

# Non-Redundant Data Clustering

David Gondek      Thomas Hofmann  
Department of Computer Science, Brown University  
Providence, RI 02912 USA  
{dcg,th}@cs.brown.edu

## Abstract

*Data clustering is a popular approach for automatically finding classes, concepts, or groups of patterns. In practice this discovery process should avoid redundancies with existing knowledge about class structures or groupings, and reveal novel, previously unknown aspects of the data. In order to deal with this problem, we present an extension of the information bottleneck framework, called coordinated conditional information bottleneck, which takes negative relevance information into account by maximizing a conditional mutual information score subject to constraints. Algorithmically, one can apply an alternating optimization scheme that can be used in conjunction with different types of numeric and non-numeric attributes. We present experimental results for applications in text mining and computer vision.*

## 1 Introduction

Data mining and knowledge discovery aim at finding concepts, patterns, relationships, regularities, and structures of interest in a given data set. However, in practice it is rather atypical to be faced with data about which nothing is known already. More typically, some knowledge has already been acquired in the past, possibly through precedent knowledge discovery processes. The goal then is more precisely to mine for concepts, relationships, etc. that will augment the existing knowledge and that are in some sense non-redundant and novel, relative to the available background knowledge. This leads to the general problem of how to *subtract* or *factor out* the background knowledge in a principled way, in order to *augment* it through additional data mining and exploration.

We address this problem in an information-theoretic framework that makes use of the concept of *conditional mutual information* as its cornerstone. We propose to quantify the information or knowledge gained by a data mining process in terms of how much *new information* it adds about

relevant aspect of our data, conditioned on the already available knowledge. It is this conditioning that will provide a well-founded basis for the subtraction process we alluded to.

In this paper, we investigate the important problem of non-redundant data clustering, which deals with finding classes that are in some sense orthogonal to existing knowledge. As far as the form of this knowledge is concerned, our focus is on problems that involve conditioning on one or more attributes or properties of instances. This includes conditioning on known classification schemes as a special case of particular relevance. There are many applications and problems that can be subsumed under this general setting, of which we enumerate a few here to provide some further motivation.

(i) Clustering a set of documents in a way that does not overlap with or recover known classification schemes, but rather discovers new ways in which to group documents. For instance, news stories may be clustered by geographic region as well as by topic. Assuming that documents are annotated by region, conditioning on this information may be valuable to enforce a clustering by topic. (ii) Clustering documents such that the found clusters are not correlated with the occurrence of certain terms. Using the above news story example, one may want to condition on the occurrence of certain geographic terms such as country and city names to introduce a bias that favors document clusters that are not based on geography. (iii) Grouping users for which transactional data has been collected in a way that is not based on certain demographic attributes. For instance, one may want to group users in ways that are not correlated with stratifications based on gender or income. (iv) Finding non-dominant clustering solutions in arbitrary data sets. By first finding the dominant grouping structure and by then conditioning on the latter, non-dominant clustering alternatives can be discovered.

While our method can be applied to these and many similar problems, our emphasis is on a common modeling framework and on the derivation of a general data mining methodology. We will only investigate specific instantia-

tions in an exemplary manner in the experiment section.

## 2 Related work

Our contribution is in line with a number of recent papers that have argued in favor of data mining techniques that are exploratory, and allow for user interaction and control (e.g. [8]): Techniques should allow users to interactively refine or modify queries based on the results of previous queries. They should furthermore allow users to specify prior knowledge and provide feedback in order to guide the search, both towards desired solutions and away from undesired solutions. Techniques have been presented for tasks ranging from constrained itemset mining [1] to constrained association rule mining [8].

For the clustering problem a variety of constrained clustering techniques exist. Constraints may take the form of *cluster aggregate constraints* [12, 15], e.g. by constraining clustering solutions to equally-sized clusters. Another line of work has focused on *instance-level constraints* [13, 6, 14]. As described in [13], these constraints are informed by prior knowledge of the desired clustering and typically take the form of relations such as *must-link* and *cannot-link* which are enforced between pairs of instances. This approach is extended by [6, 14] which infer from instance-level constraints proximity matrices and formal distance metrics respectively. The common trait of these approaches is the assumption that prior knowledge takes the form of *positive* information about certain characteristics of a desired clustering solution.

Constrained clustering techniques which take *negative* information in the form of information about undesired solutions were first formulated in [2]. Another conceptually related framework has been presented in [4]. We will postpone a detailed discussion of these methods until the end of the following section.

