CLUSTERING BASED ATTRIBUTE SUBSET SELECTION USING FAST ALGORITHM

[1] Suresh Laxman Ushalwar, [2] M.B. Nagori

[1]M.E. Student, Dept of CSE, Government College of Engineering, Aurangabad, Aurangabad (Dist), Maharashtra, India
[2] Assistant Professor, Department of CSE, Government College of Engineering, Aurangabad, Aurangabad (Dist), Maharashtra, India

ABSTRACT

In machine learning and data mining, attribute select is the practice of selecting a subset of most consequential attributes for utilize in model construction. Using an attribute select method is that the data encloses many redundant or extraneous attributes. Where redundant attributes are those which supply no supplemental information than the presently selected attributes, and impertinent attributes offer no valuable information in any context.

Among characteristics discovering a subset is the most valuable characteristics that manufacture companionable outcomes as the unique entire set of characteristics. An attribute select algorithm may be expected from efficiency as well as efficacy perspectives. In the proposed work, a FAST algorithm is proposed predicated on these principles. FAST algorithm has sundry steps. In the first step, attributes are divided into clusters by designates of graph-theoretic clustering methods. In the next step, the most representative attribute that is robustly cognate to target classes is selected from every cluster to make a subset of most germane Attributes. Adscititiously, we utilize Prim's algorithm for managing immensely colossal data set with efficacious time involution. Our proposed algorithm adscititiously deals with the Attribute interaction which is essential for efficacious attribute select. The majority of the subsisting algorithms only focus on handling impertinent and redundant attributes. As a result, simply a lesser number of discriminative attributes are selected.

We are going to compare the performance of the proposed algorithm; it will obtain the best proportion of selected features, the supreme runtime and the good relegation precision. FAST obtains the good rank for microarray data, text data, and image data .By analyzing the efficiency of the proposed work and the subsisting work, the time taken to instauration the data will be more preponderant in the proposed by abstracting all the impertinent features. It provides privacy for data and reduces the dimensionality of the data.

KEYWORDS

Attribute Selection, Subset Selection, Redundancy, Finer Cluster, and Graph-Predicated Clustering.

1. INTRODUCTION

Data mining (the analysis step of the "Knowledge Discovery in Databases" process(KDD)),is the practice of mine samples in astronomically immense data sets involving schemes at the intersection of artificial perspicacity, machine learning performances, statistics, and database systems concepts. Data mining contains six routine modules of tasks those are can be kenned as Anomaly Recognition, Association rule mining,

DOI: 10.5121/ijci.2015.4220

clustering, relegation, summarization, and regression. Whereas clustering is cognate to chore of determining groups and structures in the data that are in some way or another " cognate ", without utilizing kenned structures in the data.

Aim of selecting a subset of excellent characteristics with reverence to the aiming concepts, characteristic subset select is an efficient way for decrementing dimensionality, eliminating immaterial data, growing learning exactness, and civilizing result un-equivocalness. A lot of characteristic subset selecting methods have been proposed and machine learning applications premeditated [3]. They can be dissevered into 4 broad types the combination of Hybrid (mixture), Filter, Wrapper and Embedded approaches. The central conception of this work is to introduce an algorithm for feature select that clusters attributes utilizing a special metric and, then utilizes a hierarchical clustering for feature select. We propose a FAST algorithm, which is primarily utilized for abstraction of redundant data, reiterated data and additionally reduces the dimensionality of data. The select of estimation metric heavily effect the algorithm, and it is these estimation metrics which make a distinction between the three main categories of attribute select algorithms: wrappers, filters and embedded methods. Wrapper techniques utilize a predictive model to score attribute subsets. Every topical subset is utilized to prepare a model, which is experienced on a hold-out set. Including the amount of mistakes made on that hold-out set (the error rate of the model) gives the score for that subset. Filters are mostly fewer computationally exhaustive than wrappers, but they engender an attribute set which is not regulated to a categorical type of predictive model. Many filters supply an attribute ranking relatively than an explicit best attribute subset, and the interrupt point in the ranking is selected via cross-validation. The wrapper methods utilize the predictive exactness of a determined learning algorithm to determine the integrity of the selected subsets, the precision of the machine learning algorithms are generally high[10]. However, the majority of the selected attributes are partial and the computational involution is immensely colossal. Through the filter attribute select methods, the function of cluster analysis has been demonstrated to be extra valuable than traditional attribute select algorithms. A general framework for algorithm is shown in Fig. 1.

