Mach Learn (2015) 98:31–56
DOI 10.1007/s10994-013-5333-z

Subjectively interesting alternative clusterings

Kleanthis-Nikolaos Kontonasios · Tijl De Bie

Received: 25 May 2012 / Accepted: 10 February 2013 / Published online: 7 March 2013

Abstract We deploy a recently proposed framework for mining subjectively interesting patterns from data to the problem of alternative clustering, where patterns are sets of clusters (clusterings) in the data. This framework outlines how subjective interestingness of patterns (here, clusterings) can be quantified using sound information theoretic concepts. We demonstrate how it motivates a new objective function quantifying the interestingness of a clustering, automatically accounting for a user's prior beliefs and for redundancies between the discovered patterns.

Directly searching for the optimal set of clusterings defined in this way is hard. However, the optimization problem can be solved approximately if clusterings are generated iteratively. In this iterative scheme, each subsequent clustering is maximally interesting given the whole set of previously generated clusterings, automatically trading off interestingness with non-redundancy. Although generating each clustering in an iterative fashion is computationally hard as well, we develop an approximation technique similar to spectral clustering algorithms.

Our method can generate as many clusterings as the user requires. Subjective evaluation or the value of the objective function can guide the termination of the process. In addition our method allows varying the number of clusters in each successive clustering.

Experiments on artificial and real-world datasets show that the mined clusterings fulfill the requirements of a good clustering solution by being both non-redundant and of high compactness. Comparison with existing solutions shows that our approach compares favourably with regard to well-known objective measures of similarity and quality of clusterings, even though it is not designed to directly optimize them.

Keywords Subjective interestingness · Maximum entropy modelling · Alternative clustering

Editors: Emmanuel Müller, Ira Assent, Stephan Günnemann, Thomas Seidl, and Jennifer Dy.

K.-N. Kontonasios (✉) · T. De Bie
Intelligent Systems Laboratory, University of Bristol, Bristol, UK
e-mail: kk8232@bristol.ac.uk

1 Introduction

A main challenge in research on clustering methods and theory is that clustering is, in a way intentionally, ill-defined as a task. As a result, numerous types of syntaxes for cluster patterns have been suggested, e.g. clusters as hyperrectangles, hyperspheres, ellipsoids, decision trees, clusterings as partitions, hierarchical partitionings, etc. Additionally, even when the syntax is fixed, there are often various alternative choices for the objective function, e.g. the K-means cost function, the likelihood of a mixture of Gaussians, etc. The reader can refer to Jain et al. (1999) for a review on clustering methods.

Research in alternative clustering, i.e. the task of discovering multiple non-redundant sets of clusters in data, is structured similarly. A large number of algorithms which manipulate the original data spaces (Gondek and Hofmann 2003; Dang and Bailey 2010; Bae and Bailey 2006; Jain et al. 2008), transformed data spaces (Cui et al. 2007; Qi and Davidson 2009; Davidson and Qi 2008) or subspace projections of the data (Sequeira and Zaki 2004; Müller et al. 2009) were developed to simultaneously optimize the quality of a clustering and its diversity from the previously generated solutions. Quality and diversity are formalized using various techniques ranging from information theory (Gondek and Hofmann 2003; Dang and Bailey 2010; Qi and Davidson 2009) to constraint satisfaction (Bae and Bailey 2006; Jain et al. 2008; Davidson and Qi 2008; Qi and Davidson 2009).

Despite this variety in approaches, the goal of both traditional and alternative clustering is almost always to provide a user with insights in the structure of the data, allowing the user to conceptualize it as coming from a number of broad areas in the data space. The knowledge of such a structure can be more or less elucidating to the user, also depending on the prior beliefs the user held about the data.

Here we expand the algorithm presented in De Bie (2011) for finding alternative *clusters* to the task of finding alternative *clusterings*. More in particular, we take the perspective that a clustering is more useful if it conveys more novel information with respect to the user's prior knowledge about the dataset. We make a specific choice for a clustering syntax, and we deploy ideas from De Bie (2011) to quantify the interestingness of a clustering as the amount of information conveyed to the user when told about the clustering's presence.

Our approach attempts to quantify *subjective* interestingness (Geng and Hamilton 2006; McGarry 2005; Kontonasios et al. 2012) of clusterings, in that it takes prior beliefs held by the user into account. As a result, different clusterings might be deemed interesting to different users. One particular example is the situation where a user has already been informed about previously discovered clusterings in the data. In that case, clusterings that are individually informative while non-redundant will be the most interesting ones. Our approach naturally deals with the alternative clustering setting, by regarding already communicated clusterings as prior information.

