Semi-supervised Hidden Markov Random Fields (HMRF) Kmeans. Theory to Implementation

Dimitrios Pritsos and Efstathios Stamatatos

University of the Aegean
Karlovassi, Samos – 83200, Greece.
{dpritsos, stamatatos}@aegean.gr

1 Introduction

The objective of the Semi-supervised Clustering is to incorporate in the procedure of clusters discovery or assignment, the prior knowledge about the skeleton of the clusters schema. There are several efforts on Semi-supervised model inference in both Expectation Maximization (EM) clustering based models and in Agglomerating clustering based models. According to [3] there three main groups of EM based semi-supervised clustering methods:

1. *Constraints-based mehtods* are using the provided supervision for guiding the algorithm towards a data partitioning which is avoiding (but not prevening) the constraints violation.
2. *Distance-based approaches* in clustering method with a particular distance funciton; the distance function is parametrized and the parameters values are learned to satisfy the constraints.
3. *Semi-supervised clustering based on Hidden Markov Random Fields* where the constraint-based and distance-based approaches are combined into *a unified probabilistic model*.

In EM clustering based models there have been several efforts where the labeled data where provided in the clustering model in the form of data-labels pairs or in the initialization phase of theclustering model. In this work we present the Hidden Markov Random Fields Model(HMRF) Kmeans, where the prior knowledge about the structure or skeleton of the clusters schema has been embedded into the model in the form of constraints [1]. The HMRF Kmeans is a hard-clustering model due to the *hard* assignment of the data point to one of the a-priori fixednumber of clusters. However, the same models can be transformed into a relatively easy soft-clustering model where of each data point only Maximum a-posteriry Probability (MAP) of the point to bea member of each cluster of th final schema.

The HMRF Kmeans Semi-supervised clustering method it has been implemented, in this work, for being tested on the Web Genre Identification (WGI) Information Retrieval (IR) taxonomy problem.Therefore, here we only present the model inference procedure where the distance measure, a.k.a distortion function/measure, is the cosine similarity because in the IR literature is the distortionmeasure where in most cases maximizes performance in problems where the feature space is particularly large, as in this case. In case one would be interested in changing the model to asoft-clustering method then the probability destiny function of the final model should have been chosen to be some properly parametrized Von Mises Fisher distributions.

Since this work is focusing on the implementation of Semi-supervised HMRF-Kmeans in the IR domain framework, it has to be noted that this Semi-supervised model advantage is the interactivelearning setting where this model can be used [3], since the constraints are provided to the model in two different sets the *Must-Link* and*Cannot-Link*. These sets are not necessarily the same in size or complimentary one to the other.

What it follows is the model inference line of thought based on the there resources [1, 3, 2].

2 Model inference

The objective of a semi-supervised model like HMRF Kmeans is to drive the procedure of clustering schema taking into account the prior knowledge we have about the clusters in the form of*must-link* and *cannot-link* constraints. The main difference in the graphical model of a topical EM algorithm is the nodes of the observed labeled data over the hidden, a.k.a latent,variables as depicted in fig.1, with red colored arrows and gray shaded nodes.

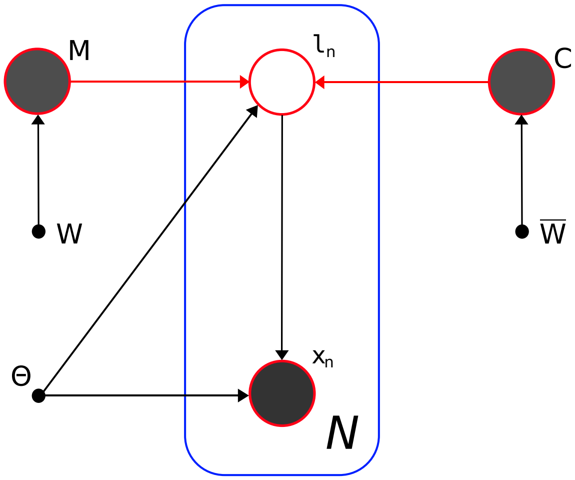

Fig. 1. Semi-supervised HMRF Expectation Maximization or Kmeans Clustering Graphical Model.

The goal of EM is to maximize the *log likelihood function* $P(X|\Theta)$ with respect of Θ as in eq.1.

$$lnP(X|\Theta) = ln\left\{\sum_{l} P(X,L|\Theta)\right\} \tag{1}$$

Where $X = \{\mathbf{x}_i\}_{i=1}^{N}$ is the set of *observable random variables* given by the conditional probability $P(X|\Theta)$ and $\mathbf{x}_i$ is a random vector (or data point) of the corpus under taxonomy. Note the boldface notation of the random vector in order not to be mixed with x_i which will a feature (or variable) of this vector. Moreover, N is the number of vector as depicted in the graphical model of fig.1.

Due to the relatively complex marginal distributions, like $P(X|\Theta)$, over observed data points where they are computationally intractable, there is a common practice to incorporate *latent or hidden variables* in order to express the conditional probability calculation more tractable over the expanded space of observed and latent variables. In eq.1 L is the set of hidden (or latent) variables over X observable data points [2].

Therefor *a hidden field* $L = \{l_i\}_{i=1}^{N}$ random variables, *whose values are unobservable.* In the clustering framework, the set of hidden variables are the unobserved cluster labels on the points, indicating cluster assignments. Every hidden variable l_i takes values from the set $1, ..., N$, which are the labels of the clusters.

Now every random data point $\mathbf{x}$ is generated from a conditional probability distribution $P(x_i|l_i)$ determined by the corresponding hidden variable l_i. The random data points X are conditionally independent given the hidden variables L. Thus $P(X|L) = \prod_{\mathbf{x}_i \in X} P(\mathbf{x}_i|l_i)$.

In eq.1 and in

References

1. Basu, S., Bilenko, M., Mooney, R.J.: A probabilistic framework for semi-supervised clustering. In: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 59–68. ACM (2004)
2. Bishop, C.: Pattern Recognition and Machine Learning, pp. 439–447. Springer (2006)
3. Chapelle, O., Scholkopf, B., Zien, A., et al.: Semi-supervised learning. pp. 73–102. The MIT Press (2006)