The Tenth International Symposium on Operations Research and Its Applications (ISORA 2011)
Dunhuang, China, August 28–31, 2011

A Flexible Convex Optimization Model for Semi-supervised Clustering with Instance-level Constraints

Xianwen Ren[1] Yong Wang[2] Xiang-Sun Zhang[2]

[1]State Key Laboratory for Molecular Virology and Genetic Engineering,
Institute of Pathogen Biology,
Chinese Academy Medical Sciences and Peking Union Medical College, Beijing, China
[2]Institute of Applied Mathematics
Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing, China

Abstract Clustering is a common task in many applications e.g. digital image processing, text mining and bioinformatics. Many techniques such as *k*-means, hierarchical clustering and spectral clustering, have been proposed. In a previous study, we proposed a quadratic programming model to address the fuzzy binary clustering problem in the unsupervised setting and then extended it to the general clustering problem. In this paper, we extend further the model in the semi-supervised setting. It has three salient characteristics. First, both the label and link information of known samples can be integrated easily. Second, it illustrates the linkage between the hard binary clustering and fuzzy binary clustering in one framework, suggesting the benefits of fuzzy binary clustering theoretically. Third, a fast iterative algorithm is proposed, which can be applied to very large data sets. Numerical experiments on two data sets suggest its practical effectiveness and efficiency.

Keywords Semi-supervised clustering; Quadratic programming; Constraint propagation; Label propagation; Fussy binary clustering

1 Introduction

Clustering is now a very common application in many disciplines such as digital image processing, text mining and bioinformatics. Many methods have been proposed to implement the task. For example, *k*-means sets up a parametric model to search *k* cluster centers and to identify the corresponding members [4, 5, 6]. Affinity propagation method uses the similarity matrix of samples and implements a message passing algorithm to reveal the underlying clustering structure of the data set [1, 2]. In a previous study, we proposed a quadratic programming model to address the fuzzy binary clustering problem in the unsupervised setting and then extended it to the general clustering question [7]. In this study, we extend the model further to the semi-supervised setting by integrating more prior information to improve the clustering accuracy.

Different from unsupervised clustering algorithms, semi-supervised clustering algorithms classify the samples not only based on the measured data for all samples, but also based on the available prior information for some samples. Because more information

can be exploited, the accuracy of semi-supervised clustering is generally assumed to be higher than that of unsupervised clustering. And it has attracted increasing interests from experts in various domains. The prior information can be summarized into two popular forms: labels for some samples and constraints for some sample pairs. The label information for some samples assigns explicitly which clusters they belong to. The constraints for sample pairs include "cannot-link" constraints and "must-link" constraints [2]. Cannot-links indicate that the two samples can not be in the same cluster. Must-links mean that the two samples must be in the same cluster. We will show that our model can integrate both the label information and the two types of constraint information and form a convex optimization problem.

Unlike k-means and other hard clustering algorithms, our model is based on fuzzy binary clustering. Binary clustering is a special case of the general clustering. However it is also an atomic operation to implement the general clustering. In the previous study, we demonstrated how binary clustering can be used to do the general clustering. We denote binary clusters by zero and one, respectively. Fussy labels between zero and one were assigned to the samples to indicate how close they are to zero. In this study, we extend the idea by replacing the zero-one cluster notations by "$+1$" and "-1" in this study so that the constraint information can be integrated easily. A parameter is added for users to control the fuzziness of the labeling. By altering this parameter, the linkage between hard binary clustering and fuzzy binary clustering is illustrated in one framework, which further suggests the benefits of fuzzy binary clustering.

In many applications such as text mining and bioinformatics, the data amount is very huge. Fast and efficient algorithms are of great demand to deal with these situations. Based on our model, we derive a very fast iterative algorithm for clustering. Given the similarity matrix of the samples and the prior information, the algorithm only requires matrix multiplication and the convergence is guaranteed. So it is very useful when dealing with large data sets.

