Partially-supervised context-specific independence mixture modeling

Benjamin Georgi and Alexander Schliep

Max Planck Institute for Molecular Genetics, Dept. of Computational Molecular Biology, Ihnestrasse 73, 14195 Berlin, Germany

Abstract. Partially supervised or semi-supervised learning refers to machine learning methods which fall between clustering and classification. In the context of clustering, labels can specify *link* and *do-not-link* constraints between data points in different ways and constrain the resulting clustering solutions. This is a very natural framework for many biological applications as some labels are often available and even very few label greatly improve clustering results.

Context-specific independence models constitute a framework for simultaneous mixture estimation and model structure determination to obtain meaningful models for high-dimensional data with many, possibly uninformative, variables. Here we present the first approach for partial learning of CSI models and demonstrate the effectiveness of modest amounts of labels for simulated data and for protein sub-family determination.

1 Introduction

Historically, clustering and classification or learning from unlabeled data and learning from labeled data were considered antipodes in machine learning with little common ground. For several application areas however, problems occupy a middle ground between them: we will focus on examples from molecular biology and on improving clustering approaches. For example, disease sub-types are often defined by clustering patients based on clinical data; clusters and their representatives are subsequently used for predicting disease outcome or choosing optimal treatment strategies (e.g., [1]). A pure unsupervised approach has to ignore information about known sub-types, which otherwise, even if incomplete, at least provides a lower bound on the number of sub-types. Moreover, it will violate known *positive* links between patients diagnosed and confirmed to suffer from the same sub-type and *negative* links between patients diagnosed and confirmed to be afflicted by distinct sub-types. The incomplete set of sub-type labels provides constraints which should not be violated in the final clustering solution.

The same general considerations about clustering and partial information apply, if we replace patients by genes and disease sub-type by cell cycle phase [2], or if we replace patients by proteins and disease sub-type by functionally related sub-group [3]. Generally speaking, pretending complete ignorance about cluster

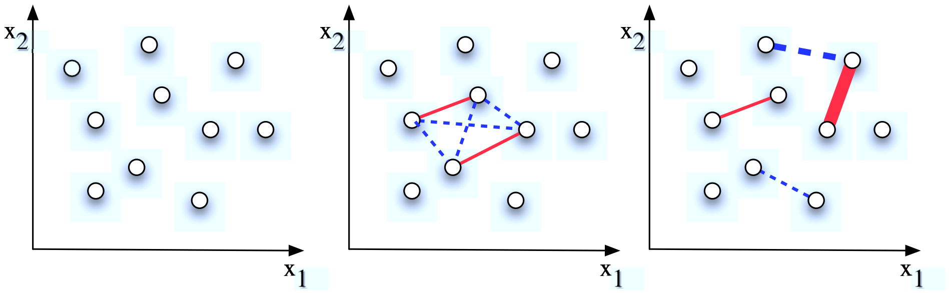

Fig. 1. Variants of partially-supervised clustering: The clustering instance of bivariate data (left) becomes easier once labels are introduced (middle). Here data points connected with a red solid line (positive constraint) share the same label. Negative constraints, indicated by dashed blue lines, result implicitly from positive constraints. A more flexible formulation (right) allows *explicit* specification of positive and negative constraints and allows to specify weights, indicated by edge weights, for the pair-wise constraints.

structure is not reflective of the availability of unlabeled mass data and sparse, labeled high quality data for a wide range of biological settings.

A recent book [4] presents a nice overview of semi-supervised learning. A lot of the literature concentrates on improving classification motivated by the observation that decrease in classification error is exponential in the proportion of labeled data [5]. Since then, a number of approaches followed the same general idea. They range from classifying text documents by constructing weighted graphs [6], partitioning graphs by min-cuts controlled by labeled examples [7], or inferring the (minimal) sub-manifold from labeled and unlabeled data and using the labeled samples for classification [8]. Cozman [9] studied how supervised mixtures get corrupted by unlabeled examples, which can also be interpreted in the framework of transductive learning [10]. More recently, a framework for integrating labeled data when learning Hidden Markov Random Fields [11] was introduced.