## 3 Coordinated Conditional Information Bottleneck

### 3.1 Preliminaries

The mutual information  $I(A; B)$  between two (discrete) random variables  $A, B$  is defined as

$$I(A; B) \equiv \sum_a \sum_b P_{AB}(a, b) \log \frac{P_{AB}(a, b)}{P_A(a)P_B(b)}, \quad (1)$$

where the sums are over the respective sample spaces for  $A$  and  $B$ . We have utilized a generic notation for probability mass functions using subscripts involving random variables. Alternatively and equivalently one may define

mutual information via conditional entropies as  $I(A; B) = H(A) - H(A|B)$ .

The conditional mutual information  $I(A; B|C)$  between random variables  $A, B$  given a random variable  $C$  can be defined as

$$\begin{aligned} I(A; B|C) &\equiv H(A|C) - H(A|B, C) \\ &= I(A; B, C) - I(A; C). \end{aligned} \quad (2)$$

### 3.2 Setting and notation

We use the following notation:  $x$  refers to objects or items, such as documents, that should be clustered,  $y$  to features that are considered relevant, e.g. word occurrences in documents,  $c$  to clusters of objects, and  $z$  to available background knowledge. Uppercase letters  $X, Y, C, Z$  are used to denote the corresponding random variables. To simplify the presentation, we assume that background knowledge and features depend deterministically on the object, i.e.  $Y = Y(X)$  and  $Z = Z(X)$ .

We work in a probabilistic setting, where objects are probabilistically assigned to clusters. The goal of data clustering is thus to find a stochastic mapping  $P_{C|X}$  of objects  $x$  to clusters  $c \in \{1, \dots, k\}$ , where the number of clusters  $k$  is assumed to be given. Here  $P_{C|X}$  refers to the conditional distribution of  $C$  given  $X$ , i.e.  $P_{C|X}(c|x)$  – or  $P(c|x)$  for short – denotes the probability of assigning object  $x$  to cluster  $c$ .

### 3.3 Conditional Information Bottleneck

Given a particular choice for  $P_{C|X}$ , we would like to quantify the amount of information preserved in the clustering about the relevant features  $Y$ . However, we also need to take into account that we assume to have access to the background knowledge  $Z$ . A natural quantity to consider is the conditional mutual information  $I(C; Y|Z)$ . It describes how much information  $C, Z$  convey jointly about relevant features  $Y$  compared to the information provided by  $Z$  alone. Finding an optimal clustering solution should involve maximizing  $I(C; Y|Z)$ .

In addition, we would like to avoid over-confidence in grouping objects together. Cluster assignment probabilities should reflect the uncertainty with which objects are assigned to clusters. One way to accomplish this is to explicitly control the fuzziness of the stochastic mapping  $P_{C|X}$ . The latter can be measured by the mutual information  $I(C; X)$  between cluster and object identities. Here  $I(C; X) = 0$ , if objects are assigned to clusters completely at random, whereas  $I(C; X)$  becomes maximal for non-stochastic mappings  $P_{C|X}$ .  $I(C; X)$  also can be given a well-known interpretation in terms of the channel capacity required for transmitting probabilistic cluster assignments over a communication channel [11].

Combining both aspects, we define the optimal clustering as the solution to the following constrained optimization problem, the Conditional Information Bottleneck (CIB), first introduced in [5]:

$$\text{(CIB)} \quad P_{C|X}^* = \operatorname{argmax}_{P_{C|X} \in \mathcal{P}} I(C; Y|Z), \quad \text{where} \quad (3a)$$

$$\mathcal{P} \equiv \{P_{C|X} : I(C; X) \leq C_{\max}\}. \quad (3b)$$

Stated in plain English, we are looking for probabilistic cluster assignments with a minimal fuzziness such that the relevant information jointly encoded in  $C, Z$  is maximal.

### 3.4 Coordinated CIB

While the conditional information reflects much of the intuition behind non-redundant data mining, there still is a potential caveat in using Eq. (3): the definition of  $C$  may lack global coordination. That is, clustering solutions obtained for different values  $z$  may not be in correspondence. For instance, if  $Z$  can take a finite number of possible values, then the meaning of each cluster  $c$  is relative to a particular value  $z$ . The reason for this is that  $I(C; Y|Z)$  only measures the information conveyed by  $C$  and  $Z$  in conjunction, but does not reflect how much relevant information  $C$  provides on its own, i.e. without knowing  $Z$ . We call this problem the *cluster coordination problem*.

One way to formally illustrate that the CIB does not address the coordination problem is via the following proposition which we state here without proof.

