Semi-supervised cross-entropy clustering with information bottleneck constraint

M. Śmieja[a,*], B. C. Geiger[b]

[a] *Faculty of Mathematics and Computer Science*
Jagiellonian University
Łojasiewicza 6, 30-348 Krakow, Poland
[b] *Institute for Communications Engineering*
Technical University of Munich
Theresienstr. 90, D-80333 Munich, Germany

Abstract

In this paper, we propose a semi-supervised clustering method, CEC-IB, that models data with a set of Gaussian distributions and that retrieves clusters based on a partial labeling provided by the user (partition-level side information). By combining the ideas from cross-entropy clustering (CEC) with those from the information bottleneck method (IB), our method trades between three conflicting goals: the accuracy with which the data set is modeled, the simplicity of the model, and the consistency of the clustering with side information. Experiments demonstrate that CEC-IB has a performance comparable to Gaussian mixture models (GMM) in a classical semi-supervised scenario, but is faster, more robust to noisy labels, automatically determines the optimal number of clusters, and performs well when not all classes are present in the side information. Moreover, in contrast to other semi-supervised models, it can be successfully applied in discovering natural subgroups if the partition-level side information is derived from the top levels of a hierarchical clustering.

Keywords: semi-supervised clustering, partition-level side information, model-based clustering, cross-entropy, information bottleneck

*Corresponding author
Email address: marek.smieja@ii.uj.edu.pl (M. Śmieja)

1. Introduction

Clustering is one of the core techniques of machine learning and data analysis, and aims at partitioning data sets based on, e.g., the internal similarity of the resulting clusters. While clustering is an unsupervised technique, one can improve its performance by introducing additional knowledge as side information. This is the field of semi-supervised or constrained clustering.

One classical type of side information in clustering are pairwise constraints: human experts determine whether a given pair of data points belongs to the same (must-link) or to different clusters (cannot-link) [1]. Although this approach received high attention in the last decade, the latest reports [2] suggest that in real-life problems it is difficult to answer whether or not two objects belong to the same group without a deeper knowledge of data set. This is even more problematic as erroneous pairwise constraints can easily lead to contradictory side information [3].

A possible remedy is to let experts categorize a set of data points rather than specifying pairwise constraints. This *partition-level side information* was proposed in [4] and recently considered in [5]. The concept is related to partial labeling applied in semi-supervised classification and assumes that a small portion of data is labeled. In contrast to semi-supervised classification [6, 7], the number of categories is not limited to the true number of classes; in semi-supervised clustering one may discover several clusters among unlabeled data points. Another advantage of partition-level side information is that, in contrast to pairwise constraints, it does not become self-contradictory if some data points are mislabeled.

In this paper, we introduce a semi-supervised clustering method, CEC-IB, based on partition-level side information. CEC-IB combines Cross-Entropy Clustering (CEC) [8, 9, 10], a model-based clustering technique, with the Information Bottleneck (IB) method [11, 12] to build **the smallest model that preserves the side information and provides a good model of the data distribution**. In other words, CEC-IB automatically determines the required number of clusters to trade between model complexity, model accuracy, and consistency with the side information.

Consistency with side information is ensured by penalizing solutions in which data points from different categories are put in the same cluster. Since modeling a category by multiple clusters is not penalized, one can apply CEC-IB to obtain a fine clustering even if the the human expert categorizes the data into only few basic groups, see Figure 1. Although this type of side

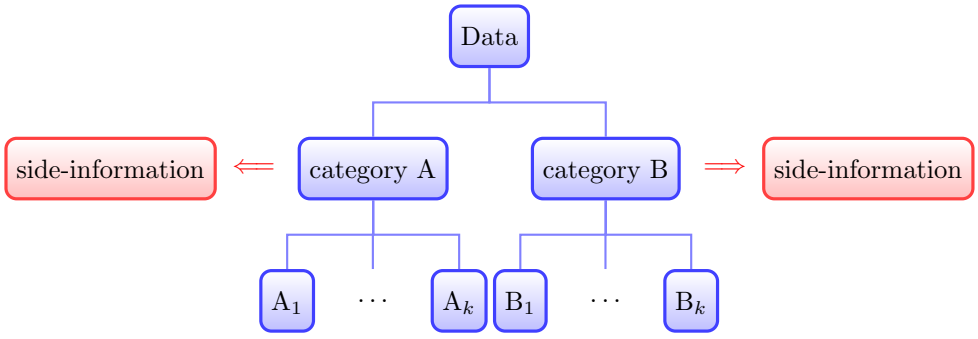


Figure 1: Subgroups discovery task. The expert provides side information by dividing a data set into two categories. Making use of this knowledge, the algorithm discovers natural subgroups more reliably than in the unsupervised case.

information seems to be a perfect use case for cannot-link constraints, the computational cost of introducing side information to CEC-IB is negligible while the incorporation of cannot-link constraints to similar Gaussian mixture model (GMM) approaches requires the use of graphical models, which involves high computational cost. CEC-IB thus combines the flexibility of cannot-link constraints with an efficient implementation.

We summarize the main contributions and the outline of our paper:

1. We combine ideas from model-based CEC and from the information bottleneck method to formulate our clustering method CEC-IB for both complete and partial side information (Sections 3.2 and 3.3). The proposed method does not require the true number of clusters as an input.
2. We propose a modified Hartigan algorithm to optimize the CEC-IB cost function (Section 3.4). The algorithm has a complexity that is linear in the number of data points in each iteration, and it usually requires less iterations than the expectation-maximization (EM) algorithm used for fitting GMMs (Appendix C).
3. We provide a theoretical analysis of the trade-off between the CEC and the IB cost function in Section 4. This places the parameter selection problem on a solid mathematical ground (see Theorems 4.1 and 4.2).
4. We perform extensive experiments demonstrating that CEC-IB is more robust to noisy side information (i.e., miscategorized data points) than state-of-the-art approaches to semi-supervised clustering (Section 5.4). Moreover, CEC-IB performs well when not all categories are present

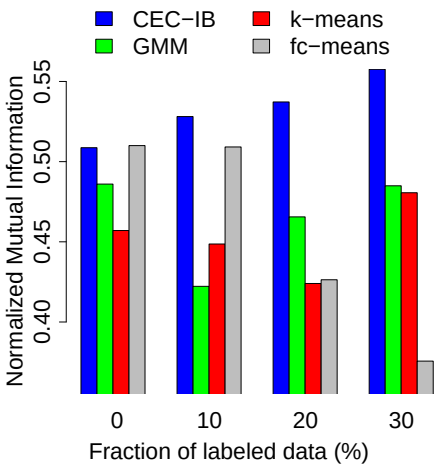


Figure 2: Detection of chemical subgroups when 10% of side information was erroneous. The results of CEC-IB, GMM with cannot-link constraints, constrained k-means and fuzzy c-means (fc-means) were measured by normalized mutual information.

in the side information (Section 5.3), even though the true number of clusters is not specified.

5. We perform two case studies: In Section 5.6, a human expert provided a partition-level side information about the division of chemical compounds into two basic groups (as in Figure 1); CEC-IB discovers natural chemical subgroups more reliably than other semi-supervised methods, even if some labels are misspecified by the expert (Figure 2). The second case study in Section 5.7 applies CEC-IB to image segmentation.

2. Related work

Clustering has been an important topic in machine learning and data analysis for a long time. Various methods were introduced for splitting data into groups, including model-based, distance-based, spectral, fuzzy, and hierarchical methods (see [13, 14] for a survey).

Adding to this diversity of techniques, a large number of specialized types of clustering have been developed. One example is multi-view clustering, which considers gathering information coming from different domains [15].

As another example, complementary or alternative clustering aims at finding groups which provide a perspective on the data that expands on what can be inferred from previous clusterings [16]. Finally, semi-supervised clustering – the problem investigated in this work – makes use of side information to achieve better clustering results or to provide robustness against noisy side information [1].

The traditional approach to incorporate side information into clustering is based on pairwise constraints. The authors of [17] suggested reducing distances between data points with a must-link constraint and adding a dimension for each cannot-link constraint. After updating all other distances to, e.g., satisfy the triangle inequality, the thus obtained pairwise distance matrix can be used for unsupervised clustering. Kamvar et al. [18] considered a similar procedure, taking the pairwise affinity matrix and setting must-links and cannot-links to predefined maximum and minimum values, respectively. Instead of clustering, they applied eigenvector-based classification taking the labeled data as training set. Another spectral technique, proposed in [19], relies on solving a generalized eigenvalue problem. Qian et al. [20] developed a framework for spectral clustering that allows using side information in the form of pairwise constraints, partial labeling, and grouping information. An information-theoretic cost function, squared mutual information, was proposed for semi-supervised clustering in [21]. Also clustering techniques based on non-negative matrix or concept factorization can incorporate pairwise constraints as regularizers [22].

As mentioned in the introduction, partition-level side information refers to a partial labeling of the data points that need not necessarily consider all classes – the categories provided as side information may be only a subset of classes, or, as in Figure 1, be of a hierarchical nature. In consequence, clustering with partition-level side information differs significantly from a typical semi-supervised classification task, as the clustering algorithm should detect clusters within categories and/or within unlabeled data points. A recent paper using partition-level side information is [5], where the authors add additional dimensions to feature vectors and propose a modification of k-means to cluster data points. In [4], partition-level side information was used to design a better initialization strategy for k-means. Similarly, partition-level side information was used to propose a semi-supervised version of fuzzy c-means [23, 24]. The authors added a regularization term to the fuzzy c-means cost function that penalizes fuzzy clusterings that are inconsistent with the side information. This technique was later combined with feature

selection methods [25]. Finally, partition-level side information can be used in density-based clustering such as DBSCAN. Specifically, in [26] the authors proposed an algorithm that sets the parameter defining the neighborhood radius of a data point based on partial labeling.

GMMs can be easily adapted to make us of partition-level side information by combining the classical unsupervised GMM with a supervised one [27, 6]. This approach can be extended to labels with reliability information [28, 29, 30]. Various statistical and machine learning libraries, such as mixmod [31] or bgmm [32], provide implementations of GMMs with partition-level side information.

Also pairwise constraints can be incorporated into GMMs, where dependencies between the hidden cluster indicator variables are then usually modeled by a hidden Markov random field. This procedure was adopted, for example, in [33] to account for cannot-link constraints. Must-link constraints were considered by treating all involved data points as a single data point with a higher weight. The parameters of the GMM, which was used for hard or soft clustering, are obtained by a generalized expectation-maximization procedure that requires simplifications or approximations [34, 35, 36]. An overview of GMM-based methods with pairwise constraints can be found in [37].

In contrast to most GMM approaches, our method does not require knowledge of the correct number of clusters; initialized with any (larger) number, CEC-IB reduces the number of clusters for an optimal trade-off between model accuracy, model complexity (i.e., number of clusters), and consistency with the side information.

Our method is closely related to the information bottleneck method, which focuses on lossy compression of data preserving the information of a stochastically related random variable [11, 38]. Modifications of IB were used in consensus clustering [39] or alternative clustering [16]. The mutual information between data points and its clusters, which describes the cost of (lossy) data compression in IB, is replaced in our model by the cross-entropy – see Section 3.2 for more details. Thus, while IB focuses model simplicity and consistency with side information, CEC-IB adds model accuracy to the cost.

3. Cross-Entropy Clustering with an Information Bottleneck Constraint

We now pave the way for our CEC-IB method. Since our model is related to CEC, we first review its basics in Section 3.1. For completely labeled data, i.e., for the case where all data points are labeled, we then introduce our CEC-IB model based on ideas from IB in Section 3.2. Section 3.3 extends the analysis to deal with the case where only some data points are labeled. We conclude this section by presenting and analyzing a clustering algorithm that finds a local optimum of our CEC-IB cost function.

3.1. Cross-entropy clustering

CEC is a model-based clustering method that minimizes the empirical cross-entropy between a finite data set $X \subset \mathbb{R}^N$ and a parametric mixture of densities [8]. This parametric mixture is a subdensity[1] given by

$$f = \max(p_1 f_1, \ldots, p_k f_k)$$

where p_1 through p_k are non-negative weights summing to one and where f_1 through f_k are densities from the Gaussian family $\mathcal{G}$ of probability distributions on $\mathbb{R}^N$. The empirical cross-entropy between X and subdensity f is

$$H^{\times}(X\|f) = -\frac{1}{|X|}\sum_{x\in X}\log f(x) = -\frac{1}{|X|}\sum_{i=1}^{k}\sum_{x\in Y_i}\log(p_i f_i(x))$$

where

$$\mathcal{Y} = \{Y_1, \ldots, Y_k\}, \quad Y_i := \{x \in X\colon p_i f_i(x) = \max_j p_j f_j(x)\} \tag{1}$$

is a partition of X induced by the subdensity f. Letting

$$\mu_{Y_i} = \frac{1}{|Y_i|}\sum_{x\in Y_i} x,$$
$$\Sigma_{Y_i} = \frac{1}{|Y_i|}\sum_{x\in Y_i}(x-\mu_{Y_i})(x-\mu_{Y_i})^T$$

[1] i.e., $f(x) \geq 0$ and $\int_{\mathbb{R}^N} f(x)dx \leq 1$.

be the sample mean vector and sample covariance matrix of cluster Y_i, we show in Appendix A that CEC looks for a clustering $\mathcal{Y}$ such that the following cost is minimized:

$$H^{\times}(X\|f) = H(\mathcal{Y}) + \sum_{i=1}^{k} \frac{|Y_i|}{|X|} H(\mathcal{N}(\mu_{Y_i}, \Sigma_{Y_i})), \tag{2}$$

where the model complexity is measured by the Shannon entropy of the partition $\mathcal{Y}$,

$$H(\mathcal{Y}) := -\sum_{i=1}^{k} \frac{|Y_i|}{|X|} \log \frac{|Y_i|}{|X|},$$

and where the model accuracy, i.e., accuracy of density estimation in cluster Y_i, is measured by the differential entropy of the Gaussian density f_i,

$$H(\mathcal{N}(\mu_{Y_i}, \Sigma_{Y_i})) = \tfrac{N}{2} \ln(2\pi e) + \tfrac{1}{2} \ln \det(\Sigma_{Y_i}) = \min_{f_i \in \mathcal{G}} H^{\times}(Y_i\|f_i).$$

The main difference between CEC and GMM-based clustering lies in substituting a mixture density $f = p_1 f_1 + \cdots + p_k f_k$ by a subdensity $f = \max(p_1 f_1, \ldots, p_k f_k)$. This modification allows to obtain a closed form solution for the mixture density given a fixed partition $\mathcal{Y}$, while for a fixed mixture density the partition $\mathcal{Y}$ is given in (1). This suggests a heuristic similar to the k-means method. In consequence, CEC might yield a slightly worse density estimation of data than GMM, but converges faster (see Section 3.4 and Appendix Appendix C) while the experimental results show that the clustering effects are similar.