For clustering several variants under several names—partially supervised, semi-supervised learning, respectively constrained clustering—have been proposed. We will concentrate on clustering with mixture models [12], as mixtures have been identified as the model of choice for complex data such as gene-expression time-courses [13] and provide a sound statistical framework for extensions. The first bioinformatics application for which partial learning was proposed was concerned with improving clustering of gene expression time-courses [14]. A mixture with hidden Markov model components was trained with a variant of the expectation-maximization (EM) algorithm which essentially implemented a hard assignment of genes to clusters. The two steps of the EM are, first, computing posterior probabilities for component models given the data based on current model parameter estimates and second, estimating updated parameters from the data where the posteriors specify the influence a particular data point has in the estimation of the parameters (see [15] for details). Recall that unlike the k-means algorithm all the data points contribute to the estimation of every component; the weighting by posterior means that ill-fitting data points contribute less. The label can be effectively used in the EM by *setting* the posterior of data points with the same label to unity for the same designated component. These explicit positive constraints (i.e., link these data points, cf. Fig. 1) do not say anything about the parameters of the designated component, they just make sure that the labeled points assigned contribute maximally to the estimation of its parameters. While data points can have distinct labels, each label corresponding to one specific component, negative constraints only arise implicitly between all pairs of data points with distinct labels. For example, it is not possible to specify two negative constraints between two pairs of data points. The advantages are an easy implementation and that the local convergence result of the EM still apply [14]. Noteworthy is the very large positive effect on clustering quality even small quantities (less than 1%) of labels.

The hard assignment can be relaxed to soft assignment by specifying posterior distributions which do not put all the mass on one component. Both implementation and theory remain unchanged. However, even for the soft assignment, it is not possible to directly use information about pair-wise similarity or dissimilarity of data points, a type of information often abundant in bioinformatics, in the EM. In other words the constraints are not weighted and a reformulation in terms of the posteriors is likely cumbersome. Recently [16, 17] a new approach was proposed to use additional soft constraints for observations in the form of pair-wise positive (link) respectively negative (do-not-link) constraints w_{ij}^{+} respectively $w_{ij}^{-} \in [0, 1]$, which reflect the degree of linking for each pair of observations; cf. Fig. 1 (right).

In parallel to this development several approaches and many applications were introduced which essentially combine mixture estimation and model structure determination to improve learning on instances with many, possibly uninformative variables, with sparse data and, ultimately, arrive at more meaningful models for high-dimensional data. The central idea of these approaches is to automatically adapt model complexity to the degree of variability present in a given data set. This notion of *context-specific independence* (CSI) arose in the Bayesian network community [18–20] and has been successfully applied in mixture model framework for application such as clustering of gene expression data [21], transcription factor binding site detection [22], subtype discovery in complex genetic disease data [23] or clustering and functional annotation of protein families [3].

In the following we propose the first approach to combine CSI structure learning with the integration of prior knowledge in a partially supervised learning setup, using hard constraints on the component posteriors for labeled data.

2 Methods

2.1 CSI Mixture Models

Let $X_1, ..., X_p$ be random variables. Given a data set D with N samples, $D = x_1, ..., x_N$ with $x_i = (x_{i1}, ..., x_{ip})$ a conventional mixture density is defined as

$$P(x_i) = \sum_{k=1}^{K} \pi_k \ f_k(x_i|\theta_k), \tag{1}$$

the non-negative π_k are the mixture coefficients, $\sum_{k=1}^{K} \pi_k = 1$ and each component distribution f_k is a product of distributions over each of the X_i,(i = 1, ..., p) parameterized by parameters $\theta_k = (\theta_{k1}, ..., \theta_{kp})$,

$$f_k(x_i|\theta_k) = \prod_{j=1}^{p} P_j(x_{ij}|\theta_{kj}). \tag{2}$$

The full parameterization of the mixture is then given by $\theta = (\pi, \theta_1, ..., \theta_k)$. For a data set D of N samples the likelihood under mixture M is simply the product of the mixture densities of each sample

