Semi-supervised Local Clustering

ANH PHAM THE, YOUNG-KOO LEE, SUNGYOUNG LEE

Department of Computer Engineering

Kyung Hee University

KOREA

{phamtheanh, sylee}@oslab.khu.ac.kr, yklee@khu.ac.kr

Abstract

Local learning for clustering (LLC) is getting a lot of attention these days in machine learning community. Local clustering can archive very good result in which a sample label can be smoothly estimated by its neighbor's label. The advantage of LLC is it can outperform global learning based techniques in accuracy point of view when dealing with the high dimensional data on nonlinear manifold. In this paper, we propose a novel semi-supervised framework which utilizes a few prior constraints for samples from expert to improve unsupervised local algorithm. The optimization problem is not only to minimize the smooth factor in local learning but also to avoid violating of must-link and cannot-link sample constraints factor from expert. Experiment shows the significant improvement of our approach when compare to some base line semi supervised methods.

***Keywords*:** Local clustering, Semi-supervised clustering, Constraint propagation.

1 Introduction

Data clustering is an important topic in machine learning, image analysis, information retrieval, and bioinformatics and it has a lot of real application in healthcare, surveillance, and manufacture domain. It indentifies the objects in the same group and the objects in different group by using the predefined similarity measure. There are some clustering algorithms have been proposed such as K-mean, spectral clustering [3] but they focus on the global structure of data without noticing the local structure of data. In these algorithms, one sample S's label has to be estimated based on '*all'* samples in dataset, even these samples are far away from S, or they are noisy and so may have different property with S. Recently, local learning for clustering [1] overcomes above limitation by using supervised idea for unsupervised problem. LLC tries to predict the sample's label based on its '*neighbors'* samples. So, the cluster result of LLC is as smooth as possible for all data samples. It is reported that local learning can get better result than global one.

In fact, local clustering is unsupervised method. It means that the algorithm indentifies the label of data samples by itself without the help of expert. Recently, many researchers show that, by using very small amount of expert information, the result of clustering algorithm can be improved a lot. This research direction is called semi supervised clustering. The additional expert information can be some labels for a few samples in dataset, or must link and cannot link constraints between some pairs of samples. In this paper, we use pair-wise constraint as expert information because it is cheaper to obtain from expert than label information. Furthermore, it can be, in some lucky situation, automatically derived from unlabeled data [4].

By seeing the advantages of local clustering and constraint based semi supervised clustering, we want to build an algorithm inherit the strength of both clustering framework. The clustering result must be smooth over dataset and the pair-wise constraint provided by expert can be spread out to local neighbors of each node in the pair. We provide the optimization based method which satisfies both smooth requirement of unsupervised local clustering and link constraints provided by expert. Experiment with digit data, news data shows the significant improvement of our method.

2 Problem Formulation

Assume that we have n data points $X = \{x_i\}_{i=1}^{n}$ .Every x_i is in R^d space and y_i is label of x_i.

In local learning clustering [1], assume that we have linear relation between label y_i and the information of sample x_i itself. So we want to minimize the difference between a samples label and their nearest neighbor's label:

$$\underset{w_i}{\text{Minimize}} J_i^* = \frac{1}{n_i} \sum_{x_j \epsilon N_i} (w_i^T x_j - y_j)^2 + \lambda \|w_i\|^2 \qquad (1)$$

In (1), we focus on N_i neighbors of i rather than all n samples in dataset at the same time. It makes local clustering work well with nonlinear manifold data as well as noisy data. The solution of (3) is for each i is:

$$w_i = (X_i X_i^T + \lambda\, n_i I)^{-1} X_i \boldsymbol{y_i} \qquad (2)$$

$\boldsymbol{y_i}$ is the real labels vector of neighbors of sample i. Now, come back to unsupervised problem when we don't know $\boldsymbol{y_i}$. We will predict y_i^* based on:

$$y_i^* = x_i^T w_i = x_i^T (X_i^T X_i + \lambda n_i \mathrm{I})^{-1} X_i \boldsymbol{y_i} \qquad (3)$$

Note that y_i^* is the single *'predicted'* label of data sample i *'itself'* and $\boldsymbol{y_i}$ is the *'true'* vector labels of *'neighbors'* of sample i. It means that $\boldsymbol{y_i}$ is a vector of label and y_i^* is a label. Based on this, we want to find y which is a vector label of all n samples in order to minimize the sum of the difference between predicted label y_i^* and true label y_i for each i.

$$J = \sum_{i=1}^{n} (x_i^T (X_i^T X_i + \lambda n_i \mathrm{I})^{-1} X_i \boldsymbol{y_i} - y_i)^2$$
$$= \|M\boldsymbol{y} - \boldsymbol{y}\|^2 = \boldsymbol{y^T}(M-I)^T(M-I)\boldsymbol{y}$$
$$= \mathbf{y}^{\mathrm{T}} \mathrm{A}\, \mathbf{y}$$

