ISSN:0975-9646

An Approach for Hybrid Cluster Based Feature Selection on High Dimensional Data

[1]P. Ramasita, [2] S. Rama Sree

[1]*PG Scholar, Dept of CS, Aditya Engineering College, Surampalem, East Godavari(Dt), AP, India,*

[2]*Vice Principal, Dept of CSE, Aditya Engineering College, Surampalem, East Godavari(Dt), AP, India,*

Abstract: Data Mining is a term that refers to searching a large datasets in an attempt to detect hidden or low-level patterns. Feature selection is also called variable selection or else attributes selection. Feature selection is an algorithm, used as a preprocessing step in machine learning task. It is a method of selecting best subset of exclusive features, as a result that features gap is optimally reduced. In existing system, they failed to remove irrelevant data because computational complication is high and, the result of the datasets is not assured. In proposed method, removing irrelevant data can be done by using the T-relevance and F-correlation metrics, and then build the spanning tree by using greedy algorithm is a graph theory that finds a minimum spanning tree for a connected weighted graph. They can effectively and efficiently remove both irrelevant and redundant features to find a good feature subset. So, high dimensional data can be used for offline and online dataset. In future, Time and Space complexity can be reduced using highly developed algorithms which can be enhanced in cloud.

Key words: Feature subset selection, filter method, dimensionality reduction.

1. INTRODUCTION

Data Mining is a procedure used by companies to turn raw data into valuable information. Data Mining used in the business region about their customer to develop extra efficient selling strategies as well as enhance sales and decrease costs. .

Feature selection is also called changeable selection or else attributes selection. It is a process of selecting a subset of appropriate features. Features Selection is different from dimensionality reducation.Two methods search for reducing, the number of attribute in the datasets, but a dimensionality reduction method is introduce few combinations of attribute, where as features selection method attribute contain and eliminate attribute current in the data without changing them. It is the simplest algorithm to test the feature subsets finding the one

Reduce the error rate. And the feature selection is usually used as data preprocessing steps in device learn task. Feature selection is a method of selecting top subsets of single features [1],[2]. There are two types for feature selection methods: Filter method, the classifiers are used in which it separates the of features and in the Wrapper method, the features are chosen by the classifier.

The filter model is used due to its computational effectiveness and also improves the capability. The wrapper methods calculate the variables which have, dissimilar like filter model, to identify the achievable interaction between the variables. The two main disadvantages of these models are 1.The increases' ended appropriate threat when the number of explanation is insufficient, 2.The significant calculation time when the number of variables is big. A hybrid model has newly projected to deal with high dimensional data. In this model, first it compute the features subsets depends on their given precedence and then cross justification is taken for ultimate best subset diagonally different precedence. These algorithms mainly focus to combine filter and wrapper model to reach best possible datasets with the minimum relevance and maximum redundancy and they can be obtained by using Entropy and Gain.

Entropy (Z) is used for separate arbitrary variables Z and W (Z) is the collecting the all values of Z, W (Z) is defined by

$$G(W) = -\sum_{z \in Z} w(z) log 2 w(z). \quad (1)$$

Gain :(W/T) are the total using the entropy of Z decrease. If reflect the extra in sequence on the T and provided the T is in sequence gain [4]

$$\text{Gain } (Z/T) = W(T) - W(T/W)$$
$$= R(G) - R(T/W). \quad (2)$$

The wrapper model and filter model are not effectively dealing with removing the irrelevant features in datasets. Using Fast Algorithm can overcome the problem and getting applicable features. Clustering can be done for applicable features and then constructing the Minimum Spanning Tree (MST) for relatives ones.

The respite of the paper is offered as follow: Section 2 consists of literature survey. The planned method is telling in Section3.Section 4 demonstrate the experimental results and section 5 finish the thesis.