$$P(D|M) = \prod_{i=1}^{N} P(x_i). \tag{3}$$

The central idea of the CSI extension to the mixture framework is that it is unnecessary to have unique parameters θ_{kj} for *all* components in *each* feature. Rather the number of parameters should be adapted to the degree of variability observed in the data. This means that multiple components share parameters for features where there is no discriminatory information for the induced grouping of the data. The CSI principle is visualized in Fig. 2. On the left side the model structure of a conventional mixture is visualized. Each cell of the matrix represents an uniquely parameterized distribution and there is a unique distribution for each component in each feature. The matrix on the right shows one possible CSI structure. Here cells spanning multiple rows represent which components share parameters in each feature. For instance for feature X_1 and X_3 components C_4 and C_5 share parameters, for feature X_2, C_1 is uniquely parameterized and for feature X_4 all components share a parameterization.
Formally the CSI mixture model is defined as follows: For the set of K component indexes $\mathcal{C} = \{1, .., K\}$ and features $X_1, ..., X_p$ let $G = \{g_j\}_{(j=1,...,p)}$ be the CSI structure of the model M. Then $g_j = (g_{j1}, ...g_{jZ_j})$ with Z_j given by the number of subgroups for X_j and each $g_{jr}, r = 1, ..., Z_j$ is a subset of component indexes from $\mathcal{C}$. That means, each g_j is a partition of $\mathcal{C}$ into disjunct subsets where each g_{jr} represents a subgroup of components with the same distribution for X_j. The CSI mixture distribution is then obtained by replacing $f_{kj}(x_{ij}; \theta_{kj})$ with $f_{kj}(x_{ij}; \theta_{g_j(k)j})$ in (2) where $g_j(k) = r$ such that $k \in g_{jr}$. Accordingly $\theta_M = (\pi, \theta_{X_1|g_{1r}}, ..., \theta_{X_p|g_{pr}})$ is the full model parameterization and $\theta_{X_j|g_{jr}}$ denotes the different parameter sets in the structure for feature j. The complete CSI model M is then given by $M = (G, \theta_M)$.

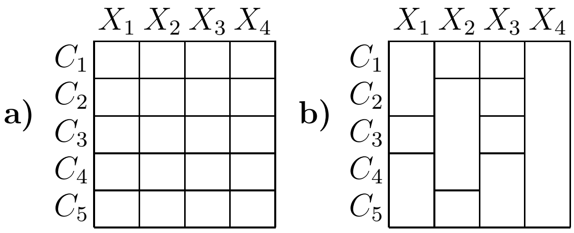

Fig. 2. Model structure matrices for a) conventional mixture model with five components over four features and b) corresponding CSI mixture model.

2.2 Partially supervised learning

The learning task in the CSI setup consist of inferring the parameterization of the mixture Θ and the CSI structure G. For the former, the standard technique is the *Expectation Maximization* (EM) algorithm [24], for the latter we apply a Bayesian approach in the structural EM framework [25, 22]. One central quantity for both of these algorithms is the posterior of component membership given by

$$\tau_{ik} = \frac{\pi_k \, f_k(x_i|\theta_k)}{\sum_{k=1}^{K} \pi_k \;\, f_k(x_i|\theta_k)}, \tag{4}$$

i.e., τ_{ik} is the probability that a sample x_i was generated by component k. In the EM algorithm the posterior is essentially a weight that determines the contribution of a sample to the parameters of a component. In the structure learning the posterior is used to compute the expected sufficient statistics of candidate structures, which then can be evaluated by the model posterior in an efficient manner (see [22] for details).
For the partially supervised case, a number of samples is assigned to components *a priori* by the labels. For a labeled sample x_i with label l this means $\tau_{ik} = 1$ for $k = l$ and 0 for all other k. This binds the contribution of the sample to parameter estimation and structure learning to a specific component. In the same way that this modification of the posterior implements partially supervised learning for the parametric EM, it gives rise to the partially supervised Structural EM algorithm in the CSI structure learning framework [21, 25, 22].

