IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-3, Issue-11 E-ISSN: 2347-2693

Comparative Analysis of Linked Unsupervised based Feature Selection Framework for Social Media Data

Pradeepa.T[1*] and Shanmugapriya.B[2]

[1*,3] *Department of Computer science, Sri Ramakrishna college of Arts and Science for women,*

www.ijcseonline.org

Received: Oct/22/2015 Revised: Nov /05/2015 Accepted: Nov/20/2015 Published: Nov/30/2015

***Abstract*— The explosive usage of social media produces massive amount of unlabeled and high- dimensional data. Feature selection has been proven to be effective in dealing with high-dimensional data for efficient learning and data mining. Unsupervised learning has been proven to be a powerful technique in unsupervised feature selection, which allows embedding feature selection into the classification (or regression) problem. In literature several numbers of feature selection methods such as supervised feature selection algorithms and unsupervised feature selection methods have been proposed to select dimensional feature in the social network. When compare to supervised methods ,unsupervised feature selection methods performs well since it perform operation without label information .But unsupervised feature selection is particularly difficult due to the definition of relevancy of features becomes unclear. To solve this problem , in this paper study a unsupervised feature selection algorithm the concept of pseudo-class labels to guide extracting constraints from link information and attribute- value information, resulting in a new Linked Unsupervised based feature selection framework (LUFS), for linked social media data. LUFS examine the differences between social media data and traditional attribute value data; investigate how the relations extracted from linked data can be exploited to help select relevant features for linked social media data. Furthermore, social theories are developed by sociologists to explain the formation of links in social media. Experimental results on various social media datasets demonstrate the effectiveness of the proposed framework LUFS is compared with existing schemas in terms of accuracy and Normalized Mutual Information (NMI). Design and conduct systemic experiments to evaluate the proposed framework on data sets from real-world social media websites.**

Keywords—***Unsupervised Feature Selection, Linked Data, Social Media, Pseudo Labels, Social Dimension Regularization***.

I. INTRODUCTION

In recent years, the rapid emergence of social media services such as Facebook and Twitter allows more and more users to participate in online social activities such as posting blogs or microblogs, uploading photos and connecting with other like-minded users. The explosive popularity of social media produces massive data at an unprecedented speed. For example, 250 million tweets are posted per day;1 3,000 photos are uploaded per minute to Flickr; and the number of Facebook users has increased from 100 million in 2008 to 800 million in 2011. The massive and high-dimensional social media data challenges traditional data mining tasks such as classification and clustering due to curse of dimensionality and scalability issues. One traditional and effective approach to handle high-dimensional data is feature selection [1-2], which aims to select a subset of relevant features from high-dimensional feature space that minimize redundancy and maximize relevance to the targets (e.g., class label). Feature selection helps improve the performance of learning models by alleviating the curse of dimensionality, speeding up the learning process, and improving the generalization capability of a learning model [3].

Data with high dimensionality not only significantly increases the time and memory requirements of the algorithms, but also degenerates many algorithms' performance due to the curse of dimensionality and the existence of irrelevant, redundant and noisy dimensions [4]. Feature selection, which reduces the dimensionality by selecting a subset of most relevant features, has been proven to be an effective and efficient way to handle high-dimensional data. In terms of the label availability, feature selection methods can be broadly classified into supervised methods and unsupervised methods. The availability of the class label allows supervised feature selection algorithms [5-6] to effectively select discriminative features to distinguish samples from different classes. Sparse learning has been proven to be a powerful technique in supervised feature selection [7] which enables feature selection to be embedded in the classification (or regression) problem. As most data is unlabeled and it is very expensive to label the data, unsupervised feature selection attracts more and more attentions in recent years [8-9].

Without label information to define feature relevance, a number of alternative criteria have been proposed for unsupervised feature selection. One commonly used criterion is to select features that can preserve the data

similarity or manifold structure constructed from the whole feature space [10]. In recent years, applying sparse learning in unsupervised feature selection has attracted increasing attention. These methods usually generate cluster labels via clustering algorithms and then transform unsupervised feature selection into sparse learning based supervised feature selection generated cluster labels such as Multi with these generated cluster labels such as Multi-cluster feature selection (MCFS) [11], Nonnegative Discriminative Feature Selection (NDFS), and Robust Unsupervised Feature Selection (RUFS). Furthermore, with high-dimensional data, it is likely to find many sets of features that seem equally good without considering additional constraints [10], [11]. Most existing feature selection algorithms work only with attribute-value data while social media data is inherently linked; adding further challenges to feature selection. Since linked data provides link information beyond attribute value data. The availability of link information presents unprecedented opportunities to advanced research for feature selection. Linked data in social media presents both challenges as well as new opportunities for unsupervised feature selection. To address the challenges of unsupervised feature selection for linked social media data, propose a novel framework of Linked Unsupervised ABC based feature selection framework (LUAFS). The major contributions of the work are summarized below:

- Introducing the concept of pseudo-class labels, enabling us to extract constraints from linked data, i.e., link information and attribute-value information, for unsupervised feature selection.
- Developing two approaches to exploit the individual and group behaviors of linked instances via graph regularization and social dimension regularization (SDR), separately.
- Proposing a novel inked Unsupervised based feature selection framework (LUFS) for linked data in social media to exploit linked information in selecting features;
- The availability of link information can provide more constraints for unsupervised feature selection and can potentially improve its performance.

II. BACKGROUND STUDY

Many researchers paid great attention to developing unsupervised feature selection [12]. In [12] propose a "filter" method for feature selection which is independent of any learning algorithm. This method can be performed in either supervised or unsupervised fashion. The proposed method is based on the observation that, in many real world classification problems, data from the same class are often close to each other. The importance of a feature is evaluated by its power of locality preserving, or, Laplacian Score. Compare proposed method with data variance (unsupervised) and Fisher score (supervised) on two data sets.

Unsupervised feature selection [13], [14] is a less constrained search problem without class labels, depending on clustering quality measures and can eventuate many equally valid feature subsets. Identify two issues involved in developing an automated feature subset selection algorithm [13] for unlabeled data: the need for finding the number of clusters in conjunction with feature selection, and the need for normalizing the bias of feature selection criteria with respect to dimension. Explore the feature selection problem and these issues through FSSEM (Feature Subset Selection using Expectation-Maximization (EM) clustering) and through two different performance criteria for evaluating candidate feature subsets: scatter separability and maximum likelihood. Present proofs on the dimensionality biases of these feature criteria, and present a cross-projection normalization scheme that can be applied to any criterion to ameliorate these biases.

Feature selection aims to reduce dimensionality for building comprehensible learning models with good generalization performance. Feature selection algorithms are largely studied separately according to the type of learning: supervised or unsupervised. This work [14] exploits intrinsic properties underlying supervised and unsupervised feature selection algorithms, and proposes a unified framework for feature selection based on spectral graph theory. The proposed framework is able to generate families of algorithms for both supervised and unsupervised feature selection. And it show that existing powerful algorithms such as ReliefF (supervised) and Laplacian Score (unsupervised) are special cases of the proposed framework. To the best of our knowledge, this work is the first attempt to unify supervised and unsupervised feature selection, and enable their joint study under a general framework. Experiments demonstrated the efficacy of the novel algorithms derived from the framework. With high-dimensional data, it is likely to find many sets of features that seem equally good without considering additional constraints. Another key difficulty is how to objectively measure the results of feature selection. A wrapper model is proposed in [13] to use a clustering algorithm in evaluating the quality of feature selection.

Present a Bayesian method [15] for mixture model training that simultaneously treats the feature selection and the model selection problem. The method is based on the integration of a mixture model formulation that takes into account the saliency of the features and a Bayesian approach to mixture learning that can be used to estimate the number of mixture components. The proposed learning algorithm follows the variational framework and can simultaneously

optimize over the number of components, the saliency of the features, and the parameters of the mixture model.

Discriminative unsupervised feature selection [16] feature selection can be embedded in the learning process. In [17], the authors proposed a supervised feature selection framework FSNet to select features for networked data. It adopts linear regression to fit the content information, and graph regularization to capture the link information. Traditional feature selection methods assume that the data are independent and identically distributed (i.i.d.). However, in real world, there are tremendous amount of data which are distributing in a network. This motivates us to study feature selection in a network. Present a supervised feature selection method based on Laplacian Regularized Least Squares (LapRLS) [17] for networked data. In detail, use linear regression to utilize the content information, and adopt graph regularization to consider the link information. The proposed feature selection method aims at selecting a subset of features such that the empirical error of LapRLS is minimized. The resultant optimization problem is a mixed integer programming, which is difficult to solve. It is relaxed into a $L_{2,1}$-norm constrained LapRLS problem and solved by accelerated proximal gradient descent algorithm. Experiments on benchmark networked data sets show that the proposed feature selection method outperforms traditional feature selection method and the state of the art learning in network approaches.