2. LITERATURE SURVEY

In 1999, Mark A. Hall [7] the trouble of selection for machine learning is solved by using an approach of relationship based feature selection that measure appropriate correlation to search strategy by evaluating the formula. Van Hulle M.M [8] implements a hybrid filter/wrapper feature subset selection algorithm for regression. In this features are filtered by means of redundancy and significance filter using common data between target variables and regression, in wrapper it search for top aspirant feature subsets by the copy of regression. Lei Yu [1] introduced the idea of dynamic

feature selection and also look at the careful sampling approach in a filter model setting for dynamic feature selection. In 2005, Huan Liu [6] describes the procedure of algorithms based on feature selection for cluster and categorization computes different algorithms with a structure related to search process and data mining mission that declare and provide the plan to select feature selection algorithm. Feng tan [11] implements approach that combines the two selection criteria by a Generic Algorithm. This method is used to find the feature subsets with good performance and small size. Kononenko I [10] identified the problem of superiority estimation of attributes with and without dependencies among them and it is also evaluate Features selection algorithm which deals with nonstop and distinct attributes, which is extended to transaction with strident, partial and multi class data sets.

3. PROPOSED METHODOLOGY

The proposed system is FAST (Feature Based Selection algorithm). Which are use for remove irrelevant features and the construct the Minimum Spanning Tree (MST) for the virtual ones and MST can be use for selecting courier feature

1. Removing unrelated features by using two methods are: T-Relevance estimate, F-Correlation estimate.

T-Relevance estimate: The significance among the feature S_i€R with objective D know as the T-Relevance of Si and D, with denote by RU (S_i, D) is larger than a fixed threshold, that S_i is tough T-Relevance characteristic.

RU (G, T) =2*Gain (g/t)/R(g) +R(t)

After result the significance value, the unnecessary element will be impassive with esteem to threshold velocity.

F-Correlation estimate: The correlation between any couple of features $S_{i\text{-and}}$ S_j (S_i, Sj€ ∧ S∧I ! j) is called F-correlation between any couple of features S_j,are denoted by RU(S_i,S_J).

2. Minimum Spanning Tree (MST) construction

Greedy Algorithm is a graph premise that finds a Minimum Spanning Tree for a related subjective diagram. It finds a subset of the boundaries in the tree is minimized. It finds a minimum spanning tree for every linked part.

Explanation:

1. Construct a jungle J a set of trees, where each Vertex in the diagram is a part of tree.
2. Construct a set w contain in the borders in the diagram.
3. Whereas W is a nonempty and G is not so far across.

Eliminate a border with lowest weight from W if that border connect two dissimilar trees, then put in to the new tree, join two trees into a single tree.

SYSTEM ARCHITECTURE:

The architecture of proposed system in fig.1 shows the datasets, FAST Algorithm and T-relevance and F-relevance, clustering, Minimum Spanning Tree (MST) Construction.

1. Datasets can be taken from database.
2. Using the FAST Algorithm eliminate the irrelevant feature by use the methods T-relevance and F-relevance.
3. T-relevance: The significance among the feature F_i∈ *F* and objective notion D, is known as the T-relevance of a mark subset-correlation: the connection among some pair of features S_i plus S_j, and denoted by RU (S_i,Sj).
   F-Correlationcal: The correlation between any couple of features $S_{i\text{-and}}$ S_j (S_i, Sj€ ∧ S∧I ! j) is called S-correlation
4. Clustering the features subsets
5. Minimum Spanning Tree can be generated and shows the results

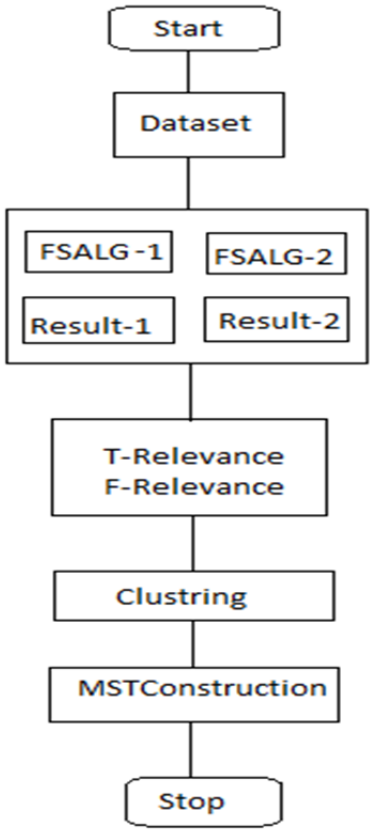

Figure 1: FAST ARCHITECTURE

4. ALGORITHM:

The proposed FAST algorithms have three steps:

Input: D (F_1,F_2…..,F_n)= the given data set

0=the T-relevance threshold.

