Maximum Margin Clustering without Nonconvex Optimization: an Equivalent Transformation

Y. Kang[1,2] , Z.Y. Liu[1,3], W. P. Wang[1] & D. Meng[1]

[1]*Institute of Information Engineering, Chinese Academy of Sciences, China*

[2]*University of Chinese Academy of Sciences*

[3]*School of Software, Beijing Institute of Technology, China*

ABSTRACT: On account of the promising performance in accuracy, maximum margin clustering (MMC) has attracted attentions from many research domains. MMC derived from the extension of support vector machine (SVM). But due to the undetermined labeling of samples in dataset, the original optimization is a nonconvex problem which is time-consuming to solve. Based on another high-quality nonlinear clustering technique–spectral clustering, this paper discusses an equivalent transformation of MMC into spectral clustering. By virtue of the establishment of equivalent relation between MMC and spectral clustering, we search for a simplified spectral clustering based method to solve the optimization problem of MMC efficiently, reducing its computational complexity. Experimental results on real world datasets show that the clustering results of MMC from the equivalent transformed spectral clustering method are better than any other baseline algorithms in comparison, and the reduced time consuming makes this advanced MMC more scalable.

KEYWORD: Maximum margin clustering; Spectral clustering; Kernel machine

1 INTRODUCTION

Clustering aims at mining dominant structures of a given dataset and grouping identical samples into the same cluster [1]. In past years, lots of clustering algorithms have emerged, of which maximum margin clustering (MMC) and spectral clustering are two methods with higher computational accuracy. MMC is an extended clustering technique which originated from the basic concept of support vector machine (SVM). In essence, MMC is to find the maximum margin hyperplanes between any two divisive clusters. Moreover, by the projection of feature space based on kernel machine, MMC has the same capability of analyzing nonlinear divisible problems with spectral clustering. But the label identification of samples makes the optimization of MMC to be a time-consuming nonconvex problem [1], which limits the scalability of MMC.

Inspired by some relevance analysis between SVM and spectral clustering, this paper firstly discusses an equivalence relation between two seemingly different clustering methods: kernel based maximum margin clustering (KMMC or MMC) and spectral clustering. By virtue of this equivalent transformation, this paper proposes a simplified spectral clustering based method to search for a global optimal labeling of data samples for MMC, reducing the entire computational complexity.

2 RELATED WORK

Maximum margin clustering based on kernel machine is a generalization of the MMC algorithm. By implicitly mapping data to a high-dimensional feature space, MMC can find nonlinear divisible clusters in the input space. MMC was proposed by Xu et al. [1] to apply to the unsupervised scenario Initially. Different from SVM, the optimization of MMC is a nonconvex integer optimization problem. Thus Xu et al. [2] used a semidefinite programming to tackle this problem. But the huge computational complexity limits the application of MMC. To transform the nonconvex property of MMC, Wang et al. [3] adopted an cutting-lane for MMC and use a series of constrained convex-concave procedures to relax the nonconvex limitation (CPM3C). However, CPM3C is prone to get stuck in local minima.

On the other hand, spectral clustering has been efficiently used for many nonlinear clustering. But due to the discrete optimization, spectral clustering is intractable. Relaxation and rounding strategy is utilized to convert the original problem to a feasible and simplified optimization problem [4]. Moreover, to extend spectral clustering to the application in large-data analysis, the approximate methods and spectral embedding skills are proposed successively.

Based on the advantages of spectral clustering like global optimal solution, this paper generalizes a

result from Suykens [5] and firstly discusses a unifying mathematical connection between MMC and spectral clustering. By virtue of the equivalent reformulation, the optimizing process of MMC can be transformed into spectral clustering, avoiding the local minima and promoting the computational velocity of MMC without loss of accuracy.

