A Hierarchical Information Theoretic Technique for the Discovery of Non Linear Alternative Clusterings

Xuan Hong Dang and James Bailey
Department of Computer Science and Software Engineering
The University of Melbourne, Australia
{xdang,jbailey}@csse.unimelb.edu.au

ABSTRACT

Discovery of alternative clusterings is an important method for exploring complex datasets. It provides the capability for the user to view clustering behaviour from different perspectives and thus explore new hypotheses. However, current algorithms for alternative clustering have focused mainly on linear scenarios and may not perform as desired for datasets containing clusters with non linear shapes. Our goal in this paper is to address this challenge of non linearity.[1] In particular, we propose a novel algorithm to uncover an alternative clustering that is distinctively different from an existing, reference clustering. Our technique is information theory based and aims to ensure alternative clustering quality by maximizing the mutual information between clustering labels and data observations, whilst at the same time ensuring alternative clustering distinctiveness by minimizing the information sharing between the two clusterings. We perform experiments to assess our method against a large range of alternative clustering algorithms in the literature. We show our technique's performance is generally better for non-linear scenarios and furthermore, is highly competitive even for simpler, linear scenarios.

1. INTRODUCTION

Data clustering aims at discovering novel patterns and structures from data. Its objective is to categorize similar data instances into the same classes (or clusters). However, while it may be reasonable to refer to a "best" model in supervised classification, it is less useful to make statements about a single, best clustering [5]. When exploring complex data, different clusterings can exist and they may each be reasonable. For example when analyzing a document dataset, one may find that it is possible to categorize according to either topics or writing styles; or when clustering a gene dataset, it is found that grouping genes based on their functions or structures is equally useful [6]. This challenge has recently stimulated the growing research area of alternative clustering, where the goal is to generate different, yet distinct clusterings or groupings of a given dataset.

[1]By non-linearity, we mean that either clusters may have unusual and non-Gaussian shapes or the border between them may not be linearly separable.

Several algorithms have been developed for the task of alternative clustering. Given an input, the reference clustering, the task is to generate another clustering, which is dissimilar from the reference one, yet it is still plausible (i.e. has high quality). It is this dual objective of achieving both dissimilarity and quality that makes the task challenging.

In this paper, we explore another aspect of the alternative clustering problem. We focus on the scenario where the dataset may contain non linear shapes or groupings. We show that current algorithms for alternative clustering tend to underperform in this scenario. This motivates us to develop a new algorithm, called NACI (an acronym stands for Non-linear Alternative Clustering with Information theory). NACI is a hierarchical technique which uses information theoretic methods to optimse the dual objective functions of both quality and dissimilarity.

In particular, given a reference clustering, the NACI algorithm aims to discover an alternative clustering for which i) the mutual information between its cluster labels and data observations is maximized, whereas at the same time ii) the mutual information between the alternative clustering and the reference clustering is minimized. Objective i) helps to reduce the uncertainty within each cluster of the alternative clustering, by ensuring there is a strong (probabilistic) relationship between the cluster labels and the data instances. We later motivate this clustering objective through the use of Fano's inequality. On the other hand, objective ii) helps to ensure that the alternative clustering is independent (different) from the reference clustering.

The principal technical contribution of our work is the formulation of a well founded alternative clustering objective function, that is purely information theoretic. The advantage of using an information theoretic approach is that it can adapt well to the presence of non linearities. However, the technical development is not straightforward, requiring the use of Parzen windows for probability density estimations, as well as approximatons based on quadratic mutual information.

Through an experimental analysis, we show that NACI performs particularly strongly when finding alternative clusterings for non-linear datasets, improving over the state of the art. Furthermore, even for simpler, more linear datasets, NACI is able to discover desirable alternative clusterings, possessing high quality and high dissimilarity to the reference clustering.

2. RELATED WORK

There are several works related to our research. The closest are those developed in [13, 2, 8, 10], which exploit various forms of negative information toward the desired clustering (as opposed to those using prior knowledge to improve clustering results [17, 22, 4]). In [13], the authors proposed a conditional information bottleneck (CIB) method, which treats class labels of a given clustering as negative information in seeking an alternative clustering. The new data partition is found by maximizing the information sharing between the cluster labels and the data features (describing for data objects), but conditioned on the given reference clustering. This method, though similar to our work in that both address the problem from an information theory viewpoint, is rather different in two important aspects. First, our approach makes no assumption regarding the data density distribution, whilst CIB requires the availability of the joint distribution information between cluster labels and the features, which is known as being hard to formalize[10]. Second, while our algorithm directly minimizes the mutual information between two clusterings to ensure their independence, the CIB approach only conditions on the reference clustering in the process of encoding properties of the data features into the new clustering. In other words, it uses mutual information in a completely different way to our approach. Another approach, which exploits the effectiveness of pairwise constraints for data clustering [22, 3], is the COALA technique [2]. From the reference clustering, COALA generates a set of cannot-link constraints between pairs of data samples and attempts to find a different clustering that satisfies as many as possible of these constraints. On the other hand, a line of work developed in [8, 10] takes a rather different approach to alternative clustering by relying on the notion of orthogonality. In [8], the authors develop two techniques to find an alternative clustering using orthogonal projections. In the first one, data is projected onto a space that is orthogonal to the space spanned by the set of mean vectors in the given clustering, while in the second technique, such a representative vectors are replaced by the feature space. An alternative clustering is then found by simply applying a clustering algorithm on this new transformed data. A similar approach is developed in [10] by which the transformation is applied on the distance matrix learnt from the provided clustering. In comparison, this work has an advantage compared to[8], since it avoids problems when the data dimension is smaller than the number of clusters (e.g., spatial datasets).

Another series of works addressing the alternative clustering problems are those developed in [15] and [9]. Unlike the work above, these methods attempt to seek two alternative clusterings at the same time. In the first work, two algorithms named Dec-kmeans and ConvEM are developed which attempt to derive two sets of mean vectors that are pairwise orthogonal, whereas in the second work [9], an algorithm, known as CAMI seeks two clusterings that share minimal mutual information. A clear distinction between these algorithms and ours in this paper is that our algorithm is not limited to spherically shaped clusterings, i.e. it is able to seek alternative clusterings for non linear datasets. We provide experimental comparisons with all the above algorithms in Section 5 of the paper.