III. PROPOSED METHODOLOGY

In this research work, introducing the concept of pseudo-class labels, enabling us to extract constraints from linked data, i.e., link information and attribute-value information, for unsupervised feature selection. Developing two approaches to exploit the individual and group behaviors of linked instances via graph regularization and social dimension regularization (SDR), separately. Let(0,..1) where $\pi(.)$ is the permutation function and k is the number of features to select where $\pi(0,\ldots 1)$ indicates that the i^{th} feature is selected. The original data can be represented as diag(s)X with k selected features, where diag(s) is a diagonal matrix. The difficulty faced with unsupervised feature selection is due to lack of class labels. Hence, introduce the concept of pseudo-class label to guide unsupervised feature selection. Assume that there is a mapping matrix W € R^{mxc}, which assigns each data point with a pseudo-class label where c is the number of pseudo-class labels. The pseudo-class label indicator matrix is Y=W>diag(s) X € R^{cxn} Y_{ji} indicates the likelihood of the i^{th} instance belonging to the j^{th} class. Following the widely adopted constraints on the cluster indicator matrix, add orthogonal constraints on Y. Since X is centered, it is easy to verify that Y is also centered

Both supervised and unsupervised feature selection tasks aim to solve the same problem: selecting features consistent with given constraints. In the supervised setting, label information plays the role of constraints; in the unsupervised setting, pseudo-class label information is tantamount to the provided label information and seeks pseudo-class label information from linked social media data. In particular, constraints from both link information R and attribute-value part X to fit pseudo class label. For each type of information, provide both an intuitive way and an advanced way to extract constraints for feature selection in this paper. The reasons are two-fold. First, feature selection for linked data is a relatively novel problem; try to explore it more widely to seek a better and deeper understanding of this problem. Second, we try to investigate both intuitive and recently proposed techniques in the studied problem to help us understand their advantages and disadvantages in the proposed problem. For link information, a widely adopted assumption is that linked instances are likely to share similar labels, which can be explained by social correlation theories such as homophily and social influence. Based on this intuition, propose graph regularization [20] to capture link information based on social correlation theories. Users in social networks are likely to form groups and users within groups are similar while users from different groups are dissimilar, which is supported by social dimension assumption. An advanced way social dimension regularization based on social dimension assumption is proposed to extract constraints from link information as well.

Capturing Link Information

In general, linked instances are correlated and allow us to exploit their correlations for feature selection. In following sections, by considering their individual and group behaviors, introduce graph regularization and social dimension regularization to capture the dependency among linked instances as individuals and groups, respectively.

Graph Regularization for Link Information

Two linked instances are distinct from two instances of attribute-value data due to the correlations between them [17]. For example, two linked users in Twitter are more likely to have similar interests than two randomly picked users. These correlations can be explained by social correlation theories such as homophily [18] and social influence [19]. Homophily suggests that two instances with similar topics are more likely to be linked while social influence theory indicates that two linked instances are more likely to have similar topics.

Social Dimensions for Link Information

In social dimension is introduced as a means to integrate the interdependency among linked data and attribute- value data. Social dimension can capture group behaviors of linked instances: instances from different social dimensions are dissimilar while instances in the same social dimension are similar. Flattening linked data with social dimensions, traditional classification methods such as SVM obtain better performance for relational learning.

Capturing Attribute-Value Information

With the pseudo-class label, further allowed to capture information from attribute-value part in a supervised manner. By introducing the concept of pseudo-class labels, constraints from both link information and unlabeled attribute- value data are ready for unsupervised feature selection. Spectral analysis indicates that similar data instances should have similar labels. Assume that S € R ^{nxn} is the similarity matrix based on X, obtained through a RBF kernel in this work as,

$$S_{i,j} = e^{\frac{-\|x_i - x_j\|^2}{\sigma^2}}$$

Force the pseudo-class labels of similar data instances to be close to each other. Therefore the constraint from attribute-value part can be formulated to minimize the following term:

$$\frac{1}{2}\sum_i\sum_j S(i,j)\left\|Y(:,i) - Y(:,j)\right\|_2^2 =$$

$$\frac{1}{2}\sum_k\sum_i\sum_j S(i,j)\left(Y(k,i) - Y(k,j)\right)^2$$

$$= \sum_k Y(k,:)(D_s - S)Y(k,:)$$

$$= Tr(YL_sY^T)$$

Where $L_S=D_S-S$ is a laplacian matrix and is a diagonal matrix with its element defined as

$$D_s(i,i) = \sum_{j=1}^{n} S(i,j)$$

Discriminative Analysis for Attribute-Value Information

Another advantage from the introducing of the pseudo class label is that it allows us to exploit local discriminative information for attribute-value part. For each data point p_j obtain its K_0 nearest neighbors, denoted as $U^j=\{j_1,j_2,\ldots j_n\}$. Let $N^j=[x_j,x_{j1},\ldots..,x_{jk}]$ and $Y^j=[Y(:;j),\ldots\ldots Y(:;j_k)]$ be the local data and local label matrix, respectively.

