Open access Journal International Journal of Emerging Trends in Science and Technology

Feature Selection using Clustering Algorithms: FAST and LUFS

Authors
Neha V. Dharmale[1], Santosh N. Shelke[2]
[1]Pune University, Sinhgad Academy of Engineering, Kondhwa, Pune, India
nehadharmale311@gmail.com
[2]Pune University, Sinhgad Academy of Engineering, Kondhwa, Pune, India
santo.shelke@gmail.com

Abstract
Feature selection is used to reduce the number of features in many applications where hundreds or thousands of features are present in data. Many feature selection methods are proposed which mainly focus on finding relevant features. High dimensional data becomes very common with the emerging growth of applications. Thus, there is a need of mining High dimensional data very effectively and efficiently. Clustering is widely used data mining model that partitions data into a set of groups, each of which is called a cluster. To reduce the dimensionality of the data and to select a subset of useful features from this clusters is the main goal of feature subset selection. In dealing with high-dimensional data for efficient data mining, feature selection has been shown very effective. Popular social media data nowadays increasingly presents new challenges to feature selection. Social media data consists of data such as posts, comments, images, tweets, and linked data which describes the relationships between users of social media and the users who post the posts. The nature of social media increases the already challenging problem of feature selection because the social media data is massive, noisy, and incomplete. There are several algorithms applied to find the efficiency and effectiveness of the features. Here we are using the combination of FAST and Linked Unsupervised feature selection algorithm for the linked high dimensional data.

Keywords:*Clustering, Feature selection, FAST, Linked data, LUFS.*

1. Introduction

Feature selection as a preprocessing step to data mining is effective in removing irrelevant data , reducing dimensionality, and improving result comprehensibility. In some applications Clustering is also called data segmentation because it makes the partitions of large data sets into groups based on their similarity. A database or a data warehouse can contain several attributes or dimensions. Handling high dimensional data is enormous issue in data mining applications. Clustering methodis used for grouping together data that are similar to each other and dissimilar to the data belonging to other clusters. Many clustering algorithms can effectively manage low-dimensional data, which involves only two to three dimensions. Human eyes can judge the quality of clustering effectively upto three dimensions. In high-dimensional space finding clusters of data objects is very challenging. Irrelevant features provide useless data which is not related to the target classes.

The objectives of feature selection includes: improving data mining performance, building simpler and more comprehensible models and preparing clean data which is easy to understand. The Feature selection technique is used for the data containing many irrelevant and replicated features. In successful data mining applications, Feature selection is an essential step which can effectively reduce data dimensionality by removing the irrelevant features and feature redundancy. Many

feature subset methods are used for this purpose which are Wrapper, filter, Embedded and hybrid which are classified on the basis of their generality, accuracy and computational complexity. The wrapper method requires predefined algorithm to select the features.

Wrapper methods uses a predictive model to score feature subsets. As wrapper methods trains a new model for each subset, they are computationally very expensive. The wrapper methods are beneficial but requires more time to select the features. The selected features are limited and computational complexity is large in wrapper methods. When the number of features are very high filter method can be a good choice. The computational complexity is low but there is no guarantee of accuracy. To achieve the better performance, combination of filter and wrapper can be used which is known as hybrid method. Two fundamental problems for feature selection on social media data are relation extraction and mathematical representation and their associated challenges are: what are different types of relations among data instances and how to capture them and how to model these relations for feature selection.

Here, we propose a framework Linked feature selection of social media data that naturally integrates different relations into a state-of the-art formulation of feature selection, and turn the integrated formulations to an optimization problem with convergence analysis when developing its corresponding feature selection algorithm. We propose an algorithm which is the combination of Linked feature selection and FAST algorithm. Which works in three steps. In the first step linking between the features takes place using Linked FS algorithm and then the features which are linked with each other are only get processed in the next step. In the second step features are divided into clusters using clustering method and then in the last step most representative features are selected which are strongly related with the target class to get the final subset of features.