3. PRELIMINARIES

3.1 Entropy and Mutual Information

In information theory, the quantity entropy plays a central role and is a measure of the uncertainty of a random variable. Mathematically, let X be a continuous random variable characterized by the probability distribution $p(x)$, the Shannon entropy of X is:

$$H(X) = -\int p(x)\log p(x)dx \tag{1}$$

When a certain variable is known and another is not, the remaining uncertainty is measured by the conditional entropy:

$$H(Y|X) = -\int\int p(x,y)\log p(y|x)dxdy \tag{2}$$

A related concept to the entropy is the Kullback-Leibler divergence. It is a measure of the distance between two distributions $p(x)$ and $q(x)$ and is defined by:

$$KL(p\|q) = \int p(x)\log\frac{p(x)}{q(x)}\,dx \tag{3}$$

Mutual information, which is of importance in our work, turns out to be a special case of the KL divergence. It measures the information shared between two objects or in other words, it accounts for the amount of information that one random variable contains about another variable. In specific form, let X and Y be two random variables with a joint probability density function $p(x,y)$ and marginal probability density functions $p(x)$ and $p(y)$, the mutual information $I(X;Y)$ is the KL distance between the joint distribution and the product of two marginal distributions $p(x)$ and $p(y)$:

$$\begin{aligned} I(X;Y) &= \iint p(x,y)\log\frac{p(x,y)}{p(x)p(y)}\,dx\,dy \\ &= KL(p(x,y)\|p(x)p(y)) \end{aligned} \tag{4}$$

which is obviously symmetric and non-negative. Importantly, two random variables have zero mutual information if and only if they are statistically independent.

3.2 Problem Definition

Given the above definitions, an intuitive problem statement is as follows:

Definition 1. Given a dataset $X = \{x_1, x_2, ..., x_n\}$ where each x_i is in d-dimensional space and for which there is an existing reference clustering C^- (found by any clustering algorithm), find a new alternative clustering C^+ from D, such that C^+ is as independent (i.e., different) from C^- as possible. The independence between two clusterings can be quantified by the information sharing between them.

The number of clusters within each clustering C^+ and C^- may be different. In our work, for easy of presentation, we assume that they are the same and use k to denote this number. However our methods can be easily adapted to the more general case.

4. CLUSTERING METHOD

4.1 Criterion Function Optimization

For any learning algorithm, the learning process should ultimately transfer the information carried in the data samples into the system's parameter [20]. It is therefore natural to find an objective function that directly manipulates information. Since mutual information is an essential tool to quantify the statistical relationship between any two random variables, it is intuitive to create a clustering cost function

that relies on it. In the particular case of cluster analysis, we would like to find a clustering solution that has a strong probabilistic relationship with the data observation X. This implies that the clustering label variable C^+ has little uncertainty given the data observation X or in other words, the observations contain much of the information about the clustering label C^+ and we can infer the value of C^+ from the observations with small error. Theoretically, this objective can be justified by the well-known Fano's theorem from information theory, which provides a lower bound for the probability of error when guessing discrete values of a random variable C^+ from the random variable X. More specifically, let $H(C^+|X)$ be the conditional entropy of C^+ given X, then Fano's inequality states that:

$$Pr(c^+ \neq \widehat{c^+}) \geq \frac{H(C^+|X) - 1}{\log(|C^+|)} = \frac{H(C^+) - I(C^+; X)}{\log(|C^+|)} \quad (5)$$

in which c^+ and $\widehat{c^+}$ are respectively the true and guessed cluster labels of C^+, after observing a sample of X. Thus, the lower bound on error probability is minimized when the mutual information between the cluster label C^+ and the data observation X is maximized. In this respect, it is possible to say that mutual information corresponds to the amount by which knowledge provided by the data observation X decreases the uncertainty about the cluster.

Therefore, combining the use of information theory to ensure cluster quality, with our objective of minimizing the mutual information between two clustering solutions, it is possible to form an alternative clustering objective function as follows:

$$C^+ = \arg\max_{C^+} \{I(C^+; X) - \eta I(C^+; C^-)\} \quad (6)$$

The parameter η regulates the relative importance of each of the clustering objectives.

4.2 Alternative Clustering with Quadratic Mutual Information

When a data instance is assigned to one of several clusters, it incurs a variation on the mutual information cost. Optimizing this variation could be used as an evaluation function for clustering. In working toward optimizing the objective function in Eq.(6), one possible way is to employ an agglomerative hierarchical clustering technique. This type of algorithm typically begins by placing each data sample into its own cluster and then successively merges pairs of clusters until all samples are grouped into a single cluster. However, unlike most of the existing agglomerative techniques, where the merging between two clusters is decided based on their similarity (e.g., Euclidean distance), in our work, two clusters in C^+ are merged if such a combination makes the global mutual information $I(C^+; X)$ maximally increased, while at the same time it minimizes the amount of $I(C^+; C^-)$. This evaluation requires the estimation of the mutual information at each merging step of the algorithm. In the following, we present an approach that can help to estimate mutual information directly from data, while making no assumption about the data density distribution.

As mentioned in Section 3.1, mutual information defined in Shannon's entropy can be viewed as the KL divergence between the joint distribution and the product of the two marginal distributions of two variables. However, computing this divergence is not an easy task in practice since it requires the availability of all variables' probability density functions. Furthermore, the numerical integration of these functions also leads to very high computational complexity.

Fortunately, notice that our clustering objective is to optimize the mutual information, rather than computing it exactly. It has been shown in [16, 20] that as long as a learning process does not require to compute an exact value of the mutual information, but rather to maximize or minimize it, then other practical divergences can be used. Importantly, the extrema of these divergences are also coincident with those of the KL divergence and therefore, the objectives of optimization are not compromised. One of such divergence is presented in [16]:

$$D(p||q) = \frac{1}{\alpha(\alpha-1)} \sum_{i=1}^{n} (p^\alpha(x_i) - \alpha \frac{p(x_i)}{q^{1-\alpha}(x_i)} + (\alpha - 1) q^\alpha(x_i))$$

where $\alpha \neq 0, 1$.

Based on this, a quadratic form of mutual information can be derived by selecting $\alpha = 2$ and extending the equation to continuous densities (the first constant term can be omitted):

$$I_{R_2}(X; Y) = \int \int (p(x, y) - p(x)p(y))^2 \, dxdy \quad (7)$$

It is easy to prove that all essential properties of the divergence are preserved, i.e., I_{R_2} is always non-negative, symmetric and equal to 0 if and only if $p(x, y) = p(x)p(y)$.