IV. EXPERIMENTATION RESULTS

In this section, present experiment details to verify the effectiveness of the proposed framework, LUFS. After introducing real-world social media data sets, first compare the four variants of LUFS, and evaluate the effectiveness of LUFS in terms of clustering performance, then study the effects of parameters on performance and finally further verify the constraint extracted from link information by social dimension. The clustering quality is evaluated by two commonly used metrics, accuracy and Normalized Mutual Information (NMI).

DATASETS

Collect two data sets from real-world social media websites, i.e., BlogCatalog and Flickr, which are the subsets of two public available data sets to uncover overlapping groups in social media. Both websites allow users to provide tags to indicate their interests, considered as the features; and users can also subscribe to some interest groups, used as the ground truth as the class labels in this work. In addition, users can follow other users if they share similar interests, forming link information. Compute the number of links for each instance and the distributions. The first observation is that most instances have a few links, while a few instances have an extremely high number of links. These distributions suggest a power law distribution that is typical in social networks. Some statistics of the data sets are shown in Table 1: Flickr has a denser network, indicating by its higher average degree and higher clustering coefficient. Note that "# of Classes" in Table 1 denotes the number of ground-truth classes and the ground-truth classes are only used for the evaluation purpose.

Table 1
Statistics of the datasets

Size	BlogCatalog	Flickr
# of features	5198	7575
# of classes	8189	12047
# of links	27965	47344
# Average of degree	5.38	6.25
Clustering coefficient	0.1224	0.3301

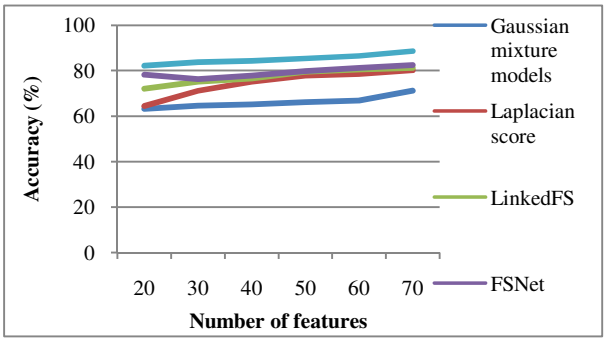

Fig. 1 The comparison of variants of LUFS and methods in Flickr

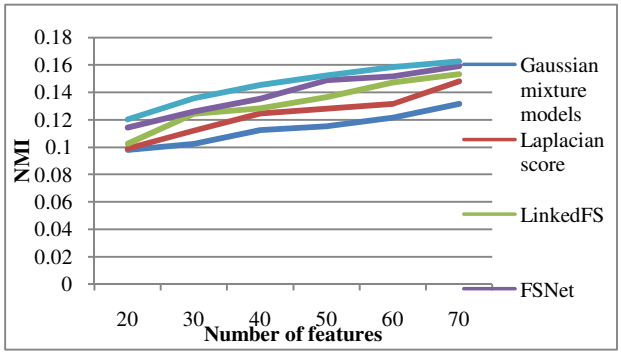

Fig. 2 The comparison of variants of LUFS in Flickr

For the parameters in the proposed algorithms, try different parameter values and report the best performance. More details about parameter analysis will be discussed in the later sections. The comparison results are demonstrated in Fig. 1 and 2 for Flickr with accuracy and NMI values, respectively. In general, with the increase of the number of selected features, the performance trends to increase first and then decrease. When the number of selected features is smaller, lose too much information, while the number of selected features is larger, may introduce noisy features. Algorithms reach their best performance with smaller numbers of selected features in Flickr.

