A survey of Semi-Supervised K-Means Clustering

Ankita Naik

Abstract

Semi-supervised clustering (SSC) or clustering using side information has received substantial attention from researchers because of its success in many applications like document, image, utterance clustering, etc. It has shown to demonstrate significant improvement over traditional unsupervised clustering methods. Most of the SSC algorithms are driven either by labeled data or cluster-level constraints as side information. This survey will cover two such semi-supervised variants of K-Means algorithms : Constrained K-Means and Seeded K-Means algorithm. The survey also incorporates a variant of the distance metric which when coupled with the above mentioned methods produces a better performance.

1. Introduction

In some situations, the cluster assignments may be known for a subset of the available data. The objective of the exercise then is to cluster the unlabeled observations into appropriate clusters using known cluster assignments. In certain sense this problem sounds very similar to a supervised classification problem. However, traditional supervised classification methods may be inefficient when only a relatively small subset of data is labeled. Basu et al. [1] thus developed a generalization of K-means algorithm for situations where labels are available for few of the data points.

1.1. Constrained K-Means

Constrained K-Means clustering is identical to conventional K-means procedure except for the following two conditions:

1. K-means clustering initializes the initial cluster centers through a random process whereas in constrained K-means clustering the initialization is based on labeled data. The initial cluster center of the lth cluster is initialized with the mean of all the labeled data points belonging to the lth cluster.
2. During the update step, the cluster membership of labeled observations is kept unchanged.

Email address: naik.ankita1210@gmail.com (Ankita Naik)

June 4, 2018

Let x_u and x_l be observations in a dataset with p features, where x_u represents unlabeled data points and x_l represents labeled data points, where $X = \{x_i : x_i \in x_l \text{ or } x_i \in x_u\}$. Also, x_{ij} represent the value of the jth feature for observation i. There exist subsets S_1,S_2,S_3,S_4,....,S_K of the x_l's such that $x_l \in S_k$ implies that observation l is known to belong to cluster k. (Here, K denotes the number of clusters which is already known and S_k denotes the cluster formed by the initially available labeled data points). Let $|S_k|$ denote number of x_l's in S_k. Also, let $S = \cup_{k=1}^{K} S_k$.

Algorithm:

1. For each feature j and cluster k, calculate the initial cluster means as follows :

$$\bar{x}_{kj} = \frac{1}{|S_k|} \sum_{x_l \in S_k} x_{lj} \tag{1}$$

2. Assign each observation i to new cluster C_i with the assignment for labeled data point remaining unchanged.

$$C_i = \begin{cases} S_k & x_i \in \text{S and if } x_i \in S_k \\ \text{arg min}_\text{k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2 & x_i \notin S \end{cases} \tag{2}$$

3. For each feature j and cluster k, calculate $\bar{x}_{kj}$, the mean for feature j in cluster k.
4. Repeat step 2 and 3 until algorithm convergence.

The constrained k-means clustering algorithm thus assumes that none of the labeled observations are misclassified. Using the constrained k-means clustering procedure, if a labeled observation is misclassified, this misclassification can never be corrected, since this observation will be assigned to the same cluster in step 2 in every iteration of the algorithm. Thus, the algorithm is very susceptible to noise in the initially assigned labels.

1.2. Seeded K-Means

Seeded K-Means clustering algorithm was recommended by Basu et al. [1] to deal with the inherent problem of misclassified labels in constrained K-Means. The seeded K-Means algorithm always assigns the observations to the nearest cluster using equation 2 even if the observation is labeled. Thus, if an observation is initially mislabeled it may be corrected during the re-assignmnet step.

Thus, seeded K-Means clustering is exactly identical to the conventional K-Means clustering with the exception of the first step in the procedure where the initialization of cluster centers is based on labeled data rather than random. As, the only difference is the initialization process seeded K-Means still struggles with the inherent issues of K-Means and is not able to completely utilizes the labeled data information. The later problem could be tackled by utilizing a more appropriate distance measure.

1.3. Learning distance metric

Eric P. Xing et al. [2] in their research have tried leveraging the similarity and dissimilarity information present in partially labeled data by incorporating the leanings into the distance metrics.

Suppose we have some set of points $\{x_i\}_{i=1}^{m} \subseteq \mathbb{R}^n$, and are given information that certain pair of them are "similar":

$$S : (x_i, x_j) \in \text{S} \quad \text{if } x_i \text{ and } x_j \text{ are similar} \tag{3}$$

A distance metric can be learned in the following form :

$$d(x, y) = d_A(x, y) = ||x - y||_A = \sqrt{(x - y)^T A(x - y)} \tag{4}$$

To ensure that this metric satisfies non-negativity and triangle inequality we require A to be a positive semi-definite. Setting $A = I$ gives us the Euclidean distance; if A is restricted to be a diagonal metric, this corresponds to learning a metric in which the different axes are given different "weights".

A simple way of defining a criterion for the desired metric is to expect similar points to have smallest possible sum of square distance between them i.e. $minimize_A \sum_{(x_i,x_j)\in S} ||x_i - x_j||_A^2$. But a trivial solution is possible for the previous equation thus the optimization is defined using constraints based on dissimilarity in data points as well. Thus, the final optimization setup is as follows:

$$\begin{aligned} minimize_A & \sum_{(x_i,x_j)\in S} ||x_i - x_j||_A^2 \\ s.t. & \sum_{(x_i,x_j)\in D} ||x_i - x_j||_A \geq 1, \\ & A \succeq 0 \end{aligned} \tag{5}$$

The A distance metric learned from the labeled data points can be further used in Constrained K-Means to calculate the distances at the update steps in place of normal Euclidean distance. In most problems, using a learned metric with constrained K-Means outperforms using constrained K-Means alone [2].

For data with misclassified labels, the above mentioned distance measure coupled with seeded K-Means clustering could produce superior performance and combat the effect of noisy labeled data seen on constrained K-Means. Also, as more and more side-information S is provided to the problem typically it leads to metrics giving better clustering [2].

References

[1] S. Basu, A. Banerjee, R. Mooney, Semi-supervised clustering by seeding, in: In Proceedings of 19th International Conference on Machine Learning (ICML-2002), Citeseer, 2002.

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

[3] E. Bair, Semi-supervised clustering methods, Wiley Interdisciplinary Reviews: Computational Statistics 5 (5) (2013) 349–361.