2 Methods

2.1 Unsupervised model for fuzzy binary clustering

We begin with a short review of the unsupervised model for fuzzy binary clustering we proposed previously. Given a data set X, the similarity matrix S is first calculated based on the domain knowledge. Then a quadratic programming model is built as follows:

$$\min_{f} \quad \sum_{i=1}^{N}\sum_{j=1}^{N} s_{ij}(f_i - f_j)^2 \tag{1}$$

$$\text{subject to} \quad f_a = 0 \tag{2}$$

$$f_b = 1 \tag{3}$$

$$f_i \le 1 \qquad i \in \{1,2,\cdots,N\} \tag{4}$$

$$f_i \ge 0 \qquad i \in \{1,2,\cdots,N\} \tag{5}$$

where N is the total number of data points; s_{ij} is the similarity score of data points x_i and x_j; and f_i is the label of data point x_i to be determined. a and b are the most dissimilar two data points in the N data points, i. e., $s_{ab} = \min\{s_{ij} : i, j \in \{1 \cdots N\}\}$. The objective

function (1) requires the similar data points have similar labels. Constraints (2) and (3) assign data points a and b to two different clusters. Constraints (4) and (5) restrict the labels f_i to be between 0 and 1. The objective function (1) can be further written in the vector form as $f^T Lf$, where L is the Laplacian matrix of S, i.e., $L = D - S$ and D is a diagonal matrix with $d_{ii} = \sum_{j=1}^{N} s_{ji}$. If s_{ij} are all non-negative, L is positive semi-definite. Then the model (1-5) is a convex optimization problem and the global optimal solution is guaranteed.

In this model, the binary labels are denoted by zero and one. The objective function implements the clustering task while the constraints (4) and (5)prevent the trivial result that is all the samples belong to one cluster. Constraints (4) and (5) are added by an assumption that the most dissimilar two samples belong to different clusters. Replacing the zero-one notations by $+1$ and -1, a scalable quadratic programming model for semi-supervised clustering with instance-level constraints can be obtained readily in the next section.

2.2 From unsupervised to semi-supervised

Replacing the zero-one notations by $+1$ and -1, we develop the new model for fuzzy binary clustering as follows:

$$\min_{f} \quad \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} s_{ij}(f_i - f_j)^2 \tag{6}$$

$$\text{subject to} \quad f_i = 1 \qquad i \in I_+ \tag{7}$$

$$f_i = -1 \qquad i \in I_- \tag{8}$$

$$f_i * f_j = 1 \qquad (i,j) \in I_{must} \tag{9}$$

$$f_i * f_j = -1 \qquad (i,j) \in I_{cannot} \tag{10}$$

$$f_i \leq 1 \qquad i \in \{1,2,\cdots,N\} \tag{11}$$

$$f_i \geq -1 \qquad i \in \{1,2,\cdots,N\} \tag{12}$$

Here s_{ij} and f_i are still the similarity scores and the real-valued labels to be determined. I_+ means the known positive sample set and I_- means the known negative sample set. I_{must} denotes the available must link information whereas I_{cannot} specifies the cannot link information.

Because $+1 * +1 = 1$ and $-1 * -1 = 1$, we can add easily the must link information into the model as (9). Because $+1 * -1 = -1$, we can add easily the cannot link information into the model as (10). We also extend (2) and (3) to (8) and (7), respectively, to include available label information. So we extend the binary clustering model from the unsupervised setting to the semi-supervised setting, which can integrate both the label information and the constraint information.

2.3 From nonconvex to convex

If all s_{ij} are positive, the objective function (6) is a convex quadratic form. But the constraints are not all convex due to (9) and (10). Because the global optimum solutions of convex programming problems are easy to obtained, we adopt approximations to convert (9) and (10) to convex constraints. First, the constraints of (9) and (10) suggest that the norms of f_i and f_j equal to one. Subtracting 2-fold (9) by $f_i * f_i + f_j * f_j$, we get (13).

Subtracting 2-fold (10) by $f_i * f_i + f_j * f_j$, we get (14). (13) and (14) are convex so that the whole model is convex.

$$(f_i - f_j)^2 = 0 \quad (i,j) \in I_{must} \tag{13}$$
$$(f_i + f_j)^2 = 0 \quad (i,j) \in I_{cannot} \tag{14}$$

Moving (13) and (14) to the subject function, we finally get the following model:

$$\min_f \quad \tfrac{1}{2} f^T (L + \mu C) f \tag{15}$$
$$\text{subject to} \quad f_i = 1 \qquad i \in I_+ \tag{16}$$
$$f_i = -1 \qquad i \in I_- \tag{17}$$
$$f_i \leq 1 \qquad i \in \{1,2,\cdots,N\} \tag{18}$$
$$f_i \geq -1 \qquad i \in \{1,2,\cdots,N\} \tag{19}$$

where L is the Laplacian transform of the similarity matrix. C is the constraint matrix constructed by I_{must} and I_{cannot}, in which $c_{ij} = -1, c_{ii} += 1 \text{and} c_{jj} += 1$ if $(i,j) \in I_{must}$ and $c_{ij} = 1, c_{ii} += 1 \text{and} c_{jj} += 1$ if $((i,j) \in I_{cannot}$. By altering μ, the prior constraint information is absorbed into the model smoothly. When μ is large, the constraints must be followed. Otherwise the constraints can be modified in the solution to fit the data better. This is especially useful when the confidence of the constraint is not strong.

