CLUSTERING BASED FEATURE SELECTION AND IDENTIFICATION OF SUBSET FOR HIGH DIMENSIONAL DATA

Neha V. Dharmale[1], Santosh N. Shelke[2]

[1,2] *Department of Computer Engineering, Sinhgad Academy of Engineering, Pune, (India).*

ABSTRACT

Feature selection is widely used in preparing high dimensional data for effective data mining. Increasingly popular social media data presents new challenges to feature selection. Social media data consists of traditional high-dimensional, attribute value data such as posts, tweets, comments, and images, and linked data that describes the relationships between social media users as well as who post the posts. The nature of social media also determines that its data is massive, noisy, and incomplete, which exacerbates the already challenging problem of feature selection. In this paper, using SVM Algorithm we illustrate the differences between attribute value data and social media data, investigate if linked data can be exploited in a new feature selection framework by taking advantage of social science theories, extensively evaluate the effects of user-user and user-post relationships manifested in linked data on feature selection.

Keywords: Cluster, Feature selection, Linked Data, SVM

1. INTRODUCTION

The myriads of social media services are emerging in recent years that allow people to communicate and express themselves conveniently and easily, e.g. Facebook and Twitter. The pervasive use of social media generates massive data in an unprecedented rate. For example, users on Twitter are sending 200 million tweets per day, which is about 200 percent growth in a year; more than 3,000 photos are uploaded to Flicker per minute and more than 153 million blogs are posted per year. The massive, high-dimensional social media data poses new challenges to data mining tasks such as classification and clustering. One conventional approach to handling large-scale, high-dimensional data is feature selection [1].

Feature selection aims to select relevant features from the high dimensional data for a compact and accurate data representation. It can alleviate the curse of dimensionality, speed up the learning process, and improve the generalization capability of a learning model [2]. The vast majority of existing feature selection algorithms work with data containing uniform entities or attribute-value data points that are typically assumed to be independent and identically distributed. However, social media data differs as its data points or instances are inherently connected to each other.

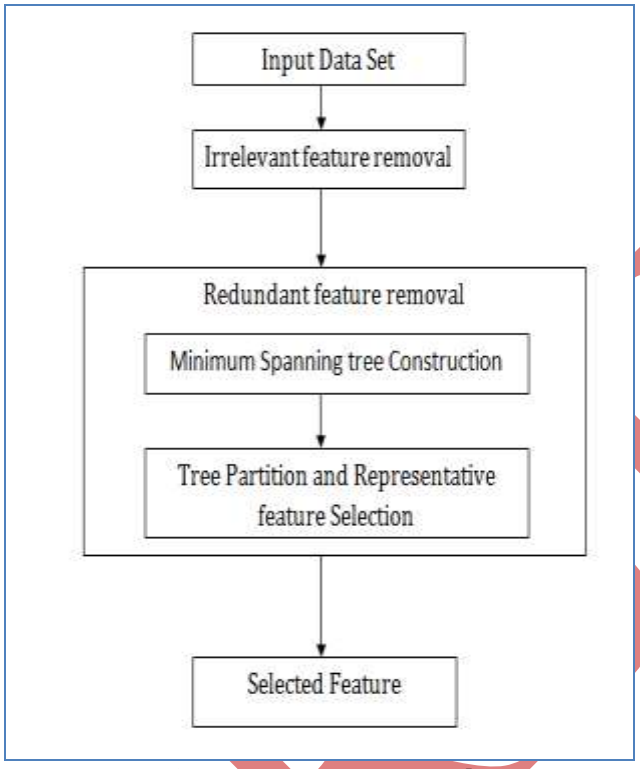

Fig.1: System Framework

Fast algorithm employs the clustering-based method to choose features. General framework as shown in Fig. 1 in which irrelevant features are removed first and then to remove redundant features minimum spanning tree is constructed and then tree partitioning is used to obtain the selected features. Fast Algorithm can eliminate the irrelevant features effectively but it is ineffective at removing redundant features which affect the speed and accuracy of algorithm, thus it should be eliminated as well.

II LITERATURE SURVEY

For High Dimensional Data several researchers have done the Fast Clustering-Based Feature Subset Selection Algorithm . Many new techniques are introduced by means of different performance and using different techniques for this purpose.

To identify and to remove as many irrelevant and redundant features as possible Feature subset selection is used. Because predictive accuracy is not contributed by irrelevant features and redundant features provides the information which is already present in other features[3]. Some feature subset selection algorithms eliminate irrelevant features efficiently but fail to handle redundant features and some of others can eliminate the irrelevant features but fails to handle redundant features.

Many feature subset selection methods have studied which are divided into four broad categories: Embedded, Wrapper, Filter, and Hybrid approaches. Some feature selection algorithms are relief which weighs each feature

based on distance-based criteria function according to its ability to discriminate instances under different targets[4]. In removing redundant features Relief is ineffective as two predictive but highly correlated features are likely both to be highly weighted. Relief-F is the extension of Relief; it enables relief to work with noisy and incomplete data sets and to deal with multi-class problems, but fails to identify redundant features[5]. CFS is achieved by the hypothesis that a good feature subset is one that contains features highly correlated with the target, yet uncorrelated with each other[6]. For retrieving high-dimensional data CFS is not efficient. FCBF is a fast filter method in which without pairwise correlation analysis it can identify relevant features as well as redundancy among relevant features[7]. CMIM picks the features iteratively which maximize their mutual information with the class to predict, conditionally to the response of any feature already picked [8]. FAST algorithm employs the clustering-based method to choose features [9].

III PROPOSED SYSTEM

Accuracy of the learning machines is severely affected by the redundant and irrelevant features. Thus Feature subset selection should identify and remove the irrelevant and redundant features.