V. CONCLUSION AND FUTURE WORK

Linked data in social media presents new challenges to traditional feature selection algorithms, which assume the data instances to be independent and identically distributed. In particular, social dimension regularization is based on a recent developed concept of social dimensions from social network analysis to extract relations among linked data as groups and defined to mathematically model these relations. Utilize graph regularization and social dimension regularization to capture the individual and group behaviors of linked instances, separately algorithm the concept of pseudo-class labels to guide extracting constraints from link information and attribute- value information, resulting in a new Linked Unsupervised based feature selection framework (LUFS), for linked social media data. LUFS significantly improves the performance of feature selection by incorporating these relations into feature selection. There are many issues needing further investigation for linked data such as handling noise, incomplete and uncertain linked social media data. When compared to all of methods proposed framework LUFS are distinctively different from methods Gaussian mixture models, Laplacian score, LinkedFS and FSNet. First, Gaussian mixture models, Laplacian score, LinkedFS and FSNet are semi supervised and supervised methods, respectively and they use label information, while LUFS is a unsupervised feature selection algorithm. Second, Gaussian mixture models, Laplacian score, LinkedFS and FSNet exploit relations individually while LUFS employs relations as groups via social dimensions.

REFERENCES

[1]. Tang, J., & Liu, H. (2014). An unsupervised feature selection framework for social media data. Knowledge and Data Engineering, IEEE Transactions on, 26(12), 2914-2927.

[2]. H. Liu and H. Motoda,(2008) Computational Methods of Feature Selection. London, U.K. Chapman & Hall

[3]. H. Liu and L. Yu, (Apr. 2005) "Toward integrating feature selection algorithms for classification and clustering," IEEE Trans. Knowl. Data Eng., vol. 17, no. 4, pp. 491–502.

[4]. Huan Liu and Hiroshi Motoda. (2007) Computational methods of feature selection. CRC Press.

[5]. Zheng Zhao, Lei Wang, and Huan Liu. (July 11-15, 2010) Efficient spectral feature selection with minimum redundancy. In Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2010, Atlanta, Georgia, USA.

[6]. Jiliang Tang, Salem Alelyani, and Huan Liu(2014) Feature selection for classification: A review. In Data Classification: Algorithms and Applications.

[7]. Jiliang Tang and Huan Liu.(2012) Unsupervised feature selection for linked social media data. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 904–912. ACM,

[8]. Salem Alelyani, Jiliang Tang, and Huan Liu.(2013) Feature selection for clustering: A review. In Data Clustering: Algorithms and Applications, pages 29–60. CRC Press.

[9]. Mingjie Qian and Chengxiang Zhai. Robust unsupervised feature selection. In Proceedings of the Twenty-Third international joint conference on Artificial Intelligence, pages 1621–1627. AAAI Press.

[10]. Zheng Zhao and Huan Liu(2013) Spectral feature selection for supervised and unsupervised learning. In Proceedings of the 24th international conference on Machine learning, pages 1151–1157. ACM.

[11]. Deng Cai, Chiyuan Zhang, and Xiaofei(2010) The Unsupervised feature selection for multi-cluster data. In Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, pages 333–342. ACM.

[12]. X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in Proc. Adv. Neural Inf. Process. Syst, pp. 507–514.

[13]. J. Dy and C. Brodley,(2006) "Feature selection for unsupervised learning," J. Mach. Learn. Res., vol. 5, pp. 845–889, 2004.
[14]. Z. Zhao and H. Liu,(2007) "Spectral feature selection for supervised and unsupervised learning," in Proc. 24th Int. Conf. Mach. Learn., pp. 1151–1157.
[15]. Constantinopoulos, C., Titsias, M. K., & Likas, A. (2006). Bayesian feature and model selection for Gaussian mixture models. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(6), 1013-1018.
[16]. Y. Yang, H. Shen, Z. Ma, Z. Huang, and X. Zhou, (2011)"L21-norm regularized discriminative feature selection for unsupervised learning," in Proc. 22nd Int. Joint Conf. Artif. Intell., pp. 1589–1594.
[17]. J. Tang and H. Liu,(2012) "Feature selection with linked data in social media," in Proc. 13th SIAM Int. Conf. Data Mining, pp. 118– 128.
[18]. P. Marsden and N. Friedkin, (1993) "Network studies of social influence," Sociol. Methods Res., vol. 22, no. 1, pp. 127–151.
[19]. M. McPherson, L. S. Lovin, and J. M. Cook,(2001) "Birds of a feather: Homophily in social networks," Annu. Rev. Sociol., vol. 27, pp. 415–444.
[20]. Tang, J., & Liu, H. (2014). An unsupervised feature selection framework for social media data. Knowledge and Data Engineering, IEEE Transactions on,26(12), 2914-2927.

Author Profile

T.Pradeepa received the MSC degree in Software Systems from Bharathiyar University, Coimbtaore, in 2013. Currently, she is working towards M.Phil Degree in the Department of Computer Science, Sri Ramakrishna college of Arts and Science for women. Her research interests include uncertain data management and data mining.