2.4 From fuzzy binary clustering to hard binary clustering

Hard binary clustering can be obtained easily from our fuzzy binary clustering results. We model the binary clustering as a process to assign samples with fuzzy labels between $+1$ and -1. $+1$ denotes the positive cluster and -1 represents the negative cluster. From the fuzzy labels, we can observe the deviation of one sample from the positive cluster or the negative cluster easily by calculating the distance from the corresponding label to $+1$ or -1. If the similarity matrix is given properly, we can convert the fuzzy labels to hard labels easily based on their symbols. For example, if one sample was assigned a fuzzy label of -0.5, it would have -1 as its hard label. If the similarity matrix is not given properly, the symbols may not reflect their genuine labels. Then the maximum-gap criterium can be applied by sorting the fuzzy labels from the smallest to the largest first and then set the cutoff at where the difference of the nearest labels is maximal. For example, given a series of fuzzy labels (0.1,0.11,0.22,0.23), the cutoff should be between 0.11 and 0.22.

We introduce a parameter, λ, into the the objective function to control the fuzziness of the resultant labels. Then, the objective function becomes

$$\frac{1}{2} f^T (L + \mu C - \lambda I) f \tag{20}$$

Through this parameter, our model forms a continuum between hard binary clustering and the trivial one cluster. If λ is small or even a negative large number, the model tends to produce the trivial one-cluster result. That is, all f_i equal to zero. The larger λ is, the more likely f_i approaches to $+1$ or -1. If λ is large enough, the objective function (20) will not be convex, indicating why hard binary clustering is hard to obtained from a new perspective.

2.5 From primary constraints to constraint propagation

If sample a is assumed to belong to the same class with sample b and b is assumed to belong to the same class as sample c, then a and c should also belong to the same class. If a and b are assumed to belong the the same class but b and c are assumed to belong to different classes, then a and c should also belong to different classes. To exploit this type of prior information, we introduce an additional parameter, v, into the objective function. The objective function turns out to be

$$\frac{1}{2}f^T(L+\mu C^v-\lambda I)f \tag{21}$$

where v should be an integer more than or equivalent to one. It is reported that constraint propagation can improve the classification. This is carried out very easily in our model through v.

2.6 A fast algorithm to deal with large data sets

The proposed convex model can be solved by the general quadratic programming solver. Or, just as the fast algorithm proposed in the unsupervised setting, a similar algorithm also exists to solve our model efficiently on very large data sets. The Lagrangian function of our final model can be written as:

$$L(f)=\frac{1}{2}f^T(L+\mu C^v-\lambda I)f+\sum_{i\in I_+}\alpha_i(f_i-1)+\sum_{i\in I_-}\beta_i(f_i+1)+\sum_i\gamma_i(f_i-1)+\sum_i\delta_i(f_i+1) \tag{22}$$

The Karush-Kuhn-Tucker (KKT) conditions are:

$$(L+\mu C^v-\lambda I)f+\alpha+\beta+\gamma+\delta=0 \tag{23}$$
$$\gamma_i(f_i-1)=0 \quad i\in\{1,2,\cdots,N\} \tag{24}$$
$$\delta_i(f_i+1)=0 \quad i\in\{1,2,\cdots,N\} \tag{25}$$
$$f_i=1 \qquad i\in I_+ \tag{26}$$
$$f_i=-1 \qquad i\in I_- \tag{27}$$
$$f_i\geq -1 \quad i\in\{1,2,\cdots,N\} \tag{28}$$
$$f_i\leq 1 \quad i\in\{1,2,\cdots,N\} \tag{29}$$
$$\gamma_i\geq 0 \quad i\in\{1,2,\cdots,N\} \tag{30}$$
$$\delta_i\leq 0 \quad i\in\{1,2,\cdots,N\} \tag{31}$$