**Proposition 1.** *Suppose  $Z$  and  $C$  are finite random variables and define pre-image sets of  $Z$  by  $\mathcal{X}_z = \{x : Z(x) = z\}$ . Assume that  $P_{C|X}^*$  has been obtained according to Eq. (3). Then one can choose arbitrary permutations  $\pi^z$  over  $C$ , one for every value  $z$  of  $Z$ , and define permuted cluster assignments  $P_{C|X}^\pi(c|x) \equiv P_{C|X}^*(\pi^{Z(x)}(c)|x)$  such that  $P_{C|X}^\pi$  is also optimal for CIB.*

Intuitively this proposition states that by independently renumbering (i.e. permuting) cluster labels within each set  $\mathcal{X}_z$ , the optimality of the solution is not affected.

A solution to the CIB problem will effectively correspond to a subcategorization or a *local* refinement of the partition induced by  $Z$ . Generally, however, one is more interested in concepts or annotations  $C$  that are consistent across the whole domain of objects. We propose to address this problem by introducing an additional constraint involving  $I(C; Y)$ . This yields the following *Coordinated Conditional Information Bottleneck (CCIB)* formulation:

$$\text{(CCIB)} \quad P_{C|X}^* = \operatorname{argmax}_{P_{C|X} \in \mathcal{P}} I(C; Y|Z), \quad \text{where} \quad (4a)$$

$$\mathcal{P} \equiv \{P_{C|X} : I(C; X) \leq C_{\max}, I(C; Y) \geq I_{\min}\}. \quad (4b)$$

With  $I_{\min} > 0$  the CCIB favors clustering solutions that obey some global consistency across the sets  $\mathcal{X}_z$ .

### 3.5 Alternating optimization

The formal derivation of an alternation scheme to compute an approximate solution for the CCIB is somewhat involved, but leads to very intuitive re-estimation equations. As shown in the appendix, one can compute probabilistic cluster assignments according to the following formula:

$$P(c|x) \propto P(c) \exp \left[ \frac{\lambda}{\rho} \sum_y P(y|x) \log P(y|c) \right] \quad (5)$$

$$\times \exp \left[ \frac{1}{\rho} \sum_z P(z|x) \sum_y P(y|x, z) \log P(y|c, z) \right],$$

where we have dropped all subscripts, since the meaning of the probability mass functions is clear from the naming convention for the arguments. The scalars  $\rho \geq 0$  and  $\lambda \geq 0$  are Lagrange multipliers enforcing the two inequality constraints; their values depend on  $C_{\max}$  and  $I_{\min}$ . Notice that  $P(y|c)$  and  $P(y|c, z)$  appearing on the right-hand side of Eq. (5) implicitly depend on  $P(c|x)$ . However iterating this equation is guaranteed to reach a fixed point corresponding to a local maximum of the CCIB criterion.

### 3.6 Relation to related work

The CCIB formulation is an extension of the seminal work by Tishby et al. [11] on the information bottleneck (IB) framework. Among the different generalizations of IB proposed so far, our approach is most closely related to the IB with side information [2]. One way to formulate the latter is as the problem of minimizing  $I(C; Z)$  subject to constraints on  $I(C; Y)$  and  $I(C; X)$ . The main disadvantage that we see in this procedure is the difficulty in adjusting the trade-off between minimizing redundancy between  $C$  and  $Z$  expressed by  $I(C; Z)$ , and maximizing relevant information as expressed by the lower bound on  $I(C; Y)$ . Notice that the latter is problematic, since  $I(C; Y)$  and  $I(C; Z)$  may live on very different scales, e.g.  $I(C; Y)$  may scale with the number of relevant features, while  $I(C; Z)$  may scale with the cardinality of the state space of  $Z$ . In the CCIB formulation, this is taken into account by conditioning on the side information  $Z$  in  $I(C; Y|Z)$ , which enforces non-redundancy without the need for an explicit term  $I(C; Z)$  to penalize redundancy.

The CIB method from [5] is less general than CCIB as it requires a seed set of labeled data in order to address the coordination problem. The CIB objective in Eq. (3a) may also be justified using the multivariate information bottleneck (MIB) framework [4]. Using the so-called  $\mathcal{L}^{(1)}$  principle, one may directly derive (3a). The derivation follows that of the parallel information bottleneck, only using the assumption that  $Z$  is known a priori. However, the coordination problem is not addressed in [4]. In fact the MIB

formulation has a distinctly different goal from the one pursued with the CCIB method.

## 4 Finite sample considerations

So far we have tacitly assumed that the joint distribution  $P_{XYZ}$  is given. However, since this will rarely be the case in practice, it is crucial to be able to effectively deal with the finite sample case. Let us denote a sample set drawn i.i.d. by  $\mathcal{X}_n = \{(x_i, y_i, z_i) : i = 1, \dots, n\}$ . We will first clarify the relationship between CCIB and likelihood maximization and then investigate particular parametric forms for the approximating distribution, leading to specific instantiations of the general scheme presented in the previous section.