Unlike most of the methods in the literature, we do not deploy ad-hoc heuristic algorithms for optimizing certain quality criteria for clusterings. Instead, we instantiate the general information-theoretic framework proposed in De Bie (2011) and optimize a quality function derived directly from this model.

The problem of generating more than one alternative clustering, one of the major challenges in alternative clustering methods (Müller et al. 2010), is naturally addressed by our method. Our algorithm can generate as many clusters as the user requires, taking into account the whole set of previously discovered clusterings. Loosely speaking we can say that our method can memorize a set of clusterings, instead of a single clustering as most of other methods do, and provide a non-redundant alternative clustering to all of them. Most interestingly, the computational cost for every new clustering remains the same for each iteration.

In addition, generating clusterings with varying number of clusters can be easily dealt with using our framework.

Outline The next section (Sect. 2) recapitulates the main ideas of the general data mining framework presented in De Bie (2011). In Sect. 3 we suggest using the first and second order cumulants of the data points as initial prior knowledge of the data miner, and we demonstrate how this prior knowledge can be formalized by means of a Maximum Entropy background distribution as required for the framework from Sect. 2. In this section we also introduce a syntax of a pattern that represents a single cluster, as well as pattern syntax to represent a set of clusters (e.g. a cluster*ing*, or more specifically a partition of the data). We subsequently derive analytical expressions for the subjective interestingness of such pattern types (quantified by means of their self-information in the framework from Sect. 2), given initial prior knowledge on the first and second cumulants of the data. These results allow us in Sect. 4 to present our iterative scheme for mining the most interesting clustering subject to prior knowledge that may also include the knowledge of previously discovered clustering patterns—i.e. to find the most interesting *alternative* clustering. Although finding the best alternative clustering in this sense is a hard problem, in Sect. 5 we present an efficient approximation algorithm similar to spectral clustering methods. We verify our approach experimentally in Sect. 6 and discuss related work in Sect. 7.

Notation Throughout this paper $\mathbf{x} \in \mathbb{R}^d$ denotes a d-dimensional data point, and $\mathbf{X} = (\mathbf{x}_1' \ \mathbf{x}_2' \ \cdots \ \mathbf{x}_n')'$ denotes the data matrix containing n data points as its rows. The space the data matrix belongs to is denoted as $\mathcal{X} = \mathbb{R}^{n \times d}$. The pseudoinverse of a matrix $\mathbf{A}$ is denoted as $\mathbf{A}^\dagger$.

2 A unified framework for data mining

For completeness, we here provide a short overview of a framework for data mining that was introduced in De Bie (2011), and readers familiar with this paper can skip this section. Earlier and more limited versions of this framework, as well as its application to other machine learning and data mining tasks, can be found in De Bie (2010, 2011), Kontonasios and De Bie (2010), Kontonasios et al. (2011), Spyropoulou and De Bie (2011). For concreteness, here we specialize the short overview of the framework to the case where the data is a data set, summarized in the data matrix $\mathbf{X}$.

The framework aims to formalize data mining as a process of information exchange between the data and the data miner (the user). The goal of the data miner is to get as good an understanding about the data as possible, i.e. to reduce his uncertainty as much as possible. To be able to do this, the degree of uncertainty must be quantified. To this end we use a probability distribution P (referred to as the *background distribution*) to model the prior beliefs of the user about the data $\mathbf{X}$, in combination with ideas from information theory.

More specifically, the framework deals with the setting where the prior beliefs specify that the background distribution belongs to a set $\mathcal{P}$ of possible distributions. The more prior beliefs, the smaller this set will be. For example, the data miner may have a set of prior beliefs that can be formalized in the form of constraints the background distribution P satisfies:

$$\int_{\mathbf{X} \in \mathbb{R}^{n \times d}} f_i(\mathbf{X}) P(\mathbf{X}) = c_i, \quad \forall i.$$

Such constraints represent the fact that the expected value of certain statistics (the functions f_i) are equal to a given number (the constants c_i). The set $\mathcal{P}$ is defined as the set of distributions satisfying these constraints. (Note that the framework is not limited to such prior beliefs, although they are convenient from a practical viewpoint.)