3 MAXIMUM MARGIN CLUSTERING

Extending the theories of SVM to unsupervised learning scenarios, we obtain the two-class maximum margin clustering as follows [1]:

$$\min_{\boldsymbol{w},b,\boldsymbol{\xi},\boldsymbol{Y}} \ \frac{1}{2}\|\boldsymbol{w}\|^2 + C\boldsymbol{\xi}^T\boldsymbol{e} \tag{1}$$

Subject to: $y_i(W^T\phi(x_i)+b) \geq 1-\xi_i,\ \xi=(\xi_1,\xi_2,...,\xi_n),$ $\xi_i \geq 0,\ Y=(y_1,y_2,...,y_n),\ y_i=\{\mp 1\},\ -l \leq e^TY \leq l, i=1,...,n$

MMC seeks an optimal labeling of *Y*, making the distance from support vectors to a hyperplane is maximum. The dual form of Eq.(1) is:

$$max_{\alpha,Y\gamma}(\alpha) = -\frac{1}{2}\sum_{i,\,j=1}^{n} y_i y_j K_{i,j}\alpha_i\alpha_j + \sum_{i=1}^{n}\alpha_i \tag{2}$$

Subject to: $\sum_{i=1}^{n}\alpha_i y_i = 0,\quad 0 \leq \alpha_i \leq C,\ y_i=\{\mp 1\},$ $-l \leq e^TY \leq l, i=1,...,n$

Where $-l \leq e^TY \leq l$ is a balance constraint to prevent all samples from assigning to one cluster. Because $\boldsymbol{Y}$ is uncertain, the binary integer programming induces MMC to be a non-convex optimization problem which is hard to settle.

4 EQUIVALENT TRANSFORMATION OF MMC INTO SPECTRAL CLUSTERING

4.1 *Transformation of MMC*

We consider to add extra n slack variables $\{\eta_i\}_{i-1}^{n}$ into the inequality constraint in Eq.(1) to obtain:

$$y_i(W^T\phi(x_i)+b) = 1-\xi_i+\eta_i$$

$$\xi_i \geq 0, \eta_i \geq 0, i=1,...,n \tag{3}$$

Let $\theta=(\theta_1,\theta_2,...,\theta_n)$, $\theta_i=\xi_i-\eta_i$ are composite slack variables. When encountering the equality constraint in Eq.(3), we can solve the optimization of two-class MMC in linear systems [5]:

$$\mathcal{H}(\boldsymbol{w},b,\boldsymbol{\theta},\boldsymbol{Y}) = \min_{\boldsymbol{w},b,\boldsymbol{\theta},\boldsymbol{Y}} \ \tfrac{1}{2}\|\boldsymbol{w}\|^2 + \tfrac{C}{2}\textstyle\sum_{i=1}^{n}\theta_i{}^2 \tag{4}$$

Subject to: $y_i(W^T\phi(x_i)+b) = 1-\theta_i$, $\theta=(\theta_1,\theta_2,...,\theta_n)$, $Y=(y_1,y_2,...,y_n)$, $y_i=\{\mp 1\}$, $-l \leq e^TY \leq l$, $i=1,...,n$

We solve this optimization by partial derivation on constructed Lagrangian, the obtained matrix is:

$$\begin{bmatrix} I & 0 & 0 & -X^T \\ 0 & 0 & 0 & -Y^T \\ 0 & 0 & CI & -I \\ X & Y & I & 0 \end{bmatrix} \begin{bmatrix} w \\ b \\ \theta \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ \vec{1} \end{bmatrix} \tag{5}$$

Where *w* is normal vector of cutting hyperplane $f(x)=W^T\phi(x)+b$, $X=(\phi(x_i)^T y_i)_{i=1}^{n}$, $\vec{1}=(1,1,...,1)$, and $\alpha=(\alpha_1,\alpha_2,...,\alpha_n)$ is Lagrangian multipliers. Eliminate *w* and θ , Eq.(5) is converted to:

$$\begin{bmatrix} 0 & Y^T \\ Y & k+I/C \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ \vec{1} \end{bmatrix} \tag{6}$$

This is the dual form of Eq.(4), where $k=XX^T=y_iy_j\phi(x_i)^T\phi(x_j)$ is the result of application of Mercer's condition.

4.2 *Equivalence to Spectral Clustering*

Motivated by the theories of kernel PCA, there is a link between two-class spectral clustering and MMC in Eq.(6). Let's consider lemma 1:

Lemma 1 [6]. *Based on a given positive semi-definite kernel matrix* $K(x_i,x_j)=\phi(x_i)^T\phi(x_j)$, *the Karush-Kuhn-Tucker optimality conditions of the original optimization in Eq.(6) are satisfied by each Lagrangian-multiplier vector* α *(or eigenvector):*

$$K_\alpha = \lambda_\alpha \tag{7}$$

where $\lambda = n/C$ *is corresponding eigenvalue, C is a regulation constant, the bias term* $b=-\frac{1}{n}\vec{1}^T K\alpha$ *induces the kernel matrix to be centered.*

Lemma 1 constructs an equivalent transformation of MMC into spectral. We search for an optimal two-class labeling of samples by spectral clustering technique implicitly, which induces the maximum margin between support vectors and hyperplane.