Output:W=select feature subset

//==== step1: unrelated feature elimination====

1 for i=2 to n do

2 T-Relevance= SU(F_i,C)

3if T-Relevance>0 next

4 S=SU{F_i}

//===step2:Minimum spanning Tree creation===

5 G=void;// G be total graph

6 for each pair of feature $\{F_i^{'},F_j^{'}\}$ C S do

7 F-correlation =SU$\{F_i^{'},F_j^{'}\}$

8 Add $F_i^{'}$ and/or $F_j^{'}$ to G with f-correlation as the load of the parallel border;

9 minspantree =Prim(G);//using prim Algorithm to create the minimum spanning tree

//===step3:Tree division and illustration feature selection===

10 MST=minspantree

11 for each border $E_{i\ j\ belongs\ to}$ border do

12 if SU $(F_I^{'},F_j^{'})$<SU $(F_i^{'},C)$ and S $(F_i^{'},F_j^{'})$<SU$(F_j^{'},C)$ then

13 Edge=Border-E_{ij}

14 S=0

15 for every tree T_i go to border do

16 S=SU$\{F_R^i\}$

17 Returns S

Fast algorithm involve the evaluation of SU values
For T-Relevance and F- Correlate they have same complication in period of the amount of instance in known information set. The initial division of the algorithm has a linear time complication O (m) in the conditions of numeral of n features. The Fast algorithm shows the better runtime illustrate by high dimensional data.
Minimum Spanning Tree is a sub-chart that compasses over all the vertices of the given figure with no sequence and has smallest amount aggregate of weights over all the integrated corner and ends is measured as the Euclidean separation between the end focus frame that border. Any border that unites two sub-trees in MST must be the in detail. The remaining boundaries of the MST acquire by uprooting these boundaries dealt with as the grouping. uproot of longest bounder result into two-gathering bunch

5. RESULTS

Generally a datasets is taken and which is numeric form apply preprocessing on it. Then it appears as in figure2.

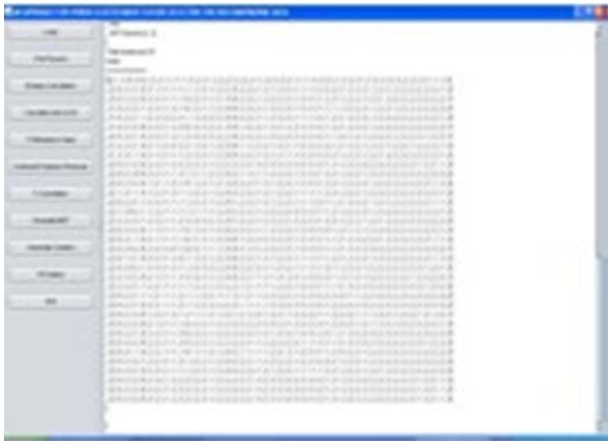

Figure2: Datasets

In figure3 gain values can be obtained and subsets can be shown that are.

