Asian Journal of Engineering and Applied Technology (AJEAT)
Vol.3.No.1 2015 pp1-5.
available at: www.goniv.com
Paper Received : 08-04-2015
Paper Accepted: 20-04-2015
Paper Reviewed by: 1. R. Venkatakrishnan 2. R. Marimuthu
Editor : Prof. P.Muthukumar

FAST CLUSTERING ALGORITHM INTEGRATES CLUSTER ANALYSIS AND SPARSE STRUCTURAL LEARNING- AN EFFECTIVE UNSUPERVISED FEATURE SELECTION

[1]G.MANJULA., [2]A.KUMARESAN., [3]K.VIJAYA KUMAR
Pg Scholar(CSE)., Hod In Dept Of CSE., Asst. Professor (CSE)
S.K.P.Engineering College, Tiruvannamalai, Tamilnadu

ABSTRACT
The representation of high dimensional data in data mining and pattern analysis is often accompanied by noise and redundancy. Hence Feature selection is the best technique for dimensionality reduction.The proposed unsupervised algorithm,clustering-guided sparse structural learning (CGSSL), integrates cluster analysis and sparse structural analysis.The development of Nonnegative spectral clustering produce more accurate cluster labels of the input samples. Prediction of the cluster labels by exploiting the latent structure shared by different features, uncovers feature correlations and is reliable.Row-wise sparse models are leveraged to make the proposed model suitable for feature selection, along with an iterative algorithm. Finally, extensive experiments are conducted on 12 diverse benchmarks, including face data, handwritten digit data, document data, and biomedical data which improves the efficiency and effectiveness of the feature selection.
Index Terms—Feature selection, nonnegative spectral clusteringsparsity,latent structure, row-sparsity.

INTRODUCTION

The number of features is probably high in domains, such as image and videounderstanding, and data mining [1].Not all the features are important and discriminative,since most of them are often interrelatedor redundant to each other, and sometimes noisy . And results in over-fitting, low-efficiency and poor performance to the traditional learning models [2]. The chore of selecting the "best" feature subset is known as*feature selection*, awidely used techniques for pattern analysis and data mining [3].

These algorithms can be categorized as supervised algorithms,semi-supervised algorithms and unsupervisedalgorithms according to the utilizinglabel information. Supervised algorithms usually fail with eitherinadvertentlyremoving many relevant features or selecting irrelevant features. Therefore, semi-supervised feature selection is developed to exploit both labeled and unlabeled data simultaneously. Since labels are expensive it is quite demanding to develop unsupervised feature selection techniques [4]. In this paper, we propose a novel unsupervised feature selection algorithm, namely Clustering-Guided Sparse Structural Learning (CGSSL), which integrates clusteranalysis and structural analysis into a joint framework. To select discriminative features, nonnegative spectral clustering is proposed.

We propose an unsupervised feature selection framework by exploiting the cluster analysis andstructural analysis with sparsity simultaneously. An effective and efficient algorithm is developed to solvethe proposed formulation.We develop nonnegative spectral analysis to learn more accurate cluster indicators by magnificent nonnegative and orthogonal constraints.We exploit the hidden structure shared by different features to predict the cluster indicators.

To facilitatefeature selection, the sparse feature selection models are exerted on the regularization ter

THE PROPOSED FRAMEWORK

Consider an arbitarymatrix $A \in R^{r \times t}$, a^i means the *i*-th row vector of **A**, A_{ij} denotes the *(i, j)*-th entry of **A**, $\|\mathbf{A}\|_F$ is Frobenius norm of **A** and Tr[**A**] is the trace of **A** if **A** is square. The $l_{2,1}$-norm is defined as;

$$\|\mathbf{A}\|_{2,1} = \sum_{i=1}^{r} \sqrt{\sum_{j=1}^{t} A_{ij}^2}.$$

Assume that we have n samples $X = \{\mathbf{x}_i\}_{i=1}^n$. Let $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$ denote the data matrix, in which $\mathbf{x}_i \in R^d$ is thefeature descriptor of the *i*-th sample. Suppose these n samplesare sampled from c classes. Denote $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]^T \in \{0, 1\}^{n \times c}$, where $\mathbf{y}_i \in \{0, 1\}^{c \times 1}$ is the cluster indicator vectorfor $\mathbf{x}_i$. That is, $Y_{ij} = 1$ if the sample $\mathbf{x}_i$ is assigned to the *j*-thcluster, and $Y_{ij} = 0$ otherwise. Clustering techniques are used to guide the process of structural learning.