When applying this quadratic form of mutual information to our alternative clustering problem, it is possible to derive the information sharing between two discrete variables C^+ and C^- as follows:

$$\begin{aligned} I_{R_2}(C^+; C^-) &= \sum_{c_i^+} \sum_{c_j^-} \left(p(c_i^+, c_j^-) - p(c_i^+)p(c_j^-)\right)^2 \\ &= \sum_{c_i^+} \sum_{c_j^-} p(c_i^+, c_j^-)^2 + \sum_{c_i^+} \sum_{c_j^-} p(c_i^+)^2 p(c_j^-)^2 \\ &\quad - 2 \sum_{c_i^+} \sum_{c_j^-} p(c_i^+, c_j^-) p(c_i^+) p(c_j^-) \end{aligned} \quad (8)$$

and the quadratic mutual information between the continuous variable X and the discrete variable C^+:

$$\begin{aligned} I_{R_2}(C^+; X) &= \sum_{c_i^+} \int_x \left(p(c_i^+, x) - p(c_i^+)p(x)\right)^2 dx \\ &= \sum_{c_i^+} \int_x p(c_i^+, x)^2 dx + \sum_{c_i^+} \int_x p(c_i^+)^2 p(x)^2 dx \\ &\quad - 2 \sum_{c_i^+} \int_x p(c_i^+, x) p(c_i^+) p(x) dx \end{aligned} \quad (9)$$

where the prior probabilities $p(c_i^+)$ and $p(c_j^-)$ are estimated by the number of data samples in each cluster (over n), i.e., respectively n_i/n and n_j/n. Similarly, the joint probability between c_i^+ and c_j^- is estimated by the number of data samples belonging to both c_i^+ and c_j^-, i.e., n_{ij}/n.

Notice that computing the mutual information in Eq.(9) requires the estimation of variables' probability density function. Nonetheless, an appealing property of the quadratic mutual information is that it is possible to combine it with the Parzen window method, an effective non-parametric density estimation technique, to simplify the computation. This involves placing a kernel function at each data sample. The

density is accordingly evaluated by the sum of kernels. When using a Gaussian function for the kernel, it follows that:

$$p(x) = \frac{1}{n}\sum_{i=1}^{n} G(x - x_i, \sigma^2) \tag{10}$$

where $$G(x - x_i, \sigma^2) = \frac{1}{(2\pi\sigma)^{d/2}} \exp\left\{-\frac{||x - x_i||^2}{2\sigma^2}\right\}$$

is the Gaussian in a d-dimensional space. An important property with this kernel is that the convolution of two Gaussians remains a Gaussian function:

$$\int_x G(x - x_i, \sigma^2) G(x - x_j, \sigma^2) dx = G(x_i - x_j, 2\sigma^2) \tag{11}$$

This equation can be interpreted as the *information potential* between x_i and x_j. Therefore, when combining the non-parametric density estimation method with the quadratic mutual information, the computational complexity can be greatly reduced. Specifically, according to the Parzen window method:

$$p(x|c_i^+) = \frac{1}{n_i}\sum_{\ell=1}^{n_i} G(x - x_\ell, \sigma^2) \tag{12}$$

where n_i is the number of data points belonging to cluster c_i^+. Followed by the Bayesian theorem, the joint probability between x and c_i^+ is:

$$p(c_i^+, x) = p(c_i^+)p(x|c_i^+) = \frac{1}{n}\sum_{\ell=1}^{n_i} G(x - x_\ell, \sigma^2) \tag{13}$$

The quadratic mutual information in Eq.(9) thus can be decomposed into three terms:

$$T_{in} = \sum_{c_i^+}\int_x p(c_i^+, x)^2 dx = \frac{1}{n^2}\sum_{c_i^+}\sum_{\ell=1}^{n_i}\sum_{m=1}^{n_i} G(x_\ell - x_m, 2\sigma^2)$$

T_{in} is interpreted as the sum of all information potentials within each of clusters.

$$\begin{aligned} T_{all} &= \sum_{c_i^+}\int_x p(c_i^+)^2 p(x)^2 dx \\ &= \left(\sum_{c_i^+}\left(\frac{n_i}{n}\right)^2\right)\left(\frac{1}{n^2}\sum_{\ell=1}^{n}\sum_{m=1}^{n} G(x_\ell - x_m, 2\sigma^2)\right) \end{aligned}$$

T_{all} can be described as the sum of all information potentials, regardless of their cluster and weighted by the cluster prior. And

$$\begin{aligned} T_{btw} &= \sum_{c_i^+}\int_x p(c_i^+, x) p(c_i^+) p(x) dx \\ &= \frac{1}{n^2}\sum_{c_i^+}\frac{n_i}{n}\left(\sum_{\ell=1}^{n_i}\sum_{m=1}^{n} G(x_\ell - x_m, 2\sigma^2)\right) \end{aligned}$$

T_{btw} is seen as the interaction of data points within each cluster with respect to all data points, weighted by the cluster prior.

Given the computations above, it is clear that the local interaction, as defined by the kernel in Eq.(11), between any two data instances needs to be computed. Therefore, a matrix $\mathcal{G}$ having size of $n \times n$ is generated in which at row i column j, the information potential term $G(x_i - x_j, 2\sigma^2)$ is computed. Notice that, in practice, only half the number of these interactions need to be evaluated due to the symmetry.

In addition to the $\mathcal{G}$ matrix, two other matrices D_{in} and D_{btw}, which respectively account for the variation in $I_{R_2}(C^+; X)$ and $I_{R_2}(C^+; C^-)$ incurred by merging any pairs of clusters in C^+, are utilized. Indeed, the combination of these two matrices acts as the similarity matrix in a classical agglomerative clustering technique. Notice that, different from $\mathcal{G}$ whose elements do not change during the clustering process, elements of these two matrices are updated regularly and their size is reduced one upon each merging step of the algorithm. These two matrices are initialized by computing the variation in mutual information when grouping any pair of data samples.

At each iterative step of the agglomerative clustering technique, the maximum element from the combined matrix is selected:

$$(\alpha, \beta) = \arg\max_{i,j}\{D_{in} - \eta D_{btw}\} \tag{14}$$

This means that grouping cluster c_β^+ to cluster c_α^+ leads to the maximum variation in our global objective function. Upon this merging, elements located at column and row β will be removed from D_{in} and D_{btw} and it is necessary to update elements in column and row α of these two matrices. That means a new variation on the mutual information is computed if the updated cluster c_α^+ is combined with any of the rest clusters in C^+.

