Hidden Markov Random Fields Based LSI Text Semi-supervised Clustering

Kerui Min
School of Computer Science, Fudan University, Shanghai, China

Gang Liu
Software School, Zhangjiang Branch of Fudan University, Shanghai, China

Xin Chen
Department of Computer Science and Technology, Nanjing University, Nanjing, China

Shengqi Lu
Software School, Zhangjiang Branch of Fudan University, Shanghai, China

Abstract

Semi-supervised learning is an active research field. Previous results shown that unite background information into the original unsupervised clustering problem could archive higher accuracy. In this paper, we explore the cooperation between the pairwise constrains given by the user and the sematic information in natural language. In addition, we reduce the time complexity to make the algorithm feasible for large quantities of data. Experiments on different scales of corpus show the robustness and effectiveness of the proposed algorithm, which the F-measure archives 20% higher than previous algorithms.

Introduction

Clustering is an important analytic tool in fields such as machine learning, natural language processing, and computer vision. Generally, clustering is widely used as a kind of unsupervised learning for its low human participation and computational cost. However, since the assessments are different for different tasks, the expected clustering result of one data set may vary substantially in pursuit of different aims.

As application of information technology increases worldwide, huge amounts of digital text, images, and other data have been accumulated. When trying to cluster such data, we find that although unsupervised clustering can hardly yield satisfactory results, semi-supervised learning is viable because at least some human-generated information about the data tend to be available or can be obtained at an reasonable cost.

In this paper, we adopt the conditional model presented in (Wagstaff & Cardie 2000). Given a data set $\mathscr{X}$ and a pair $(\mathbf{x}_i, \mathbf{x}_j) \in \mathscr{X}^2$, the user can put either Must-link or Cannot-link constraints on it. $(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{M}$ means $\mathbf{x}_i$ and $\mathbf{x}_j$ should be in the same class, and $(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{C}$ stipulate that $\mathbf{x}_i$ and $\mathbf{x}_j$ must be in different classes. Besides, we assume that $\mathcal{M}$ is symmetric, reflexive and transitive while $\mathcal{C}$ is symmetric and anti-reflexive.

Experiments show that the Hidden Markov Random Field(HMRF) clustering algorithm, in cooperation with Latent Semantic Analysis(LSA), could significantly increase the accuracy of the clustering results. Moreover, we employ a heuristic strategy to reduce the time complexity of the proposed algorithm, which makes it feasible for large quantities of data.

LSA and HMRF Clustering

In this section, we give a brief introduction of LSA and HMRF clustering.

LSA, which uses matrix SVD of linear algebra, is widely applied in fields such as dimensionality reduction, information retrival, and multi-language inquiry (Deerwester *et al.* 1990). For a $m \times n$ matrix A, it is well-known that A could be rewritten as $A = U\Sigma V^T$, where $\Sigma = diag(\sigma_1, \sigma_2, \ldots, \sigma_n)$ consists of the singular values of matrix A. Let A_k be the optimal approximation matrix of all matrices of rank k with least square error with respect to A. Research shows that if A is the Term×Document matrix whose weights are calculated by the $TF \cdot IDF$ formula, we can decompose it using SVD into A_k, which can be written as

$$A_k = U_k \Sigma_k V_k^T \tag{1}$$

Given $U \in \mathbb{R}^{m \times m}$, and $U_k \in \mathbb{R}^{m \times k}$ after dimensionality reduction, the new matrix U_k can be regard as a mapping from original term space to semantic space, which improves the accuracy of similarity calculation. In the clustering algorithm, instead of using the term feature $d \in \mathbb{R}^m$, we adopt the semantic feature vector $\hat{d} = dU_k\Sigma_k^{-1}$.

The HMRF consists of a observable random field $\mathscr{X}$ and a hidden random field $\mathscr{Y}$. Intuitively, $\mathbf{x}_i \in \mathscr{X}$ is the feature vector of a given text, and the $\mathbf{y}_i \in \mathscr{Y}$ is the corresponding cluster ID, $\mathbf{y}_i \in \{1, \cdots, K\}$.

Definition 1 $[\![\cdot]\!]$ *is a indicator function, which is 1 when its condition is true and 0 otherwise.*

Definition 2 *Let $d(\mathbf{x}, \mathbf{y})$ be the distance between $\mathbf{x}$ and $\mathbf{y}$, $\mathbf{x}, \mathbf{y} \in \mathbb{R}^m$. Their Mahalanobis distance is*

$$d(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|_A = \sqrt{(\mathbf{x} - \mathbf{y})^T A (\mathbf{x} - \mathbf{y})} \tag{2}$$

where A is a semi-positive definite real matrix, $A \succeq 0$.