3.2. CEC-IB with completely labeled data

We now introduce CEC-IB for completely labeled data (i.e., all data points are labeled) by combining the ideas from model-based clustering with those from the information bottleneck method. We also show that under some assumptions CEC-IB admits an alternative derivation based on conditional cross-entropy given the side information.

Definition 3.1. Let X be a finite data set and let $X_\ell \subseteq X$ denote the set of labeled data points. The *partition-level side information* is a partition $\mathcal{Z} = \{Z_1, \ldots, Z_m\}$ of X_ℓ, where every $Z_j \in \mathcal{Z}$ contains all elements of X_ℓ with the same label.

To make this definition clear, suppose that $\mathcal{X} = \{X_1, X_2, \ldots, X_l\}$ is the *true* partition of the data that we want to recover, i.e., we want to obtain $\mathcal{Y} = \mathcal{X}$. The partition-level side information $\mathcal{Z}$ can take several possible forms, including:

- $|\mathcal{Z}| = l$, and $Z_j \subseteq X_j$ for $j = 1, \ldots, l$. This is equivalent to the notion of partial labeling in semi-supervised classification.
- $|\mathcal{Z}| = m < l$ and for every $j = 1, \ldots, m$ there is a different i such that $Z_j \subseteq X_i$. This is the case where only some of the true clusters are labeled.
- $|\mathcal{Z}| = m < l$ and there are m disjoint sets $I_j \subset \{1, \ldots, l\}$ such that $Z_j \subset \bigcup_{i \in I_j} X_i$. This is the case where the labeling is derived from a higher level of the hierarchical true clustering (cf. Figure 1).

For the remainder of this subsection, we assume that the side information is complete, i.e., that each data point $x \in X$ is labeled with exactly one category. In other words, $X_\ell = X$ and $\mathcal{Z}$ is a partition of X. We drop this assumption in Section 3.3, where we consider partial labeling, i.e., $X_\ell \subseteq X$.

Our effort focuses on finding a partition that is consistent with side information:

Definition 3.2. Let X be a finite data set and let $X_\ell \subseteq X$ be the set of labeled data points that is partitioned into $\mathcal{Z} = \{Z_1, \ldots, Z_m\}$. We say that a partition $\mathcal{Y} = \{Y_1, \ldots, Y_k\}$ of X is *consistent* with $\mathcal{Z}$, if for every Y_i there exists at most one Z_j such that $Z_j \cap Y_i \neq \emptyset$.

The definition of consistency generalizes the refinement relation between partitions of the same set. If, as in this section, $X_\ell = X$, then $\mathcal{Y}$ is consistent with $\mathcal{Z}$ if and only if $\mathcal{Y}$ is a refinement of $\mathcal{Z}$. In other words, a clustering $\mathcal{Y}$ is consistent with $\mathcal{Z}$ if every $Y_i \in \mathcal{Y}$ contains elements from at most one category $Z_j \in \mathcal{Z}$. Mathematically, for a clustering $\mathcal{Y}$ consistent with $\mathcal{Z}$ we have

$$\forall Y_i \in \mathcal{Y}\colon \exists! j' = j'(i)\colon Z_j \cap Y_i = \begin{cases} Y_i & j = j' \\ 0 & \text{else.} \end{cases} \tag{3}$$

Thus, for a consistent clustering $\mathcal{Y}$ the conditional entropy $H(\mathcal{Z}|\mathcal{Y})$ vanishes:

$$H(\mathcal{Z}|\mathcal{Y}) = \sum_{i=1}^{k} \frac{|Y_i|}{|X|} H(\mathcal{Z}|Y_i) = -\sum_{i=1}^{k}\sum_{j=1}^{m} \frac{|Z_j \cap Y_i|}{|X|} \log\left(\frac{|Z_j \cap Y_i|}{|Y_i|}\right)$$
$$\stackrel{(a)}{=} -\sum_{i=1}^{k} \frac{|Z_{j'} \cap Y_i|}{|X|} \log\left(\frac{|Y_i|}{|Y_i|}\right) = 0$$

where (a) is due to (1).

The conditional entropy $H(\mathcal{Z}|\mathcal{Y})$ therefore is a measure for consistency with side information: the smaller the conditional entropy, the higher is the consistency. We thus propose the following cost function for CEC-IB in the case of complete side information, i.e., when $X_\ell = X$ and $\mathcal{Z}$ is a partition of X:

$$\mathrm{E}_\beta(X, \mathcal{Z}; \mathcal{Y}) := H(\mathcal{Y}) + \sum_{i=1}^{k} \frac{|Y_i|}{|X|} H(\mathcal{N}(\mu_{Y_i}, \Sigma_{Y_i})) + \beta H(\mathcal{Z}|\mathcal{Y}), \text{ where } \beta \geq 0. \tag{4}$$

The first two terms are the CEC cost function (2), and the last term $H(\mathcal{Z}|\mathcal{Y})$ penalizes clusterings $\mathcal{Y}$ that are not consistent with the side information $\mathcal{Z}$. Thus CEC-IB aims at finding the minimal number of clusters needed to model the data set distribution and to preserve the consistency with the side information. The weight parameter β trades between these objectives; we will analyze rationales for selecting this parameter in Section 4.

Our cost function (4) is intricately connected to the IB and related methods. In the notation of this work, i.e., in terms of partitions rather than random variables, the IB cost function is given as [11]

$$I(X;\mathcal{Y}) - \beta I(\mathcal{Y};\mathcal{Z}) = H(\mathcal{Y}) - H(\mathcal{Y}|X) - \beta H(\mathcal{Z}) + \beta H(\mathcal{Z}|\mathcal{Y}).$$

Noticing that $H(\mathcal{Z})$ does not depend on the clustering $\mathcal{Y}$, the main difference between IB and CEC-IB is that CEC-IB incorporates a term accounting for the modeling accuracy in each cluster, while IB adds a term related to the "softness" of the clustering: Since $H(\mathcal{Y}|X)$ is minimized for deterministic, i.e., hard clusters, IB implicitly encourages soft clusters. A version of IB ensuring deterministic clusters was recently introduced in [40]. The cost function of this method dispenses with the term related to the softness of the clusters leading to a clustering method minimizing

$$H(\mathcal{Y}) + \beta H(\mathcal{Z}|\mathcal{Y}).$$

Our CEC-IB method can thus be seen as deterministic IB with an additional term accounting for model accuracy. CEC-IB can therefore be considered as a model-based version of the information bottleneck method.

We end this subsection by showing that under some assumptions, one can arrive at the CEC-IB cost function (with $\beta = 1$) in a slightly different way, by minimizing the conditional cross-entropy function:

Theorem 3.1. *Let X be a finite data set that is partitioned into $\mathcal{Z} = \{Z_1, \ldots, Z_m\}$. Minimizing the* CEC-IB *cost function* (4)*, for $\beta = 1$, is equivalent to minimizing the conditional cross-entropy function:*

$$H^{\times}((X\|f)|\mathcal{Z}) := \sum_{j=1}^{m} \frac{|Z_j|}{|X|} H^{\times}(Z_j\|f_{|j}),$$

where

$$f_{|Z_j} := f_{|j} = \max(p_1(j)f_1, \ldots, p_k(j)f_k)$$

is the conditional density f given j-th category and where $p_1(j), \ldots, p_k(j)$ are non-negative weights summing to one.

The proof of this theorem is given in Appendix B. It essentially states that our cost function ensures that each category Z_j is modeled by a parametric mixture of densities that is both simple and accurate. We believe that this view on the problem can lead to the development of a clustering algorithm slightly different from what is presented in this paper.

3.3. CEC-IB with partially labeled data

The previous section assumed that all data points in X were labeled, i.e., the partition-level side information $\mathcal{Z} = \{Z_1, \ldots, Z_m\}$ was a partition of X. In this section, we relax this assumption and assume that only a subset $X_\ell \subseteq X$ is labeled. In this case $\mathcal{Z}$ is a partition only of X_ℓ, and in consequence, the conditional entropy $H(\mathcal{Z}|\mathcal{Y})$ from the previous subsection is undefined.

To deal with this problem, let $\mathcal{L} = \{X_\ell, X \setminus X_\ell\}$ denote the partition of X into labeled and unlabeled data. We decompose the conditional entropy of $\mathcal{Z}$ given partitions $\mathcal{Y}$ and $\mathcal{L}$ as

$$H(\mathcal{Z}|\mathcal{Y},\mathcal{L}) = \frac{|X_\ell|}{|X|} H(\mathcal{Z}|\mathcal{Y}, X_\ell) + \frac{|X \setminus X_\ell|}{|X|} H(\mathcal{Z}|\mathcal{Y}, X \setminus X_\ell), \tag{5}$$

where

$$H(\mathcal{Z}|\mathcal{Y}, X_\ell) = \sum_{i=1}^{k} \frac{|Y_i \cap X_\ell|}{|X_\ell|} H(\mathcal{Z}|Y_i \cap X_\ell)$$
$$= \sum_{i=1}^{k} \frac{|Y_i \cap X_\ell|}{|X_\ell|} \sum_{j=1}^{m} \frac{|Y_i \cap Z_j|}{|Y_i \cap X_\ell|} \left(-\log \frac{|Y_i \cap Z_j|}{|Y_i \cap X_\ell|} \right).$$

Let us now assume that the partition-level side information is a representative sample of true categories. In other words, assume that the probability for a category of an unlabeled data point given the cluster equals the empirical probability of this category for labeled data points in this cluster. To formalize this reasoning, we view the partition $\mathcal{Z}$ as a random variable that takes values in $\{1, \ldots, m\}$. Our labeled data set X_ℓ corresponds to realizations of this random variable, i.e., for every $x \in X_\ell$, the corresponding random variable $\mathcal{Z}$ assumes the value indicated by the labeling. Since the side information was assumed to be representative, the relative fraction of data points in cluster Y_i assigned to category Z_j gives us an estimate of the true underlying probability; we extrapolate this estimate to unlabeled data points and put

$$\mathbf{P}(\mathcal{Z} = j|Y_i \cap (X \setminus X_\ell)) = \mathbf{P}(\mathcal{Z} = j|Y_i \cap X_\ell) = \frac{|Y_i \cap Z_j|}{|Y_i \cap X_\ell|} = \mathbf{P}(\mathcal{Z} = j|Y_i).$$

Hence, $H(\mathcal{Z}|Y_i \cap (X \setminus X_\ell)) = H(\mathcal{Z}|Y_i \cap X_\ell)$ for every Y_i, and we get for (5):

$$H(\mathcal{Z}|\mathcal{Y}) = H(\mathcal{Z}|\mathcal{Y}, \mathcal{L}) \tag{6}$$
$$= \frac{|X_\ell|}{|X|} H(\mathcal{Z}|\mathcal{Y}, X_\ell) + \frac{|X \setminus X_\ell|}{|X|} H(\mathcal{Z}|\mathcal{Y}, X \setminus X_\ell) \tag{7}$$
$$= \frac{|X_\ell|}{|X|} \sum_{i=1}^{k} \frac{|Y_i \cap X_\ell|}{|X_\ell|} H(\mathcal{Z}|Y_i \cap X_\ell)$$
$$+ \frac{|X \setminus X_\ell|}{|X|} \sum_{i=1}^{k} \frac{|Y_i \cap (X \setminus X_\ell)|}{|X \setminus X_\ell|} H(\mathcal{Z}|Y_i \cap X_\ell) \tag{8}$$
$$= \sum_{i=1}^{k} \frac{|Y_i|}{|X|} H(\mathcal{Z}|Y_i \cap X_\ell). \tag{9}$$

where the first equality follows because the conditional entropy $H(\mathcal{Z}|\mathcal{Y},\mathcal{L})$ does not depend on the partition $\mathcal{L}$.

With this, we define the CEC-IB cost function for a model with partition-level side information:

Definition 3.3. (CEC-IB cost function) Let X be a finite data set and let $X_\ell \subseteq X$ be the set of labeled data points that is partitioned into $\mathcal{Z} = \{Z_1, \ldots, Z_m\}$. The cost of clustering X into the partition $\mathcal{Y} = \{Y_1, \ldots, Y_k\}$ for a given parameter $\beta \geq 0$ equals

$$\mathrm{E}_\beta(X, \mathcal{Z}; \mathcal{Y}) := H(\mathcal{Y}) + \sum_{i=1}^{k} \frac{|Y_i|}{|X|} \left(H(\mathcal{N}(\mu_{Y_i}, \Sigma_{Y_i})) + \beta H(\mathcal{Z}|Y_i \cap X_\ell) \right). \quad (10)$$

To shorten the notation we sometimes write $\mathrm{E}_\beta(\mathcal{Y})$ assuming that X and $\mathcal{Z}$ are fixed.

Note that for a complete side information, i.e., for $X_\ell = X$, we get precisely the cost function (4) obtained in the previous subsection.

3.4. Optimization algorithm

The optimization of CEC-IB cost function can be performed similarly as in the classical CEC method, in which the Hartigan approach [41] is used.

Let X be a finite data set and let $X_\ell \subseteq X$ be the set of labeled data points that is partitioned into $\mathcal{Z} = \{Z_1, \ldots, Z_m\}$. The entire procedure consists of two steps: initialization and iteration. In the initialization step, a partition $\mathcal{Y}$ is created randomly; $f_i = \mathcal{N}(\mu_{Y_i}, \Sigma_{Y_i})$ are Gaussian maximum likelihood estimators on Y_i. In the iteration stage, we go over all data points and reassign each of them to the cluster that decreases the CEC-IB cost (10) the most. After each reassignment, the clusters densities f_i are re-parameterized by the maximum likelihood estimators of new clusters and the cardinalities of categories $Z_j \cap Y_i$ are recalculated. If no cluster membership changed then the method terminates with a partition $\mathcal{Y}$.

Note that this procedure automatically removes unnecessary clusters by introducing the term $H(\mathcal{Y})$, which is the cost of cluster identification. If the method is initialized with more clusters than necessary, some clusters will lose data points to other clusters in order to reduce $H(\mathcal{Y})$, and the corresponding clusters may finally disappear altogether (e.g., by the number of data points contained in this cluster falling below a predefined threshold).