### 4.1 Likelihood maximization

A natural measure for the predictive performance of a model is the average conditional log-likelihood function

$$L(\mathcal{X}_n) = \frac{1}{n} \sum_{i=1}^n \sum_c P_{C|X}(c|x_i) \log \hat{P}_{Y|C,Z}(y_i|c_i, z_i). \quad (6)$$

Here  $\hat{P}_{Y|C,Z}$  is some approximation to the true distribution. This amounts to a two-stage prediction process, where  $x_i$  is first assigned to one of the clusters according to  $P_{C|X}(c|x_i)$  and then features are predicted using the distribution  $\hat{P}_{Y|C,Z}(y_i|c, z_i)$ . Asymptotically one gets

$$\begin{aligned} L(\mathcal{X}_n) &\xrightarrow{n \rightarrow \infty} \sum_{y,z} P_{C,Y,Z}(c, y, z) \log \hat{P}_{Y|C,Z}(y|c, z) \\ &= -H(Y|C, Z) - \mathbf{E}_{C,Z} \left[ KL(P_{Y|C,Z} || \hat{P}_{Y|C,Z}) \right]. \quad (7) \end{aligned}$$

Provided that the estimation error (represented by the expected Kullback-Leibler divergence) vanishes as  $n \rightarrow \infty$ , maximizing the log-likelihood with respect to  $P_{C|X}$  will thus asymptotically be equivalent to minimizing  $H(Y|C, Z)$  and thus to maximizing the conditional information  $I(C; Y|Z)$ .

The practical relevance of the above considerations is that one can use the likelihood function Eq. (6) as the basis for computing a suitable approximation  $\hat{P}_{Y|C,Z}$ . For instance, if the latter is parameterized by some parameter  $\theta$ , then one may compute the optimal parameter  $\theta^*$  as the one that maximizes  $L(\mathcal{X}_n)$ .

### 4.2 Categorical background knowledge

Let us focus on the simplest case first, where  $Z$  is finite and its cardinality is small enough to allow estimating separate conditional feature distributions  $P_{Y|C,Z}$  and  $P_{Y|Z}$  for

every combination  $c, z$  and every  $z$ , respectively. As discussed before, we can estimate the above probabilities by conditional maximum likelihood estimation. For concreteness, we present and discuss the resulting update equations and algorithm for the special case of a multinomial sampling model for  $Y$ , which is of some importance, for example, in the context of text mining application. It is simple to derive similar algorithms for other sampling models such as Bernoulli, normal, or Poisson.

Denote by  $n_{ij}$  observed feature frequencies for the  $i$ -th object and the  $j$ -th possible  $Y$ -value and by  $n_i$  the total number of observations for  $x_i$ . For instance,  $n_{ij}$  may denote the number of times the  $j$ -th term occurs in the  $i$ -th document in the context of text mining. Then we can define the relevant empirical distributions by

$$P(y_j|x_i) = P(y_j|x_i, z_i) \equiv \frac{n_{ij}}{n_i}, \quad P(z|x_i) \equiv \delta(z, z_i). \quad (8)$$

The maximum likelihood estimates for given probabilistic cluster assignments can be computed according to

$$P(y_j|c) = \frac{\sum_{i=1}^n P(c|x_i) n_{ij}}{\sum_{i=1}^n P(c|x_i) n_i}, \quad (9a)$$

$$P(y_j|c, z) = \frac{\sum_{i=1}^n P(z|x_i) P(c|x_i) n_{ij}}{\sum_{i=1}^n P(z|x_i) P(c|x_i) n_i}, \quad (9b)$$

$$P(c) = \frac{1}{n} \sum_{i=1}^n P(c|x_i). \quad (9c)$$

These equations need to be iterated in alternation with the re-estimation equation in Eq. (5), where the sum over  $y$  is replaced by a sum over the feature index  $j$ .

### 4.3 Continuous-valued background knowledge

A more general case involves background knowledge consisting of a vector  $z \in \mathbb{R}^d$ . This includes situations where  $Z$  might be a function of  $Y$  or might consist of a subset of the features (cf. Section 1). In order to obtain an estimate for  $P(y|c, z)$  one has to fit a regression model that predicts the relevant features  $Y$  from the background knowledge  $Z$  for every cluster. If the response variable  $Y$  is itself vector-valued, then we propose to fit regression models for every feature dimension separately.