2. Related Work

Set Predictive accuracy is not contributed by irrelevant features and redundant features which provides the information which is already present in other features.

FAST algorithm[1] employs the clustering-based method to choose effective and efficient features. **C**onsist method[2] uses best first strategy for searching the minimal subset.Some feature subset selection algorithms[3] eliminate irrelevant features [4] efficiently but fail to handle redundant features and some of others can eliminate the irrelevant features but fails to handle redundant features.

For High Dimensional Data several researchers have done the Fast Clustering-Based Feature Subset Selection Algorithm.Many feature subset selection methods have studied which are divided into four broad categories . In Wrapper method [5] generality of the selected features is limited and complexity in computation is very high. Computational complexity of the Filter method is very low with good generality. Hybrid method [6] is the combination of wrapper and filter method. To achieve best performance it is used. Embedded method is more efficient than above three categories. Decision tree and artificial neural networks are the examples of embedded method.

CMIM[7] picks the features iteratively which maximizes their mutual information with the class to predict. Conditional Mutual Information Maximization criterion does not select a feature similar to already picked features, even if it is individually powerful, as it does not carry additional information about the class to predict. Relief[8] is a weight-based algorithm inspired by instance-based learning algorithms for feature selection. Some feature selection algorithms uses relief which weighs each feature based on distance-based criteria function according to its ability to discriminate instances under different targets.The problem in using Relief is, two or more predictive but highly correlated features are likely to be highly weighted. Therefore in removing redundant features Relief is ineffective and hence generates non-optimal feature set size in the presence of redundant features. Relief-F[9] extends the Relief. Relief–F can work with

incomplete and noisy datasets and It can deal with multi-class problems, but fails to identify redundant features. It includes the search for nearest neighbors of different classes. CFS [10],[11]uses best first search which quickly identifies and screens irrelevant, redundant, and noisy features and identifies relevant features[12] as long as their relevance does not strongly depends on other features. CFS is inefficient at retrieving high-dimensional data FCBF [13] is a very similar solution called Fast Correlation-Based Filter. It selects features which are highly correlated with the class to predict if they are less correlated to any feature already selected. It can identify relevant features as well as redundancy among relevant features. It is a fast filter method**.**

The FOCUS algorithm[14]keep features in the queue which may contain a solution. Initially, the queue contains only the element which represents the whole set. In each iteration the queue is partitioned into disjoint subspaces, and those subspaces that cannot contain solutions are pruned from the search. The FOCUS only prunes subspaces that cannot be complete, and it will not miss any sufficient feature subsets. FOCUS-SF [15] reduces the exhaustive search in the FOCUS by using sequential forward selection.

3. Proposed System

3.1 Framework

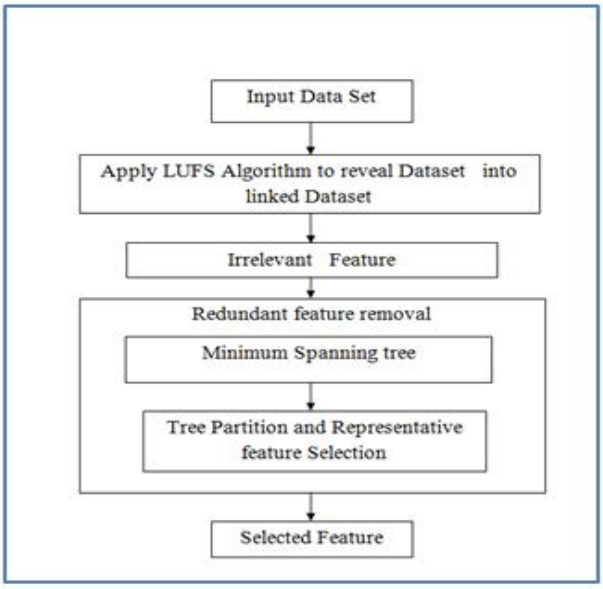

Figure 1: Proposed System

The combination of FAST [1] Algorithm and Linked feature selection algorithm [2] is our proposed system. Main goal of using the combination of LUFS and FAST algorithm is to select a subset of useful features by using the linked dataset which should not contain the irrelevant and redundant features. The final features in the subset should contain the features which are not correlated with each other but highly correlated with the class.