For simplicity, we re-use the cluster index notations and denote c_γ^+ for the new cluster by merging c_α^+ with any cluster c_β^+ in C^+, then it is clear that:

$$\begin{aligned} p(c_\gamma^+) &= p(c_\alpha^+) + p(c_\beta^+) \\ p(c_\gamma^+, c_j^-) &= p(c_\alpha^+, c_j^-) + p(c_\beta^+, c_j^-) \text{ and} \\ p(c_\gamma^+, x) &= p(c_\alpha^+, x) + p(c_\beta^+, x) \end{aligned}$$

Upon these, the variation in the mutual information can be simply computed. Specifically, the variation of mutual information with respect to each cluster c_j^- in C^- is:

$$\begin{aligned} &\Delta I_{R_2}(C^+; c_j^-) \\ &= \left(p(c_\gamma^+, c_j^-) - p(c_\gamma^+)p(c_j^-)\right)^2 - \\ &\quad \left(p(c_\alpha^+, c_j^-) - p(c_\alpha^+)p(c_j^-)\right)^2 - \left(p(c_\beta^+, c_j^-) - p(c_\beta^+)p(c_j^-)\right)^2 \\ &= 2\left(p(c_\alpha^+, c_j^-)p(c_\beta^+, c_j^-) + p(c_\alpha^+)p(c_\beta^+)p(c_j^-)^2 - \right. \\ &\quad \left. p(c_\alpha^+)p(c_\beta^+, c_j^-)p(c_j^-) - p(c_\beta^+)p(c_\alpha^+, c_j^-)p(c_j^-)\right) \end{aligned}$$

The entire variation on $I_{R_2}(C^+; C^-)$ is therefore:

$$\Delta I_{R_2}(C^+; C^-) = \sum_j \Delta I_{R_2}(C^+; c_j^-) \tag{15}$$

Analogously, the variation on $I_{R_2}(C^+; X)$ can be derived by interchanging summations and integrations:

$$\begin{aligned} &\Delta I_{R_2}(C^+; X) \\ &= 2\left(\int p(c_\alpha^+, x)p(c_\beta^+, x)dx + \int p(c_\alpha^+)p(c_\beta^+)p(x)^2 dx - \right. \\ &\quad \left. \int p(c_\alpha^+)p(c_\beta^+, x)p(x)dx - \int p(c_\beta^+)p(c_\alpha^+, x)p(x)dx\right) \end{aligned} \tag{16}$$

in which:

$$\int p(c_\alpha^+, x)p(c_\beta^+, x)dx = \frac{1}{n^2}\sum_{k=1}^{n_\alpha}\sum_{\ell=1}^{n_\beta} G(x_k - x_\ell, 2\sigma^2)$$

$$\int p(c_\alpha^+)p(c_\beta^+)p(x)^2dx = \frac{n_\alpha n_\beta}{n^4}\sum_{k=1}^{n}\sum_{\ell=1}^{n} G(x_k - x_\ell, 2\sigma^2)$$

and

$$\int p(c_\alpha^+)p(c_\beta^+, x)p(x)dx = \frac{n_\alpha}{n^3}\sum_{k=\overline{n_\alpha}1}^{n_\beta}\sum_{\ell=\overline{n}1}^{n} G(x_k - x_\ell, 2\sigma^2)$$
$$\int p(c_\beta^+)p(c_\alpha^+, x)p(x)dx = \frac{n_\beta}{n^3}\sum_{k=1}^{n}\sum_{\ell=1}^{n} G(x_k - x_\ell, 2\sigma^2)$$

Notice that all these summations can be easily obtained from the $\mathcal{G}$ matrix.

Finally, it can be observed from the above that the variation in the two mutual information values may be in different units. This is because $I_{R_2}(C^+; X)$ is calculated between a clustering solution and the data (i.e. a discrete and continuous variables), whereas $I_{R_2}(C^+; C^-)$ is computed between two clustering solutions (i.e. two discrete variables). Therefore, in order to avoid this difficulty, the variation with respect to each mutual information is normalized by dividing it for the corresponding quadratic mutual information. This also makes it easier when regularizing the trade-off factor η between these two information's quantities.

4.3 Kernel Parameter Setting

One of the key advantages in our algorithm is that it makes no prior assumption about the probability density functions and these functions are approximated directly from data using the non-parametric method. However, it should be noticed that the success of this approach is dependent on an appropriate selection for the kernel parameter, that is the standard deviation σ. It is shown in [19] that if the kernel width σ is annealed toward zero at a sufficiently low rate as n tends to infinity, then the Parzen window density estimator will be asymptotically unbiased and consistent. However, for most practical applications where data are finite, the kernel size should be selected in such a way that it balances out the bias and variance, which essentially being derived from the optimization of the mean integrated squared error between an estimator $\widehat{p}(x)$ and the true density $p(x)$: $MISE\Big\{\widehat{p}(x)\Big\} = \int_x E\Big\{[\widehat{p}(x) - p(x)]^2\Big\}dx$.

In our work, we choose $\sigma = \widehat{\sigma}\left(\frac{4}{n(2d+1)}\right)^{\frac{1}{d+4}}$, [2] which indeed is found by applying the least square cross-validation and the normal reference rule [23] to minimize the generalization error above. It is also further experimentally observed that by setting σ at this value, the interaction between samples that are far distant is still considered, while the interaction between close data samples remains emphasized. As shown in our experimental section, this σ selection often results in good clusterings for most of the datasets examined.

4.4 Algorithm Complexity

The proposed algorithm requires to compute the matrix $\mathcal{G}$ of information potentials between any pair of data samples. The complexity of this operation requires $O(dn^2)$ where d is the cost of calculating the interaction according to Eq.(11). The calculation of mutual information's variation when merging any two data samples takes $O(n^2)$. At each merging step, the maximum value is selected from the combination of two matrices D_{in} and D_{btw}. By using the priority queue data structure [7] that supports the search and deletion in $O(\log n)$ from this matrix, this step thus takes $O(n \log n)$. The calculation of updating information variation according to Eqs.(15) and (16) takes constant time given the availability of information potential matrix $\mathcal{G}$. Since there are $(n-1)$ merging steps for the agglomerative clustering, the computation is $O(n^2 \log n)$. The overall complexity of our proposed algorithm is therefore $O(n^2 \log n + dn^2)$. If considering d is usually smaller than $\log n$, it is possible to say that the final complexity is $O(n^2 \log n)$, which is the same as that of a conventional hierarchial clustering using group-average similarity.