Meanwhile, the pseudo class labels are also predicted by the structural learning with predictive functions, which compare the samples and the pseudo class labels. To conduct effective feature selection, we impose the sparse feature selection models on the regularization term. By our framework;

$$\min_{\mathrm{F},h} \mathcal{J}(\mathrm{F}) + \sum_{i=1}^{c} \left(\alpha \sum_{j=1}^{n} l(h_i(\mathrm{x}_j), \mathrm{f}_i) + \Omega(h_i) \right) \quad \text{s.t.} \quad \mathrm{F} = \mathrm{Y}(\mathrm{Y}^T\mathrm{Y})^{-\frac{1}{2}},$$

Algorithm 1: CGSSL for Feature Selection

Input:

Data matrix $X \in R^{d \times n}$; Parameters α, β, γ, λ, *k*, *c*, *r* and *p*

1: Construct the *k*-nearest neighbor graph and calculate L;

2: The iteration step $t = 0$; Initialize $F_0 \in R^{n \times c}$ and set $D_0 \in R^{d \times d}$ as an identity matrix;

3: repeat

4: $G_t = \alpha XX^T + \beta D_t + \gamma I_d$;

5: $N_t = I_d - \gamma G_t^{-1}$;

6: Tt= G−t 1XFtFTt XTG−t 1;

7:9:8:Obtain

MHtt==GLtQ+−t+αγ1IQby−t+α1theQ2XTt+

eigen-decompositionT1H; −t 1X;of N_t^{-1}Tt;n(λFt)i

10:(Ft+1)=ij= (F1tXF)ij(tM+1tF;t+λFtFtTFt)ij;

11: W_{t+1} H−t

12: Update the diagonal matrix D as

$$D_{t+1} = \begin{bmatrix} \frac{1}{2\|(\mathbf{w}_{t+1})_1\|_2} & & \\ & \cdots & \\ & & \frac{1}{2\|(\mathbf{w}_{t+1})_d\|_2} \end{bmatrix};$$

13: t=t+1;

14: until Convergence criterion satisfied Output: Sort all *d* features according to $\|(w_t)_i\|_2$ in descending order and select the top *p* ranked features.

Nonnegative Spectral Clustering

From various graph-theoretic methods, **spectral clustering** has been verified to be effective to detect the cluster structure ofdata and has received significant research attention . The local geometrical structure can be exploited by

$$\min_{\mathbf{F}} \frac{1}{2} \sum_{i,j=1}^{n} S_{ij} \| \frac{\mathbf{f}_i}{\sqrt{E_{ii}}} - \frac{\mathbf{f}_j}{\sqrt{E_{jj}}} \|_2^2 = \mathrm{Tr}[\mathbf{F}^T \mathbf{L} \mathbf{F}],$$

According to the definition of **F**, its elements are constrained to be discrete values, making the problem anNP-hard problem. A well-known solution

$$\min_{\mathbf{F},h} \mathrm{Tr}[\mathbf{F}^T \mathbf{L} \mathbf{F}] + \sum_{i=1}^{c} \left(\alpha \sum_{j=1}^{n} l(h_i(\mathbf{x}_j), \mathbf{f}_i) + \Omega(h_i) \right) \quad \text{s.t.} \quad \mathbf{F}^T \mathbf{F} = \mathbf{I}_c.$$

When both nonnegative and orthogonal constraints are satisfied, only one element in each row of **F** is greater than zero and all of the others are zeros, which makes the results more appropriate for clustering.

Sparse Structural Analysis

The experiments are conducted on 12 publicly available datasets.