It was argued in De Bie (2011) that among all distributions $P \in \mathcal{P}$, the 'best' choice for P is the one of maximum entropy (Jaynes 1982) given these constraints. This background distribution is the least biased one, thus not introducing any other undue constraints on the background distribution. A further game-theoretic argument in favour of using the distribution of maximum entropy is given in De Bie (2011).

In the framework, a pattern is defined as any piece of knowledge about the data that reduces the set of possible values it may take from the original data space $\mathcal{X} = \mathbb{R}^{n \times d}$ to a subset $\mathcal{X}' \subseteq \mathcal{X}$. It was then argued that the subjective interestingness of such a pattern can be adequately formalized as the *self-information* of the pattern, i.e. the negative logarithm of the probability that the pattern is present in the data, i.e. by $-\log(P(\mathbf{X} \in \mathcal{X}'))$. Thus, patterns are deemed more interesting if their probability is smaller under the background model, and thus if the user is more surprised by their observation.

After observing a pattern, a user instinctively adapts his beliefs. In De Bie (2011) it was argued that a natural and robust way to model this is by updating the background distribution to a new distribution P' defined as P conditioned on the pattern's presence. The self-information of subsequent patterns can thus be evaluated by referring to the new background distribution P', and so on in an iterative fashion.

In De Bie (2011) it was demonstrated that mining the most informative *set* of patterns formally corresponds to a weighted set coverage problem, attempting to cover as many elements from the set $\mathcal{X}$ that have a small probability under the initial background distribution P. This problem is NP-hard, but it can be approximated well by a greedy approach. An iterative data mining approach is equivalent to such a greedy approximation. Selecting and encoding patterns in an iterative manner, ensures that at any time the patterns generated are approximately the most informative set of patterns of that size.

In the next section we instantiate this general framework for discovering clusterings in the data.

3 Subjective interestingness of (a set of) cluster(s)

3.1 Prior beliefs and the maximum entropy background distribution

Here we consider two types of initial prior beliefs, expressed as constraints on the first and second order cumulants of the data points. It is conceptually easy to extend the results from this paper to other types of prior beliefs, although the computational cost will vary. The background distribution incorporating these constraints is the maximum entropy distribution that has the prescribed first and second order cumulants. It is easy to show that for data with unbounded domain, this distribution is the multivariate Gaussian distribution with mean and covariance matrix equal to the prescribed cumulants:

$$P(\mathbf{X}) = \frac{1}{\sqrt{(2\pi)^{nd}|\boldsymbol{\Sigma}|^n}} \exp\left(-\frac{1}{2}\text{trace}\left[\left(\mathbf{X} - \mathbf{e}\boldsymbol{\mu}'\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{X} - \mathbf{e}\boldsymbol{\mu}'\right)'\right]\right), \tag{1}$$

where $\boldsymbol{\mu}$ is a $d \times 1$ vector containing the means of every row in the dataset, $\boldsymbol{\Sigma}$ is a $n \times n$ covariance matrix and $\mathbf{e}$ is a $n \times 1$ containing ones.

We note that the prescribed cumulants may be computed from the actual data at the request of the data miner, such that they are indeed part of the prior knowledge. However, they may also be beliefs, in the sense that they may be based on external information or assumptions that may be right or wrong.

3.2 A syntax for cluster patterns

The framework from De Bie (2011) was developed for patterns generally defined as properties of the data. Thus, a pattern's presence in the data constrains the set of possible values the data may have, and in this sense the knowledge of the presence of a pattern reduces the uncertainty about the data and conveys information.

In this paper we restrict our attention to one specific type of cluster pattern. The pattern type we consider is parameterized by a set of indices to the data points and a vector in the data space. The pattern is then the fact that the mean of the data points for these indices is equal to the specified vector.

More formally, let $\mathbf{e}_I \in \{0,1\}^n$ be defined as an indicator vector containing zeros at positions $i \notin I$ and ones at positions $i \in I$, and let $n_I = |I| = \mathbf{e}_I' \mathbf{e}_I$ denote the number of elements in I. Then, our patterns are constraints of the form:

$$\begin{aligned} &\frac{1}{n_I} \sum_{i \in I} \mathbf{x}_i = \boldsymbol{\mu}_I, \\ &\iff \quad \mathbf{X}' \mathbf{e}_I = n_I \boldsymbol{\mu}_I. \end{aligned}$$

Such a constraint restricts the possible values of the data set $\mathbf{X}$, in that the mean of a subset of the data points is required to have a prescribed value $\boldsymbol{\mu}_I$.