[2] where $\widehat{\sigma} = \frac{1}{d}\sum_i \sigma_i$ and σ_i's are the diagonal elements of the sample covariance matrix

5. EXPERIMENTAL RESULTS

In this section, we provide experimental results on both synthetic and real-world benchmark datasets. The proposed NACI algorithm is compared against eight methods, including five semi-supervised alternative clustering algorithms: the CIB method [13], COALA [2], two methods from [8] denoted by Algo1 and Algo2, and the ADFT algorithm[10]; and three unsupervised alternative clustering algorithms: the Dec-kmeans, ConvEM from [15], and our previous algorithm CAMI [9]. Otherwise indicated, we set $\eta = 0.2$ as the default value for NACI. For the CIB method, we implement the iterative version [12, 13] and its outputs are post-processed by assigning each data point to the cluster to which it has the highest probability. For ADFT, we implement the gradient descent method integrated with the iterative projection technique (in learning the full family of the Mahalanobis distance matrix) [24, 25]. We also use the EM technique as the background clustering technique for the approaches developed in [8, 10]. For the Dec-kmeans, ConvEM and CAMI, we follow the heuristic method described in their work to set the trade-off factor between the clustering quality and dissimilarity. Since both NACI and COALA are developed based on agglomerative hierarchical clusterings, which are not sensitive to the initial parameters but possibly to the data instances' order. Therefore, when running them, we randomly swap the order amongst data samples. We run each algorithm 20 times and report the average results.

5.1 Clustering Evaluation

We evaluate the clustering results based on both clustering dissimilarity and clustering quality measures. For measuring dissimilarity between two clusterings, we report the values of two different measures. The first and also the most popular one is the normalized mutual information[18, 21, 14, 11]: $NMI(C^+; C^-) = I(C^+; C^-)/(H(C^+)H(C^-))$, where $I(C^+; C^-) = \sum_{i=1}^{M^+}\sum_{j=1}^{M^-}\frac{n_{ij}}{n}\log\left(\frac{n.n_{ij}}{n_i^+.n_j^-}\right)$ with n_{ij} denoting the number of shared instances between clusters $c_i^+ \in C^+$ and $c_j^- \in C^-$. The second is the Jaccard index (JI) [2, 10]: $J(C^+; C^-) = \frac{n_{11}}{n_{11}+n_{01}+n_{10}}$ in which n_{11} is the number of pairs of samples in the same cluster for both C^+ and C^-, n_{01} and n_{10} are the number of samples' pairs belonging to the same cluster in one solution, but not in the other.

For measuring clustering quality we divide into two cases: if true class labels are known, the agreement between clustering results and the correct labels is calculated by the F-

	NMI	JI	F	NMI	JI	F	NMI	JI	F	NMI	JI	F
Methods	Syn1			Syn2			Syn3			Syn4		
Algo1	0.25	0.41	0.83	0.42	0.41	0.62	0.2	0.42	0.78	0.28	0.34	0.63
Algo2	0.26	0.43	0.81	NA	NA	NA	0.21	0.44	0.76	0.28	0.34	0.63
ADFT	0.12	0.39	0.92	0.62	0.61	0.57	0.14	0.36	0.86	0.30	0.36	0.62
COALA	0	0.33	1	0.38	0.35	0.53	0.18	0.41	0.79	0.25	0.37	0.58
CIB	0.12	0.4	0.91	0.41	0.39	0.72	0.24	0.46	0.69	0.37	0.39	0.53
Dec-kmeans	0.12	0.39	0.93	0.39	0.34	0.68	0.12	0.35	0.87	0.22	0.34	0.62
ConvEM	0.12	0.4	0.92	0.4	0.36	0.66	0.12	0.36	0.85	0.22	0.35	0.62
CAMI	0.1	0.38	0.95	0.37	0.33	0.92	0.11	0.35	0.88	0.21	0.34	0.63
NACI	0	0.33	1	0.35	0.32	1	0	0.33	1	0	0.33	1

Table 1: The clustering performances of all algorithms on four synthetic datasets. Results are not available for Algo2 on Syn2 since its transformation is undefined when number of clusters greater than data dimensionality.

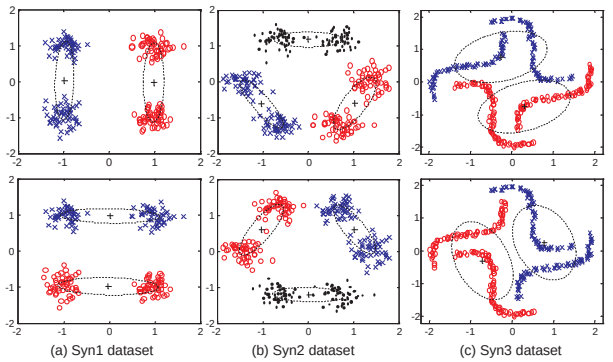


Figure 1: Alternative Clustering returned by NACI on Syn1, 2 and 3 datasets

measure: $F = 2P \times R/(P+R)$, in which P and R are respectively the precision and recall. If true class labels are not known, we use the Dunn Index, similar to [2, 10]: $DI(C) = \frac{\min_{i \neq j}\{\ \delta(c_i,c_j)\}}{\max_{1 \leq \ell \leq k}\{\triangle(c_\ell)\}}$ where C is a clustering, δ: $C \times C \rightarrow \mathbb{R}_0^+$ is the cluster-to-cluster distance and $\triangle$: $C \rightarrow \mathbb{R}_0^+$ is the cluster diameter measure.

Note that for the NMI and JI measures, *a smaller value is desirable*, indicating higher dissimilarity between clusterings, while for the F-measure and Dunn Index, *a larger value is desirable*, indicating a better clustering quality. Also, since methods like Dec-kmeans, ConvEM and CAMI do not require existing clusterings to be provided and instead seek two alternative clusterings at the same time, we try to achieve a fair comparison with them by reporting the higher values of F-measure in the case true labels are available, and averaging the Dunn Index when the class labels are not known.