The parametric form of the parametric regression function depends on the type and sampling model of the feature variable  $Y$ . For instance,  $Y$  may be a multivariate normal, a multinomial variable, or a vector of Bernoulli variables. In order to cover most cases of interest in a generic way, we propose to use of the framework of *generalized linear models* (GLMs) [7]. Since a detailed presentation of GLMs is beyond the scope of this paper, we only provide a brief outline of what is involved in this process: We assume that the conditional mean of  $Y$  can be

written as a function of  $C$  and  $Z$  in the following way  $\mathbb{E}[Y|C, Z] = \mu(C, Z) = h(\langle \theta, \phi(C, Z) \rangle)$ , where  $h$  is the inverse link function and  $\phi$  is a vector of predictor variables. Taking  $h = \text{id}$  results in standard linear regression based on the independent variables  $\phi(C, Z)$ , but a variety of other (inverse) link functions can be used dependent on the application.

In this general case, computing the quantities  $P(y|c) = P(y|c; \eta)$  and  $P(y|c, z) = P(y|c, z; \theta)$  requires to estimate  $\eta$  and  $\theta$  by maximizing the log-likelihood criterion in Eq. (6). The latter can be accomplished by standard model fitting algorithms for GLMs, which may themselves be iterative in nature.

#### 4.4 Deterministic annealing

We now address the issue of how to deal with the free parameters  $C_{\max}$  and  $I_{\min}$  of the CCIB or – equivalently – the Lagrange multipliers  $\rho$  and  $\lambda$ . Notice that the constraint  $I(C; Y) \geq I_{\min}$  leads to a Lagrangian function that additively combines two (conditional) mutual informations  $I(C; Y|Z) + \lambda I(C; Y)$ . It is often more natural to directly set  $\lambda$  which controls the trade-off between conditional and unconditional information maximization. Since the  $I(C; Y)$  term has been added to address the coordination problem, we will in practice typically chose  $\lambda \leq 1$ .

The  $\rho$  parameter in Eq. (5) on the other hand directly controls the fuzziness of the assignments such that hard clusterings are computed in the limit of  $\rho \rightarrow 0$ . We propose to adjust  $\rho$  using a well-known continuation method called deterministic annealing [10]. This has two advantages: Conceptually, non-zero values for  $\rho$  avoid over-confidence in assigning objects to clusters and thus addresses the crucial problem of overfitting in learning from finite data. For instance, we may chose to select a value for  $\rho$  that maximizes the predictive performance on some held-out data set.

The second advantage is algorithmic in nature. The proposed alternating scheme is sensitive with respect to the choice of initial values. As a result of that, convergence to poor local optima may be a nuisance in practice, a problem that plagues many similar alternating schemes such as k-means and mixture modeling. However, a simple control strategy that starts with high entropy cluster assignments and then successively lowers the entropy of the assignments has proven to be a simple, yet effective tool in practice to improve the quality of the solutions obtained (cf. [10, 11]). In analogy of a physical system, one may think of  $\rho$  in terms of a *computational temperature*.

We thus propose the following scheme: Starting with a large enough value for  $\rho = \rho_0$ , one alternates the update equations until convergence. Then one lowers  $\rho$  according to some schedule, for instance an exponential schedule

Algorithm	mean		opt	
	$Prec_B$	$Prec_C$	$Prec_B$	$Prec_C$
CIB	0.540	0.773	0.540	0.773
CCIB1	0.566	0.970	0.566	0.970
CCIB2	0.562	0.919	0.590	0.939

**Table 1. Synthetic data results with 50 sample sets and 10 random initializations on each.**

$\rho \leftarrow b\rho$  with  $b < 1$ . The process terminates, if the chosen  $\rho$  leads to a value for  $I(C; X)$  that is close to the desired bound  $I_{\max}$  or if cluster assignments numerically reach hard assignments.

## 5 Experimental results

### 5.1 Synthetic data

We have generated test data sets with binary features  $y \in \{0, 1\}^m$  where  $m = 12$ . Two independent partitionings  $B$  and  $C$  with  $k = 2$  are embedded in the data by associating  $B$  with 8 of the features and  $C$  with 4 of the features, making  $B$  the dominant partitioning. Noise is introduced by randomly flipping each binary feature with probability  $p_{noise} = 0.1$ .

In the experiments on synthetic data, we investigate two types of background knowledge  $Z$ : the dominant classification itself  $Z = B$  (CCIB1) and the features associated with the dominant classification (CCIB2). For comparison, we also consider  $Z = B$  when the coordination term is not used (CIB). In all cases, a Bernoulli distribution is assumed for the features  $Y$ . The results are summarized in Table 1. Here 'mean' denotes the average precision of individual runs, whereas 'opt' is the precision obtained by selecting the best out of 10 solutions based on the CCIB criterion. Discovered and target clustering have been aligned using an optimal matching.

Both algorithms, CCIB1 and CCIB2, recover the target clustering with high accuracy, significantly outperforming CIB. The CCIB1 version using categorical side information slightly outperforms the GLM version CCIB2, which is due to the feature noise.