To describe the algorithm in detail, let us denote the cost of a single cluster $Y \subset X$ by

$$\mathrm{E}_\beta(Y) := \frac{|Y|}{|X|}\left(-\ln\frac{|Y|}{|X|} + H(\mathcal{N}(\mu_Y, \Sigma_Y)) + \beta H(\mathcal{Z}|Y \cap X_\ell)\right), \qquad (11)$$

assuming that X, $\mathcal{Z}$, and β are fixed. Then, for a given partition $\mathcal{Y}$ of X the minimal value of CEC-IB cost function equals:

$$\mathrm{E}_\beta(X, \mathcal{Z}; \mathcal{Y}) = \sum_{i=1}^{k} \mathrm{E}_\beta(Y_i).$$

Making use of the above notation, the algorithm can be written as follows:

INPUT:
$X \subset \mathbb{R}^N$ – data set
$\mathcal{Z} = \{Z_1, \ldots, Z_m\}$ – partition-level side information
k – initial number of clusters
β – weight parameter
$\varepsilon > 0$ – cluster reduction parameter
OUTPUT:
Partition $\mathcal{Y}$ of X
INITIALIZATION:
$\mathcal{Y} = \{Y_1, \ldots, Y_k\}$ – random partition of X into k groups
ITERATION:
repeat
  for all $x \in X$ **do**
    $Y_x \leftarrow$ get cluster of x
    $Y \leftarrow \arg\max_{Y \in \mathcal{Y}} \{\mathrm{E}_\beta(Y_x) + \mathrm{E}_\beta(Y) - \mathrm{E}_\beta(Y_x \setminus \{x\}) - \mathrm{E}_\beta(Y \cup \{x\})\}$
    if $Y \neq Y_x$ **then**
      move x from Y_x to Y
      update density models of Y_x and Y
      if $|Y_x| < \varepsilon \cdot |X|$ **then**
        delete cluster Y_x and assign its elements to these clusters which minimize the CEC-IB cost function
      end if
    end if
  end for
until no switch for all subsequent elements of X

The outlined algorithm is not deterministic and its results depend on the randomly chosen initial partition. Therefore, the algorithm can be restarted multiple times to avoid getting stuck in bad local minima.

One may think that the recalculation of the models and the evaluation of the cost in lines 15 and 18 is computationally complex. Looking at (11), one can see that evaluating the cost for a given cluster requires recomputing the sample mean vector and sample covariance matrix, which, according to [8, Theorem 4.3.], has a complexity quadratic in the dimension N of the dataset. Computing the determinant of the sample covariance matrix can be done with cubic complexity. Moreover, computing the conditional entropy of $\mathcal{Z}$ given the current cluster Y is linear in the number m of categories; if the selected data point x is not labeled, then there is no cost at all for computing these terms, since they cancel in the difference in line 15. Since in each iteration, all data points have to be visited and, for each data point, all clusters have to be tested, one arrives at a computational complexity in the order of $\mathcal{O}(nk(N^3 + m))$. In comparison, Lloyd's algorithm for k-means has a complexity of $\mathcal{O}(nkN)$ in each iteration and the expectation maximization (EM) algorithm to fit a GMM has a complexity of $\mathcal{O}(nkN^2)$ [42, p. 232]. Note, however, that neither classical k-means nor EM is designed to deal with side information, hence, the complexity of semi-supervised algorithms is in general larger. In particular, the addition of cannot-link constraints to GMM can involve a high computational cost. Moreover, in Appendix C we provide experimental evidence that the proposed Hartigan algorithm converges faster than Lloyd's or EM, because the model is re-parametrized after each switch.

4. Selection of the weight parameter

In this section we discuss the selection of the weight parameter β, trading between model complexity, model accuracy, and consistency with side information. Trivially, for $\beta = 0$ we obtain pure model-based clustering, i.e., the CEC method while, for $\beta \to \infty$, model fitting becomes irrelevant and the obtained clustering is fully consistent with reference labeling.

Our first theoretical result states that for $\beta = 1$ the algorithm tends to create clusters that are fully consistent with the side-information. Before proceeding, we introduce the following definitions:

Definition 4.1. Let X be a data set and let $\mathcal{Y} = \{Y_1, \ldots, Y_k\}$ be a partition of X. Let further $\mathcal{Z} = \{Z_1, \ldots, Z_m\}$ be a partition of $X_\ell \subseteq X$. We say that $\mathcal{Y}$ is a *coarsening* of $\mathcal{Z}$, if for every Z_j there exists a cluster Y_i such that $Z_j \subseteq Y_i$.

We say that the partition $\mathcal{Y}$ is *proportional* to $\mathcal{Z}$, if the fraction of data points in each cluster equals the fraction of *labeled* data points in this cluster, i.e., if $\frac{|Y_i|}{|X|} = \sum_{j=1}^{m} \frac{|Z_j \cap Y_i|}{|X_\ell|} = \frac{|Y_i \cap X_\ell|}{|X_\ell|}$.

Proportionality is required in the proofs below since it admits applying the chain rule of entropy to $H(\mathcal{Z}|\mathcal{Y})$ even in the case where $X_\ell \subset X$. In other words, if $\mathcal{Y}$ is proportional to $\mathcal{Z}$, then (see Appendix D for the proof):

$$H(\mathcal{Z}|\mathcal{Y}) + H(\mathcal{Y}) = H(\mathcal{Z}, \mathcal{Y}).$$

Every coarsening of a proportional partition $\mathcal{Y}$ is proportional. Trivially, if $X_\ell = X$, then every partition $\mathcal{Y}$ is proportional to $\mathcal{Z}$. Note, however, that for finite data sets and if $X_\ell \subset X$, it may happen that there exists no partition $\mathcal{Y}$ proportional to the side information $\mathcal{Z}$ (e.g., if all but one data points are labeled). Nevertheless, the following theorems remain valid as guidelines for parameter selection.

Finally, note that consistency as in Definition 3.2 and coarsening as in Definition 4.1 are, loosely speaking, opposites of each other. In fact, if $X_\ell = X$ and if $\mathcal{Z}$ is a partition of X, then $\mathcal{Y}$ is a coarsening of $\mathcal{Z}$ if and only of $\mathcal{Z}$ is consistent with $\mathcal{Y}$. Although we are interested in partitions $\mathcal{Y}$ consistent with $\mathcal{Z}$, we use the concept of a coarsening to derive results for parameter selection. Moreover, note that a partition $\mathcal{Y}$ can be both consistent with and a coarsening of the side information $\mathcal{Z}$. This is the case where every Y_i contains exactly one Z_j and every Z_j is contained in exactly one Y_i (i.e., $\mathcal{Y}$ has the same number of elements as $\mathcal{Z}$).

Theorem 4.1. *Let $X \subset \mathbb{R}^N$ be a finite data set and $X_\ell \subseteq X$ be the set of labeled data points that is partitioned into $\mathcal{Z} = \{Z_1, \ldots, Z_m\}$. Let $\mathcal{Y} = \{Y_1, \ldots, Y_k\}$ be a proportional coarsening of $\mathcal{Z}$, and suppose that the sample covariance matrices Σ_i of Y_i are positive definite.*

If $\tilde{\mathcal{Y}} = \{\tilde{Y}_1, \ldots, \tilde{Y}_{k'}\}$ is a coarsening of $\mathcal{Y}$, then

$$\mathrm{E}_1(\tilde{\mathcal{Y}}) \geq \mathrm{E}_1(\mathcal{Y}). \tag{12}$$

Proof. See Appendix E. □

An immediate consequence of Theorem 4.1 is that, for $\beta = 1$, CEC-IB tends to put elements with different labels in different clusters. Note, however, that there might be partitions $\mathcal{Y}$ that are consistent with $\mathcal{Z}$ that have an even lower cost (10): Since every consistent partition $\mathcal{Y}$ satisfies

$H(\mathcal{Z}|\mathcal{Y}) = 0$, any further refinement of $\mathcal{Y}$ reduces the cost whenever the cost for model complexity, $H(\mathcal{Y})$, is outweighed by the modeling inaccuracy $\sum_{i=1}^{k} \frac{|Y_i|}{|X|} H(\mathcal{N}(\mu_{Y_i}, \Sigma_{Y_i}))$.

Theorem 4.1 does not assume that the side information induces a partition of X that fits our intuition of clusters: the Z_j need not be a connected set, but could result from, say, a random labeling of the data set X. Then, for $\beta = 1$, splitting X into clusters that are consistent with the side information will be at least as good as creating a single cluster. Interestingly, if the labeling is completely random, any $\beta < 1$ will prevent dividing the data set into clusters:

Remark 4.1. Suppose a completely random labeling for the setting of Theorem 4.1. More precisely, we assume that the set of labeled data X_ℓ is divided into $\mathcal{Z} = \{Z_1, \ldots, Z_m\}$ and that the partition $\mathcal{Y}$ is a proportional coarsening of $\mathcal{Z}$. If sufficiently many data points are labeled, we may assume that the sample covariance matrix Σ_{Y_i} of Y_i is close to the covariance matrix Σ_X of X, i.e. $\Sigma_{Y_i} \approx \Sigma_X$. For any coarsening $\tilde{\mathcal{Y}}$ of $\mathcal{Y}$, the cross-entropies for $\mathcal{Y}$ and $\tilde{\mathcal{Y}}$ are approximately equal:

$$\sum_i \frac{|Y_i|}{|X|} H(\mathcal{N}(\mu_{Y_i}, \Sigma_{Y_i})) \approx \sum_j \frac{|\tilde{Y}_j|}{|X|} H(\mathcal{N}(\mu_{\tilde{Y}_j}, \Sigma_{\tilde{Y}_j})) \approx H(\mathcal{N}(\mu_X, \Sigma_X))$$

because the sample covariance matrices of $\tilde{Y} \in \tilde{\mathcal{Y}}$ are also close to Σ_X.

If we compare the remaining parts of cost function (10), then with $\beta < 1$ we obtain:

$$\begin{aligned} H(\mathcal{Y}) + \beta H(\mathcal{Z}|\mathcal{Y}) &= (1-\beta)H(\mathcal{Y}) + \beta(H(\mathcal{Y}) + H(\mathcal{Z}|\mathcal{Y})) \\ &= (1-\beta)H(\mathcal{Y}) + \beta(H(\tilde{\mathcal{Y}}) + H(\mathcal{Z}|\tilde{\mathcal{Y}})) > H(\tilde{\mathcal{Y}}) + \beta H(\mathcal{Z}|\tilde{\mathcal{Y}}). \quad (13) \end{aligned}$$

The last inequality follows from the fact that $H(\mathcal{Y}) > H(\tilde{\mathcal{Y}})$. Therefore, CEC-IB with $\beta < 1$ is robust on random labeling.

Our second result is a critical threshold β_0, above which splitting a given cluster $\tilde{Y}_1$ into smaller clusters $Y_1, \ldots, Y_l$ reduces the cost. This threshold β_0 depends on the data set and on the side information. For example, as Remark 4.1 shows, for a completely random labeling we get $\beta_0 = 1$. To derive the threshold in the general case, we combine the proof of Theorem 4.1 with (13):

Theorem 4.2. *Let $X \subset \mathbb{R}^N$ be a finite data set and $X_\ell \subseteq X$ be the set of labeled data points that is partitioned into $\mathcal{Z} = \{Z_1, \ldots, Z_m\}$. Let $\mathcal{Y} = \{Y_1, \ldots, Y_k\}$ be a proportional coarsening of $\mathcal{Z}$, and suppose that the sample covariance matrices Σ_i of Y_i are positive definite. Suppose that $\tilde{\mathcal{Y}} = \{Y_1, \ldots, Y_{k'-1}, (Y_{k'} \cup \cdots \cup Y_k)\}$, for $1 < k' < k$, is a coarsening of $\mathcal{Y}$, and let μ and Σ be the sample mean vector and sample covariance matrix of $Y_{k'} \cup \cdots \cup Y_k$. Let $q_i = p_i$ for $i = 1, \ldots, k'-1$ and $q_k = \sum_{i=k'}^{k} p_i$. We put*

$$\beta_0 = 1 + \frac{\sum_{i=k'}^{k} \frac{p_i}{2q_{k'}} \ln\left(\frac{\det \Sigma_i}{\det \Sigma}\right)}{H\left(\frac{p_{k'}}{q_{k'}}, \ldots, \frac{p_k}{q_{k'}}\right)}. \tag{14}$$

If $\beta \geq \beta_0$, then

$$\mathrm{E}_\beta(\tilde{\mathcal{Y}}) \geq \mathrm{E}_\beta(\mathcal{Y}).$$

Proof. See Appendix F. □

We now evaluate a practically relevant instance of the above theorem, where the data follows a Gaussian distribution and the partition-level side information is "reasonable":

Example 4.1. Let $X \subset \mathbb{R}$ be a data set generated by a one-dimensional Gaussian distribution $f = \mathcal{N}(\mu, \sigma^2)$, and suppose that the data set is large enough such that the sample mean μ_X and sample variance σ_X^2 are close to μ and σ^2, respectively. A classical unsupervised model-based clustering technique, such as CEC or GMM, terminates with a single cluster.

Now suppose that $Z_1 \subset (-\infty, \mu)$ and $Z_2 \subset [\mu, +\infty)$ are equally-sized sets, which suggests that $\mathcal{Y} = \{Y_1, Y_2\} = \{(-\infty, \mu) \cap X, [\mu, +\infty) \cap X\}$ is the expected clustering. Consequently, on one hand, the data distribution indicates that a single cluster should be created while, on the other hand, the side information suggests splitting the data set into two clusters. At the threshold β_0 these two conflicting goals are balanced, while for $\beta > \beta_0$ a consistent clustering is obtained.

To calculate the critical value β_0, let $\mathcal{Y} = \{Y_1, Y_2\}$ be proportional to $\mathcal{Z}$, and let $f_i = \mathcal{N}(\mu_{Y_i}, \sigma_{Y_i}^2)$ be the optimal fit for cluster Y_i. Since the data in Y_i can be well approximated by a truncated Gaussian distribution, we can calculate:

$$\sigma_{Y_1}^2 \approx \sigma_{Y_2}^2 \approx \sigma^2 \left(1 - \frac{2}{\pi}\right).$$

Making use of the previous theorem, $\mathrm{E}_\beta(\{X\}) = \mathrm{E}_\beta(\mathcal{Y})$ for

$$\beta = \beta_0 \approx 1 + \frac{\ln \sqrt{1 - \frac{2}{\pi}}}{H(\frac{1}{2}, \frac{1}{2})} \approx 0.269. \tag{15}$$

Continuing this example, in some cases the side information might be noisy, i.e., data points are labeled wrongly. Consider the labeling $\mathcal{Z}$ that satisfies $Z_1 \subset (-\infty, \mu + c)$ and $Z_2 \subset [\mu - c, +\infty)$, for some $c > 0$. In other words, the human experts did not agree on the labeling at the boundary between the clusters. If we choose $\mathcal{Y}$ proportional to this noisy side information $\mathcal{Z}$, then one has reason to suppose that the sample variances of Y_1 and Y_2 are larger than in the noiseless case, hence leading to a larger threshold β_0 according to Theorem 4.1. Setting β to a value only slightly higher than the threshold β_0 for the noiseless case thus ensures a partition $\mathcal{Y}$ that is consistent with the noiseless labeling, but that is robust to noise. In summary, one should choose β large (i.e., close to 1), if one believes that the side information is correct, but small if one has to expect noisy side information.