These conditions can be further reduced as:

$$f_i=1 \tag{32}$$

or

$$f_i=-1 \tag{33}$$

or

$$f_i=-\frac{1}{\overline{L}_{ii}}\sum_{j\neq i}\overline{L}_{ij}f_j \tag{34}$$

for $i\notin I_+\bigcup I_-$, where $\overline{L}=L+\mu C^v-\lambda I$. So we design the following algorithm to solve our model efficiently on large data sets:

- **Step 1**: Let $t = 0$, initialize $f_i = 1$ if $i \in I_+$, $f_i = -1$ if $i \in I_-$ and $f_i = 0$ if $i \notin I_+ \bigcup I_-$.
- **Step 2**: Let $t = t + 1$, calculate $f_i^t = -\frac{1}{\overline{L}_{ii}} \sum_{j \neq i} \overline{L}_{ij} f_j^{t-1}$. If $f_i^t > 1$, let $f_i^t = 1$; if $f_i^t < -1$, let $f_i^t = -1$.
- **Step 3**: If $\max_i |f_i^t - f_i^{t-1}| < \varepsilon$ where ε is a predefined stopping criterion, then stop. Otherwise go to **Step 2**.

If $\overline{L}$ is positive definite, the whole model is convex and the global optimum can be reached by this algorithm.

2.7 From binary clustering to multiple clustering

Our model is initially proposed for fuzzy binary clustering. It can be extended to multiple clustering easily by adopting the one-vs-all strategy that can be stated as follows

- **Step 1**: Select one class randomly, denote it as G and the other classes as $\overline{G}$. Apply our model to classify the unknown samples to G or $\overline{G}$.
- **Step 2**: If there are more than two classes in $\overline{G}$, repeat **Step 1** on $\overline{G}$.

3 Experiments

We evaluated our model on two data sets. One is Fisher's Iris data in which there are three classes of samples. Class One can be separated linearly from Class Two and Three [8]. Class Two can not be separated linearly from Class Three. The other data set is the gene expression data of a series of leukemia patients [3]. There are two classes. One is acute myeloid leukemia (AML) and the other is acute lymphoblastic leukemia (ALL). The ALL samples can be further divided into two groups based on the sample sources. One is from T cells and the other is from B cells. Evaluations on these data sets suggest the effectiveness of our model and the efficiency of our algorithm. The optimization model is implemented and solved by MATLAB on a PC with 2.4G Hz Pentium 4 processor.

3.1 Fisher's Iris data

The construction of the similarity matrix is the first step to use our method for clustering samples. We first calculated the Pearson correlation coefficients of samples by exploiting the given four features for each sample. Most of the samples are highly correlated. The minimum correlation coefficient is 0.3574. To highlight the underlying cluster structure, we set a cutoff (0.9) to convert the smaller values to zero. The similarity matrix is shown in Figure 1. Randomly selecting three must links and three cannot links, the samples were assigned fuzzy labels between -1 and $+1$ (Figure 2) and Class One was discriminated from Class two and Class Three accurately. This is an easy task and we use it to show how our model assigns fuzzy labels based on the similarity matrix and prior constraint and label information.

After Class One was identified, we applied our model to the other 100 samples to discriminate Class Two from Class Three that can not be classified linearly. We constructed the similarity matrix by first calculating the Pearson correlation coefficients between samples and then set a cutoff (0.998) to convert the similarity matrix to a sparse graph (Figure 3. Then we evaluated the impacts of λ, μ and ν on the classification. Randomly selecting ten must links and cannot links ($n = 10$), we repeated classification for 100 times for each combination of $\lambda(\lambda \in \{-1, -2, -3\})$ and $\mu(\mu \in \{0, 1, 10\})$ with $\nu = 1$. The accuracy

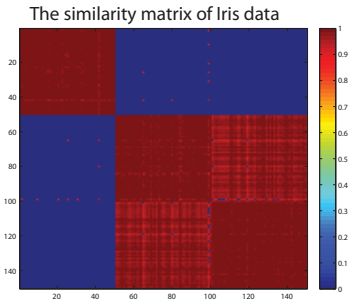

Figure 1: The similarity matrix of Iris data. Class One: 1-50; Class Two: 51-100; Class Three: 101-150.

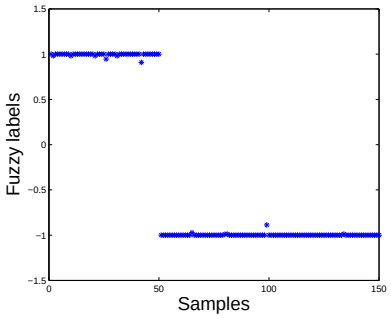

Figure 2: The fussy labels for each sample in Iris data.