### 5.2 Face database

We consider a set of 369 face images with  $40 \times 40$  grayscale pixels and gender annotations. We performed clustering with  $k = 2$  clusters and a Gaussian noise model for the features. Initially, no background knowledge was used. All of 20 trials converged to the same clustering, suggesting that this clustering is the dominant structure in



**Figure 1. Centroids from initial clustering with no side information.**



**Figure 2. Centroids from second clustering using initial clustering as side information.**

the data set. The precision score between the discovered clustering and the gender classification was 0.5122, i.e. the overlap is close to random. Examining the centroids of each cluster in Figure 1 shows the clustering which was obtained partitions the data into face-and-shoulder views and face-only views.

We then introduce the clustering generated from the first attempt as background knowledge and perform a second attempt. The resulting precision score is substantially higher, at 0.7805. Confusion matrices for both clusterings are in Table 2. Centroids for this clustering are in Figure 2 and confirm that the dominant structure found in the previous attempt has been avoided, revealing lower-order structure that is more informative with respect to gender.

initial clustering			second clustering		
	female	male		female	male
$c_1$	140	144	$c_1$	105	1
$c_2$	45	40	$c_2$	80	183
Precision = 0.5122			Precision = 0.7805		

**Table 2. Confusion matrices for face data.**

### 5.3 Text mining

We evaluate performance on several real-world text data sets. Each may be partitioned according to either one of two independent classification schemes. Experiments are performed using either one of these classification schemes as background knowledge. An algorithm is considered successful if it finds a clustering not associated with the background knowledge, that is similar to the other classification scheme. Documents are represented by term frequency vectors that are assumed to follow a multinomial distribution. For all experiments described,  $\lambda = 0.3$  is used and  $k$  is set to the cardinality of the target categorization.

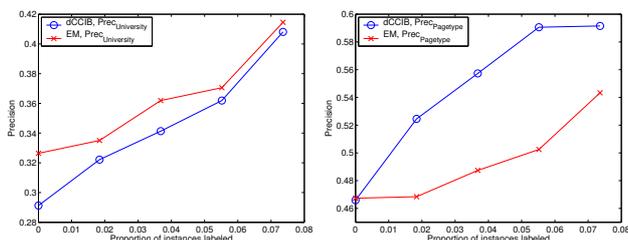
We use the CMU 4 Universities WebKB data set as described in [3] which consists of webpages collected from computer science departments and has a classification scheme based on page type: ('course', 'faculty', 'project', 'staff', 'student') as well source university: ('Cornell', 'Texas', 'Washington', 'Wisconsin'). Documents belonging to the 'misc' and 'other' categories, as well as the 'department' category which contained only 4 members, were removed, leaving 1087 pages remaining. Stopwords were removed, numbers were tokenized, and only terms appearing in more than one document were retained, leaving 5650 terms.

Additional data sets were derived from the Reuters RCV1 news corpus which contains multiple labels for each document. We first select a number of topic labels  $Topic$  and region labels  $Region$  and then sample documents from the set of documents having labels  $\{Topic \times Region\}$ . We take up to  $n$  documents from each combination of labels. For ease of evaluation, those documents which contain multiple labels from  $Topic$  or multiple labels from  $Region$  were excluded. We selected labels which would produce high numbers of eligible documents and generated the following sets: (i)  $RCV1-gmcat2x2$ :  $Topic = \{MCAT (Markets), GCAT (Government/Social)\}$  and  $Region = \{UK, INDIA\}$ : 1600 documents and 3295 terms. (ii)  $RCV1-ec5x6$ : 5 of the most frequent subcategories of the ECAT (Economics) and 6 of the most frequent country codes were chosen: 5362 documents and 4052 terms. (iii)  $RCV1-top7x9$ : ECAT (Economics), MCAT (Markets) and the 5 most frequent subcategories of the GCAT (General) topic and the 9 most frequent country codes were chosen: 4345 documents and 4178 terms. As with the WebKB set, stopwords were removed, numbers were tokenized and a term frequency cut-off was used to remove low-frequency terms.

Results of the CCIB method on the text data sets are shown in Table 3. It is interesting to note that results improve as sets with more categories are considered. In all cases, the solutions found are more similar to the target classification than the known classification although the task appears to be more difficult for the RCV1-ec5x6 set.