3 Results

3.1 Simulation study

In order to demonstrate the impact of a small number of labels on parameter estimation and structure learning we compared the performance of models trained with and without labels on simulated data. The generating model G was a Gaussian mixture with uniform weights on the three components over 12 features. The first two features were informative for the discrimination of the components, the remaining ten were uninformative with equal randomly chosen parameters for

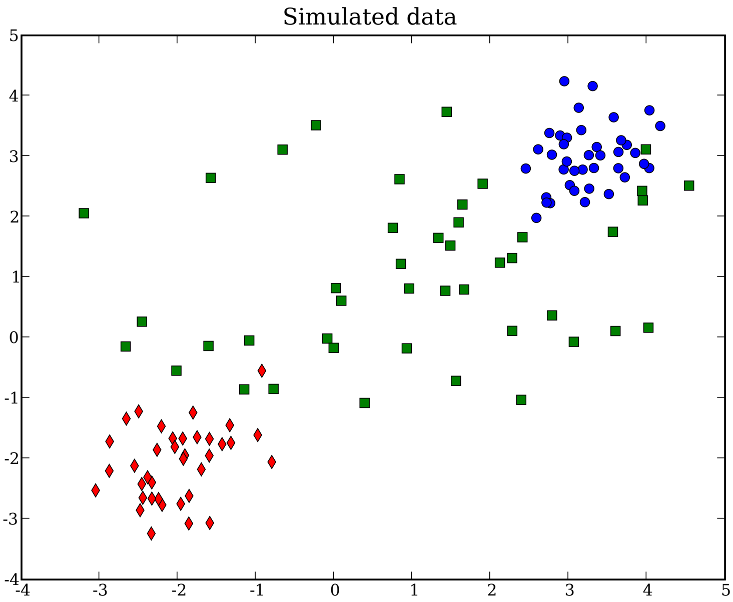

Fig. 3. Example simulated data for the first two informative features. The distinct classes are indicated by carets, rectangles and circles respectively.

all components. An example data set for the informative components is shown in Fig. 3. Two components were rather compact with diagonal covariance matrices and diagonal entries 0.5, the other component was more spread out (diagonal covariance with diagonal entries 1.5). The components with smaller variance each overlapped to a degree with the central large-variance component. The ten uninformative features provided the opportunity for the structure learning to adapt model complexity in the learned models.

We sampled 30 data sets of size 120 from G and trained CSI mixtures with and without labels. For the latter three labels were used for each component. The average performance of the models over the 30 data sets with respect to the true component labels is summarized in Tab. 1

	Unlabeled		Partially Labeled	
Sensitivity	92.76%	SD 3.96%	92.10%	SD 4.13%
Specificity	71.47%	SD 15.13%	91.56%	SD 4.34%

Table 1. Average sensitivity and specificity for labeled and unlabeled data over 30 simulated data sets.

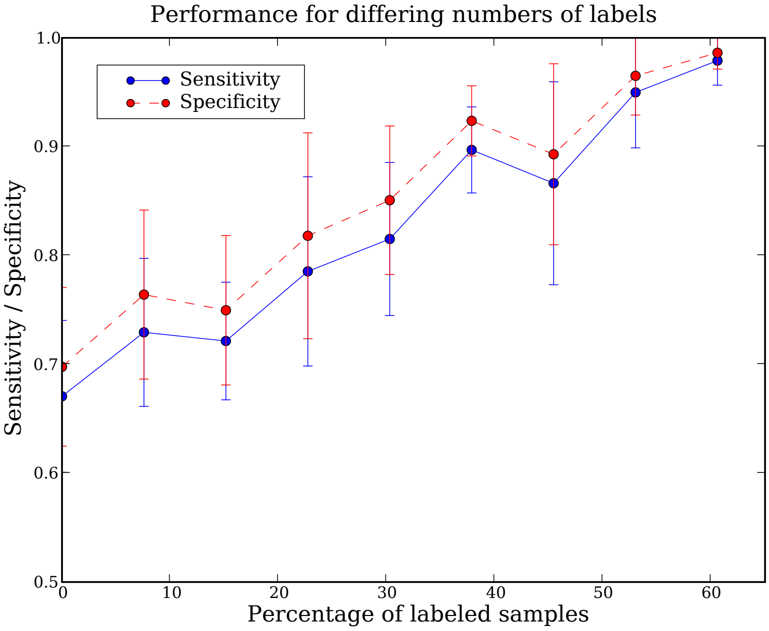

Fig. 4. Average sensitivity and specificity of clustering of the nucleotidyl cyclase data for different numbers of labels. Standard deviations are shown by error bars.

