Unsupervised Feature Selection Algorithms: A Survey

P.Miruthula[1], S.Nithya Roopa[2]

[1]M.E – Computer Science and Engineering, Kumaraguru College of Technology, Coimbatore-641049

[2]Assistant Professor,Department Computer Science and Engineering, Kumaraguru College of Technology,Coimbatore-641049

Abstract: ***The prodigious usage of features and variables are very high in most of the domains which are often unwanted and noisy .Feature selection is a method which is used to handle high dimensional data into dataset with less dimensions by eliminating most unwanted or repetitive features or attributes . Unsupervised feature selection means that the best features are selected among the large set of unlabelled data. some of the unsupervised feature selection algorithms namely, Clustering guided sparse structural learning(CGSSL), The Linked unsupervised feature selection(LUFS), Unsupervised spatial-spectral feature selection method, Unsupervised feature selection via optic diffraction principle, Joint embedded learning and sparse regression for unsupervised feature selection(JELSR). CGSSL is an iterative approach which integrates cluster analysis and sparse structural analysis and experimentally results are examined. The LUFS focuses on linked data to achieve linked information in selecting features and results are analyzed. The unsupervised spatial-spectral feature selection method bands are represented in prototype space and data are represented in pixel space and optimal features are obtained. Unsupervised feature selection method via optic diffraction principle based on the property of Fourier transform of probability density distribution. In JELSR embedding learning and sparse regression are fused and implemented. In this review paper, survey of above unsupervised feature selection algorithms are discussed***.

Keywords: Data mining**,** unsupervised learning, feature selection, clustering.

1. Introduction

Data mining is the discovery and evaluation of large sets of data in order to explore structured data and rules. Its main focus is to seek efficient way to unite the computer's capability to process the data with the human eye's competence to examine patterns. Data mining is the self contained part of extensive process called knowledge discovery from database. In data mining datasets are extracted using two learning methods (ie) supervised or unsupervised learning.

1.1 ***Supervised learning:*** It is the machine learning task of concluding a function from labelled training data. This method is otherwise called as directed data mining technique. In this technique the values of the dependent variable should be known for adequately large part of dataset.

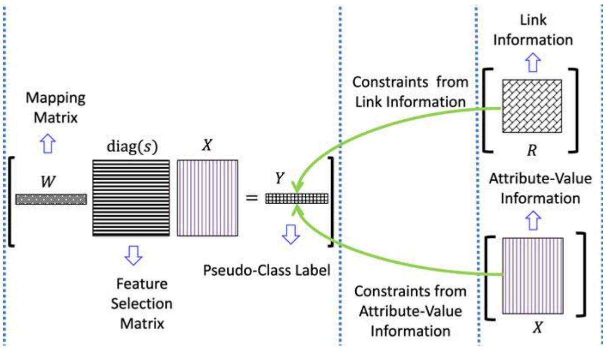
1.2 ***Unsupervised learning:*** It is the process of finding the hidden structure in the data which is not labelled. This technique is otherwise called undirected data mining. In this technique the target is achieved mostly by clustering process.

1.3 ***Unsupervised feature selection****:* It is the task of selecting the most relevant and best feature [6] among large sets of data which are not labelled. There are several algorithms used to extract features among unlabelled data. A good unsupervised selection algorithm gives the most relevant feature as a result.

2. Related Work

2.1 Linked Unsupervised Feature Selection (LUFS)

The vast usage of social media [1] leads to very large sets of unlabelled and datasets with high dimensions. Here we examine how the connections are originated from linked data can be used to select relevant features, for this purpose a new unsupervised feature selection algorithm LUFS for linked social media is used. Here we have to capture the linked information so that the linked objects are correlated and make us to use the correlation for future selection. In order to capture the dependency among linked information we use graph regularization and social dimension regularization. In graph regularization we have to make the labels of two linked data objects close to each other by minimizing the terms in graph regularization. In social dimension regularization here we integrate the linked data and attribute value data. Next is that we have to capture the attribute-value information using spectral analysis and discrimination analysis. Finally the set of most relevant features are obtained. The working of LUFS algorithm for unsupervised feature selection is diagrammatically shown below:



2.2 Clustering Guided Sparse Structural Learning (CGSSL):

The main theme is to select the most relevant or related features among huge sets of unlabelled data. The method is an iterative approach which used to obtain the above theme is by integrating cluster analysis and sparse structural analysis [2] into a combined framework. CGSSL Algorithm for feature selection:

1) Input data is given in the type of matrix called data matrix along with the procedure parameters
2) We have to construct the k-nearest neighbour graph using clustering and normalized matrix is also calculated.
3). Now the identity matrices are initialized and the iterative step is proceeded.
4) After each processing step the diagonal matrix is updated and this iteration is repeated until the convergence criteria is satisfied
5) Now all the relevant features are sorted in descending order and the features with top rank is given the first priority.

Finally the clustering results of various feature selection techniques are compared using datasets namely text data, handwritten data etc and proved that best result is obtained using CGSSL method. Hence the below equations gives the approximate optimization of this unsupervised feature selection algorithm.

$$\max \mathrm{Tr}[\mathbf{F}T\mathbf{X}T\mathbf{G}-1\mathbf{Q}(\mathbf{I}r-\gamma\mathbf{Q}T\mathbf{G}-1\mathbf{Q})-1\mathbf{Q}T\mathbf{G}-1\mathbf{XF}]$$
$$\Leftrightarrow \max \mathrm{Tr}[(\mathbf{I}r-\gamma\mathbf{Q}T\mathbf{G}-1\mathbf{Q})-1\mathbf{Q}T\mathbf{G}-1\mathbf{XFF}T\mathbf{X}T\mathbf{G}-1\mathbf{Q}]$$
$$\Leftrightarrow \max \mathrm{Tr}[(\mathbf{Q}T(\mathbf{I}d-\gamma\mathbf{G}-1)\mathbf{Q})-1\mathbf{Q}T\mathbf{TQ}] \Leftrightarrow \max \mathrm{Tr}[\mathbf{Q}T\mathbf{N}-1\mathbf{TQ}]$$

2.3 Unsupervised Spatial-Spectral Feature Selection Method

The main focus of this paper is to select the features that maintain the multi cluster structure of the multi spatial-spectral features. This information about multi cluster is gained through spectral clustering by utilizing the weighted fusion of multiple features. Each group of features is weighted by a fusion scheme in clustering. The multiple features are the spectral features and spatial features. In this design the group sparsity -based feature selection [3] model is evaluated and it is prolonged to the unsupervised case and this case is extended to multiple features of hyperspectral images.
Following steps are the procedure for the above method:
Input: large number of unlabelled samples

1) Different types of features are extracted from the hyperspectral images.
2) Laplacian matrix [7] , is calculated for each type of features
3) Weights of each feature is calculated
4) Diagonal matrix is calculated and updated for each iteration
5) As soon the convergence criteria is satisfied the most relevant selected features are shown as output

The kth feature can be selected by using the following equation
$\mathbf{f}k = \mathbf{argmax}\{c_2^j\}$ j= 1, 2, . . . ,B.After selecting the kth

feature further selection of best feature process is carried out.

2.4 Unsupervised Feature Selection Via Optic Diffraction Principle

The main theme of this algorithm is based on the property of Fourier transform of the probability density distribution. Each feature is calculated based on the concept of optic diffraction [4] which does not change under feature scaling. For each dataset the following steps are performed for 10 times

1) The dataset is randomly divided into five subsamples. Each subsample contains roughly same ratio of different classes.
2) Four subsamples are used as training data and the remaining as test data
3) Feature selection is performed using the dataset to be trained.
4) Each feature subset of the training data is classified for selecting features with high accuracy
5) Classification accuracy on the test set is measured using the selected features
6) The above steps are repeated to obtain the relevant features to be selected

2.4.1 Issues Regarding Discrimination Analysis Procedure

1) Connection between optic diffraction and discrimination analysis
2) Discrimination metric
3) Aperture representation
4) Basis orientation and computation analysis

Thus the relevant features are selected and further process can be carried out.

2.5 Joint Embedding Learning and Sparse Regression for Unsupervised Feature Selection *(JELSR)*

In this algorithm, a method is provided which uses the weight via local linear approximation and adding the normalization method [5] which jointly provides the effective algorithm to solve the optimization problem and some sightful discussions are made on convergence analysis, computational complexity and parameter determination, here two separate procedures namely manifold characterization and feature selection.