Data Sets

TABLE 1
Dataset Description

Domain	Dataset	n	d	c
Image, Face	UMIST	575	644	20
	JAFFE	213	676	10
	Poingting4	2790	1120	15
Image, Handwritten Digits	MNIST	5000	784	10
	BA	1404	320	36
	USPS	400	256	10
Text	WebKB	814	4029	7
	tr11	414	6429	9
	oh15	913	3100	10
Microarray, Bio	TOX-171	171	5748	4
	Tumors9	60	5726	9
	Leukemia1	72	5327	3

Compared Scheme

The compared algorithms are enumerated as follows.

1) Baseline: All original features are adopted;

In our framework, the features which are most discriminative to the pseudo class labels are selected. For simplicity, we assume that the shared structure is a concealed low-dimensional subspace in this work. Therefore, the original data features together with the features in the low-dimensional subspace are both used to predict the pseudo labels. To make the problem tractable, the orthogonal constraint **Q***T***Q** = **I***r*is imposed . Denote **V** =[**v**1, . . . , **v***c*] ∈**R***d*×*c*and **P** = [**p**1, . . . , **p***c*] ∈**R***r*×*c*. Thus our formulation becomes;

$$\min_{\mathbf{V},\mathbf{W},\mathbf{Q},\mathbf{F}} \mathrm{Tr}[\mathbf{F}^T \mathbf{L} \mathbf{F}] + \alpha l(\mathbf{W}^T \mathbf{X}, \mathbf{F}) + \Omega(\mathbf{V}, \mathbf{W}) \quad \text{s.t.} \quad \mathbf{F}^T \mathbf{F} = \mathbf{I}_c,\ \mathbf{F} \geq 0;\ \mathbf{Q}^T \mathbf{Q} = \mathbf{I}_r.$$

EXPERIMENTS

The performance of the proposed formulation, which can be applied to many applications, such as clustering and classification. We first select the top *p* features and then utilize Kmeans algorithm to cluster samples based on the selected features.

2) *MaxVar:* Features corresponding to the maximum variance are selected to obtain the best expressive features;

3*) LS* [5]: Features consistent with Gaussian Laplacian matrix are selected to best preserve the local manifold structure [21];

4) *SPEC* [6]: Features are selected using spectral regression;

5) *SPFS-SFS* [7]: The traditional forward search strategy is utilized for similarity preserving feature selection in the SPFS framework.

6) *MCFS* [8]: Features are selected based on spectral analysis and sparse regression problem;

7) *UDFS* [9]: Features are selected by a joint framework of discriminative analysis and 2,1-norm minimization.

8) *NDFS* [10]: Discriminative features are selected by a joint framework of nonnegative spectral analysis and linear regression with 2,1-norm regularization.

9) *CGSSL:* The proposed Cluster-Guided Sparse

Structural learning framework.

Table:2

Clustering Results Comparison on the Biomedical Data Sets

Dataset	ACC± std (%)								
	Baseline	MaxVar	LS	SPFS-SFS	SPEC	MCFS	UDFS	NDFS	CGSSL
TOX-171	40.9±3.6	41.8±2.7	41.8±1.5	41.7±5.0	44.8±2.7	42.7±3.1	45.7±2.0	47.4±2.5	**48.6±3.4**
Tumors9	40.2±3.3	41.8±5.2	42.3±3.6	42.6±4.6	41.8±3.4	41.6±4.9	42.9±4.3	45.4±3.7	**46.8±4.6**
Leukemia1	56.7±8.2	56.5±11.6	70.2±7.2	70.3±6.4	60.2±4.7	70.6±10.1	71.3±10.6	72.1±7.9	**74.7±6.7**
	NMI± std (%)								
TOX-171	12.7±4.0	12.4±2.6	15.5±0.8	12.2±5.4	18.1±1.8	13.4±2.6	19.3±2.2	22.6±3.3	**25.7±2.8**
Tumors9	38.2±4.9	40.3±2.8	41.9±3.3	42.0±3.4	38.7±2.5	40.3±6.0	42.3±4.2	43.9±3.8	**44.4±3.4**
Leukemia1	21.6±1.6	27.2±18.3	34.0±11.1	34.0±11.7	31.1±3.2	30.0±11.6	40.8±12.0	43.6±13.8	**48.5±8.9**