5.2 Synthetic Datasets

Four synthetic datasets are used to evaluate the performance of our proposed clustering technique against other algorithms. For the first dataset Syn1, we use a popular one from [2, 8, 10, 15], which consists of 4 Gaussian sub-classes. Each Gaussian contains 200 points in 2-dimensional data space. The goal of using this dataset, when setting $k = 2$, is to test whether our algorithm can discover an alternative clustering that is orthogonal to the existing one. For the second synthetic dataset Syn2, we use a more complicated one in which 6 Gaussian sub-classes are located in a ring shape. Different from Syn1, this dataset consists two equally important clusterings (with $k = 3$) that are not orthogonal and it is unable to find them by simply projecting the data on either of the subspaces. We generate the third and fourth datasets by replacing a non-Gaussian shape for each subclass in the first synthetic dataset. By using these datasets, we aim to test the ability of our algorithm in uncovering non-linear clusterings.

Figures 1(a), (b) and (c) show clustering solutions respectively for the first three synthetic datasets. Clusterings in the top graphs of each figure are provided as pre-existing reference solutions to each semi-supervised algorithm and in the bottom graphs, we demonstrate the alternative clusterings returned by the NACI, which exactly are the second important clusterings included in each dataset. We compare the performances of all algorithms via the results summarized in Table 1. As can be seen from the table, like other alternative clustering methods, our proposed algorithm can easily identify the orthogonal clustering for the first simple synthetic dataset. Its performance on the second dataset is also accurate, although the two alternative clusterings solutions have been deliberately designed to be non-orthogonality. Notice that, unlike Syn1, there are no dominant features that fully support any of the clusterings in this dataset. Therefore, the methods developed based on orthogonal space transformation [8, 10] or orthogonal clusters' mean projection [15] are usually less successful in discovering the second alternative clustering. Only CAMI and NACI are able to achieve high accurate results since they both adopt the approach of mutual information minimization. However, different from CAMI, which slightly suffers from the problem of initial parameters sensitivity, NACI can completely avoid this since it is designed based on the hierarchical clustering technique. In addition, though we tested the strategy at which data samples were randomly swapped before each running, it was still found that NACI's performance was more consistent across all trials.[3]

We provide the clustering output of NACI for the Syn3 dataset in Figure 1(c) and all algorithms for the Syn4 in Figure 2 (the clustering outputs of Algo2 and ConvEM are omitted since they are very much similar to those of Algo1 and Dec-kmeans, respectively). For these two synthetic datasets, which aim to test the algorithms' ability in uncovering nonlinear clusterings, it is clear that the methods like Algo1 and Algo2, or Dec-kmeans, ConvEM and CAMI are unable to identify the correct alternative clustering since their core algorithms are tied to a particular spherical clustering technique (e.g., k-means, EM). Moreover, the data pro-

[3]We refer the reader to our previous work [9] for detailed justification on the clustering outputs of the other algorithms.

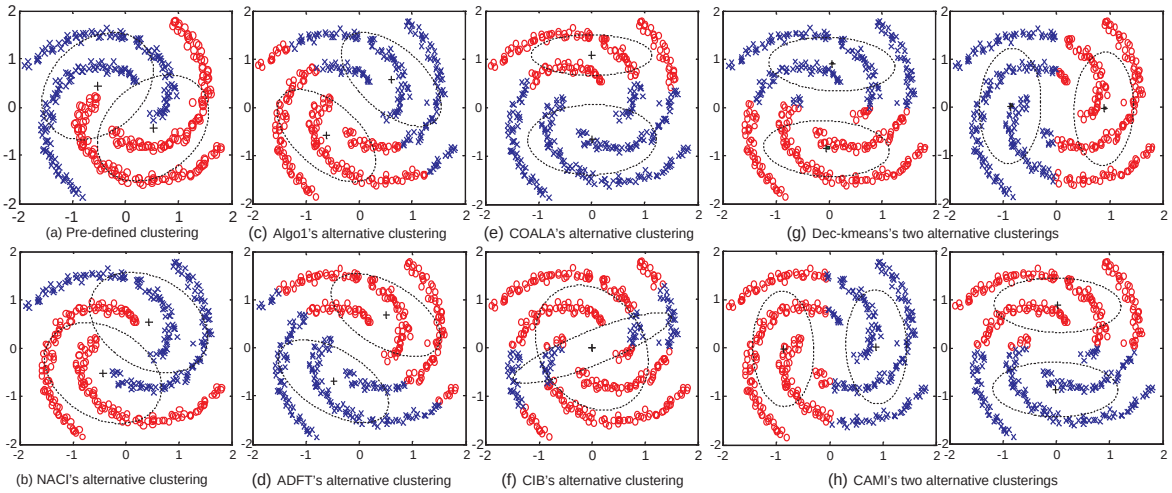


Figure 2: Alternative clusterings returned by all algorithms on Syn4 dataset. The first graph presents for the pre-defined clustering.

jection on a space orthogonal to the set of mean vectors used in Algo1 or analogically to the feature subspace used in Algo2 does not help them uncover the second alternative clustering, since the data is distorted and seems to be more overlapping by these transformations; whereas the clustering approaches used in ConvEM and CAMI only ensure a low value of decorrelation between two alternative clusterings, but cannot guarantee that accurate clusterings can be found for these nonlinear clustering shapes due to the EM technique. Similarly, by inverting the stretcher matrix, ADFT is also unable to deduce the hidden clustering structure since notice that elements in this diagonal matrix actually are the stretching factors along each dimension. Thus, varying any of the elements (corresponding to dimensions) does not make the alternative clusterings easier to be discovered, especially for the Syn4 dataset where both cluster shapes and clustering boundary are strongly nonlinear.

COALA, although it is essentially an average linkage algorithm, its merging criteria is still based on the Euclidean distance between clusters. This is in contrast to our algorithm, which exploits the similarity between clusters based on the information embedded in the data. Furthermore, the dissimilarity between two clusterings is optimized by NACI via the global quantity of the clusterings' mutual information, rather than the *local* pairwise cannot-link constraints between any two data points as used in COALA.

The CIB method also approaches the problem based on information theory. However, different from our approach in which we directly minimize the mutual information between the alternative and the existing clustering, it can be noticed that CIB only conditions on this provided clustering in its process of maximizing the information between the new clustering and the set of data features. Its resultant alternative clustering therefore look somewhat unnatural (as observed in Figure 2). Moreover, our approach in utilizing the mutual information is also rather different from the CIB's approach. In particular, while we approach the problem by making no assumption regarding the data distribution and exploit computational advantages of quadratic mutual information combined with the reliable and non-parametric density estimation technique, CIB still relies on the mutual information using Shannon's entropy and explicitly assumes the availability of the variables' joint distribution. This certainly cannot be guaranteed, especially for the datasets with limited sizes, which might explain for its inferior performance.