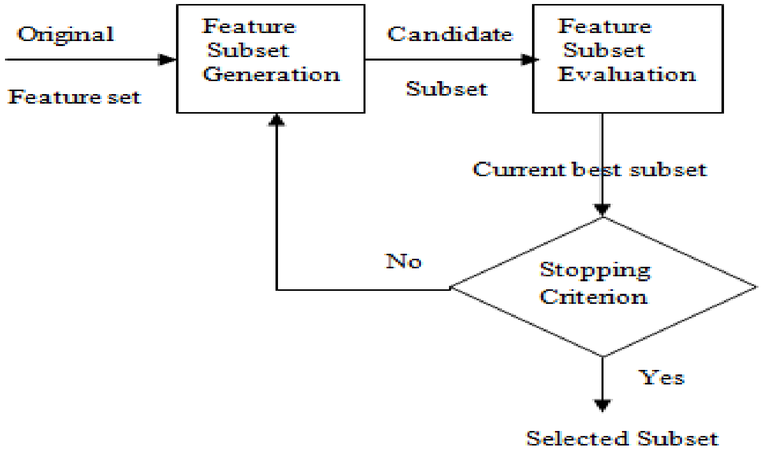

Fig. 1: A Framework for algorithm

FAST algorithm consummates in two steps. First of all, attributes are dissevered into a variety of clusters. After that the most efficient attribute is selected from each cluster

.

2. PROBLEM STATEMENT

2.1 Previous system

The process of detecting and eliminating the impertinent and redundant attributes is possible in attribute subset select. Because of 1) impertinent attributes do not involve to the expected precision and 2) redundant attributes getting information which is antecedently subsists [9][8]. Many attribute subset select algorithm can prosperously abstracts extraneous attributes but does not control on redundant attributes. But our proposed FAST algorithm can abstract extraneous attributes by taking care of the redundant attributes.

Traditional algorithms of machine learning like artificial neural networks or decision trees are embedded approaches examples [5]. The methods of wrapper utilize the analytical precision of a algorithm (as shown in Fig. 2) kenned as predetermined learning to decide the munificence of the selected subsets, the exactness of the cognition algorithms is generally high. Though, the generality of the selected characteristics is restricted and the computational arduousness is astronomically immense. The methods of filter are not dependent of learning algorithms, with fine generality. Their calculation involution is minute, but the correctness of the cognition algorithms is not sure. We have explored a novel characteristic subset selecting algorithm predicated upon clustering for maximum dimensional information [14]. The algorithm occupies following features.

(i) Irrelevant characteristics reduction. (ii) From relative ones minimum spanning tree construction. (iii) Minimum spanning tree partitioning and selecting representative characteristics. In the algorithm which is proposed, a cluster contains of characteristics. Every cluster is postulated as a solo characteristic and therefore dimensionality is radically decremented [4].

