduction to variable and attribute selection, Journal of Machine Learning Research, 3, pp 1157-1182, 2003.

[5] Dash M. and Liu H., Attribute Selection for Classification, Intelligent Data Analysis, 1(3), pp 131156, 1997.

[6] Pereira F., Tishby N. and Lee L., Distributional clustering of English words, In Proceedings of the 31st Annual Meeting on Association For Computational Linguistics, pp 183-190, 1993.

[7] Dhillon I.S., Mallela S. and Kumar R., A divisive information theoretic attribute clustering algorithm for text classification, J. Mach. Learn. Res., 3,pp 12651287, 2003.

[8] N.Magendiran and J.Jayaranjani, An Efficient Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data - (*ICETS'14*)

[9] Mr. M. Senthil Kumar and Ms. V. Latha Jothi, A Fast Clustering Based Feature Subset Selection Using Affinity Propagation Algorithm - (ICGICT'14)

[10] T.Jaga Priya Vathana, C. Saravanabhavan, and Dr.J. Vellingiri, A Survey On Feature Selection Algorithm For high Dimensional Data Using Fuzzy Logic - (IJES)

[11] R. Butterworth, G. Piatetsky-Shapiro, and D.A. Simovici, "On Feature Selection through Clustering," Proc. IEEE Fifth Int'l Conf.Data Mining, pp. 581-584, 2005

[12] A.Arauzo-Azofra, J.M. Benitez, and J.L. Castro, A Feature Set Measure Based on Relief, Proc. Fifth Int"l Conf. Recent Advances in Soft Computing, pp. 104-109, 2004s

[13] Saurabh Soni & Pratik Patel, "IFSS – An Improved Filter-Wrapper Algorithm for Feature Subset Selection", International Journal of Computer Application (0975-8887), Volume 95-No. 14, June 2014.