Data set	$Z = L1$				$Z = L2$			
	$Prec_{L2}$	$I(C, L1)$	$I(C, L2)$	time(s)	$Prec_{L1}$	$I(C, L1)$	$I(C, L2)$	time (s)
WebKB	0.2917	0.0067	0.0189	54.3	0.4735	0.2342	0.0085	61.5
RCV1-gmcat2x2	0.5516	0.0015	0.0107	12.5	0.9781	0.8548	0.0001	9.1
RCV1-ec5x6	0.1970	0.0792	0.3027	389.8	0.4189	0.2801	0.2296	519.6
RCV1-top7x9	0.2758	0.0224	0.1383	333.6	0.6076	0.4781	0.0074	183.2

**Table 3. Results on text data sets averaged over 10 initializations for  $L1 = \text{Topic/page type}$ ,  $L2 = \text{Region/university}$ .**



**Figure 3. WebKB results: adding labeled data**

## 5.4 Semi-supervised learning

Up to this point, all experiments have been performed subject to the assumption that no positive supervised information is available. We now address the question of whether semi-supervised learning may be enhanced by avoiding redundancy with background knowledge. We assume a small set of examples is labeled according to a desired classification. Performance is compared against EM augmented with a Naive Bayes model for labeled data, as described in [9]. Labeled data is incorporated into the CCIB framework by fixing the corresponding  $C(x_i)$  for each labeled instance  $x_i$ . Results for varying amounts of labeled data per class are shown in Figure 3. The results show that for the considered regime, using side information about undesired clusterings can significantly improve the classification performance. In the first case, where  $C = \text{'page type'}$ , for proportions of labeled instances greater than 0.04, the CCIB method has performance slightly less than that of EM in finding solutions similar to the target clustering. In the second case, where  $C = \text{'university'}$ , the CCIB method substantially outperforms EM over the entire range considered.

## 5.5 Parameter sensitivity

We have conducted a series of experiments to investigate the sensitivity of the results on the parameters, most crucially the relative weight  $\lambda$  ensuring cluster coordination. Due to lack of space, we can only report the main observations. On synthetic data as well as on the text data sets, we have found the procedure to be highly robust with respect to

variations of  $\lambda$  in a typical range of  $[0.2; 1.0]$ . As expected, too small values for  $\lambda$  lead to solutions that lack global coordination and for values that are too large, the  $I(C; Y)$  tends to dominate. In comparison, we have verified that the side information bottleneck of [2] is much more sensitive with respect to the corresponding tuning parameter, with the optimal range varying significantly across different data sets.

## 6 Conclusion

We have presented a general information-theoretic framework for non-redundant clustering based on the idea of maximizing conditional mutual information relative to given background knowledge. We have pointed out the cluster coordination problem and provided a way to deal with it in the Coordinated Conditional Information Bottleneck. A generic alternating optimization and special instantiations thereof have been derived and discussed, emphasizing the principled nature of our approach as well as its broad applicability. Experiments on synthetic and real-world data sets have underpinned that claim.

## Acknowledgment

D.G. has been supported by an NSF IGERT Ph.D. fellowship. Part of this work was completed while T.H. was at the Max-Planck Institute for Biological Cybernetics in Tübingen, Germany.

## Appendix

The Lagrangian of the CCIB problem is given by

$$F = I(C; Y|Z) - \rho I(C; X) + \lambda I(C; Y).$$

We rewrite mutual informations as (conditional) entropy differences and after dropping constant terms, we arrive at

$$F' = -H(Y|C, Z) - \rho H(C) + \rho H(C|X) - \lambda H(Y|C).$$

We introduce cross entropies with auxiliary parameters  $Q_C$ ,  $Q_{Y|C}$ , and  $Q_{Y|C,Z}$ , non-negative and normalized. Now we

define

$$\tilde{H}(C) = - \sum_c P_C(c) \log Q_C(c),$$

$$\tilde{H}(Y|C) = - \sum_{c,y} P_{CY}(c,y) \log Q_{Y|C}(y|c),$$

$$\tilde{H}(Y|C,Z) = - \sum_{c,y,z} P_{CYZ}(c,y,z) \log Q_{Y|CZ}(y|c,z),$$

and a new objective

$$\tilde{F} = -\tilde{H}(Y|Z,C) - \rho\tilde{H}(C) + \rho H(C|X) - \lambda\tilde{H}(Y|C).$$

The advantage of  $\tilde{F}$  is that it is concave in  $P_{C|X}$  for given  $Q$ , since  $H(C|X)$  is concave and the  $\tilde{H}$  terms are linear in  $P_{C|X}$ . Moreover,  $\tilde{F}$  is also concave in the  $Q$  parameters for given  $P_{C|X}$ , since the logarithm function is concave.