5. Experiments

We evaluated our method in classical semi-supervised clustering tasks on examples retrieved from the UCI repository [43] and compared its results to state-of-the-art semi-supervised clustering methods. We evaluated performance in the case of only few classes being present in the partition-level side information and for noisy labeling, and investigated the influence of the parameter β on the clustering results. We furthermore applied CEC-IB on a data set of chemical compounds [44] to discover subgroups based on partition-level side information derived from the top of a cluster hierarchy and illustrate its performance in an image segmentation task.

5.1. Experimental setting

We considered five related semi-supervised clustering methods for comparison. The first is a classical semi-supervised classification method that is based on fitting a GMM to the data set taking partial labeling into account. Since it is a classification method, it only works if all true classes are present in the categorization $\mathcal{Z}$. We used the R implementation Rmixmod [31] with default settings; we refer to this method as "mixmod".

The second method incorporates pairwise constraints as side information for a GMM-based clustering technique [33]. To transfer the partition-level side information to pairwise constraints, we went over all pairs of labeled data points in X_ℓ and generated a must-link constraint if they were in the same, and a cannot-link constraint if they were in different categories. We used the implementation from one of the authors' website[2] and refer to this method as "c-GMM". We ran c-GMM in MultiCov mode, i.e. every cluster was characterized by its own covariance matrix.

We also applied an extension of k-means to accept partition level-side information [5]. The method requires setting a weight parameter λ, that places weight on the features derived from the side information. The authors suggested $\lambda = 100$, but we found that the method performs more stable for $\lambda = 100 \cdot \mathrm{tr}(\Sigma)$, i.e., for λ being proportional to the trace of the sample covariance matrix of the data set X. We refer to this method as "k-means".

Moreover, we considered a semi-supervised variant of fuzzy c-means [23, 24], which incorporates partition-level side information. We used the Euclidean distance, set the fuzzifier parameter to 2, and chose a trade-off parameter $\alpha = \frac{|X|}{|X_\ell|}$ as suggested by the authors. To obtain a "hard" clustering $\mathcal{Y}$ from a fuzzy partition we assigned every point to its most probable cluster. This technique will be referred to as "fc-means".

Finally, we used a semi-supervised version of spectral clustering [20] (referred to as "spec"), which was claimed to achieve state-of-the-art performance among spectral algorithms. The method accepts pairwise constraints and operates on the affinity (similarity) matrix of the data set. The authors of [20] suggested setting the similarity between data points x_i and x_j to $e^{-\|x_i - x_j\|^2/2\rho^2}$, where $\|\cdot\|$ is the Euclidean distance and where $\rho > 0$ is called affinity parameter. In order to account for different variances in different dimensions, we used

$$e^{-\sum_{\ell=1}^{N} \frac{|x_i^{(\ell)} - x_j^{(\ell)}|^2}{2\rho^2 \sigma_{(l)}^2}}, \tag{16}$$

where $x_i^{(\ell)}$ is the value of the ℓ-th coordinate of x_i and where $\sigma_{(\ell)}^2$ is the variance of the ℓ-th coordinate of X. The method can be tuned with two parameters: affinity parameter ρ and trade-off factor η. The authors suggest to find the best possible combination of these parameters using a grid-search strategy. Since we did not allow for tuning any parameters of other methods

[2] `http://www.scharp.org/thertz/code.html`

Table 1: Summary of UCI datasets used in the experiments.

Data set	**# Instances**	**# Features**	**# Classes**
Ecoli$^+$	327	5	5
Glass	214	9	6
Iris	150	4	3
Segmentation$^+$	210	5	7
User Modeling	403	5	4
Vertebral	310	6	3
Wine	178	13	3

$^+$: PCA was used to reduce a dimensionality of the data set and remove dependent attributes

(including β in CEC-IB), for a fair comparison we decided to fix these two parameters. Specifically, we put $\eta = 0.7$ analyzing the results reported in [20]. We moreover set $\rho = 1$ based on the fact that the Euclidean distances are already normalized according to the variances of the respective dimensions and since [20] reports little influence of the selection of ρ. We generated must-link and cannot-link constraints as we did for c-GMM; moreover, the entries of the affinity matrix were set to one for must-link, and to zero for cannot-link constraints.

Since CEC-IB automatically determines an appropriate number of clusters by removing clusters with too few data points, we initialized CEC-IB with twice the correct number of clusters. In contrast, other methods were run with the correct numbers of clusters. In a semi-supervised clustering task with correct labels from all classes, the competing methods can thus be expected to perform better than CEC-IB.

To better illustrate the effect of the weight parameter β, we used two parameterization of CEC-IB, using $\beta = 1$ and $\beta = \beta_0 \approx 0.269$ given by (15). We refer to these two variants as CEC-IB$_1$ and CEC-IB$_0$, respectively.

The similarity between the obtained clusterings and the ground truth was evaluated using Normalized Mutual Information (NMI) [45]. For a reference grouping $\mathcal{X}$ and a clustering $\mathcal{Y}$ it is defined by

$$\mathrm{NMI}(\mathcal{Y}, \mathcal{X}) = \frac{2I(\mathcal{Y}; \mathcal{X})}{H(\mathcal{Y}) + H(\mathcal{X})}.$$

Since $I(\mathcal{Y}; \mathcal{X}) \leq \min\{H(\mathcal{Y}), H(\mathcal{X})\}$, NMI is bounded from above by 1, which is attained for identical partitions. If $\mathcal{Y}$ and $\mathcal{X}$ contain different numbers of clusters, then NMI is always below 1.

Table 2: Number of clusters returned by CEC-IB for a given amount of labeled data.

Data set	**# Classes**	**0%**	**10%**	**20%**	**30%**
Ecoli	5	7	6	6	6
Glass	6	5	6	6	6
Iris	3	5	5	5	5
Segmentation	7	8	8	7	8
User	4	7	6	6	6
Vertebral	3	4	4	4	4
Wine	3	3	3	3	3

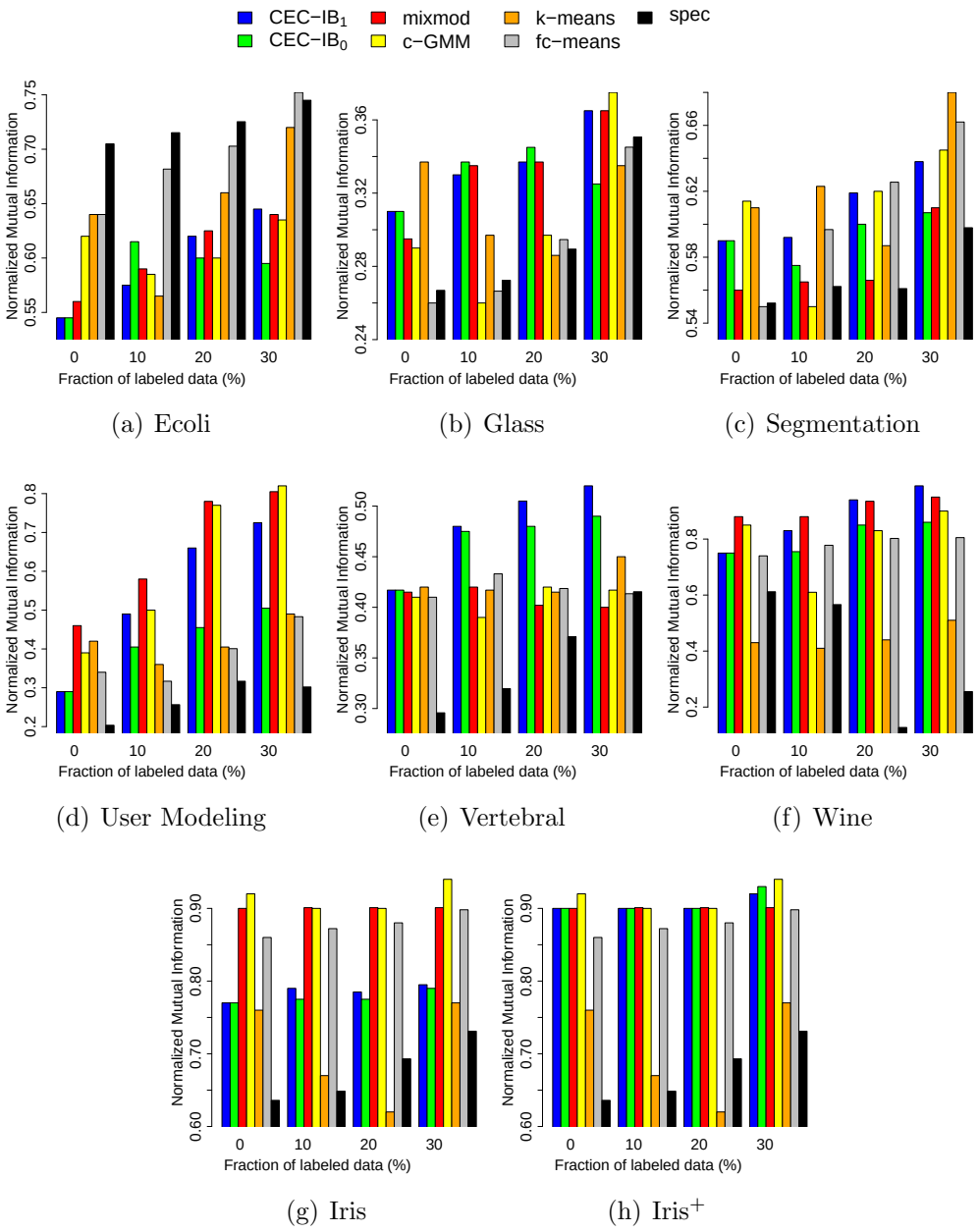
5.2. Semi-supervised clustering

We evaluated the proposed method in a classical semi-supervised clustering task, in which we aim to recover a reference partition based on a small sample of labeled data.

We used seven UCI data sets, which are summarized in Table 1. The partition-level side information was generated by choosing 0%, 10%, 20%, and 30% of the data points and labeling them according to their class. To remove effects from random initializations, we generated 10 different samples of side information for each percentage and averaged the resulting NMI values.

The clustering results presented in Figure 3 show that CEC-IB$_1$ usually achieved a higher NMI than CEC-IB$_0$: Since the partition-level side information is noise-free, i.e., agrees with the reference grouping, a larger weight parameter β leads to better performance. In general, CEC-IB$_1$ produced results similar to the two other GMM-based techniques, c-GMM and mixmod. Notable differences can be observed on Vertebral dataset, where CEC-IB performed significantly better, and on Iris and User Modeling, where the competing methods gave higher NMI. This is most likely caused by the fact that CEC-IB failed to determine the correct number of clusters (see Table 2), while the GMM implementations were given this correct number of clusters as side information. As it can be seen in Table 2, in all other cases, CEC-IB terminated with a number of clusters very close to the true value. Initializing CEC-IB with the correct number of clusters for the Iris data set, we get results that are comparable to those of mixmod and c-GMM (see Figure 3(h)).

Observe that k-means gave slightly lower NMI than fc-means. Nevertheless, both algorithms performed worse than the GMM-based methods, except for the Ecoli and Segmentation data sets. The difference in the results can be



(a) Ecoli (b) Glass (c) Segmentation

(d) User Modeling (e) Vertebral (f) Wine

(g) Iris (h) Iris^{+}

Figure 3: Normalized Mutual Information of examined methods evaluated on UCI datasets. CEC-IB was initialized with twice the true number of clusters while other methods used the correct number of clusters.
$^{+}$ CEC-IB was initialized with the correct number of clusters

explained by the fact that fc-means and k-means are distance-based methods and therefore perform differently from model-based approaches. Although the performance of spec usually increases with more labeled examples, its results are worse than the other methods.

5.3. Few labeled classes

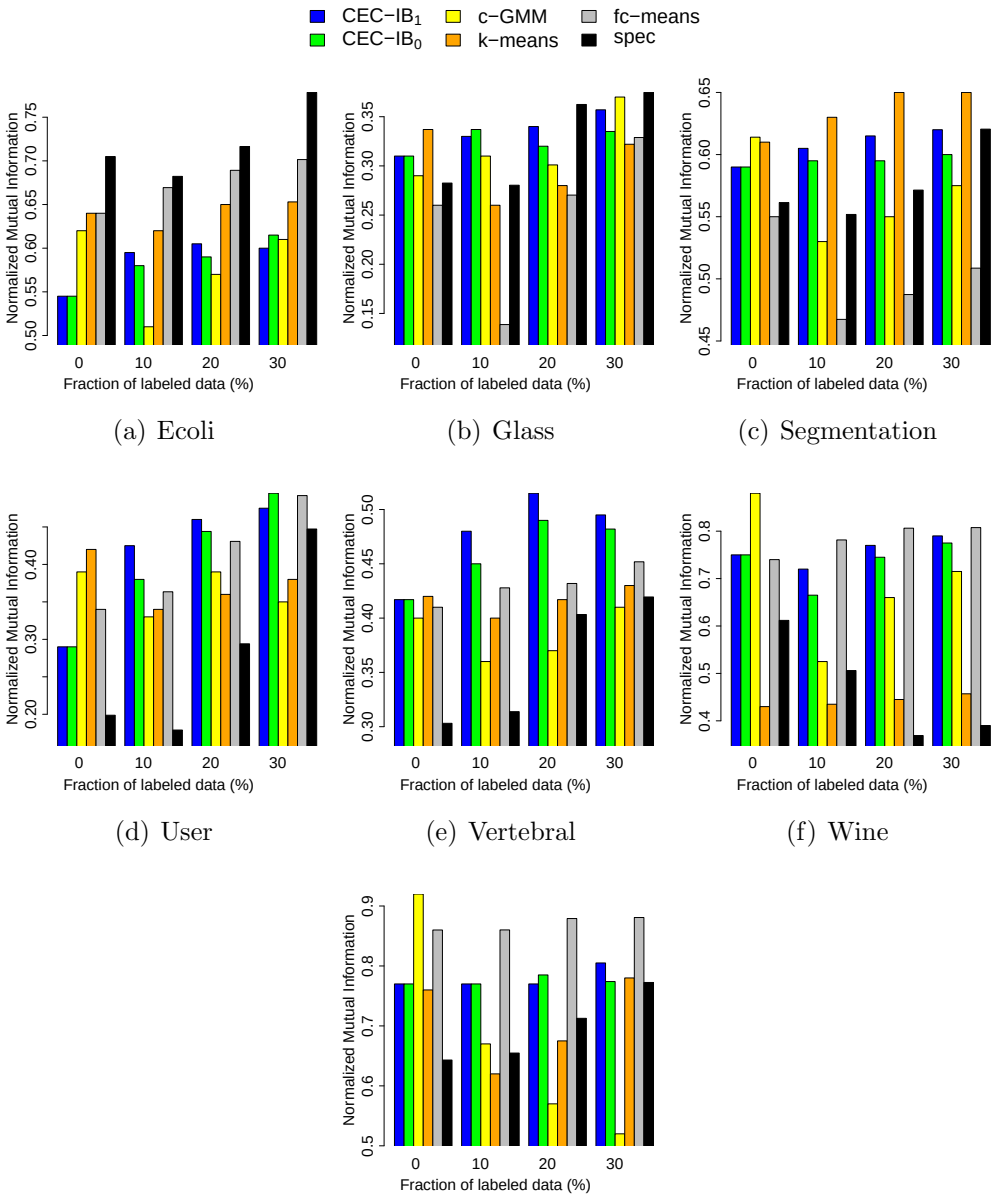
In a (semi-)supervised classification task, the model learns classes from a set of labeled data and applies this knowledge to unlabeled data points. More specifically, the classifier cannot assign class labels that were not present in a training set. In contrast, clustering with partition-level side information, can detect clusters within a labeled category or within the set of unlabeled data points.