4.3 *Multi-Class Extending Formulation*

In light of the extension of two-class MMC to multi-class MMC, it is easy to deduce an equivalent transformation from multi-class MMC to multi-class spectral clustering. The formulation is:

$$\mathcal{H}(\boldsymbol{w}_m,\theta_i^m,\boldsymbol{Y}) = \min_{\boldsymbol{w}_m,\theta_i^m,\boldsymbol{Y}} \ \tfrac{1}{2}\textstyle\sum_{m=1}^{k}\boldsymbol{w}_m\boldsymbol{w}_m^{\mathrm{T}} + \tfrac{1}{2}\sum_{m=1}^{k}\sum_{i=1}^{n}C_m\theta_i^{m2} \tag{8}$$

Subject to: $w_{yi}x_i - w_m x_i = e_i^m - \theta_i^m$, $i = 1,...,n$, $m = 1,...,k$

where w_m^T are cluster indicator vectors, $\{w_m\}_{m=1}^k$ denotes k different clusters respectively, and e_i^m is:

$$e_i^m = \begin{cases} 0 & if \ \ y_i = m \\ 1 & if \ \ y_i \neq m \end{cases} \quad (9)$$

In the sequel, let's consider lemma 2:

Lemma 2 [6]. *Based on a given positive semi-definite kernel matrix* $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$, *and the inverse of diagonal degree matrix* D^{-1}, *the Karush-Kuhn-Tucker optimality conditions of the original optimization in Eq.(8) are satisfied by the eigenvector* α^m *in the following eigen-function :*

$$D^{-1}F_D K\alpha^m = \lambda_m \alpha^m, \ \ m = 1,...,k \quad (10)$$

where C_m are regulation constants, $\lambda_m = 1/C_m$ are eigenvalue of α^m, and the definition of F_D is:

$$F_D = I - \frac{1}{\vec{1}^T D^{-1} \vec{1}} \vec{1}^T \vec{1}^T D^{-1} \quad (11)$$

The equivalence relation between multi-class MMC and spectral clustering is exhibited in Eq.(10), Identically, the outputted real value clustering results from Eq.(10) also need to be rounded into integer value for final cluster indicating. We rewrite the formulized multiclass MMC in Eq.(8) as follows:

$$\mathcal{H}(\boldsymbol{W}, \boldsymbol{\theta}^m, \boldsymbol{Y}) = \min_{\boldsymbol{W}, \boldsymbol{\theta}^m, \boldsymbol{Y}} \frac{1}{2} tr(\boldsymbol{W}\boldsymbol{W}^{\mathrm{T}}) + \frac{1}{2}\sum_{m=1}^k C_m f_m(\boldsymbol{\theta}^m) \quad (12)$$

Subject to: $(W_Y - W)X = e^m - \theta^m$, $m = 1,...,k$

where W_m and W_{yi} in Eq.(8) are the rows of $\boldsymbol{W}$ and W_Y respectively. Assume that we have the eigenvector solutions $E = (E_1, E_2, ..., E_l)$ of Eq.(10) in hand, according to Fan's theorem [7], the relaxed eigenvectors $\{E_t\}_{t=1}^l$ are arbitrary rotated which makes the solving procedure complicated. Back to the initial viewpoint, the aim of MMC is to minimize the classification loss [2] depicted as:

$$\epsilon(\boldsymbol{W}) = \frac{1}{k}\sum_{m=1}^k \langle \mathcal{C}_{\boldsymbol{W}}(\boldsymbol{x}) \neq y \rangle \quad (13)$$