2.5.1 Procedure for JELSR:

1) Input: dataset, balanced parameter, neighbourhood size, dimensionality of embedding parameter, selected feature number.
2) Output: selected feature index set
3) Stage one: graph construction
    - Nearest neighbourhood graph is constructed
    - Similarity matrix and graph laplacian is constructed
4) Stage two: alternative optimization is done
5) Stage three: feature selection is done here

2.6 Comparison Table

S. no	*Name of the algorithm*	*Pros*	*Cons*
1.	LUFS	Provides linked information of the social media data in efficient way	less efficient for multi source datasets
2.	CGSSL	Provides label information for the structured learning in optimized form	Feature correlations are not investigated explicitly
3.	Unsupervised Spatial-Spectral Feature Selection Method	Best relevant features from hyper spectral image dataset are obtained with approximation	Not applicable for large datasets
4.	Unsupervised Feature Selection Via Optic Diffraction Principle	The notion of physical optics is used effectively for discrimination calculation of distribution	Sometimes depends on probability density estimation which requires future search for finding optimal solution
5.	JELSR	Provides standard perspective for understanding and comparison of many famous unsupervised feature selection algorithms	All possible extensions of various algorithm comparisons are not discussed

3. Conclusion

The complete aim of this unsupervised feature selection method is to select the best relevant feature set from large sets of unlabelled dataset. Different types of unsupervised feature selection algorithms are discussed along with their pros and cons. Various unsupervised feature selection algorithms are linked unsupervised feature selection (LUFS), clustering guided sparse structural learning (CGSSL), unsupervised spatial-structural feature selection method, unsupervised feature selection via optic diffraction principle, joint embedding learning and sparse regression for unsupervised feature selection. Each algorithms mentioned above are use in different scenarios to achieve the same target.

References

[1] Jiliang Tang and Huan Liu"Unsupervised feature selection framework for social media data"IEEE trans on knowledge engg and datamining., vol 26, no.12,Dec 2014.
[2] Zechao Li, Jing Liu, Yi Yang, Xiaofang Zhou, *Senior Member*, *IEEE*, and Hanqing Lu, *Senior Member,IEEE"* Clustering-Guided Sparse Structural Learning For Unsupervised Feature Selection"IEEE trans on knowledge engg and data mining., vol 26,sept 2014.
[3] Haichang Li, Shiming Xiang, Zisha Zhong, Kun Ding, and Chunhong Pan" Multicluster Spatial–Spectral Unsupervised Feature Selection for Hyperspectral Image Classification" IEEE Geoscience and remote sensing letters mar 2014.
[4] Praisan Padungweang, Chidchanok Lursinsap, and Khamron Sunat" A Discrimination Analysis for UnsupervisedFeature Selection via Optic Diffraction Principle" IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, VOL. 23, NO. 10, OCT 2012.
[5] Chenping Hou, *Member*, *IEEE*, Feiping Nie, Xuelong Li, *Fellow*, *IEEE*, Dongyun Yi, and Yi Wu" Joint Embedding Learning and Sparse Regression:A Framework for Unsupervised Feature Selection" IEEE TRANSACTIONS ON CYBERNETICS, VOL. 44, NO. 6, JUNE 2014
[6] Mihaela Breaban _, HenriLuchian Faculty of ComputerScience,AlexandruIoanCuzaUniversity,Iasi, Romania**"**A unifying criterion for unsupervised clustering and feature selection" Pattern Recognition 44 (2011) 854–865
[7] Satoshi Niijima and Yasushi Okuno" Laplacian Linear Discriminant Analysis Approach to Unsupervised Feature Selection" IEEE/ACM TRANSACTIONS ON COMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 6, NO. 4, OCTOBER-DECEMBER 2009.

Author Profile

P. Miruthula, ME (Computer Science & Engineering) Kumaraguru College Of Engineering and Technology, India.

S. Nithya Roopa, Assistant Proffessor (Department of Computer Science & Engineering) Kumaraguru College of Engineering and Technology, India