So at the end the objective function becomes:

$$\underset{y}{\text{Minimize}}\ J = \boldsymbol{y}^T A \boldsymbol{y} \quad (4)$$

In above equation, $\boldsymbol{y}$ is label vector for all data samples. M and A are matrixes which satisfies:

$$M_{ij} = \alpha_j^i \ if\ x_j \epsilon N_i\ ,\ \alpha^{\mathrm{i}} = x_i^T (X_i^T X_i + \lambda n_i \mathrm{I})^{-1} X_i \text{ for each i}$$
$$= 0\ otherwise$$
$$A = (M-I)^T (M-I).$$

3 Problem Solution

Now, our contribution is to build a semi-supervised local clustering method not only satisfy the smoothness of unsupervised clustering mentioned above, but also satisfy the constraint requirement by the expert.

The expert information is under the link form. Must link constraint between two nodes means that two nodes must be in the same cluster can cannot link means that two nodes cannot be in the same cluster. In our work, we represent the must link and cannot link constraint under matrix form.

We build matrix M_m is matrix for must-link constraint.

- $M_m[i,j] = 1$ If i and j have must link constraint.
- $M_m[i,j] = 0$ Otherwise.

We build matrix M_c is matrix for cannot-link constraint.

- $M_c[i,j] = 1$ If i and j have cannot-link constraint.
- $M_c[i,j] = 0$ Otherwise.

Now, we build two graphs from these two matrixes. Every node i,j connect by an edge whose weight is $M_m[i,j]$ for must link graph and is $M_c[i,j]$ for cannot link graph.

To minimize the risk to violate constraint, we see that if two nodes in the graph are in must link constraint the value of $M_m[i,j] = 1$ is big compare to when these two nodes are not in any constraints $M_m[i,j] = 0$. This value is also the weight between nodes. So if we divide the graph into two sub-graphs in order to satisfy two nodes in the must link constraint must be in the same sub-graph, the cut will be small because it only contain the edges whose nodes in these edges have no must link connection $M_m[i,j] = 0$. So the minimizing constraint violation problem is changed to minimum cut problem with the weight of graph is defined above. The explanation is the same for cannot link constraint but in maximize way.

In order to minimize the risk to violate constraint, we should find the minimum cut of the graph. On the other hand, we want to cut the graph to maximize the can-not link weight. So the optimization term is to minimize:

$$J_c = \sum_{i,j} (y_i - y_j)^2 (\mathrm{M_m[i,j]} - \mathrm{M_c[i,j]}) \quad (5)$$

We can change (5) into Laplacian form:

$$\underset{y}{\text{Minimize}}\ \mathrm{J_c} = \boldsymbol{y^T} L_m \boldsymbol{y} - \boldsymbol{y^T} L_c \boldsymbol{y} \quad (6)$$

Using (4) and (6), our new objective function is:

$$\underset{y}{\text{Minimize}}\ J = \boldsymbol{y^T}(A + L_m - L_c)\boldsymbol{y} \quad (7)$$

The new optimization can be solved efficiently by eigenvector-solver. Experiment with real data which is used in [1], show significant result of our approach.

	Our solution	LLC	Spectral Clustering	CCSR
USPS 3568	**0.0256**	0.0357	0.0334	0.0866
USPS 49	**0.0335**	0.0801	0.0705	0.0634
U-Mist	**0.2139**	0.3601	0.2887	0.5335
Umist5	**0**	0	0.0214	0.0214
News4a	**0.0484**	0.0797	0.2753	no result
News4b	**0.0521**	0.0971	0.1211	no result

Table 1. Average error of our semi local clustering when compare with original LLC algorithm [1], spectral clustering [3] and constraint based semi-supervised clustering [2]. Due to high time complexity using semi definite solver, CCSR cannot give result for two big datasets News4a and News4b.

4 Conclusion

In this paper, we propose a novel semi supervised local clustering method. The challenge when applying semi supervised idea into local clustering is how to efficiently utilize user provided pair-wise constraint for two samples and propagate this information to their neighbor's label. The experiment show the better performance of our work when compare to well known other semi supervised methods. There are many avenues for future work. Noisy data and overlapping clusters in data are two difficult problems in manifold learning today. We plan to deal with noise constraints from expert and deal with overlapping cluster in order to make local clustering algorithm become more robust in imperfect environment.

References

[1]. M Wu, Bernhard Schölkopf. A local learning approach for clustering. Advances in Neural Information Processing Systems 2007.

[2]. Zhenguo Li, Jianzhuang Liu, Xiaoou Tang. Constrained Clustering via Spectral Regularization. IEEE Conference on Computer Vision and Pattern Recognition 2009.

[3]. S.X. Yu and J. Shi. Multiclass spectral clustering. International Conference on Computer Vision 2003.

[4]. Xiaojin Zhu, Andrew B. Goldberg, Introduction to Semi-Supervised Learning. Synthesis Lectures on Artificial Intelligence and Machine Learning 2009.