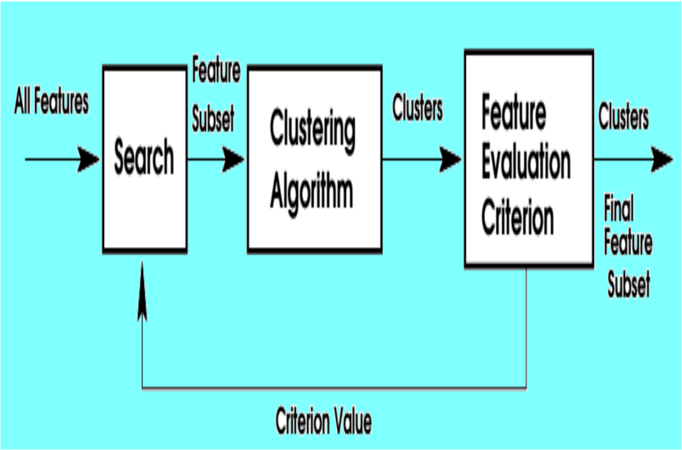

Fig. 2: Wrapper approach for unsupervised learning

2.2 Proposed System

According to the Precedent System, Incongruous attributes, along with dispensable attributes, astringently influence the precision of the learning machines[7]. Hence, attribute subset select algorithm should be capable to discover and eliminate as much of the inopportune and dispensable information as possible. In integration, to this “superior attribute subsets include attributes highly interrelated with (predictive of) the class, still uncorrelated with (not predictive of) each other.” For the above mentioned arduousness, we develop a unique algorithm which can proficiently and prosperously deal with both infelicitous and nonessential attributes, and acquire a good attribute subset. We achieve this via a recent attribute select framework (as shown in Fig. 3) which composed of the two associated mechanism of infelicitous attribute abstraction and dispensable attribute abstraction[11]. The antecedent gained attributes cognate to the intention concept by abstracting inopportune ones, and the later abstracts nonessential attributes from cognate ones through selecting representatives from unlike attribute clusters, and therefore constructs the final germane subset[13][6].

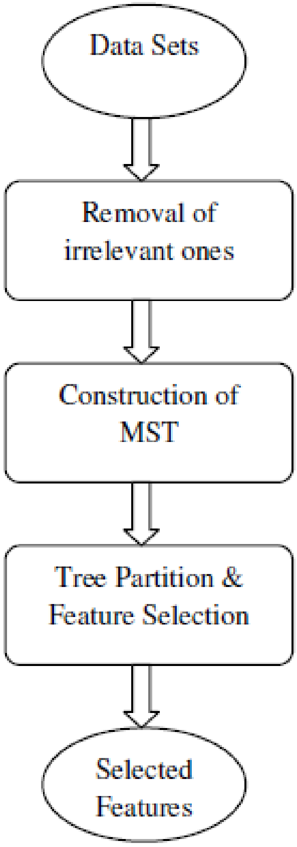
Data Sets
Removal of irrelevant ones
Construction of MST
Tree Partition & Feature Selection
Selected Features

Fig. 3: Proposed subset selection algorithm framework

3. SYSTEM DEVELOPMENT

- User Module
- Generate Subset Partition
- Impertinent Attribute Removal
- MST Construction
- Redundant & Relevant attributes List
- Time Complexity

3.1 User Module

In this module, Users are having authentication and security to entrée the detail which is presented in the ontology system. Before accessing or searching the details user should have the account in that otherwise they should register first.

3.2 Create Subset Partition

Create Partition is the step to divide the whole dataset into partitions, will be able to relegate and identify for extraneous & redundant attributes. An approach for analyzing the quality of model simplification is to division the data source.

3.3 Impertinent Attribute Removal

Utilizable way for sinking dimensionality, eliminating inopportune data, elevating learning precision, and civilizing result comprehensibility. The impertinent attribute abstraction is directly once the right pertinence quantify is defined or selected, while the nonessential attribute elimination is remotely of intricate[1]. We firstly present the symmetric uncertainty (SU), Symmetric uncertainty pleasures a couple of variables symmetrically, it give back for information gain's inequitableness in the direction of variables with more preponderant values and regulates its value to the range [0, 1].

$$SU(X,Y) = \frac{2 \times \ Gain(X|Y)}{H(X) \mid H(Y)}$$

3.4 MST Construction

This can be able to shown by an example, suppose the Minimum Spanning Tree shown in Fig.4 is constructed from a consummate graph G. In organize to cluster the attributes, we initially go across all the six edges, and then determined to remove the edge ($F0$, $F4$) because its weight ($F0$,4)=0.3 is lesser than both $SU(F0,C)$=0.5 and$SU(F4,C)$=0.7.This constructs the MST is grouped into two clusters denoted as ($T1$) and ($T2$).To construct MST we used Prim's Algorithm.

