

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 methods* are using the provided supervision for guiding the algorithm towards a data partitioning which is avoiding (but not preventing) the constraints violation.
2. *Distance-based approaches* in clustering method with a particular distance function; 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 were provided in the clustering model in the form of data-labels pairs or in the initialization phase of the clustering 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 fixed number of clusters. However, the same models can be transformed into a relatively easy soft-clustering model where of each data point only Maximum a-posteriori Probability (MAP) of the point to be a member of each cluster of the 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 distortion measure 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 a soft-clustering method then the probability density 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 interactive learning 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 complementary 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.

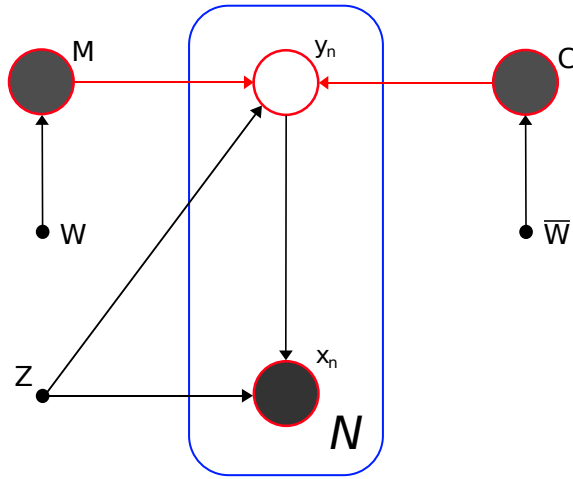


Fig. 1. Semi-supervised HMRF Expectation Maximization or Kmeans Clustering Graphical Model.

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. Springer (2006)
3. Chapelle, O., Scholkopf, B., Zien, A., et al.: Semi-supervised learning. The MIT Press (2006)