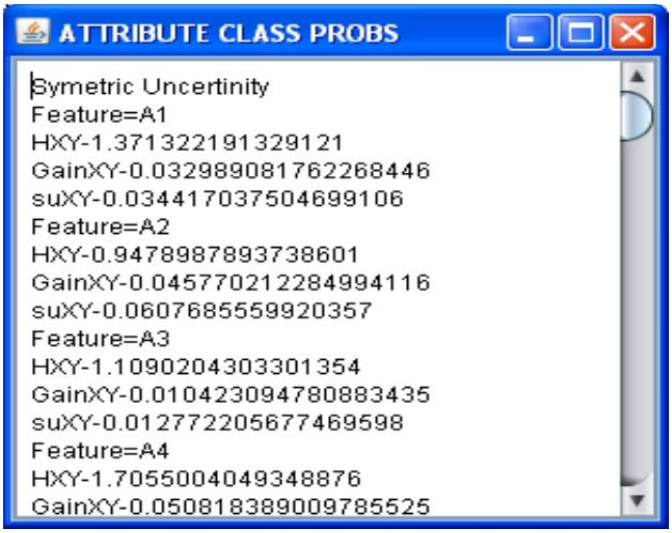

Figure3: Gain&SU calculation

Above figure4 has irrelevant feature and relevant features that are obtained by using the entropy and gain formulas**.**

Figure4: irrelevant and relevant features

Figure5 has only relevant feature subsets which have tough association with objective concept which is constantly required for best subsets.

Figure5: Relevant Feature

8. CONCLUSION

Clustering base feature subset describe about the functionalities of data mining and also regarding the algorithm of feature subset selection. The identification of suitable data is also very simple by using the subset selection algorithm. This is used to remove the feature base on cluster which is used for grouping the virtual ones, High dimensional data can use for offline and online data sets. In future, Time and Space complexity can be compact using highly developed algorithms which can be enhanced in cloud.

REFERENCES

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

2 Molina L.c.,Belanche L. and Nebot A., Feature selection algorithms: A survey and experimental evaluation, in Proc.IEEE Int. Conf.Data Mining, pp 306-313,2002.

3 GuYon I .and Elisseff A. ,An introduction to variable and feature selection, journal of Machine learing Researc,3, pp 1157-1182,2003

4 Quinlan J.R, C4.5: program for Machine learing. San Mates, Calif: Morgan Kaufman, 1992.

5 Das, S. (201).Filters, Wrappers and boosting-based hybrid for feature selection proceedings of the Eighteenth International Conference on Machine learing (pp 74-81)

6 Human Liu and Lei yu. Toward Integration Feature Selection Algorithm for classification and clustering, IEEE transaction on Knowledge and data engineering, volume 17, issue, pages: 491-502, 2005.

7 Hall, M.A, (1999), "correlation based feature selection for machine learning", Doctoral dissertation, university of Waikato, Dept of Computer Science

8 Van Dij K g.and van Hulle M.M., Speeding up the Wrapper Feature Subset selection in Regression by Mutual Information Relevance and Redundancy Analysis, International Conference On AI, pp 38-45, 1992

9 Almuallim h and Dietterich T.G., algorithms for Identifying Relevant feature, in proceedings of the 9th Canadian Conference on AI, pp 38-45, 1992

10 Kononenko I., Estimation Attributes: Analysis and extension of RELIEF, in proceedings of the 1994 European Conference on Machine learning, pp 171-182, 1994

11 Feng tan, XueZhengFug, YanquingZhang, andAnuG.Bourgeois.A genetic algorithm-based method for feature subset selection. Soft computing-A Fusion of foundation, Methodologies and Application, volume 12, issue 2, page: 111-120, Springer-Verlag, 2007.

Dr.S.Rama Sree obtained her B.Tech. Degree in Computer Science& Engineering form Koneru Lakshmaiah College of Engineering, affiliated to Achary Nagarjuna University in year 2001. and M.Tech Degree in Computer Science from Jawaharlal Nehru Technological University Kakinada in the year 2006. She is currently working as Professor in CSE and Vice Principal in Aditya Engineering College, Surampalem, India. She has 22 International Journal Papers and 7 National/International Conferences to her credit. Her Research intrests include Software Engineering,Cost Estimation, Fuzzy Logic, Neural Networks and Neuro Fuzzy Systems.

P.Ramasita Obtained B.Tech degree in Computer Science & Engineering,Sri Sai Madhavi institute of science & technology Affiliated to Jawaharlal Nehru Technological University Kakinada in the year 2013 and M.Tech pursuing Degree in Computer Science in Aditya Engineering College,Affiliated to Jawaharalal Nehru Technological unversity Kakinada,India