In this section, we apply CEC-IB to a data set for which the partition-level side information contains labels of only two classes from the reference grouping. As before we considered 0%, 10%, 20% and 30% of labeled data. For each of the 10 runs we randomly selected two classes from a reference grouping that covered at least 30% of data in total and generated the partition-level side information from these two categories (the same classes were used for all percentages of side information). It was not possible to run mixmod in this case because this package does not allow to use a number of clusters different from the categories given in the side information.

Figure 4 shows that CEC-IB was able to consistently improve its clustering performance with an increasing size of the labeled data set[3]. Surprisingly, c-GMM sometimes dropped in performance when adding side information. This effect was already visible in Figure 3, but seems to be more pronounced here. While a deeper analysis of this effect is out of scope of this work, we believe that it is due to the simplification made in [33] to facilitate applying a generalized EM scheme. This simplification is valid if pairs of points with cannot-link constraints are disjoint, an assumption that is clearly violated by the way we generate cannot-link constraints (see Section 5.1).

Contrary to c-GMM, the results of fc-means and k-means were far more stable. In most cases both algorithms increased their performance having access to more labeled data. Interestingly, spec performed in general better when only two classes were labeled than in the previous experiment where

[3]Although the results for 0% of labeled data should be identical with the ones reported in Section 5.2, some minor differences might follow from a random initialization of the methods, see Section 3.4.



(a) Ecoli (b) Glass (c) Segmentation

(d) User (e) Vertebral (f) Wine

(g) Iris

Figure 4: Normalized Mutual Information of examined methods evaluated on UCI datasets when the partition-level side information covered only two classes.

all classes were labeled. In consequence, its results were often comparable to or sometimes even better than other methods.

5.4. Noisy side information

In real-world applications, the side information usually comes from human experts, who label training samples. Depending on the expertise of these workers, some part of this side information might be noisy or erroneous. Therefore, the clustering algorithm needs to be robust w.r.t. noisy side information.

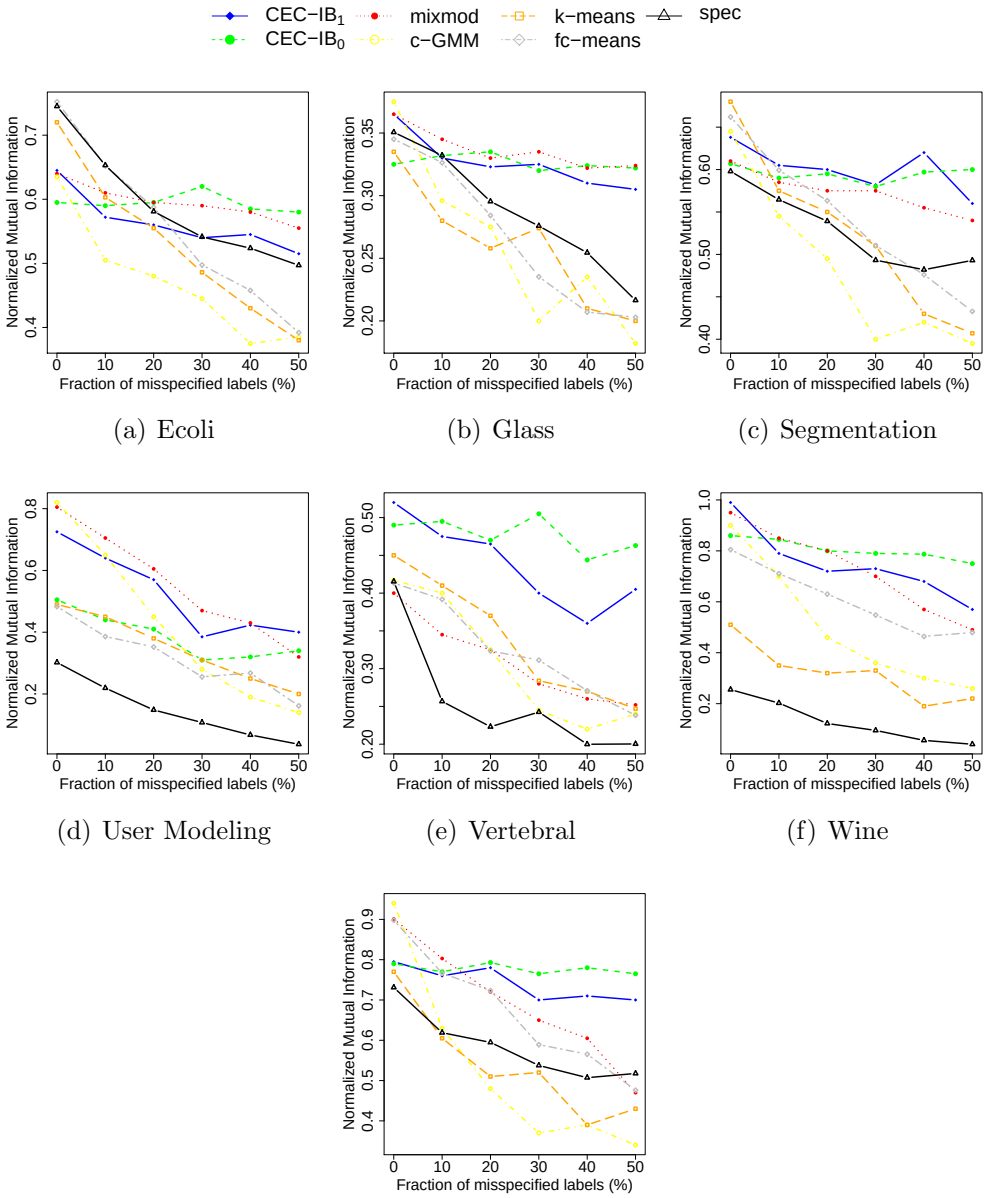
To simulate the above scenario, we randomly selected 30% of the data points as side information, as in Section 5.2, and assign incorrect labels for a fixed percentage of them (0%, 10%, 20%, 30%, 40%, 50% of misspecified labels). All methods were run in the same manner as in the previous experiments.

One can see in Figure 5 that CEC-IB$_0$ showed the highest robustness to noisy labels among all competing methods, i.e., the NMI deteriorated the least with increasing noise. Although CEC-IB$_1$ achieved higher NMI than CEC-IB$_0$ for correctly labeled data (without noise), its performance is usually worse than CEC-IB$_0$ when at least 30% of labels are misspecified. The robustness of mixmod and spec is acceptable; their results vary with the used data set, but on average they cope with incorrect labels comparably to CEC-IB$_1$. In contrast, c-GMM, k-means and fc-means are very sensitive to noisy side information. Since their performance falls drastically below the results returned for strictly unsupervised case, they should not be used if there is a risk of unreliable side information.

5.5. Influence of weight parameter

From Figure 3 it can be seen that $\beta = \beta_0$ often seems to be too small to benefit sufficiently from partition-level side information, although it provides high robustness to noisy labels. In this experiment, we investigate the dependence between the value of β and the size of the labeled data set and the fraction of noisy labels, respectively.

First, we checked the performance of CEC-IB with different values of β in the noiseless case. We randomly selected 10%, 20% and 30% of the data points, respectively, and labeled them according to their true classes. Figure 6 shows that CEC-IB with $\beta = \beta_0$ run on 30% labels performed similarly to CEC-IB with $\beta = 1$ run on 10% labels. Therefore, we see that the lack of labeled data can be compensated with a larger value of β. Moreover, a larger

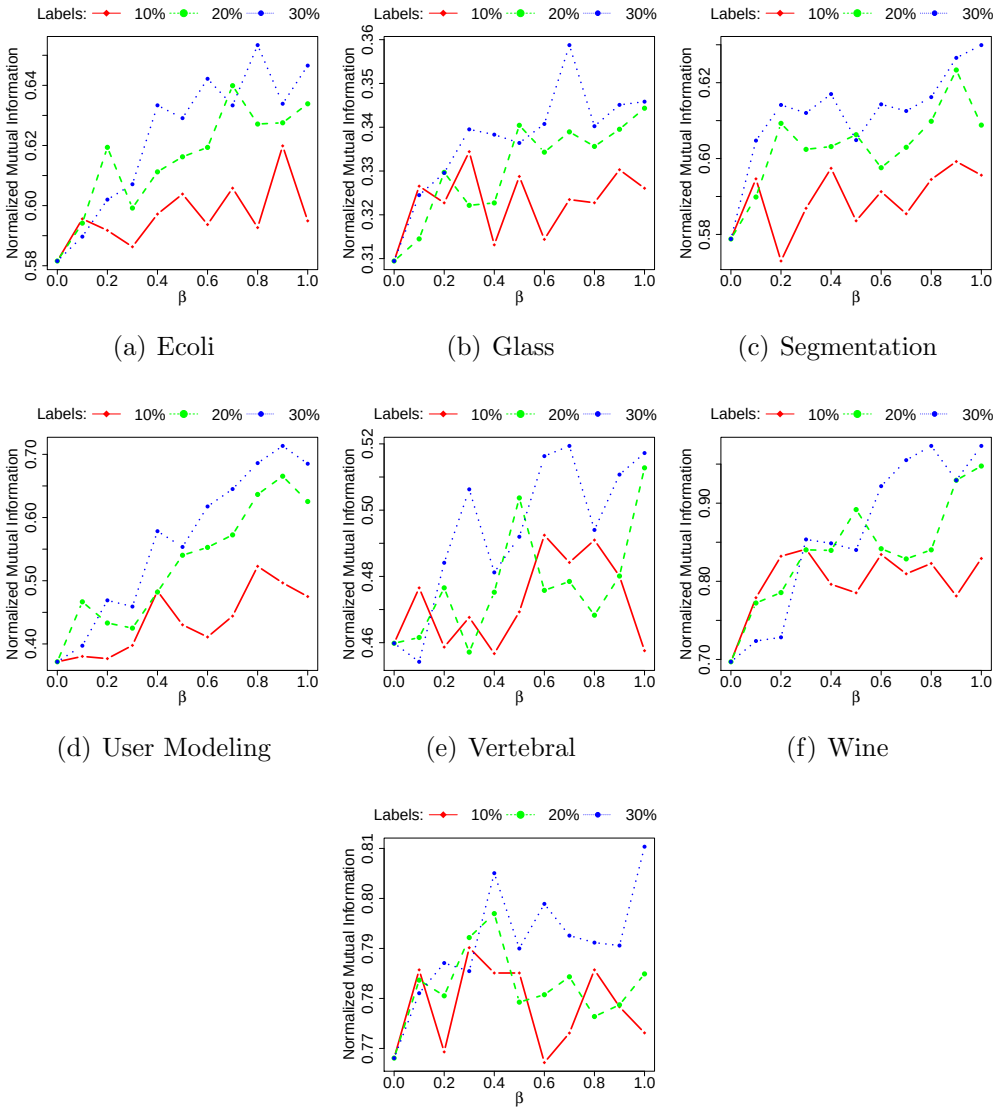


(a) Ecoli (b) Glass (c) Segmentation

(d) User Modeling (e) Vertebral (f) Wine

(g) Iris

Figure 5: Influence of misspecified labels on the clustering results. CEC-IB was run with twice the true number of clusters.

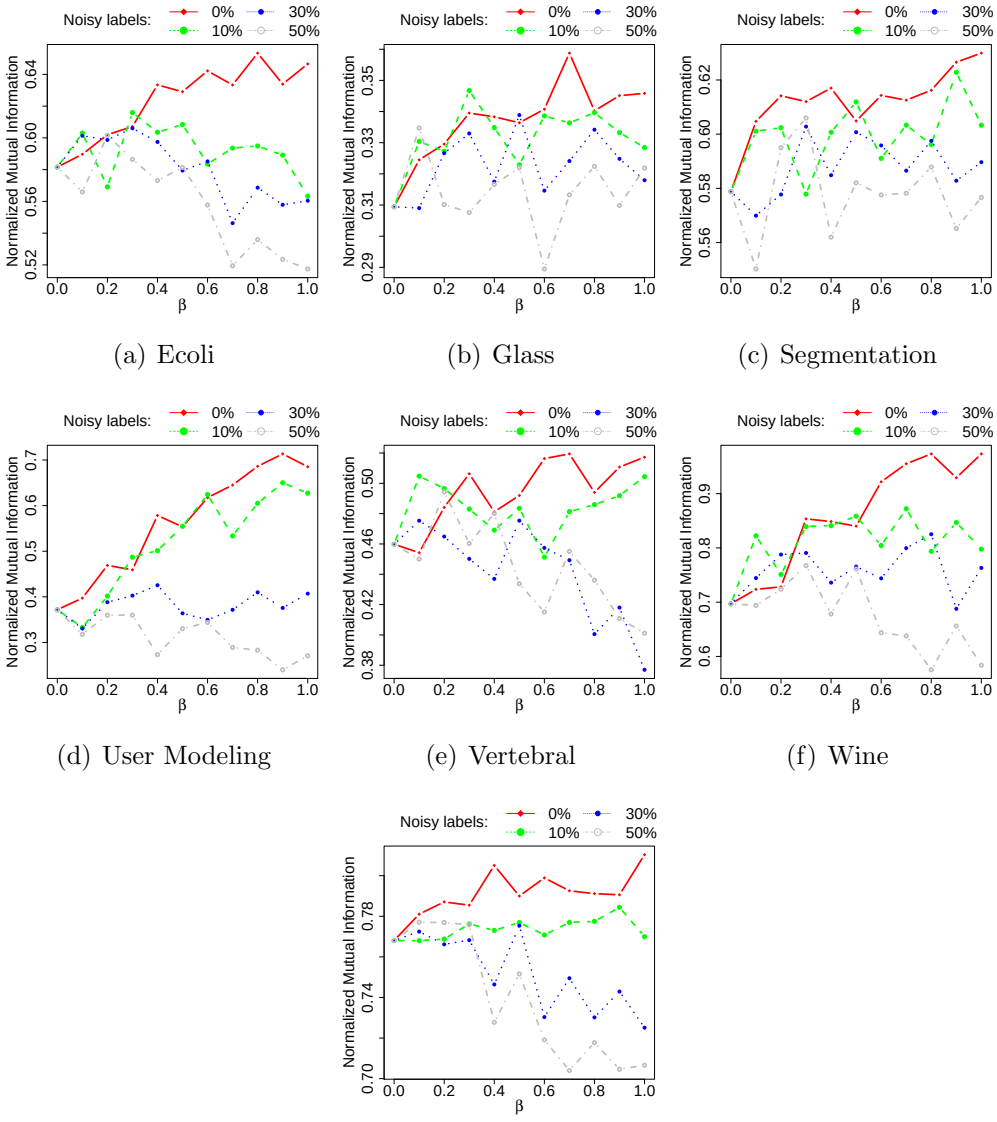


(a) Ecoli (b) Glass (c) Segmentation

(d) User Modeling (e) Vertebral (f) Wine

(g) Iris

Figure 6: Dependence between the number of labeled data points and the value of parameter β.