It can be seen that the addition of three labels for each components yields a considerable increase in specificity of the trained models. To assess the impact of the labeling on the structure learning we considered the edit distance of the learned structures to the true structure in G with respect to merge/split operations in the structure matrix. For instance the edit distance of Fig. 2a) to 2b) is nine since nine merges are needed to convert a) into b) (the same holds for splits in the other direction). The average edit distance of the models based on unlabeled data was 6.3 (SD 4.71), the labeled data yielded an average distance of 0.17 (SD 0.38). This indicates a greatly increased precision in the structure learning for the labeled data.

3.2 Protein sequence data

In order to examine the effect of labels in the data on a true data set we applied CSI mixture models on a multiple sequence alignment of nucleotidyl cyclase family protein sequences. We used the model extensions previously introduced for CSI for protein data [3]. The 132 sequences fall into subgroups of guanylyl cyclases (GC) and adenylyl cyclases (AC). We used the true classification into these subgroups as labels for the partially supervised learning. Labels were chosen randomly in equal numbers for GC and AC subgroups. The average sensitivity and specificity for different numbers of labels is shown in Fig. 4. It can be seen that qualitatively both sensitivity and specificity increase with the amount of prior knowledge considered, i.e. the number of labels assigned to the data set. It is noteworthy that for 60 labels (45% of the data set labeled) there is a drop in performance. This can probably be attributed to the random choice of labels. If by chance a poor selection of labels is chosen, for instance only labels from one boundary region of a cluster, the partially supervised approach may actually mislead the parameter estimation.

4 Discussion

The results on the simulated data indicate that a partially supervised setup even for a small number of labels greatly increases the clustering performance. While sensitivity was similar for unlabeled and labeled data, the addition of labels yielded greatly increased specificity. This was the expected result from the literature on partially supervised learning. A more interesting question was how much the CSI structure learning would be impacted by the labels. The vastly smaller structure edit distance to the true CSI structure of the generating model we observed for the partially supervised case, indicates that the structure learning can also greatly benefit from the addition of labels.

When applying the partially supervised learning on protein data the picture was somewhat more noisy, though the advantage of the labeling could still be seen. The rather high variance in results we observed can probably be attributed to the inherent noisiness of the data and the random choice of labels. Taken together the results suggest that the partially supervised learning can bring considerable improvement to both the parameter estimates and the learned CSI structure but one should be aware that in order to fulfil its potential the appraoch requires high-quality labels.

There are several open questions regarding the objective formulation for partially supervised learning of CSI models, in particular if pair-wise constraints need to be included, as the CSI structure controls cluster membership only indirectly and, more importantly, *not* variable-wise but rather by all variables simultaneously. This suggests that pair-wise constraints could negate the computational advantages gained by the independence assumption between variables. Nevertheless, the bioinformatics applications directly drive the need for partially supervised learning and our results show that non-trivial improvements can be realized on realistic instances from applications.

References

1. Hummel, M., Bentink, S., Berger, H., Klapper, W., Wessendorf, S., Barth, T.F.E., Bernd, H.W., Cogliatti, S.B., Dierlamm, J., Feller, A.C., Hansmann, M.L., Haralambieva, E., Harder, L., Hasenclever, D., Kuhn, M., Lenze, D., Lichter, P., Martin-Subero, J.I., Moller, P., Muller-Hermelink, H.K., Ott, G., Parwaresch,

R.M., Pott, C., Rosenwald, A., Rosolowski, M., Schwaenen, C., Sturzenhofecker, B., Szczepanowski, M., Trautmann, H., Wacker, H.H., Spang, R., Loeffler, M., Trumper, L., Stein, H., Siebert, R.: A biologic definition of Burkitt's lymphoma from transcriptional and genomic profiling. N Engl J Med **354**(23) (Jun 2006) 2419–2430