5.3 CMUFace Dataset

The CMUFace data obtained from the UCI KDD repository [1] is an interesting dataset, since its data samples can be partitioned in several different ways (e.g. by individual, by pose, etc.). The dataset consists of images of 20 people taken at various features such as facial expressions (neutral, happy, sad, angry), head positions (left, right or straight), and eye states (open or sunglasses). Each person has 32 images captured in every combination of these features. We randomly select 3 people along with all their images. Since it is known which image comes from which person, this forms an existing partition of the dataset. We run NACI and the other algorithms with this provided clustering. As a preprocessing step, the PCA technique is applied to reduce the number of dimensions, in which we retain the number of first principal components that cover 90% of the original data's variance.

In Figure 3, we show the mean vectors of the provided clustering (in the top graphs) and the mean vectors returned in the NACI's alternative clustering (in the bottom graphs). Pictorially, it is possible to observe that the uncovered alternative clustering returned by the NACI provides another different, yet equally important clustering on the set of image data. While pictures in the first row show that they represent for different individuals, pictures in the second row clearly reveal that images have been partitioned according to different poses. This obviously provides another meaningful interpretation about this dataset. For specific results, we report NACI's ones together with other techniques' in Table 2. As observed from this table, COALA and CIB perform slightly better than Algo1, Algo2 which attempt to find alternative clusterings in an orthogonal transformation space. For the methods like Dec-kmeans and CAMI which seek two alternative clusterings simultaneously, we found that they perform fairly well for the clustering based on individuals but achieve a very moderate accuracy on the

Figure 3: NACI's results on CMUFace dataset

Methods	NMI	JI	F(pose)	F(person)
Algo1	0.31	0.34	0.68	0.87
Algo2	0.33	0.36	0.67	0.84
ADFT	0.29	0.33	0.69	0.89
COALA	0.27	0.32	0.71	0.87
CIB	0.28	0.34	0.69	0.86
Dec-kmeans	0.26	0.32	0.72	0.9
ConvEM	0.28	0.33	0.7	0.89
CAMI	0.24	0.31	0.74	0.89
NACI	0.2	0.24	0.81	0.94

Table 2: Results on the CMUFace dataset.

clustering based on poses. Looking deeper, we also found that the clustering based on poses is quite hidden and non-linearly separable, but the configuration based on persons are very obvious and quite separated when visualizing using the first three PCs. This might explain why the methods like Dec-kmeans and CAMI work well for the first clustering, but not for the second one. Our algorithm outperforms these algorithms since its clustering objective is to maximize the probabilistic relationship between cluster labels and the data, and thus is not limited to the Gaussian shapes of the clusters. We also test another strategy by which the clustering labels based on poses are provided as background information. The clustering accuracy for the person based partitioning of all algorithms is summarized in the fourth column of the Table 2.

5.4 Other Real-World Datasets

We further compared the nine algorithms on three real-world datasets selected from the UCI repository: Segmentation, Vehicle Silhouette, and Vowel. For the Segmentation dataset, three attributes 5,7 and 9 are removed as they were reported to be repetitive with the attributes 4,6 and 8 [1]. Since these datasets already contain pre-defined class labels, we utilize them as an existing clustering provided. Also, as we do not have ground truth for alternative clusterings, the Dunn Index (instead of F-measure) will be used for clustering quality comparison amongst the nine clustering techniques. We report the results of all techniques on these datasets in Table 3.

Looking at this table, we see that NACI also performing well with all three datasets. Note that these datasets have a much higher degree of non-linearity, compared to the ones we have already examined. It can be noted that NACI typically finds more dissimilar clusterings (as measured by both NMI and JI) compared to those of other algorithms. Its clusters found in the alternative clustering are also well separated as indicated by the small values of Dunn index. This measure is only slightly larger than that of CAMI in the Vehicle dataset and ADFT in Vowel dataset. This might happen, since the Dunn Index is essentially computed by the averaging distances between pairs of points in two clusters over the maximum cluster diameter, and thus somewhat supports clusters returned by CAMI and ADFT. However, overall, one can observe that our algorithm tends to achieve more decorrelated (i.e., more different) clusterings, whereas its clustering quality is very competitive in all three datasets. Its performance in the Segmentation dataset is better than all other algorithms.

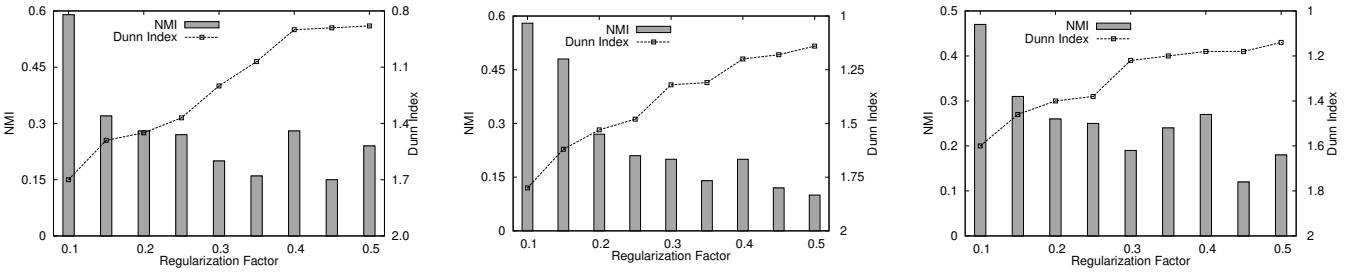
Methods	Segmentation			Vehicle			Vowel		
	NMI	JI	DI	NMI	JI	DI	NMI	JI	DI
Algo1	0.51	0.38	1.31	0.38	0.39	1.28	0.42	0.19	1.27
Algo2	0.44	0.3	1.27	0.39	0.44	1.46	0.43	0.21	1.3
ADFT	0.46	0.31	1.3	0.35	0.37	1.42	0.48	0.33	**1.41**
COALA	0.44	0.29	1.25	0.29	0.35	1.51	0.36	0.27	1.29
CIB	0.45	0.32	1.32	0.33	0.41	1.39	0.41	0.26	1.25
Deckm	0.39	0.29	1.26	0.26	0.36	1.4	0.27	0.17	1.26
ConvEM	0.41	0.3	1.27	0.25	0.34	1.41	0.31	0.19	1.23
CAMI	0.31	0.27	1.44	0.23	0.32	**1.53**	0.24	0.11	1.38
NACI	**0.26**	**0.25**	**1.46**	**0.21**	**0.28**	1.51	**0.22**	**0.11**	1.38

Table 3: Results on three real-world datasets

5.5 Parameter Sensibility

There are two parameters that may impact the performance of our NACI algorithm: the kernel width σ and the regularization factor η. We have conducted a series of experiments to examine the sensitivity of the results on these parameters. For the kernel width σ, though its variation can affect the algorithm's performance, we have found that for most of the cases, setting it to the value derived in Section 4.3 often leads to good and stable clustering results (as reported in the previous sections). Due to lack of space, we can only present here the main observations with respect to the parameter η, which regularizes for the relative importance between two quantities of mutual information.