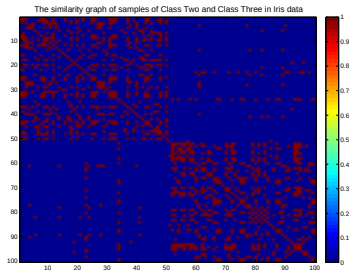

Figure 3: The similarity matrix of samples in Class Two and Three in Iris data.

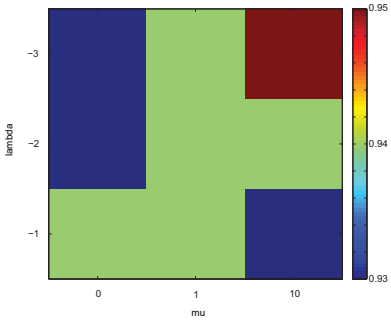

Figure 4: The accuracy with different λ and μ.

can reach 0.95 (Figure 4). Increasing the amount of the prior information ($n = 100$), the accuracy can be further improved to 0.98.

3.2 Golub's gene expression data

We further evaluated our method on the gene expression data. Because there is much noise in the gene expression data, we first set a ceiling (16,000) and a floor (100) for the intensities and then filter those genes with $\max/\min \leq 5$ or $\max - \min \leq 500$, where max and min mean the maximum and minimum gene expression values in all the samples, respectively. Then we calculated the Pearson correlation coefficients between samples. To enhance the statistical signals, we calculated the Pearson correlation coefficients among samples based their initial correlations. All except one AML samples are successfully discriminated from the ALL samples with $\lambda = 1, \mu = 1, \nu = 1$ and $n = 10$ (accuracy: $71/72 = 98.6\%$). Then we applied our method to discriminate B cell ALLs from T cell ALLs based on the same similarity matrix in which the similarity scores less than 0.4 were filtered to enhance the statistical signal. All T cell ALLs except one were correctly identified (accuracy: $46/47 = 97.8\%$). This example suggests the effectiveness of our

method further.

4 Discussions and Conclusion

In this study, we propose a flexible quadratic programming framework for semi-supervised clustering. It can integrate both label information and constraint information. It provides handy parameters for users to calibrate the fuzziness of the resultant labels, to control the confidence of the prior constraint information and to tune the propagation of the constraint links. It solves multi-class clustering through recursive binary clustering. Numerical experiments on two real data sets suggest its effectiveness. It should be a useful tool to help researchers understand the meanings underlying various types of data in many disciplines. Because it is related closely to spectral clustering, we will compare it to the available spectral clustering methods and other semi-supervised methods in future [9, 10]. Because the clustering is dependent on the original data, we will also do the sensitivity analysis of our model in future.

Acknowledgements

The authors thank the members of Zhangroup of Academy of Mathematics and Systems Science, Chinese Academy of Sciences for their valuable discussion and comments.

References

[1] Brendan J. Frey and Delbert Dueck. Clustering by Passing Messages Between Data Points. *Science*, 315(5814):972–976, 2007.

[2] Inmar E. Givoni and Brendan J. Frey. Semi-supervised affinity propagation with instance-level constraints. *Journal of Machine Learning Research*, 5:161-168, 2009.

[3] T R Golub, D K Slonim, P Tamayo, C Huard, M Gaasenbeek, J P Mesirov, H Coller, M L Loh, J R Downing, M A Caligiuri, C D Bloomfield, and E S Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, 1999.

[4] Stuart P. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28:129–137, 1982.

[5] J. B. Macqueen. Some methods of classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

[6] Hugo Steinhaus. Sur la division des corps matériels en parties. *Bull. Acad. Pol. Sci., Cl. III*, 4:801–804, 1957.

[7] Xianwen Ren, Yong Wang, Jiguang Wang and Xiang-Sun Zhang An optimization model for fuzzy binary clustering. in *Lecture Notes in Operations Research*, 12:422-432, 2010.

[8] Fisher, Ronald A. The use of multiple measurements in taxonomic problems. *Annals Eugen*,7:179–188, 1936.

[9] U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2006.

[10] Xiaojin Zhu. Semi-Supervised Learning Literature Survey. Encyclopedia entry in Claude Sammut and Geoffrey Webb, editors, Encyclopedia of Machine Learning. Springer. in press.