(a) Ecoli (b) Glass (c) Segmentation

(d) User Modeling (e) Vertebral (f) Wine

(g) Iris

Figure 7: Dependence between the fraction of incorrect labels and the value of parameter β.

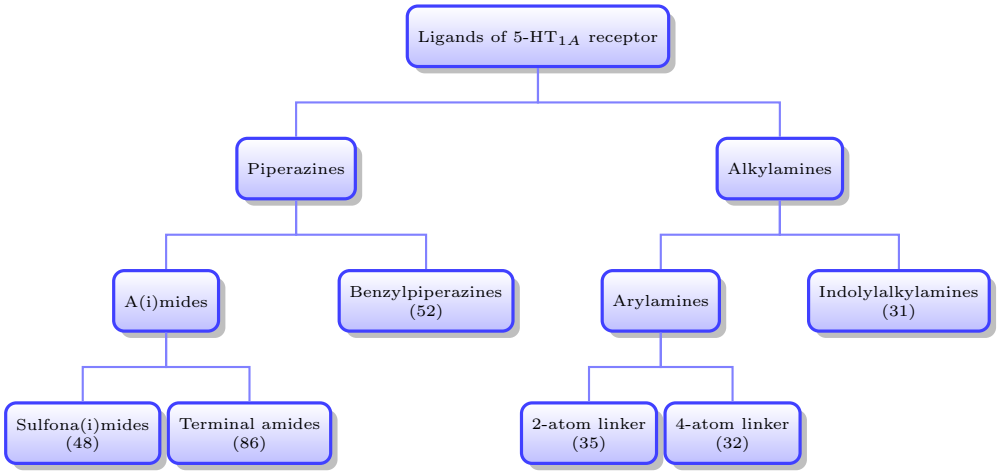


Figure 8: Hierarchy tree of chemical compounds classes. The numbers in brackets indicate the number of times the corresponding class appeared in the data set.

value of β makes CEC-IB benefit more from a larger number of correctly labeled data points.

In the second experiment we investigated the relation between the fraction of noisy side information and the weight parameter. We drew 30% of the data points and labeled 0%, 10%, 30%, and 50% of them incorrectly (the remaining selected data points were labeled with their correct class labels). We see in Figure 7 that a small noise of 10% did not have severe negative effects on the clustering results. In this case NMI was almost always higher than in the unsupervised case (i.e., for $\beta = 0$), even for $\beta = 1$. For 50% of incorrectly labeled data points, increasing β has a negative effect on the clustering performance, while $\beta = \beta_0$ provided high robustness to the large amount of noisy labels and in most cases performed at least as well as the unsupervised scenario. For the case where 30% of labels were misspecified, choosing $\beta < 0.6$ seems to produce results at least as good as when no side information is available.

5.6. Hierarchy of chemical classes – a case study

Our CEC-IB cost function only penalizes including elements from different categories into the same cluster. Covering a single category by more than one cluster is not penalized if the cost for model accuracy outweighs the cost for model complexity. In this experiment, we will show that this property is

useful in discovering subgroups from side information derived from a cluster hierarchy.

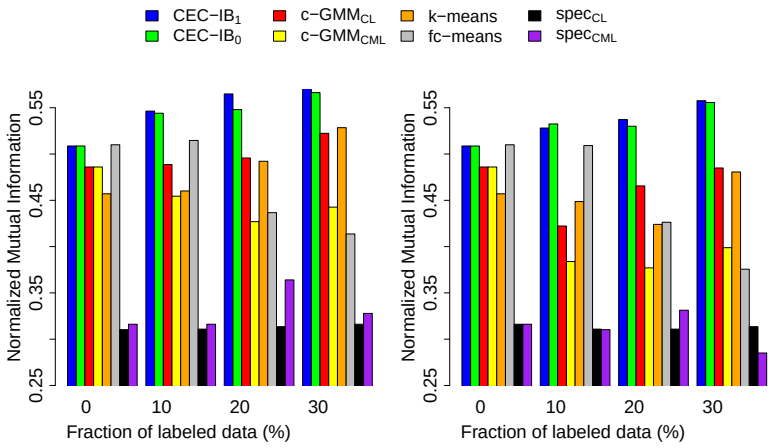
We considered a data set of chemical compounds that act on 5-HT_{1A} receptor ligands, one of the proteins responsible for the regulation of the central nervous system [46, 47]. Part of this data set was classified hierarchically by an expert [44], as shown in Figure 8. For an expert it is easier to provide a coarse categorization rather than a full hierarchical classificiation, especially if it is not clear how many subgroups exist. Therefore, in some cases, it might be easier to get a hierarchical structure based on this coarse categorization made by the expert and an automatic clustering algorithm that finds a partition corresponding to the clusters at the bottom of the hierarchy.

We used the Klekota-Roth fingerprint representation of chemical compounds [48], which describes each object by a binary vector, where "1" means presence and "0" denotes absence of a predefined chemical pattern. Since this representation contains 4860 features in total, its direct application to model-based clustering can lead to singular covariance matrices of clusters. Therefore, PCA was used to reduce its dimension to the 10 most informative components. This data set contains 284 examples in total (see Figure 8 for the cardinalities of particular chemical classes).

We generated partition-level side information from the division of chemical data set into two classes: Piperazines and Alkylamines. We considered 0%, 10%, 20% and 30% of data points to be labeled and supposed that the human expert assigns incorrect labels with probabilities: 0%, 10%, 20%, and 30% respectively. Based on the results from previous subsection we used $\beta = 0.6$ instead of $\beta = \beta_0$, which is denoted by CEC-$IB_{0.6}$. Our method was run with 10 initial groups, while the other algorithms used the knowledge of the correct number of clusters. As mentioned in Section 5.3, it is not possible to run mixmod in this case, since the desired number of clusters is larger than the number of categories.

It can be seen from Figure 9 that CEC-IB_1 gave the highest score among all methods when the expert was always assigning the correct labels and it was only slightly better than CEC-$IB_{0.6}$. In the case of 20% and 30% of misspecified labels it was slightly better to use $\beta = 0.6$, although the differences were very small. CEC-IB terminated usually with 6 or 7 groups.

One can observe that GMM with negative constraints was able to use this type of side information effectively. In the noiseless case, its results improved with the number of labeled data points, but not as much as with our method. In the noisy case, however, its performance dropped down. It is



(a) Noiseless (b) 10% of noise

(c) 20% of noise (d) 30% of noise

Figure 9: Detection of chemical subgroups.

(a) Image (b) Side Information

Figure 10: Test image and sample labeling dividing the picture into background and foreground.

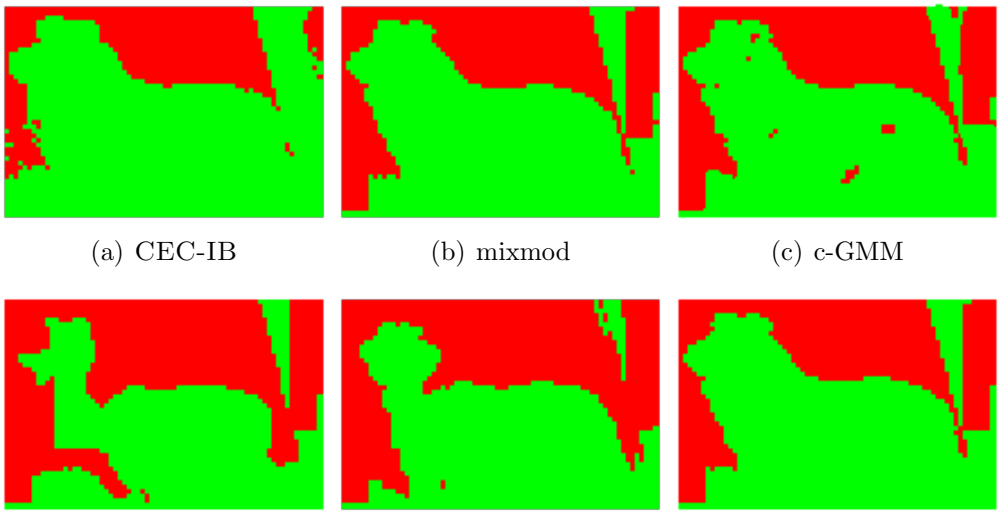
worth mentioning that the implementation of negative constraints with hidden Markov random fields is very costly, while our method is computationally efficient. k-means benefited from the side information in noiseless case, but deteriorated its performance when incorrect labels were introduced. GMM with positive and negative constraints, and fc-means were not able to use this type of knowledge to full effect. We observed that the use of negative constraints only has no effect on spec, i.e. its results were almost identical for any number of labeled data[4]. The results of spec with both types of constraints led to some improvements, but its overall performance was quite low. We were unable to provide a satisfactory explanation for this behavior.

5.7. Image segmentation

To further illustrate the performance of CEC-IB we applied it to an image segmentation task. We chose the picture of a dog retrieved from Berkeley Image Segmentation database [5] presented in Figure 10(a) (picture no. 247085 resized into 70×46 resolution) and tried to separate the shape of the dog from the background. As partition-level side information, we marked four regions by one of two labels (indicated by white and black colors, see Figure 10(b)). This information was passed to all considered clustering methods. In this example we focus on noiseless side information, thus we put $\beta = 1$ for CEC-IB.

[4]We observed similar effects for most UCI data sets when we used negative constraints only in the setting of Section 5.3. Changing the parametrization of the method did not overcome this negative behavior.

[5]`https://www2.eecs.berkeley.edu/Research/Projects/CS/vision/bsds/`



(a) CEC-IB (b) mixmod (c) c-GMM

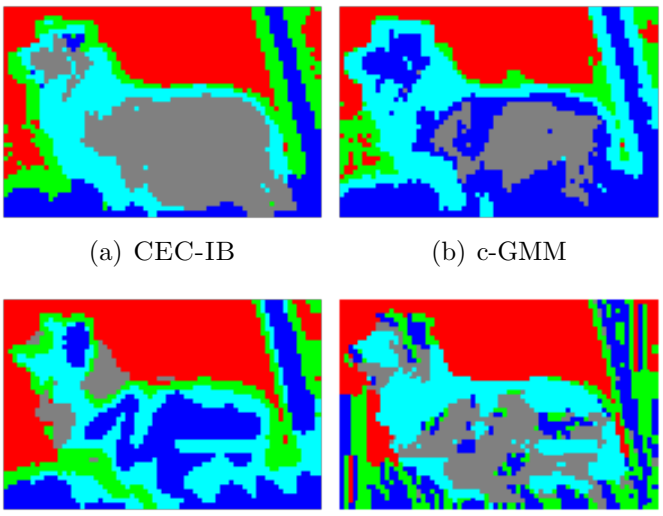
(d) k-means (e) fc-means (f) spec

Figure 11: Image segmentation into foreground and background.

To transform the image into vector data, we selected a window of size 7×7 around each pixel and used it as a feature vector of dimension 147 (3 color intensities for 49 pixels each). Then, we applied PCA to reduce the dimension of these vectors to 5 most informative components. In consequence, we obtained a data set with 3220 data points in $\mathbb{R}^5$.

Figure 11 shows the clustering results when the algorithms were run with two clusters. It can be seen that CEC-IB, mixmod, c-GMM and spec provided reasonable results. Generally though, in all cases the shape of the dog was often mixed with a part of background. This is not surprising, since 1) CEC-IB, mixmod and c-GMM are "unimodal", i.e. they try to detect compact groups described by single Gaussians, and 2) k-means and fc-means represent clusters by a single central point. Both background and foreground are too complex to be generalized by so simple patterns.

In order to take this into account, we first tried to detect what is a "natural" number of segments. For this purpose, we ran CEC-IB with 10 initial groups, which was finally reduced to 5 clusters and used this number in other algorithms. As it was shown in previous experiments using chemical compounds, must-link constraints cannot help when we have a partial labeling for two coarse classes, but we are interested in discovering their subgroups.



(a) CEC-IB (b) c-GMM

(c) k-means (d) fc-means

Figure 12: Image segmentation with five clusters.

Thus, the partition-level side information was only transformed into cannot-link constraints.

The results presented in Figure 12 show that CEC-IB separated the background from the foreground quite well. Each of these two regions was described by two clusters, while the fifth group was used for detecting the boundary between them. The creation of such an additional group is natural, because feature vectors were constructed using overlapping windows and the contrast between background and foreground was sharp. One may notice that c-GMM and k-means also allocated one group for the boundary (green colored cluster). Nevertheless, both methods created clusters which mixed some elements from the background and foreground (blue and cyan colored clusters). The result returned by fc-means separated the two main parts of the image, but also contains a lot of small artifacts. As in the chemical example, spec was not able to achieve reasonable results with cannot-link constraints only. Similarly, it was not possible to run mixmod in this case.

