Discriminative Approach to Semi-Supervised Clustering

Marek Śmieja

Faculty of Mathematics and
Computer Science
Jagiellonian University
Lojasiewicza 6, 30-348 Kraków
Email: marek.smieja@ii.uj.edu.pl

***Abstract*—We consider a semi-supervised clustering problem, where selected pairs of data points are labeled by an expert as must-links or cannot-links. Basically, must-link constraints indicate that two points should be grouped together, while those with cannot-link constraints should be grouped separately. We present a clustering algorithm, which creates a partition consistent with pairwise constraints by maximizing the probability of correct assignments. Moreover, unlabeled data are used by maximizing their prediction confidence. Preliminary experimental studies show that the proposed method gives accurate results on sample data sets. Moreover, its kernelization allows to discover clustering patterns of arbitrary shapes.**

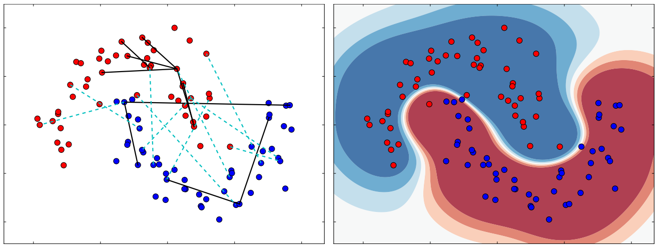
Keywords–semi-supervised clustering; pairwise constraints; discriminative model.

I. Introduction

Clustering is one of core branches of machine learning and data analysis, which aims to find homogeneous groups in data. Since cluster analysis is purely unsupervised technique, its results may be unsatisfactory for a given problem. Semi-supervised clustering allows to include side information (expert knowledge) about class labels into clustering to obtain more appropriate effects for the user [1]. Pairwise constraints (relations) are a typical form of additional class information used in semi-supervised clustering. They indicate whether two points belong to the same (must-link) or different groups (cannot-link). The aim of semi-supervised clustering is to use pairwise constraints in order to produce more accurate results [2][3].

To meet the user expectations revealed in pairwise constraints, we follow a discriminative approach, which is usually applied in classification, but is rarely used in clustering. Discriminative model is more natural and effective for semi-supervised task than typical generative approaches, such as k-means or Gaussian mixture model (GMM), because it directly focuses on the underlying classification problem. We formulate a clustering model, which maximizes the probability that pairwise relations are preserved. Unlabeled data are handled by maximizing their prediction confidence, which agrees with a typical paradigm of semi-supervised learning (cluster assumption) stating that decision boundary should fall in low density region.

Our method is easy to implement and can be optimized with use of a gradient approach. Moreover, it can be kernelized so that to fit arbitrary clustering structures, see Figure 1 for the illustration. Preliminary experimental results show that our method is promising and allows to obtain competitive results to the state-of-the-art models.



(a) (b)

Figure 1. Sample results of our method on two moons data set (b); must-links (solid black line) and cannot-link (dashed cyan) are shown in (a).

II. Model

We consider a data set $X \subset \mathbb{R}^D$, such that $N = |X|$, where every element $x \in X$ belongs to one of K unknown classes. By $\mathcal{X} = X \times X$ we denote the set of all pairs in X. Partial information about class labels is revealed in the form of pairwise constraints, which cover selected pairs of data points $\mathcal{L} \subset \mathcal{X}$. Pairwise constraints indicate whether two points originate from the same or different classes, thus $\mathcal{L}$ can be split into the sets of must-link and cannot-link constraints given by [1]:

$$\mathcal{M} = \{(x, y) \in \mathcal{L} : x \text{ and } y \text{ belong to the same class}\},$$
$$\mathcal{C} = \{(x, y) \in \mathcal{L} : x \text{ and } y \text{ belong to the different classes}\}.$$

Our clustering model follows a discriminative approach in which the assignments of data points to clusters are directly modeled by posterior probabilities. Let $p_k(x) = p(k|x)$ be a posterior probability that a data point $x \in X$ is assigned to k-th cluster, where $k = 1, \ldots, K$. Once these conditional probabilities are defined, we get a partition of X, in which a point $x \in X$ is assigned to this group that maximizes its posterior probability. More precisely, we get a partition of X into $C_1, \ldots, C_k \subset X$, where

$$C_k = \{x \in X : p_k(x) = \max_j p_j(x)\}.$$

We assumed that posterior probabilities are given by a logistic function:

$$p_k(x) = p_k(x; \mathcal{V}) \propto \exp(\langle v_k, x\rangle + b_k), \qquad (1)$$

where the set of parameters $\mathcal{V} = (v, b)$ consists of weight vectors $v = (v_1, \ldots, v_K)$ and bias values $b = (b_1, \ldots, b_K)$.

Our model focuses on maximizing a probability that pairwise constraints are satisfied. Let us first observe that the

ISBN: 978-1-61208-620-0

probability that a clustering model assigns two points $x, y \in X$ to the same cluster equals:

$$p_{\mathcal{M}}(x, y) = \sum_{k=1}^{K} p_k(x) p_k(y). \tag{2}$$

Consequently, the probability that $x, y \in X$ are classified to different groups is given by:

$$p_{\mathcal{C}}(x, y) = 1 - p_{\mathcal{M}}(x, y). \tag{3}$$

To meet the expert knowledge, we maximize both terms over all pairwise relations.