More precisely the solutions over  $Q$  can be obtained as follows: First observe that the only  $\tilde{H}(C)$  depends on  $Q_C$ , only  $\tilde{H}(Y|C)$  depends on  $Q_{Y|C}$ , and  $\tilde{H}(Y|C,Z)$  on  $Q_{Y|C,Z}$ . Differentiating  $\tilde{F}$  with respect to the  $Q$  parameters and setting to zero results in  $Q_C = P_C$ ,  $Q_{Y|C} = P_{Y|C}$ , and  $Q_{Y|C,Z} = P_{Y|C,Z}$ , as is straightforward to prove (cf. [11], Lemma 2). These correspond to maxima of the function  $\tilde{F}$  for given clustering probabilities  $P_{C|X}$ .

Finally, in order to optimize  $\tilde{F}$  explicitly over the clustering probabilities  $P_{C|X}$  one makes use of the relation

$$\frac{\partial \sum_x P_{CXYZ}(c,x,y,z)}{\partial P_{C|X}(c|x)} = P_{XYZ}(x,y,z).$$

Setting to zero and accounting for the normalization of  $P_{C|X}$  one gets

$$P(c|x) \propto Q(c) \exp \left[ \frac{\lambda}{\rho} \sum_y P(y|x) \log Q(y|c) \right] \\ \times \exp \left[ \frac{1}{\rho} \sum_z P(z|x) \sum_y P(y|x,z) \log Q(y|c,z) \right],$$

where subscripts of  $P$  and  $Q$  have been dropped for better readability.

At a stationary point (corresponding to a maximum or saddle-point of  $\tilde{F}$ ) this can again be rewritten more suggestively as a characterization of the resulting joint distribution, since there  $Q$  probabilities agree with their  $P$  counterparts. The asymptotic convergence of this scheme is a simple consequence of the fact that  $\tilde{F}$  is reduced in every update iteration and that is bounded from below. This may not correspond to a global maximum, but it will fulfill the optimality conditions over the  $P_{C|X}$  and  $Q$  subspaces separately.

## References

- [1] C. Bucila, J. Gehrke, D. Kifer, and W. White. Dualminer: A dual-pruning algorithm for itemsets with constraints. In *Proc. 8th SIGKDD Intl. Conf. on Knowledge Discovery and Data Mining*, 2002.
- [2] G. Chechik and N. Tishby. Extracting relevant structures with side information. In *Advances in Neural Information Processing Systems 15 (NIPS '02)*, 2002.
- [3] M. Craven, D. DiPasquo, D. Freitag, A. K. McCallum, T. M. Mitchell, K. Nigam, and S. Slattery. Learning to extract symbolic knowledge from the World Wide Web. In *Proc. of the 15th Conf. of the American Association for Artificial Intelligence*, pages 509–516, 1998.
- [4] N. Friedman, O. Mosenzon, N. Slonim, and N. Tishby. Multivariate information bottleneck. In *Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence*, 2001.
- [5] D. Gondek and T. Hofmann. Conditional information bottleneck clustering. In *3rd IEEE International Conference on Data Mining, Workshop on Clustering Large Data Sets*, 2003.
- [6] D. Klein, S. D. Kamvar, and C. D. Manning. From instance-level constraints to space-level constraints: Making the most of prior knowledge in data clustering. In *Proceedings of the Nineteenth International Conference on Machine Learning*, 2002.
- [7] P. McCullagh and J. Nelder. *Generalized Linear Models*. Chapman & Hall, London, U.K., 1989.
- [8] R. T. Ng, L. V. Lakshmanan, J. Han, and A. Pang. Exploratory mining and pruning optimizations of constrained association rule. In *Proceedings of ACM SIGMOND*, pages 13–24, 1998.
- [9] K. Nigam, A. K. McCallum, S. Thrun, and T. M. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2/3):103–134, 2000.
- [10] K. Rose. Deterministic annealing for clustering, compression, classification, regression, and related optimization problems, 1998.
- [11] N. Tishby, F. C. Pereira, and W. Bialek. The information bottleneck method. In *Proceedings of the 37-th Annual Allerton Conference on Communication, Control and Computing*, pages 368–377, 1999.
- [12] A. Tung, R. Ng, J. Han, and L. Lakshmanan. Constraint-based clustering in large databases. In *Proceedings 2001 International Conference on Database Theory*, pages 405–419, 2001.
- [13] K. Wagstaff and C. Cardie. Clustering with instance-level constraints. In *Proc. of the Seventeenth International Conference on Machine Learning*, pages 1103–1110, 2000.
- [14] E. P. Xing, A. Y. Ng, M. I. Jordan, and S. Russell. Distance metric learning, with application to clustering with side-information. In *Advances in Neural Information Processing Systems*, 2003.
- [15] S. Zhong and J. Ghosh. Model-based clustering with soft balancing. In *Proc. 3rd IEEE Int. Conf. Data Mining*, pages 459–466, 2003.