Since both the variations computed in Eqs.(15) and (16) have been normalized by the corresponding mutual information, we typically set the η to be within the unit interval. In Figure 4, we show the relationship between the normalized mutual information, the Dunn index and the values of η for three real-world datasets: Segmentation, Vehicle Silhouette, and Vowel. The results are reported when η is varied from 0.1 to 0.5 with a step of 0.05. As we expected, small values of η lead to compact clustering solutions (in terms of Dunn index), but such results seem to be quite overlapping with the given clusterings (shown by the high values of NMI). As the values of η increase, the clustering quality somewhat reduces but the alternative clusterings are more decorrelated from the existing solutions. However, once η is above 0.3, it was observed the fluctuations happened with the NMI's values. This can be justified by the hierarchical clustering approach of NACI, where it tends to combine clusters which overly support small values of the information between two clusterings at the beginning, but such decisions cannot be undone at a later time where it converges to a small number of final clusters. Likewise, it was seen that the resultant clusters in the alternative clusterings in this case were also very imbalance, which made the Dunn index located at small values as well. Nevertheless, as we observed from all three graphs in Figure 4, high quality and dissimilar alternative



(a) Segmentation dataset (b) Vehicle dataset (c) Vowel dataset

Figure 4: Impact of η on NACI's performance. For ideal results, NMI should be low and DI should be high.

clusterings can be achieved if η is set around 0.2 since the Dunn Index in this range is relatively high, whereas that value of NMI is also small.

6. CONCLUSIONS

In this paper we have proposed a novel algorithm called NACI, to discover alternative clusterings, which are of high quality, yet distinctively different from a provided reference clustering. The problem is addressed purely from information theory in which the clustering quality is achieved by maximizing the mutual information between cluster labels and the data observations (this implicitly ensures the strong probabilistic relationship between them), whereas the dissimilarity between two alternative clusterings is guaranteed by the minimization of the mutual information between them. To fully exploit the information embedded in the data, we employed the Parzen window method for pdfs estimation. Such a non-parametric technique does not impose any assumptions regarding the data distribution and further enables practical computations when combined with the quadratic mutual information form. These all made the NACI particularly suitable for the challenging scenario where datasets have non linear structures.

We evaluated the performance of the algorithm on a number of synthetic and real-world benchmark datasets, comparing against eight well known existing approaches. The experimental results show NACI is able to achieve excellent performance for non linear cases, and is able to obtain highly competitive performance even for simpler, more linear structures. We believe that NACI is to be a powerful tool for alternative clustering discovery and exploration.

7. REFERENCES

[1] A. Asuncion and D. Newman. UCI machine learning repository, 2007.

[2] E. Bae and J. Bailey. Coala: A novel approach for the extraction of an alternate clustering of high quality and high dissimilarity. In *ICDM Conference*, 2006.

[3] S. Basu, A. Banerjee, and R. Mooney. Active semi-supervision for pairwise constrained clustering. In *SDM Conference*, 2004.

[4] M. Bilenko, S. Basu, and R. J. Mooney. Integrating constraints and metric learning in semi-supervised clustering. In *ICML*, 2004.

[5] R. Caruana, M. Elhawary, N. Nguyen, and C. Smith. Meta clustering. In *ICDM Conference*, 2006.

[6] G. Chechik and N. Tishby. Extracting relevant structures with side information. In *NIPS*, 2002.

[7] D. M. Christopher, R. Prabhakar, and S. Hinrich. *Introduction to Information Retrieval*. Cambridge University Press, 2008.

[8] Y. Cui, X. Fern, and J. Dy. Non-redundant multi-view clustering via orthogonalization. In *ICDM*, 2007.

[9] X. Dang and J. Bailey. Generation of alternative clusterings using the cami approach. In *SDM Conference. To appear*, 2010.

[10] I. Davidson and Z. Qi. Finding alternative clusterings using constraints. In *ICDM Conference*.

[11] X. Fern and W. Lin. Cluster ensemble selection. *Stat. Anal. Data Min.*, 1(3):128–141, 2008.

[12] D. Gondek. Non-redundant clustering. In *PhD Thesis, Brown University*, 2005.

[13] D. Gondek and T. Hofmann. Conditional information bottleneck clustering. In *ICDM Conference*, 2003.

[14] D. Gondek, S. Vaithyanathan, and A. Garg. Clustering with model-level constraints. In *SDM Conference*, 2005.

[15] P. Jain, R. Meka, and I. Dhillon. Simultaneous unsupervised learning of disparate clusterings. In *SDM Conference*, 2008.

[16] J. Kapur. *Measures of Information and their Application*. John Wiley, 1994.

[17] K.Wagstaff and C.Cardie. Clustering with instance-level constraints. In *ICML*, 2000.

[18] M. Law, A. Topchy, and A. Jain. Multiobjective data clustering. In *CVPR Conference*, 2004.

[19] E. Parzen. On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, 33(3), 1962.

[20] J. Principe, D. Xu, and J. Fisher. *Information Theoretic Learning*. John Wiley & Sons, 2000.

[21] A. Topchy, A. Jain, and W. Punch. A mixture model for clustering ensembles. In *SDM Conference*, 2004.

[22] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl. Constrained k-means clustering with background knowledge. In *ICML*, 2001.

[23] M. P. Wand and M. C. Jones. *Kernel Smoothing-Monographs on Statistics and Applied Probability*. Chapman & Hall/CRC, 1994.

[24] E. Xing, A. Ng, M. Jordan, and S. Russell. Distance metric learning with application to clustering with side-information. In *NIPS*, 2002.

[25] L. Yang and R. Jin. Distance metric learning: A comprehensive survey, 2006.