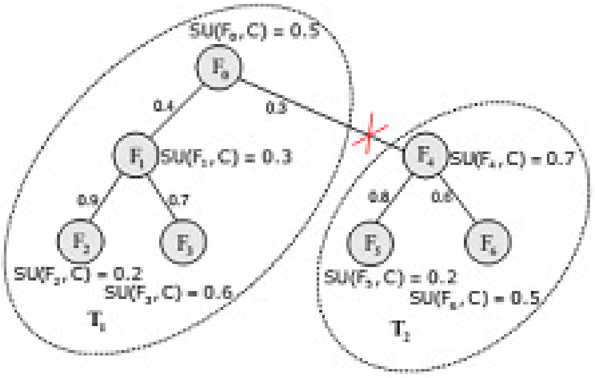

Fig. 4: Redundant & Relevant Attributes List

Ultimately it includes for final attribute subset. Then calculate the pertinent/impertinent attribute. These Attributes are pertinent and most subsidiary from the entire set of dataset.

3.5 Time Complexity

The major amount of work for Algorithm 1 involves the computation of SU values for TR relevance and F-Correlation, which has linear complexity in terms of the number of instances in a given data set. The first part of the algorithm has a linear time complexity in terms of the number of features m. Assuming features are selected as relevant ones in the first part, when k ¼ only one feature is selected.

3.6 Data Source

For evaluating the performance and effectiveness of our proposed algorithm different publically available datasets are going to be used. The number of features varies from 37 to 49152.these datasets covers range of application domains such as text, image microarray data [2][12]. Following Table 1 shows the different dataset used in paper.

Table 1: List of Data source

Data Id	Data Name	Features	Domain
1	Chess	37	text
2	mfeat-fourier	77	image, face
3	Coil2000	86	text
4	Elephant	232	microarray
5	tr12.wc	5805	text

3.6.1 Chess Dataset

This dataset contains 3196 chess end game board descriptions each end game is King+Rook versus King+Pawn on a7(one square away from queening) and it is King+Rook side(white) to move. The task is to predict if white can win on the basis of 36 features that describe the board.

3.7 Performance Evaluation

To evaluate the performance of our FAST algorithm we will compare our algorithm with four different types of representative feature selection algorithms. These algorithms are (i) FCBF (ii) ReliefF (iii) CFS (iv) FOCUS SF [3].FCBF and ReliefF evaluate features individually. For FCBF the prevalence threshold to be the SU value of the [m/logm] th rank feature of each dataset.(m is no of feature in dataset).ReliefF searches for nearest neighbor of instances of classes and weight features according to how well they differ with other classes. CFS uses best first search based on evaluation of subset which contains features highly correlated with target but uncorrelated with each other. FOCUS SF [3] uses exhaustive search in Focus with sequential forward selection.

For evaluating the performance of feature selection algorithm two metrics are used i) proportion of selected features ii) time to obtain the feature subset are going to be used. Proportion of selected feature is the ratio of number of features selected to the original no of features of the dataset.

3.8 Result & Analysis

In this section we present expected experimental results in terms of proportion of selected features and time to obtain the subset. We are going to compare results of our proposed FAST algorithm with the results of other algorithms. Fast on an average will obtain best proportion of selected features. The results will indicate that proportion of selected features of FAST is smaller than those of ReliefF, CFS, and FCBF. We are also going to compare the runtime of our proposed FAST algorithm with runtime of Relief, CFS, and FCBF each different dataset.