Where $\mathcal{C}_{\boldsymbol{W}}(\boldsymbol{x}) = arg\max_m \{\boldsymbol{w}_m \boldsymbol{x}\}$ is a classifier generated by MMC with a matrix parameter *W*, (x, y) is a test point and y is a given referential label of *x*, $\langle * \rangle$ is a sign function. The minimal loss means the maximum margin. So to reduce eigenvector's rotation, we adopt a margin-based rounding scheme [8]:

$$\mathcal{S}(\boldsymbol{E}) = \frac{1}{n}\sum_{i=1}^n \sum_{i \neq t} \exp(y_i \neq y_t)$$
$$subject \ to: \ \boldsymbol{E}\boldsymbol{E}^{\mathrm{T}} = \boldsymbol{I}_{l-1} \quad (14)$$

Where $\sum_{i \neq k} \exp(y_i \neq y_m)$ denotes a surrogate loss. Meanwhile there exists a *l-1*-dimensional subspace spanned by the *l* eigenvectors of *E*. By virtue of the Taylor unfolding formula, we expand Eq.(14) into an approximate expression as:

$$\mathcal{Q}(\boldsymbol{E}) \approx (l-1) - \frac{l}{n}\sum_{i=1}^n \boldsymbol{g}_{y_i}^{\mathrm{T}} \mathcal{C}_{\boldsymbol{W}}(\boldsymbol{x}_i) + l^2 \sum_{i=1}^n \pi_i^{-1} \quad (15)$$

Where g_j is the *j*th column of $\mathrm{G} = [\boldsymbol{I}_{l-1} - \frac{1}{l}\vec{\mathbf{1}}_{l-1}\vec{\mathbf{1}}^{\mathrm{T}}{}_{l-1} - \frac{1}{l}\vec{\mathbf{1}}_{l-1}]$, π_i are user-defined weights of samples [10]. The rounded results from Eq.(15) are the final clustering results based on an equivalent transforming computation.

5 EXPERIMENT

5.1 *Datasets and Experiment Setup*

Normalized Cuts (NCC) [9] and Kernel K-means (KKM) [9] are two clustering methods solved by equivalent transforming into spectral clustering. We compare the performances of MMC with NCC and KKM by means of the implicit analysis based on spectral clustering, meanwhile name them separately as MMCSP, NCCSP, KKMSP. We also choose another two multiclass maximum margin clustering algorithms like MMMC [2], CPM3C [3] as baseline. All experimental datasets in Table 1 are collected from real world. Experiments execute on a Linux machine with 4Core 2.6GHz CPU and 4G main memory. All algorithms are implemented in Java.

Table 1. Description of datasets

Dataset	Size	Dimensions	Classes
Iris	150	4	3
Vote	435	16	2
UMIST	575	644	20
YALE-B	2414	1024	38
CMU PIE	3329	1024	68
Gene	384	17	5
Vowel	990	10	11
NIST	1000	256	10

All datasets have their own classes, which are used as the groundtruth. We measure the clustering-result quality of all algorithms by two criterion: cluster purity (CP) and clustering error (CE) [10]. Besides, we test CPU time on all datasets to contrast the time complexity among all clustering algorithms.

5.2 *Experimental Results and Analysis*

Table 2. Accuracy comparison of all algorithms by CP and CE (%)

Dataset	NCCSP		KKMSP		MMCSP
	CP	CE	CP	CE	CP
Iris	0.675	15.82	0.513	17.19	0.861
Gene	0.809	9.89	0.691	10.84	0.926
Vote	0.837	9.06	0.705	9.97	0.953
UMIST	0.507	17.21	0.426	20.01	0.811
Vowel	0.646	15.90	0.502	18.34	0.845
NIST	0.725	13.17	0.675	14.96	0.885
YALE	0.481	19.69	0.479	23.39	0.795
CMU	0.539	16.83	0.480	18.75	0.827
Dataset	MMMC		CPM3C		MMCSP
	CP	CE	CP	CE	CE
Iris	0.760	12.11	0.817	9.05	7.71
Gene	0.855	7.56	0.896	5.82	4.07
Vote	0.892	7.15	0.925	4.93	3.16
UMIST	0.615	14.78	0.728	11.39	9.90
Vowel	0.749	12.65	0.802	10.03	8.82
NIST	0.822	10.81	0.850	7.79	6.75
YALE	0.605	15.87	0.733	12.61	10.52