Let $w_{\mathcal{M}}$ and $w_{\mathcal{C}}$ be the cost of violating single Must-Link and Cannot-Link constraint respectively, and θ_i be the center of the i-th cluster, by employing the HMRF clustering framework introduced in (Basu, Bilenko, & Mooney 2004), the following objective function is obtained

$$\begin{aligned} J_{obj} = & \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{M}} w_{\mathcal{M}} \|\mathbf{x}_i - \mathbf{x}_j\|_A [\![y_i \neq y_j]\!] \\ & + \sum_{(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{C}} w_{\mathcal{C}} (D_{max} - \|\mathbf{x}_i - \mathbf{x}_j\|_A) [\![y_i = y_j]\!] \\ & + \sum_{i=1}^{n} \|\mathbf{x}_i - \theta_i\|_A + \ln C \end{aligned} \tag{3}$$

where C is a constant and $D_{max} = \max_{i,j} \|\mathbf{x}_i - \mathbf{x}_j\|_A$. The optimization of the objective function could be solved by EM algorithm. For detailed steps, please refer to (Basu, Bilenko, & Mooney 2004).

Discussion of Two Improvements

Constrains Expansion Note that the constraint set $\mathcal{M}$ is a equivalence relation. We could deduce more constraints based on the given constraint set.

Definition 3 *Define $\Gamma(\mathbf{x}_i)$ as the equivalence set including $\mathbf{x}_i$.*

Map every data point to a graph $G = (V, E)$, which $V = \mathscr{X}$, $E = \mathcal{M}$. According to transitivity, $\forall \mathbf{x}_i, \mathbf{x}_j \in V$, $\mathbf{x}_i \mathbf{x}_j$ connected means they belong the same equivalence class. We can expand M using this method to determine $\Gamma(\mathbf{x}_i)$ for every $\mathbf{x}_i$. $\forall \mathbf{x}_j \in \Gamma(\mathbf{x}_i)$, add a constraint $(\mathbf{x}_i, \mathbf{x}_j)$ to Must-link.

Similarly, for $(\mathbf{x}_i, \mathbf{x}_j) \in \mathcal{C}$, $\forall \mathbf{x}_p \in \Gamma(\mathbf{x}_i), \mathbf{x}_q \in \Gamma(\mathbf{x}_j)$, $\mathbf{x}_p$ and $\mathbf{x}_q$ cannot belong the same equivalence class, so we add $(\mathbf{x}_p, \mathbf{x}_q)$ to Cannot-link. It can be proven that we made full use of the known information by these two expansions.

Calculation of the Most Distant Points Note that D_{max} is recalculated at a cost of $O(n^2 m)$ in time in each iteration, which is the bottleneck. Here a better heuristic iterative algorithm to calculate D_{max} is proposed.

This algorithm can actually be viewed as an application of hill climbing method, which from the current point finds the most remote point as the next stop and go ahead until it cannot continue. Although it doesn't guarantee the global optimum, it can almost always find a sufficiently good solution. In our experiments involving 1000, 5000, and 10000 random points, the algorithm converges after constant number of iteration almost without exception and find the global optimum in 41% of the all the cases.

Algorithm 1 D_{max} Calculation Algorithm

```
$p \leftarrow Random(1, n), D_{max} \leftarrow 0, q \leftarrow 0$
while ($q \neq p$) do
    $q \leftarrow p$
    for $i = 1$ to $n$ do
        if $\|x_q - x_i\|_A > D_{max}$ then
            $D_{max} \leftarrow \|x_q - x_i\|_A, p \leftarrow i$
        end if
    end for
end while
```

Experiments

The text classification data set provided by Sogou Inc[1]. is used in this experiment. The data set includes 9 categories such as "Military Affairs", "Culture", and "Education". We randomly picking 1000 articles from 3 categories as the test data for the algorithm. The test uses 10-fold validation. All test results are the average of the results of 10 executions to reduce the random deviation. The stopwords are eliminated.

Algorithms in comparison include: K-Means, m-PCKMeans (Basu, Bilenko, & Mooney 2003), HMRF-LSI-*#K*, where *#K* represents the semantic dimensions.

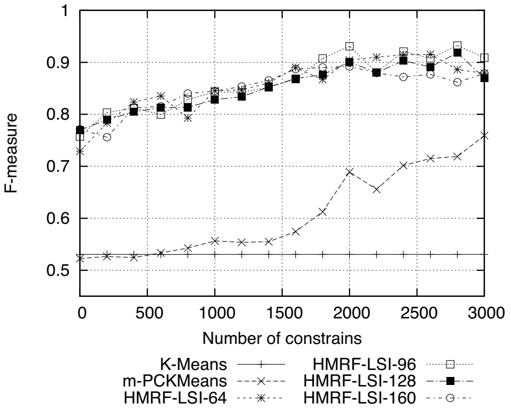

Figure 1: The experiment result for $n = 3000$

References

Basu, S.; Bilenko, M.; and Mooney, R. 2003. Comparing and unifying search-based and similarity-based approaches to semi-supervised clustering. In *ICML*, 42–49.

Basu, S.; Bilenko, M.; and Mooney, R. J. 2004. A probabilistic framework for semi-supervised clustering. In *KDD*, 59–68.

Deerwester, S.; Dumais, S. T.; Furnas, G. W.; Landauer, T. K.; and Harshman, R. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science* 41:391–407.

Wagstaff, K., and Cardie, C. 2000. Clustering with instance-level constraints. In *ICML*, 1103–1110.

[1] http://www.sogou.com/labs