Framework of proposed Combination of two algorithms is shown in fig.1. It consistof two components of irrelevant feature removal and redundant feature elimination which are connected to each other. When we input the dataset, Linked feature selection algorithm reveals that dataset into linked dataset and then by using it as a input, the first component gives output which contains the feature relevant to the target concept and eliminates the irrelevant features, and the second component removes redundant features from relevant features by choosing representatives feature from different clusters. Removal of redundant features involves minimum spanning tree construction from the linked dataset after removal of irrelevant features and thus produces the final subset of representative features. After defining the right relevance measure we can easily remove the irrelevant features, but the removal of redundant features is a bit of complicated. More calculations has to be done for removal of redundant features.Redundancy is related with feature correlation and Relevance is related with feature-target concept correlation.

Symmetric uncertainty is a measure of correlation between two features or between target concept and feature. When we have two variables, then symmetric uncertainty indicates that knowledge of one value can predict the value of other or not.If value of SU is 1 then it means that one value can completely predict the value of other and if it is 0 then it means that two variables are independent of each other.

Feature subset selection is the process that identifies and retains the strong T-Relevance features and selects representative features as output.

3.2 Algorithm

Input: D{F1,F2,…Fm,C}- the given dataset
Output: S subset of relevant features
//--------LUFS Algorithm for linking dataset-----

1. Obtain the social dimension indicator matrix H
2. Set F = $H(H^TH)^{-1/2}$
3. Construct S through Eq.6
4. Set L = D – S
5. Set t = 0 and initialize D0 as an identity matrix
6. Update the diagonal matrixD_{t+1}, where the ith diagonal element is$1/(2||W_t(i,:)||_2)$;
7. Sort each feature according to$||W_t(i,:)||_2$ in descending order and select top-k ranked ones;

//---- Minimum Spanning Tree Construction-----

8. G = NULL; //G is a complete graph
9. for each pair of features $\{F'_i,F'_j\} \subset$ S do
10. F-Correlation = SU (F'_i,F'_j)
11. $Add F'_i,nd/or F'_j,to Gwith$F-Correlation as the weight of the corresponding edge;

12.minSpanTree= Prim (G); //Using Prim Algorithmto generate the minimum spanning tree
//---Tree Partition and Representative Feature Selection
13.Forest = minSpanTree
14.for each edge $Eij \in$ Forest do
15. ifSU (F'_i,F'_j)<SU $(F'_i,C)\wedge$SU (F'_i,F'_j)< SU (F'_i,C)then
16. Forest=Forest–E_{ij}
17.S = ϕ
18.for each tree $Ti \in$ Forest do
19. F'_R= $argmax_{F'_k \in Ti}$SU(F'_k,C)
20. S = S ∪ $\{F'_k\}$;
21.return S.