In addition to pairwise constraints, we usually have the access to a large number of unlabeled data. Since there is no information about their classes, we cannot simply maximize their correct assignments. However, we can encourage the model to give the most confident answers about their classes. Let us consider a function

$$\sum_{k=1}^{K} p_k(x)^2, \text{ for every } x \in \mathcal{X}, \tag{4}$$

which attains a maximal value, if the prediction confidence is maximal, i.e., $p_l(x) = 1$ for specific l and $p_k(x) = 0$, for all $k \neq l$. On the other hand, if all classes are equally probable then its value is minimal. Thus, to maximize a prediction confidence of the model, we maximize (4) over unlabeled data.

Our clustering objective function gathers (2), (3) and (4) over all data points. Its maximization can be implemented with use of gradient approach. Moreover, due to the form of posterior probabilities (1) one can introduce kernel functions to detect arbitrary shapes of clusters.

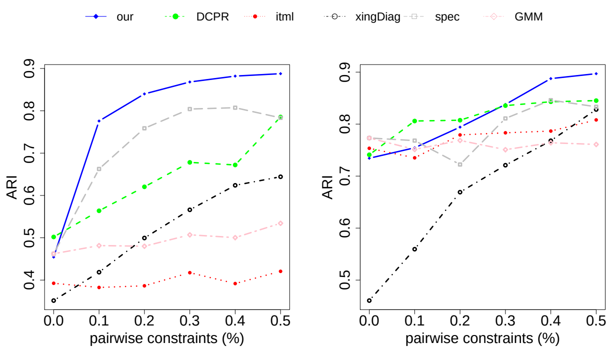
III. Experiments

We examined our method on two standard data sets retrieved from UCI repository [4]: Letter (1000 examples, 16 features, 5 classes) and Seeds (210 examples, 7 features, 3 classes). To acquire pairwise relations, we randomly selected a pair of points (x, y) and label it as must-link if both x, y belong to the same cluster or as cannot-link, otherwise. We vary the number of constraints from $0.1N$ to $0.5N$ with a $0.1N$ increment. The results were evaluated using adjusted rand index (ARI) [5]. ARI attains a maximal value 1 for a partition identical with a ground-truth, while for a random grouping gives score 0.

We compared our method with five state-of-the-art techniques:

- another discriminative framework proposed in [6], referred to as DCPR (discriminative clustering with pairwise constraints).
- recent semi-supervised spectral clustering [7], referred to as spec
- constrained GMM proposed in [2] (GMM)
- two metric learning algorithms: diag [8] and itml [9]

The results presented in Figure 2 show that our method usually obtained very high scores. Its performance gradually increases as the number of constraints grows. It can be observed that itml and DCPR also gave high resemblance with reference grouping, while the performance of GMM, spec and diag were usually worse.



(a) Letter (b) Seeds

Figure 2. Adjusted rand index of examined methods two data sets retrieved from UCI repository.

IV. Conclusion and Future Work

In this paper, we presented a new approach to clustering with pairwise constraints and demonstrated its usefulness on two data sets. In future, we plan to extend this model to handle unlabeled data in a more efficient way. In particular, we plan to use the information about data points' neighborhoods. We would also like to use different types of expert knowledge such partial labeling or relative constraints. Moreover, we will apply the proposed approach in real life problems.

Acknowledgment

This work was partially supported by the National Science Centre (Poland) grant no. 2016/21/D/ST6/00980.

References

[1] S. Basu, I. Davidson, and K. Wagstaff, Constrained clustering: Advances in algorithms, theory, and applications. CRC Press, 2008.

[2] N. Shental, A. Bar-hillel, T. Hertz, and D. Weinshall, “Computing Gaussian mixture models with EM using equivalence constraints,” in Advances in Neural Information Processing Systems (NIPS), Vancouver, British Columbia, Canada, December 2004, pp. 465–472.

[3] M. Bilenko, S. Basu, and R. J. Mooney, “Integrating constraints and metric learning in semi-supervised clustering,” in Proceedings of the twenty-first international conference on Machine learning. ACM, 2004, p. 11.

[4] M. Lichman, “UCI machine learning repository,” 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[5] L. Hubert and P. Arabie, “Comparing partitions,” Journal of Classification, vol. 2, no. 1, 1985, pp. 193–218.

[6] Y. Pei, X. Z. Fern, T. V. Tjahja, and R. Rosales, “Comparing clustering with pairwise and relative constraints: A unified framework,” ACM Transactions on Knowledge Discovery from Data (TKDD), vol. 11, no. 2, 2016, p. 22.

[7] P. Qian et al., “Affinity and penalty jointly constrained spectral clustering with all-compatibility, flexibility, and robustness,” IEEE transactions on neural networks and learning systems, vol. 28, no. 5, 2017, pp. 1123–1138.

[8] E. P. Xing, M. I. Jordan, S. J. Russell, and A. Y. Ng, “Distance metric learning with application to clustering with side-information,” in Advances in neural information processing systems, 2003, pp. 521–528.

[9] J. V. Davis, B. Kulis, P. Jain, S. Sra, and I. S. Dhillon, “Information-theoretic metric learning,” in Proceedings of the 24th international conference on Machine learning. ACM, 2007, pp. 209–216.