2. Schliep, A., Costa, I.G., Steinhoff, C., Schönhuth, A.: Analyzing gene expression time-courses. IEEE/ACM Transactions on Computational Biology and Bioinformatics **2**(3) (2005) 179–193
3. Georgi, B., Schultz, J., Schliep, A.: Context-specific independence mixture modelling for protein families. In: Proceedings of the European Conference on Principles and Practice of Knowledge Discovery in Databases, Springer (2007)
4. Chapelle, O., Schölkopf, B., Zien, A., eds.: Semi-Supervised Learning. MIT Press, Cambridge, MA (2006)
5. Castelli, V., Cover, T.M.: On the exponential value of labeled samples. Pattern Recognition Letters **16** (1994) 105–111
6. Szummer, M., Jaakkola, T.: Partially labeled classification with Markov random walks. Neural Information Processing Systems (NIPS) **14** (2002)
7. Blum, A., Chawla, S.: Learning from labeled and unlabeled data using graph mincuts. In: ICML. (2001)
8. Belkin, M.: Problems of learning on manifolds. PhD thesis, University of Chicago (2003)
9. Cozman, F.G., Cohen, I., Cirelo, M.C.: Semi-supervised learning of mixture models. In: Proceedings of the Twentieth International Conference on Machine Learning (ICML 2003), Washington DC, 2003. (2003)
10. Vapnik, V.: The Nature of Statistical Learning Theory. Wiley (1998)
11. 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 (KDD-2004), Seattle WA, August 2004. (2004)
12. McLachlan, G., Peel, D.: Finite Mixture Models. Wiley Series in Probability and Statistics. Wiley, New York (2000)
13. Bar-Joseph, Z.: Analyzing time series gene expression data. Bioinformatics **20**(16) (Nov 2004) 2493–503
14. Schliep, A., Schönhuth, A., Steinhoff, C.: Using Hidden Markov models to analyze gene expression time course data. Bioinformatics **19 Suppl 1** (Jul 2003) I255–I263
15. McLachlan, G., Krishnan, T.: The EM Algorithm and Extensions. Wiley, New York (1997)
16. Lange, T., Law, M.H.C., Jain, A.K., Buhmann, J.M.: Learning with constrained and unlabelled data. In: CVPR (1), IEEE Computer Society (2005) 731–738
17. Lu, Z., Leen, T.: Semi-supervised learning with penalized probabilistic clustering. In Saul, L.K., Weiss, Y., Bottou, L., eds.: Advances in Neural Information Processing Systems 17. MIT Press (2005) 849–856
18. Boutilier, C., Friedman, N., Goldszmidt, M., Koller, D.: Context-specific independence in Bayesian networks. In: Uncertainty in Artificial Intelligence. (1996) 115–123
19. Chickering, D.M., Heckerman, D.: Efficient approximations for the marginal likelihood of bayesian networks with hidden variables. Mach. Learn. **29**(2-3) (1997) 181–212
20. Friedman, N., Goldszmidt, M.: Learning bayesian networks with local structure. In: Proceedings of the NATO Advanced Study Institute on Learning in graphical models, Norwell, MA, USA, Kluwer Academic Publishers (1998) 421–459

21. Barash, Y., Friedman, N.: Context-specific bayesian clustering for gene expression data. J Comput Biol **9**(2) (2002) 169–91
22. Georgi, B., Schliep, A.: Context-specific independence mixture modeling for positional weight matrices. Bioinformatics **22**(14) (2006) e166–73
23. Georgi, B., Spence, M., Flodman, P., Schliep, A.: Mixture model based group inference in fused genotype and phenotype data. In: Studies in Classification, Data Analysis, and Knowledge Organization, Springer (2007)
24. Dempster, A., Laird, N., Rubin, D.: Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B (1977) 1–38
25. Friedman, N.: The Bayesian structural EM algorithm. Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence (1998) 129–138