Our proposed FAST algorithm effectively removes irrelevant features in first step. This will reduce the chances of improperly bringing the irrelevant features into subsequent analysis.in second step fast will remove the redundant features by selecting a single representative feature from each cluster of redundant features.as result only very small number of relevant features will be selected.

4. FAST ALGORITHM

```
Inputs: D ($F1$, $F2$, $Fm$, $C$) - the given data set
$\theta$- the T-Relevance threshold.
For i=1tomdo
T-Relevance=SU ($Fi$, $C$)
If T-Relevance>$\theta$then
S=SU $\{Fi\}$;
G=NULL;
For each pair of attributes $\{F'i,'j\}$⊂S do
F-Correlation=SU ($F'i$, $F'j$)
$Add\ F'i\ and/or\ F'j\ to\ G\ with$ F-Correlation$as\ the\ weight\ of\ the\ corresponding\ edge$;
MinSpanTree=Prim (G);
Forest=minSpanTree
For each edge$Eij$ ∈Forest do
if SU($F'i$,$F'j$)<SU($F'i$,$C$)∧SU($F'i$,$F'j$)<SU($F'j$,$C$) then
```

```
Forest=Forest−$Eij$
S=$\phi$
for each tree$Ti$ ∈Forest do
$FjR$= argmax$F'k∈Ti$
SU($F'k$,$C$)
S=SU{$FjR$};
Return S
```

5. CONCLUSION

In this paper, we presented a clustering predicated attribute subset select algorithm for high dimensional data. The algorithm includes (1) eliminating impertinent attributes, (2) engendering a minimum spanning tree from relative ones, and (3) partitioning the MST and selecting most utilizable attributes. (4) Formed as incipient clusters from the selected most utilizable attributes. In the proposed algorithm, a cluster consists of attributes. Every cluster is treated as a distinct attribute and thus dimensionality is radically reduced. To elongate this work, we are organizing to explore sundry categories of correlation measures, and study some felicitous properties of attribute space.

REFERENCES

[1] Appice, M. Ceci, S. Rawles, and P. Flach. Redundant feature elimination for multi-class problems. In Proceedings of the 21st International Conference on Machine learning, pages 33-40, 2004.

[2] Achuthsankar, S. Computational biology and bioinformatics – a gentle overview. Communications of Computer Society of India (2007).

[3] Almuallim H. and Dietterich T.G., Learning boolean concepts in the presence of many irrelevant features, Artificial Intelligence, 69(1-2), pp279-305, 1994.

[4] 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.

[5] Baldi, P., and Brunak, S. Bioinformatics: The Machine Learning Approach. The MIT Press, 1998.

[6] Biesiada J. and Duch W., Features election for high-dimensionaldatała Pearson redundancy based filter, Advancesin Soft Computing, 45, pp242C249,2008.

[7] Blum, A. L., and Langley, P. Selection of relevant features and examples in machine learning. Artificial Intelligence 97, 1-2 (1997), 245–271.

[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] Chikhi S. and Benhammada S., ReliefMSS: a variation on a feature ranking ReliefF algorithm. Int. J. Bus. Intell. Data Min. 4(3/4), pp375-390, 2009

[10] Dash M. and Liu H., Feature Selection for Classification, Intelligent Data Analysis, 1(3), pp131-156, 1997.

[11] Dash M., Liu H. and Motoda H., Consistency based feature Selection, In Proceedings of the Fourth Pacific Asia Conference on Knowledge Discovery and Data Mining, pp 98-109, 2000.

[12] M. Berens, H. Liu, L. Parsons, L. Yu, and Z. Zhao. Fostering biological relevance in feature selection for microarray data. IEEE Intelligent Systems, 20(6):29-32, 2005.

[13] P.Chanda, Y.Cho, A. Zhang and M. Ramanathan, "Mining of Attribute Interactions Using Information Theoretic Metrics," Proc. IEEE Int'l Conf. Data Mining Workshops, pp. 350-355, 2009.

[14] Y. Cheng. Mean shift, mode seeking, and clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 17:790-799, 1995.