4. Results

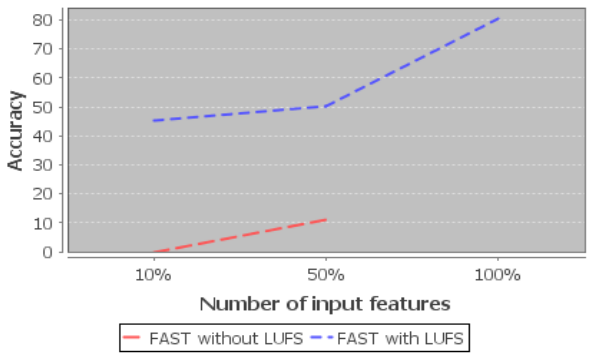

Figure 2: Accuracy vs Number of features

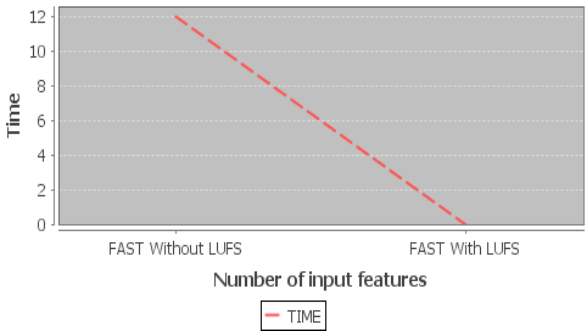

Figure 3: Time vs Number of features

For evaluating the performance and effectiveness of the proposed algorithm we are using the dataset related with social media. After applying dataset cleaning we get the connected clusters and high dimensional matrix of that clusters. In the existing FAST algorithm major amount of work involves the calculation of SU values for F-correlation and T-Relevance having linear time complexity O(m). where m indicates the number of features.

Using proposed algorithm linking between the features in the given dataset is performed using Linked feature selection algorithm. The features which are related with other features are only considered for the next procedure. So in the next step we have to take into consideration the linked features and have to calculate the SU for the linked features only. Which reduces the complexity of calculating SU for each feature to O(k). where k is the number of linked features and (k ≤ m).

References

1. AQinbao Song, Jingjie Ni and Guangtao Wang , “A fast clustering- based feature subset selection algorithm for high-dimensional data”, IEEE Transactions on Knowledge and Data Engineering, vol. 25, no. 1, 2013.
2. J. Tang and H. Liu, “Feature selection with linked data in social media”, In SIAM International Conference on Data Mining, 2012.
3. L. Yu and H. Liu, “Efficiently Handling Feature Redundancy in High-Dimensional Data,” Proc. Ninth ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (KDD ’03), pp. 685-690, 2003.
4. 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.

5. R. Kohavi and G.H. John, "Wrappers for Feature Subset Selection," Artificial Intelligence, vol. 97, nos. 1/2, pp. 273-324, 1997.
6. J. Souza, "Feature Selection with a General Hybrid Algorithm," PhD dissertation, Univ. of Ottawa, 2004.
7. F. Fleuret, "Fast Binary Feature Selection with Conditional Mutual Information," J. Machine Learning Research, v. 5, pp. 1531- 1555, 2004.
8. 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, 2004.
9. I. Kononenko, "Estimating Attributes: Analysis and Extensions of RELIEF," Proc. European Conf. Machine Learning, pp. 171-182, 1994.
10. L.C. Molina, L. Belanche, and A. Nebot, "Feature Selection Algorithms: A Survey and Experimental Evaluation," Proc. IEEE Int'l Conf. Data Mining, pp. 306-313, 2002.
11. M. Dash, H. Liu, and H. Motoda, "Consistency Based Feature Selection," Proc. Fourth Pacific Asia Conf. Knowledge Discovery and Data Mining, pp. 98-109, 2000.
12. K. Kira and L.A. Rendell, "The Feature Selection Problem: Traditional Methods and a New Algorithm," Proc. 10th Nat'l Conf. Artificial Intelligence, pp. 129-134, 1992.
13. M.A. Hall, "Correlation-Based Feature Subset Selection for Machine Learning," PhD dissertation, Univ. of Waikato, 1999.
14. L. Yu and H. Liu, "Feature Selection for High-Dimensional Data: A Fast Correlation- Based Filter Solution,"Proc. 20th Int'l Conf. Machine Leaning, vol. 20, no. 2, pp. 856-863, 2003.
15. G.H. John, R. Kohavi, and K. Pfleger, "Irrelevant Features and the Subset Selection Problem," Proc. 11th Int'l Conf. Machine Learning, pp. 121-129, 1994.