6. Conclusion

We introduced a semi-supervised clustering method that combines model-based clustering realized by CEC with the constraint used by the information bottleneck method. The proposed cost function consists of three terms: the first tries to minimize the final number of clusters, the second penalizes the model for being inconsistent with side information, and the third controls the quality of data modeling. The performance of our method can be tuned by changing a weight parameter that trades between these three conflicting goals. Our method is flexible in the sense that it can be applied to both classical semi-supervised clustering tasks, as well as to tasks in which either not all classes appear in the labels or in which subgroups should be discovered based on the labels. For the latter problems, it is difficult or computationally expensive to use existing techniques. Setting the weight parameter appropriately, for which we provide a deep theoretical analysis, makes our method robust to incorrect labels. We evaluated the performance of our method on several data sets, including a case study using chemical compounds data set and an image segmentation task.

Appendix A. Cross-Entropy Clustering

The empirical cross-entropy between the data set X and the parametric mixture f of Gaussian densities is, for a given clustering $\mathcal{Y}$,

$$\begin{aligned} H^{\times}(X\|f) &= -\frac{1}{|X|}\sum_{i=1}^{k}\sum_{x\in Y_i}\log(p_i f_i(x)) \\ &= -\sum_{i=1}^{k}\frac{|Y_i|}{|X|}\left(\log p_i + \frac{1}{|Y_i|}\sum_{x\in Y_i}\log f_i(x)\right) \\ &= -\sum_{i=1}^{k}\frac{|Y_i|}{|X|}\log p_i + \sum_{i=1}^{k}\frac{|Y_i|}{|X|}H^{\times}(Y_i\|f_i). \end{aligned}$$

The first sum is minimized by selecting $p_i = |Y_i|/|X|$, in which case the cross-entropy reduces to the entropy of the cluster partition

$$H(\mathcal{Y}) := -\sum_{i=1}^{k}\frac{|Y_i|}{|X|}\log\frac{|Y_i|}{|X|}.$$

For the second sum, recall that the cross-entropy of a Gaussian density $f = \mathcal{N}(\mu, \Sigma)$ with mean vector μ and covariance matrix Σ equals:

$$H^{\times}(X\|f) = \tfrac{N}{2}\ln(2\pi) + \tfrac{1}{2}\|\mu_X - \mu\|_{\Sigma} + \tfrac{1}{2}\mathrm{tr}(\Sigma^{-1}\Sigma_X) + \tfrac{1}{2}\ln\det(\Sigma),$$

where μ_X and Σ_X are the sample mean vector and sample covariance matrix of X, respectively, and where $\|x\|_{\Sigma}$ is the Mahalanobis norm of x with respect to Σ. The density $f \in \mathcal{G}$ minimizing the cross-entropy function is $f = \mathcal{N}(\mu_X, \Sigma_X)$, i.e., its mean equals the sample mean of X, and its covariance matrix equals the sample covariance matrix of X [8, Theorem 4.1]. In this case, the cross entropy equals the differential Shannon entropy of $\mathcal{N}(\mu_X, \Sigma_X)$, i.e.,

$$H^{\times}(X\|\mathcal{N}(\mu_X, \Sigma_X)) = \tfrac{N}{2}\ln(2\pi e) + \tfrac{1}{2}\ln\det(\Sigma_X) = H(\mathcal{N}(\mu_X, \Sigma_X)).$$

It follows that the second sum is minimized by selecting, for every i, the maximum likelihood estimator of Y_i, $f_i = \mathcal{N}(\mu_{Y_i}, \Sigma_{Y_i})$ [8, Theorem 4.1, Proposition 4.1].

Appendix B. Proof of Theorem 3.1

We consider the CEC cost function (2) separately for each category Z_j and define the conditional cross-entropy as

$$H^{\times}((X\|f)|\mathcal{Z}) = \sum_{j=1}^{m} \frac{|Z_j|}{|X|} H^{\times}(Z_j\|f_{|j})$$

where

$$f_{|j} = \max(p_1(j)f_1, \ldots, p_k(j)f_k).$$

In other words, we assume a parameterized density model in which the weights $p_i(j)$ may depend on the category, while the densities f_i may not. Rewriting above cost yields

$$\begin{aligned} H^{\times}((X\|f)|\mathcal{Z}) &= -\sum_{j=1}^{m} \frac{|Z_j|}{|X|} \sum_{x\in Z_j} \frac{1}{|Z_j|} \log f_{|j}(x) \\ &= -\frac{1}{|X|} \sum_{j=1}^{m} \sum_{i=1}^{k} \sum_{x\in Z_j\cap Y_i} \log p_i(j) f_i(x) \\ &= -\sum_{j=1}^{m} \sum_{i=1}^{k} \frac{|Z_j \cap Y_i|}{|X|} \log p_i(j) - \frac{1}{|X|} \sum_{i=1}^{k} \sum_{x\in Y_i} \log f_i(x). \end{aligned}$$

Table C.3: Mean number of iterations that Hartigan CEC, EM-based GMM and Lloyd k-means need to converge.

Data set	**Hartigan CEC**	**EM-based GMM**	**Hartigan k-means**	**Lloyd k-means**
Ecoli	6.4	18.6	3.2	10.4
Glass	5.5	15.7	3.3	9.7
Iris	5.1	19.1	2.3	7
Segmentation	4.4	16.7	3.3	9.3
User Modeling	8.5	48.2	3.9	10.9
Vertebral	7.6	17.8	2.3	9
Wine	7.6	13.6	2.3	8.6

The second sum is minimized by the maximum likelihood estimates $f_i = \mathcal{N}(\mu_{Y_i}, \Sigma_{Y_i})$, while the first term is minimized for $p_i(j) = \frac{|Z_j \cap Y_i|}{|Z_j|}$. We thus get

$$\begin{aligned} -\sum_{j=1}^{m}\sum_{i=1}^{k} \frac{|Z_j \cap Y_i|}{|X|} \log p_i(j) &= -\sum_{j=1}^{m}\sum_{i=1}^{k} \frac{|Z_j \cap Y_i|}{|X|} \log \frac{|Z_j \cap Y_i|}{|Z_j|} \\ &= H(\mathcal{Y}|\mathcal{Z}) \\ &= H(\mathcal{Z}|\mathcal{Y}) + H(\mathcal{Y}) - H(\mathcal{Z}) \end{aligned}$$

by the chain rule of entropy. Since $H(\mathcal{Z})$ does not depend on the clustering $\mathcal{Y}$, the minimization of the above conditional cross-entropy is equivalent to the minimization of

$$H(\mathcal{Y}) + \sum_{i=1}^{k} \frac{|Y_i|}{|X|} H(\mathcal{N}(\mu_{Y_i}, \Sigma_{Y_i})) + H(\mathcal{Z}|\mathcal{Y}). \tag{B.1}$$

This is exactly the cost (4) for $\beta = 1$. □

Appendix C. Convergence Speed of Hartigan-Based CEC

We compared the number of iterations that CEC, EM, and k-means required to converge to a local minimum. We used the seven UCI data sets from Table 1 and averaged the results over ten runs. Side information was not considered in these experiments. Table C.3 shows that the Hartigan heuristic applied to the CEC cost function converges faster than EM does for fitting a GMM. The same holds when comparing the Hartigan algorithm with Lloyd's

method applied to k-means. Similar results were obtained in an experimental evaluation [9]. We also found out that that CEC-IB uses a similar number of iterations as CEC; however, the convergence speed varies with the particular sample of side information, which makes a reliable comparison more difficult.

Appendix D. Chain Rule for Proportional Partitions

We now show that, for partitions $\mathcal{Y}$ proportional to $\mathcal{Z}$, the chain rule of entropy can be applied, i.e.,

$$H(\mathcal{Y}) + H(\mathcal{Z}|\mathcal{Y}) = H(\mathcal{Z}, \mathcal{Y}).$$

We have

$$\begin{aligned}
&H(\mathcal{Y}) + H(\mathcal{Z}|\mathcal{Y}) \\
&= -\sum_{i=1}^{k} \frac{|Y_i|}{|X|} \log \frac{|Y_i|}{|X|} - \sum_{i=1}^{k} \frac{|Y_i|}{|X|} \sum_{j=1}^{m} \frac{|Y_i \cap Z_j|}{|Y_i \cap X_\ell|} \log \frac{|Y_i \cap Z_j|}{|Y_i \cap X_\ell|} \\
&\overset{(a)}{=} -\sum_{i=1}^{k} \frac{|Y_i \cap X_\ell|}{|X_\ell|} \log \frac{|Y_i \cap X_\ell|}{|X_\ell|} - \sum_{i=1}^{k} \frac{|Y_i \cap X_\ell|}{|X_\ell|} \sum_{j=1}^{m} \frac{|Y_i \cap Z_j|}{|Y_i \cap X_\ell|} \log \frac{|Y_i \cap Z_j|}{|Y_i \cap X_\ell|} \\
&= -\sum_{i=1}^{k} \sum_{j=1}^{m} \frac{|Y_i \cap Z_j|}{|X_\ell|} \log \frac{|Y_i \cap X_\ell|}{|X_\ell|} - \sum_{i=1}^{k} \sum_{j=1}^{m} \frac{|Y_i \cap Z_j|}{|X_\ell|} \log \frac{|Y_i \cap Z_j|}{|Y_i \cap X_\ell|} \\
&= -\sum_{i=1}^{k} \sum_{j=1}^{m} \frac{|Y_i \cap Z_j|}{|X_\ell|} \log \frac{|Y_i \cap Z_j|}{|X_\ell|} \\
&= H(\mathcal{Z}, \mathcal{Y})
\end{aligned}$$

where (a) is because $\mathcal{Y}$ is proportional to $\mathcal{Z}$ and thus $\frac{|Y_i \cap X_\ell|}{|X_\ell|} = \frac{|Y_i|}{|X|}$. In a similar manner it can be shown that

$$H(\mathcal{Z}) + H(\mathcal{Y}|\mathcal{Z}) = H(\mathcal{Z}, \mathcal{Y})$$

where $H(\mathcal{Z}) = -\sum_{j=1}^{m} \frac{|Z_j|}{|X_\ell|} \log \frac{|Z_j|}{|X_\ell|}$ and where

$$H(\mathcal{Y}|\mathcal{Z}) = \sum_{j=1}^{m} \frac{|Z_j|}{|X_\ell|} H(\mathcal{Y}|Z_j) = -\sum_{j=1}^{m} \sum_{i=1}^{k} \frac{|Y_i \cap Z_j|}{|X_\ell|} \log \frac{|Y_i \cap Z_j|}{|Z_j|}.$$

Appendix E. Proof of Theorem 4.1

Lemma Appendix E.1. *Let the data set $X \subset \mathbb{R}^N$ be partitioned into two clusters Y_1 and Y_2 such that the sample covariance matrix Σ_i of Y_i is positive definite for $i = 1, 2$.*

Then

$$H(\mathcal{N}(\mu_X, \Sigma_X)) \geq \frac{|Y_1|}{|X|} H(\mathcal{N}(\mu_{Y_1}, \Sigma_{Y_1})) + \frac{|Y_2|}{|X|} H(\mathcal{N}(\mu_{Y_2}, \Sigma_{Y_2})).$$

Proof. Let $p = |Y_1|/|X|$. By the law of total (co-)variance, we have

$$\begin{aligned} \Sigma_X = {} & p\Sigma_{Y_1} + (1-p)\Sigma_{Y_2} \\ & + \underbrace{p(\mu_X - \mu_{Y_1})(\mu_X - \mu_{Y_1})^T + (1-p)(\mu_X - \mu_{Y_2})(\mu_X - \mu_{Y_2})^T}_{=:\tilde{\Sigma}} \end{aligned} \tag{E.1}$$

where $\tilde{\Sigma}$ is the covariance matrix obtained from the sample mean vectors μ_{Y_1} and μ_{Y_2} of Y_1 and Y_2. Consequently, we get

$$\begin{aligned} & pH(\mathcal{N}(\mu_{Y_1}, \Sigma_{Y_1})) + (1-p)H(\mathcal{N}(\mu_{Y_2}, \Sigma_{Y_2})) \\ & = \frac{Np}{2}\ln(2\pi e) + \frac{p}{2}\ln(\det \Sigma_{Y_1}) + \frac{N(1-p)}{2}\ln(2\pi e) + \frac{(1-p)}{2}\ln(\det \Sigma_{Y_2}) \\ & = \frac{N}{2}\ln(2\pi e) + \frac{1}{2}\ln\left((\det \Sigma_{Y_1})^p (\det \Sigma_{Y_2})^{(1-p)}\right) \\ & \overset{(a)}{\leq} \frac{N}{2}\ln(2\pi e) + \frac{1}{2}\ln(\det(p\Sigma_{Y_1} + (1-p)\Sigma_{Y_2})) \\ & \overset{(b)}{\leq} \frac{N}{2}\ln(2\pi e) + \frac{1}{2}\ln(\det(p\Sigma_{Y_1} + (1-p)\Sigma_{Y_2} + \tilde{\Sigma})) \\ & = \frac{N}{2}\ln(2\pi e) + \frac{1}{2}\ln(\det \Sigma) \\ & = H(\mathcal{N}(m, \Sigma_X)) \end{aligned}$$

where (a) follows because Σ_{Y_1} and Σ_{Y_2} are positive definite and from [49, Cor. 7.6.8], and where (b) follows because $\tilde{\Sigma}$ is positive semi-definite and from, e.g., [49, Cor. 4.3.12]. □

Proof of Theorem 4.1. Since $\mathcal{Y}$ is proportional to $\mathcal{Z}$ and since the coarsening of a proportional partition is proportional, we can apply the chain rule of

entropy to get (see Appendix D)

$$H(\mathcal{Y}) + H(\mathcal{Z}|\mathcal{Y}) = H(\mathcal{Z}, \mathcal{Y})$$
$$H(\tilde{\mathcal{Y}}) + H(\mathcal{Z}|\tilde{\mathcal{Y}}) = H(\mathcal{Z}, \tilde{\mathcal{Y}}).$$

We hence get

$$H(\mathcal{Y}) + H(\mathcal{Z}|\mathcal{Y}) = H(\mathcal{Z}, \mathcal{Y}) \overset{(a)}{=} H(\mathcal{Z}, \mathcal{Y}, \tilde{\mathcal{Y}}) = H(\mathcal{Z}, \tilde{\mathcal{Y}}) + H(\mathcal{Y}|\mathcal{Z}, \tilde{\mathcal{Y}})$$
$$\overset{(b)}{=} H(\mathcal{Z}, \tilde{\mathcal{Y}}) = H(\tilde{\mathcal{Y}}) + H(\mathcal{Z}|\tilde{\mathcal{Y}}) \quad \text{(E.2)}$$

where (a) is because $\tilde{\mathcal{Y}}$ is a coarsening of $\mathcal{Y}$ and (b) is because $\mathcal{Y}$ is a coarsening of $\mathcal{Z}$, respectively. In other words, for proportional coarsenings of $\mathcal{Z}$, consistency with $\mathcal{Z}$ (measured by the conditional entropy) can be freely traded for model simplicity (measured by entropy).