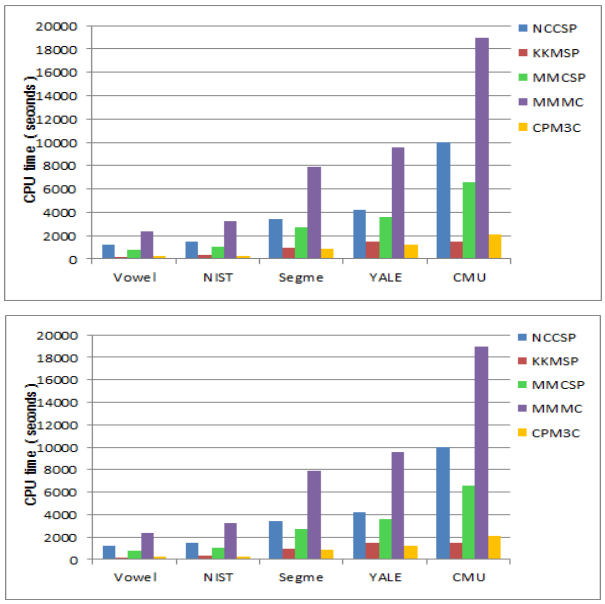

Fig. 1. Time complexity comparison of all algorithms on various datasets

The accuracy criterion CP and CE of five different algorithms are listed in Table 2. Compare the CP value of all algorithms on each dataset, the value of MMCSP is superior to any other algorithm. Meanwhile, the CE value of MMCSP is the smallest among five different algorithms on all datasets. Both of the results demonstrate that MMCSP excels any other algorithm in computational accuracy.

The CPU time of five different algorithms on all datasets is depicted in Figure 1, which consists of two divisive sub-graphs. Obviously, the CPU time of MMCSP is less than NCCSP and MMMC, but more than KMMSP and CPM3C. Without loss of accuracy, the consuming time of MMCSP has been reduced and superior to MMMC. Although the time complexity of KMMSP and CPM3C is lower than MMCSP, this advantage is gained by the expense of computational accuracy.

6 CONCLUSION

This paper discusses an equivalent transformation of MMC into spectral clustering. By virtue of the establishment of equivalence relation, the optimizing process of MMC can be transformed into a clustering process of spectral clustering. This simplified spectral clustering based MMC method avoids the local minima, and the reduced time consuming accelerates the entire computational velocity of MMC without loss of accuracy.

ACKNOWLEDGEMENTS

This work was supported by National Natural Science Foundation of China (61 272361), National Science-technology Support Plan Projects (2012BAH46B03), and "Strategic Priority Research Program" of the Chinese Academy of Sciences (XDA06030200)

REFERENCES

[1] Xu, L., Neufeld, J., Larson, B.et al.: Maximum margin clustering, Advances in NIPS, pp. 1537-1544, 2004.

[2] Xu, L., & Schuurmans, D.: Unsupervised and semi-supervised muliti-class support vector macine, Proc. NCAI, vol.2, pp.904-910, 2005.

[3] Wang, F., Zhao, B., & Zhang, C. S.: Linear time maximum margin clustering, IEEE Trans. on Neural Network, vol.21(2), pp.319-332, 2010.

[4] Bach, F. R., & Jordan, M. I.: Learning spectral clustering, with application to speech separation, Journal of Machine Learning, vol.7, pp.1963–2001, 2006.

[5] Suykens, J. A. K., & Vandewalle, J.: Least squares support vector machine classifiers, Neural Processing Letters, vol.9(3), pp.293-300, 1999.

[6] Alzate, C., & Suykens, J. A. K.: Multiway spectral clustering with out-of-sample extensions through weighted kernel PCA, PAMI, pp.335-347, 2010.

[7] Fan, K.: On a theorem of weyl concerning eigenvalues of linear transformations, Proc. Natl Acad Sci USA, vol.35(11), pp.652-655, 1949.

[8] Zhang, Z. H., & Jordon, M.: Multiway spectral clustering: a margin-based perspective, Statistical Science, vol.23(3), pp.384-403, 2008.

[9] Dhillon, I. S., Guan, Y., & Kulis, B.: Kernel k-means: spectral clustering and normalized cuts, Proc. SIGKDD, pp.551–556, 2004.

[10] Labatut, V.: Generalized measures for the evaluation of community detection methods, Journal of CoRR , 2013.