The best results are highlighted in bold

Table:3

Clustering Results Comparison on the Face Data Sets

Dataset	ACC± std (%)								
	Baseline	MaxVar	LS	SPFS-SFS	SPEC	MCFS	UDFS	NDFS	CGSSL
UMIST	41.8±2.7	45.8±1.8	45.9±2.9	44.8±3.3	47.6±3.0	46.5±3.6	48.6±3.7	51.3±3.8	**53.4±3.1**
JAFFE	72.5±9.2	67.3±5.8	70.0±7.6	71.6±9.4	76.9±7.2	78.6±9.1	76.7±7.1	82.2±8.1	**82.3±7.5**
Pointing4	35.9±2.2	44.0±2.8	37.1±1.6	37.4±1.3	38.6±2.3	46.2±2.8	45.1±3.4	48.9±3.2	**51.3±2.6**
	NMI± std (%)								
UMIST	62.3±2.3	63.5±1.5	63.9±1.8	63.2±3.3	65.3±2.0	66.7±1.9	67.3±1.0	68.7±2.3	**70.9±2.2**
JAFFE	80.8±1.7	78.3±4.2	78.4±7.8	82.1±4.9	82.6±5.8	83.4±5.8	82.3±6.5	86.3±7.1	**87.5±5.1**
Pointing4	41.7±3.4	50.8±1.8	42.7±1.2	43.4±3.4	48.5±1.8	53.1±1.1	52.4±1.7	56.4±1.3	**57.7±1.3**

Conclusion

In this paper, we propose a novel unsupervised feature selection approach, which jointly exploits nonnegative spectralanalysis and structural learning with sparsity. The nonnegative spectral clustering provide label information for the structural learning. The 2,1-norm regularization, our methods jointly selects the most discriminative features across the entire feature space. Extensive experiments on 12 real-world data sets are conducted to validate the effectiveness of theproposed method. Besides, how to select the adaptive hyper-parameters and the number of selected features are also our directions for future research.

Parameter Setting

There are some parameters to be set in advance. For LS,SPEC, MCFS, UDFS, NDFS and CGSSL, we set k = 5 for

REFERENCES:

[1] A. Jain and D. Zongker, "Feature selection: Evaluation, application, and small sample performance," IEEE Trans. Pattern Anal.Mach. Intell., vol. 19, no. 2, pp. 153–158, Feb. 1997.

[2] L. Wolf and A. Shashua, "Feature selection for unsupervised and supervised inference: The emergence of sparsity in a weightbasedapproach," J. Mach. Learn. Res., vol. 6, pp. 1855–1887, Nov. 2005.

[3] G. Forman, "An extensive empirical study of feature selection metrics for text classification," J. Mach. Learn. Res., vol. 3, pp. 1289–1305, Mar. 2003

[4] P. Mitra, C. A. Murthy, and S. K. Pal, "Unsupervised feature selection using feature similarity," IEEE Trans. Pattern Anal. Mach.Intell., vol. 24, no. 3, pp. 301–312, Mar. 2002.

[5] X. He, D. Cai, and P. Niyogi, "Laplacian score for feature selection," in Proc. Adv. NIPS, 2005.

[6] Z. Zhao and H. Liu, "Spectral feature selection for supervised and unsupervised learning," in Proc. Int. Conf. Mach. Learn., Corvallis,OR, USA, 2007.

[7] Z. Zhao, L. Wang, H. Liu, and J. Ye, "On similarity preserving feature selection," IEEE Trans. Knowl. Data Eng., vol. 25, no. 3,pp. 619–632, Mar. 2013.

[8] D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data," in Proc. ACM SIGKDD Int. Conf. KDD,Washington, DC, USA, 2010.

[9] Y. Yang, H. T. Shen, Z. Ma, Z. Huang, and X. Zhou, "_2,1-norm regularized discriminative feature selection for unsupervisedlearning," in Proc. 22nd IJCAI, 2011.

[10] Z. Li, Y. Yang, J. Liu, X. Zhou, and H. Lu, "Unsupervised feature selection using nonnegative spectral analysis," in Proc. Conf.AAAI, 2012.