In this paper, we investigate issues of feature selection for social media data. Specifically, we perform feature selection on posts (e.g., tweets, blogs, or images) in the context of social media with link information between user and user or between user and posts. Since conventional feature selection methods cannot take advantage of the additional information in linked data. We proceed to study two fundamental problems for feature selection on social media data:

- **Relation extraction** - what are distinctive relations that can be extracted from linked data, and
- **Mathematical representation** - how to represent these relations and integrate them in a state-of- the-art feature selection formulation.

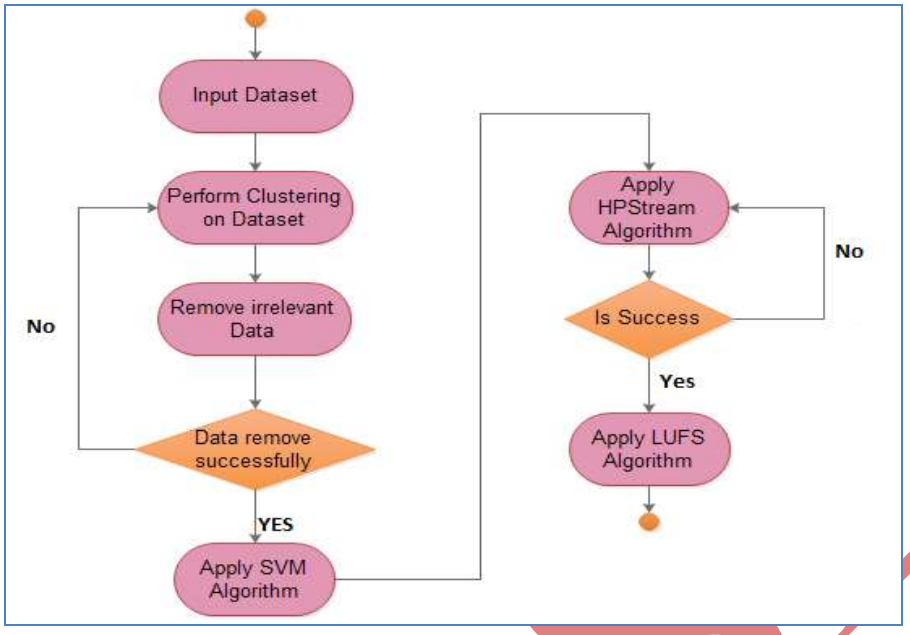

Fig.2: Data Flow Diagram.

Their associated challenge is what different types of relations among data instances are and how to capture them and how to model these relations for feature selection. To capture relations from linked data guided by social correlation theories, we propose a framework (Linked FS) 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 develop a SVM (support vector machine) algorithm which is a classifier used to predict the data. SVM algorithm identify the redundant features and remove them, It can deal with both irrelevant and redundant features efficiently and effectively , through which we can obtained a subset of useful features.

We will get the Subset of good features with the help of propoed SVM Algorithm whose data flow is given in Fig. 2. Which will be the result of eliminating redundant and irrelevant features effectively and efficiently.

- First Input the Training dataset(Linked dataset).
- Perform clustering on the training dataset which shows the connectivity between connected nodes in dataset.
- Remove Irrelevant data from the training dataset.
- After that we will do the labeling for unlabeled nodes by using SVM algorithm.
- Remove the redundant data from the training dataset

Then we will do the projected stream clustering using the High-dimensional Projected Stream Clustering (HPStream) algorithm.

We are using the feature subset selection concept for the purpose of

1. Identifying redundant features.
2. Searching for relevant features.
3. Improving the efficiency and effectiveness of data retrieval.
4. Using linked datasets as an input dataset.

IV CONCLUSION

We have presented feature subset selection algorithm for high dimensional social media data. In the proposed algorithm, a cluster consists of features. In previous system generality of the selected features was limited and the computational complexity was large. In proposed system each cluster is treated as a single feature and thus dimensionality is drastically reduced and computational complexity is also reduced .It efficiently and effectively deal with both irrelevant and redundant features and obtain a good feature subset. Good feature subsets contain features highly correlated with the class, yet uncorrelated with each other.

REFERENCES

[1] M. Dash, and H. Liu, Feature Selection for Classification, *Intelligent Data Analysis*, 1(3), 1997, 131-156.

[2] P. Langley, "Selection of Relevant Features in Machine Learning," *Proc. AAAI Fall Symp.* Relevance, pp. 1-5, 1994.

[3] G.H. John, R. Kohavi, and K. Pfleger, Irrelevant Features and the Subset Selection Problem, *Proc. 11th Int'l Conf. Machine Learning*, 1994, 121-129.

[4] I. Kononenko, Estimating Attributes: Analysis and Extensions of RELIEF, *Proc. European Conf. Machine Learning*, 78(4), 1994, 171-182.

[5] S. Chikhi and S. Benhammada, ReliefMSS: A Variation on a Feature Ranking Relieff Algorithm, *Int'l J. Business Intelligence and Data Mining*, vol. 4, nos. 3/4, pp. 375-390, 2009.

[6] M. Dash, H. Liu, and H. Motoda, Consistency Based Feature Selection, *Knowledge Discovery and Data Mining*, 18(5), 2000, 98-109.

[7] M.A. Hall, *Correlation-Based Feature Subset Selection for Machine Learning*, doctoral diss., Department of Computer Science, university of Waikato, 1999.

[8] F. Fleuret, Fast Binary Feature Selection with Conditional Mutual Information, *Journal of Machine Learning Research*, 5, 2004, 1531-1555.

[9] Q. Song, J. Ni and G. Wang, A Fast Clustering-Based Feature Subset Selection Algorithm for High Dimensional Data, *IEEE transactions on knowledge and data engineering*, 25(1), 2013,1-14.