For the remaining part of the RHS of (12), we write:

$$\sum_{i=1}^{k} \frac{|Y_i|}{|X|} H(\mathcal{N}(\mu_{Y_i}, \Sigma_{Y_i})) = \sum_{j=1}^{k'} \frac{|\tilde{Y}_j|}{|X|} \sum_{i: Y_i \subseteq \tilde{Y}_j} \frac{|Y_i|}{|\tilde{Y}_j|} H(\mathcal{N}(\mu_{Y_i}, \Sigma_{Y_i})).$$

If the inner sums on the RHS consist of at most two terms, i.e. $\tilde{Y}_j = \{Y_{i_1}, Y_{i_2}\}$ or $\tilde{Y}_j = Y_i$, the inequality is established by Lemma Appendix E.1. If the inner sum consists of more than two clusters, one needs to apply Lemma Appendix E.1 recursively. For example, if $\tilde{Y}_1 = \{Y_1, Y_2, Y_3\}$,

$$\sum_{i=1}^{3} \frac{|Y_i|}{|\tilde{Y}_1|} H(\mathcal{N}(\mu_{Y_i}, \Sigma_{Y_i}))$$
$$\leq \frac{|Y_1 \cup Y_2|}{|\tilde{Y}_1|} H(\mathcal{N}(\mu_{Y_1 \cup Y_2}, \Sigma_{Y_1 \cup Y_2})) + \frac{|Y_3|}{|\tilde{Y}_1|} H(\mathcal{N}(\mu_{Y_3}, \Sigma_{Y_3}))$$
$$\leq H(\mathcal{N}(\mu_{\tilde{Y}_1}, \Sigma_{\tilde{Y}_1}).$$

This completes the proof. □

Appendix F. Proof of Theorem 4.2

Note that, with (13) and (E.2) (since both $\mathcal{Y}$ and $\tilde{\mathcal{Y}}$ are proportional coarsenings of $\mathcal{Z}$), we obtain

$$\begin{aligned}
&H(\tilde{\mathcal{Y}}) + \beta H(\mathcal{Z}|\tilde{\mathcal{Y}}) - H(\mathcal{Y}) - \beta H(\mathcal{Z}|\mathcal{Y}) \\
&= H(\tilde{\mathcal{Y}}) + \beta H(\mathcal{Z}|\tilde{\mathcal{Y}}) - (1-\beta)H(\mathcal{Y}) - \beta(H(\tilde{\mathcal{Y}}) + H(\mathcal{Z}|\tilde{\mathcal{Y}})) \\
&= (1-\beta)\Big(H(\tilde{\mathcal{Y}}) - H(\mathcal{Y})\Big) \\
&= (\beta-1)H(\mathcal{Y}|\tilde{\mathcal{Y}}) = (\beta-1)H\left(\frac{p_{k'}}{q_{k'}}, \ldots, \frac{p_k}{q_{k'}}\right).
\end{aligned}$$

For $i = 1, \ldots, k'-1$, we have $H(\mathcal{N}(\mu_{Y_i}, \Sigma_{Y_i})) = H(\mathcal{N}(\mu_{\tilde{Y}_i}, \Sigma_{\tilde{Y}_i}))$, hence only the last term remains. We obtain with the proof of Theorem 4.1,

$$H(\mathcal{N}(\mu, \Sigma)) - \sum_{i=k'}^{k} \frac{p_i}{q_{k'}} H(\mathcal{N}(\mu_{Y_i}, \Sigma_{Y_i})) = \sum_{i=k'}^{k} \frac{p_i}{2q_{k'}} \ln\left(\frac{\det \Sigma}{\det \Sigma_{Y_i}}\right).$$

It follows that the two costs in the statement are equal for β_0 such that

$$(\beta_0 - 1)H\left(\frac{p_{k'}}{q_{k'}}, \ldots, \frac{p_k}{q_{k'}}\right) = \sum_{i=k'}^{k} \frac{p_i}{2q_{k'}} \ln\left(\frac{\det \Sigma_{Y_i}}{\det \Sigma}\right)$$

from which we get

$$\beta_0 = 1 + \frac{\sum_{i=k'}^{k} \frac{p_i}{2q_{k'}} \ln\left(\frac{\det \Sigma_{Y_i}}{\det \Sigma}\right)}{H\left(\frac{p_{k'}}{q_{k'}}, \ldots, \frac{p_k}{q_{k'}}\right)}. \tag{F.1}$$

□

Acknowledgement

The authors thank Hongfu Liu, Pengjiang Qian and Daphne Teck Ching Lai for sharing their codes implementing semi-supervised versions of k-means, spectral clustering and fuzzy clustering. We also thank Jacek Tabor for useful discussions and comments.

The work of Marek Śmieja was supported by the National Science Centre (Poland) grant no. 2016/21/D/ST6/00980. The work of Bernhard C. Geiger has been funded by the Erwin Schrödinger Fellowship J 3765 of the Austrian Science Fund and by the German Ministry of Education and Research in the framework of an Alexander von Humboldt Professorship.

References

[1] S. Basu, I. Davidson, K. Wagstaff, Constrained clustering: Advances in algorithms, theory, and applications, CRC Press, 2008.

[2] J. Yi, R. Jin, S. Jain, T. Yang, A. K. Jain, Semi-crowdsourced clustering: Generalizing crowd labeling by robust distance metric learning, in: Advances in Neural Information Processing Systems (NIPS), Lake Tahoe, 2012, pp. 1772–1780.

[3] L. Fei-Fei, P. Perona, A Bayesian hierarchical model for learning natural scene categories, in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, 2005, pp. 524–531.

[4] S. Basu, Semi-supervised clustering: Learning with limited user feedback, Ph.D. thesis, The University of Texas at Austin (2003).

[5] H. Liu, Y. Fu, Clustering with partition level side information, in: Proc. IEEE Int. Conf. on Data Mining (ICDM), Atlantic City, NJ, 2015, pp. 877–882.

[6] X. Zhu, A. B. Goldberg, Introduction to semi-supervised learning, Synthesis lectures on artificial intelligence and machine learning 3 (1) (2009) 1–130.

[7] E. Tu, Y. Zhang, L. Zhu, J. Yang, N. Kasabov, A graph-based semi-supervised k nearest-neighbor method for nonlinear manifold distributed data classification, Information Sciences 367 (2016) 673–688.

[8] J. Tabor, P. Spurek, Cross-entropy clustering, Pattern Recognition 47 (9) (2014) 3046–3059.

[9] P. Spurek, K. Kamieniecki, J. Tabor, K. Misztal, M. Śmieja, R package CEC, Neurocomputing 237 (2016) 410–413.

[10] P. Spurek, J. Tabor, K. Byrski, Active function cross-entropy clustering, Expert Systems with Applications 72 (2017) 49–66.

[11] N. Tishby, F. C. Pereira, W. Bialek, The information bottleneck method, in: Proc. Allerton Conf. on Communication, Control, and Computing, Monticello, IL, 1999, pp. 368–377.

[12] G. Chechik, A. Globerson, N. Tishby, Y. Weiss, Information bottleneck for Gaussian variables, Journal of Machine Learning Research 6 (Jan) (2005) 165–188.

[13] A. K. Jain, M. N. Murty, P. J. Flynn, Data clustering: a review, ACM computing surveys (CSUR) 31 (3) (1999) 264–323.

[14] C. C. Aggarwal, C. K. Reddy, Data clustering: algorithms and applications, Chapman and Hall/CRC, 2013.

[15] Y. Jiang, F.-L. Chung, S. Wang, Z. Deng, J. Wang, P. Qian, Collaborative fuzzy clustering from multiple weighted views, IEEE Transactions on Cybernetics 45 (4) (2015) 688–701.

[16] D. Gondek, T. Hofmann, Non-redundant data clustering, Knowledge and Information Systems 12 (1) (2007) 1–24.

[17] S. Asafi, D. Cohen-Or, Constraints as features, in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Portland, OR, 2013, pp. 1634–1641.

[18] S. Kamvar, D. Klein, C. Manning, Spectral learning, in: Proc. Int. Joint Conf. on Artificial Intelligence (IJCAI), Acapulco, Mexico, 2003, pp. 561–566.

[19] Z. Wang, I. Davidson, Flexible constrained spectral clustering, in: Proc. ACM Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD), Washington, DC, 2010, pp. 563–572.

[20] P. Qian, Y. Jiang, S. Wang, K.-H. Su, J. Wang, L. Hu, R. F. Muzic, Affinity and penalty jointly constrained spectral clustering with all-compatibility, flexibility, and robustness, IEEE Transactions on Neural Networks and Learning Systems, accepted for publication.

[21] D. Calandriello, G. Niu, M. Sugiyama, Semi-supervised information-maximization clustering, Neural Networks 57 (2014) 103–111.

[22] M. Lu, X.-J. Zhao, L. Zhang, F.-Z. Li, Semi-supervised concept factorization for document clustering, Information Sciences 331 (2016) 86–98.

[23] W. Pedrycz, J. Waletzky, Fuzzy clustering with partial supervision, IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 27 (5) (1997) 787–795.

[24] W. Pedrycz, A. Amato, V. Di Lecce, V. Piuri, Fuzzy clustering with partial supervision in organization and classification of digital images, IEEE Transactions on Fuzzy Systems 16 (4) (2008) 1008–1026.

[25] D. T. C. Lai, J. M. Garibaldi, Improving semi-supervised fuzzy c-means classification of breast cancer data using feature selection, in: Proc. IEEE Int. Conf. on Fuzzy Systems (FUZZ-IEEE), IEEE, Hyderabad, India, 2013, pp. 1–8.

[26] L. Lelis, J. Sander, Semi-supervised density-based clustering, in: Proc. IEEE Int. Conf. on Data Mining, IEEE, Miami, Florida, 2009, pp. 842–847.

[27] C. Ambroise, T. Denoeux, G. Govaert, P. Smets, Learning from an imprecise teacher: probabilistic and evidential approaches, Applied Stochastic Models and Data Analysis 1 (2001) 100–105.

[28] E. Côme, L. Oukhellou, T. Denoeux, P. Aknin, Learning from partially supervised data using mixture models and belief functions, Pattern Recognition 42 (3) (2009) 334–348.

[29] E. Hüllermeier, J. Beringer, Learning from ambiguously labeled examples, in: Proc. Int. Symposium on Intelligent Data Analysis (IDA), Springer, Madrid, Spain, 2005, pp. 168–179.

[30] C. Bouveyron, S. Girard, Robust supervised classification with mixture models: Learning from data with uncertain labels, Pattern Recognition 42 (11) (2009) 2649–2658.

[31] R. Lebret, S. Iovleff, F. Langrognet, C. Biernacki, G. Celeux, G. Govaert, Rmixmod: the R package of the model-based unsupervised, supervised and semi-supervised classification mixmod library, Journal of Statistical Software 67 (6) (2015) 241–270.

[32] P. Biecek, E. Szczurek, M. Vingron, J. Tiuryn, et al., The R package bgmm: mixture modeling with uncertain knowledge, Journal of Statistical Software 47 (3) (2012) 31.

[33] N. Shental, A. Bar-hillel, T. Hertz, D. Weinshall, Computing Gaussian mixture models with EM using equivalence constraints, in: Advances in Neural Information Processing Systems (NIPS), Vancouver, British Columbia, Canada, 2004, pp. 465–472.

[34] Z. Lu, T. K. Leen, Semi-supervised learning with penalized probabilistic clustering, in: Advances in Neural Information Processing Systems (NIPS), Vancouver, British Columbia, Canada, 2005, pp. 849–856.

[35] S. Basu, M. Bilenko, R. J. Mooney, A probabilistic framework for semi-supervised clustering, in: Proc. ACM Int. Conf. on Knowledge Discovery and Data Mining (SIGKDD), Seattle, WA, 2004, pp. 59–68.

[36] T. Lange, M. H. Law, A. K. Jain, J. M. Buhmann, Learning with constrained and unlabelled data, in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), San Diego, CA, 2005, pp. 731–738.

[37] B. Nelson, I. Cohen, Revisiting probabilistic models for clustering with pair-wise constraints, in: Proc. Int. Conf. on Machine Learning (ICML), Corvallis, OR, 2007, pp. 673–680.

[38] N. Slonim, The information bottleneck: Theory and applications, Ph.D. thesis, Hebrew University of Jerusalem (2002).

[39] A. Topchy, A. K. Jain, W. Punch, Combining multiple weak clusterings, in: Proc. IEEE Int. Conf. on Data Mining (ICDM), Melbourne, Florida, 2003, pp. 331–338.

[40] D. Strouse, D. J. Schwab, The deterministic information bottleneck, in: Proc. Conf. on Uncertainty in Artificial Intelligence (UAI), New York City, NY, 2016, pp. 696–705.

[41] J. A. Hartigan, M. A. Wong, Algorithm AS 136: A k-means clustering algorithm, Journal of the Royal Statistical Society. Series C (Applied Statistics) 28 (1) (1979) 100–108.

[42] R. A. Redner, H. F. Walker, Mixture densities, maximum likelihood and the EM algorithm, SIAM Review 26 (2) (1984) 195–239.

[43] M. Lichman, UCI machine learning repository (2013).
URL `http://archive.ics.uci.edu/ml`

[44] D. Warszycki, S. Mordalski, K. Kristiansen, R. Kafel, I. Sylte, Z. Chilmonczyk, A. J. Bojarski, A linear combination of pharmacophore hypotheses as a new tool in search of new active compounds–an application for 5-HT1A receptor ligands, PloS ONE 8 (12) (2013) e84510.

[45] L. Ana, A. K. Jain, Robust data clustering, in: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Vol. 2, 2003, pp. II–128.

[46] B. Olivier, W. Soudijn, I. van Wijngaarden, The 5-HT1A receptor and its ligands: structure and function, in: Progress in Drug Research, Vol. 52, 1999, pp. 103–165.

[47] M. Śmieja, D. Warszycki, Average information content maximization - a new approach for fingerprint hybridization and reduction, PLoS ONE 11 (1) (2016) e0146666.

[48] J. Klekota, F. P. Roth, Chemical substructures that enrich for biological activity, Bioinformatics 24 (21) (2008) 2518–2525.

[49] R. A. Horn, C. R. Johnson, Matrix Analysis, 2nd Edition, Cambridge University Press, Cambridge, 2013.