3.3 A syntax for a pattern defined as a set of clusters

The present paper is concerned with clustering data, i.e. with identifying a meaningful partition of the data. In the case of alternative clustering, we also need a way to consider different such clusterings simultaneously.

We can formalize a clustering pattern as a set of cluster patterns, one for each cluster in the clustering. A pattern describing the presence of a *set of* clusterings can also be described by a set of cluster patterns, one for each cluster present in any of the clusterings. Consequently, to describe a single clustering as well as a set of different clusterings, it suffices to define a general pattern type to describe the presence of a *set of clusters*.

Based on the syntax of cluster patterns, the syntax of *a set of* k cluster patterns can be defined as the union of k cluster patterns. We create the indicator matrix $\mathbf{E} \in \{0,1\}^{n \times k}$, which contains as its columns indicator vectors $\mathbf{e}_{I_i}$ for each one of the k clusters in the set, and the matrix $\mathbf{M} \in \mathbb{R}^{d \times k}$, a real-valued matrix which contains the scaled means, i.e. $n_{I_i} \boldsymbol{\mu}_{I_i}$, of each cluster as its columns. Then the set of constraints for a set of clusters can be written concisely as $\mathbf{X}'\mathbf{E} = \mathbf{M}$.

We reiterate that this pattern syntax can be used to describe the presence of any set of cluster patterns. The case of a clustering (in the sense of partition) is obviously subsumed by this formulation, by ensuring that the columns of $\mathbf{E}$ form a partition.

3.4 The self-information of a cluster pattern

The *self-information* of a pattern is the measure of subjective interestingness proposed in De Bie (2011). It is defined as the negative logarithm of the probability of a pattern to be present

in the data under the maximum entropy model. The following theorem shows how to assess the self-information of a cluster.

Theorem 1 *Given a background distribution of the form in Eq.* (1)*, the probability of a pattern of the form* $\mathbf{X}'\mathbf{e}_I = n_I\boldsymbol{\mu}_I$ *is given by*:

$$P\left(\mathbf{X}'\mathbf{e}_I = n_I\boldsymbol{\mu}_I\right) = \frac{1}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}|}}\exp\left(-\frac{1}{2|I|}\mathbf{e}_I'\cdot\left[\left(\mathbf{X}-\mathbf{e}\boldsymbol{\mu}'\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{X}-\mathbf{e}\boldsymbol{\mu}'\right)'\right]\cdot\mathbf{e}_I\right).$$

Thus the self-information of a cluster specified by the set I*, defined as its negative log probability under the Maximum Entropy model and denoted as* $SelfInformation_I$*, is equal to*:

$$SelfInformation_I = \frac{1}{2}\log\left((2\pi)^d|\boldsymbol{\Sigma}|\right) + \frac{1}{2}Q_I,$$
$$where\ Q_I = \frac{1}{|I|}\mathbf{e}_I'\cdot\left[\left(\mathbf{X}-\mathbf{e}\boldsymbol{\mu}'\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{X}-\mathbf{e}\boldsymbol{\mu}'\right)'\right]\cdot\mathbf{e}_I.$$

Note that the self-information depends on I only through Q_I, so we may choose to use Q_I as a quality measure for a cluster, as we will do in this paper.

This theorem can be used to quantify the self-information of any cluster given the background model based on the initial prior beliefs of the data miner. Note however that it cannot be used to assess the self-information of a cluster after other clusters have already been found, as each new cluster will affect the user's prior beliefs. How this can be accounted for will be discussed in Sect. 3.5 (and further in Sect. 4), based on a generalization of Theorem 1. As Theorem 1 directly follows from Theorem 2, we will only provide a proof for the latter in Sect. 3.5.

3.5 The self-information of a set of clusters

The following theorem is a generalization of Theorem 1 as it demonstrates the calculation of the *self-information* for a set of clusters.

Theorem 2 *Let the columns of the matrix* $\mathbf{E}$ *be the indicator vectors of the sets in* $\mathcal{I} = \{I_i\}$*, and let* $\mathbf{P}_\mathbf{E} = \mathbf{E}(\mathbf{E}'\mathbf{E})^\dagger\mathbf{E}'$*, the projection matrix onto the column space of* $\mathbf{E}$*. Then, the probability of the composite pattern* $\mathbf{X}'\mathbf{E} = \mathbf{M}$ *is given by*:

$$P\left(\mathbf{X}'\mathbf{E} = \mathbf{M}\right) = \frac{1}{\sqrt{(2\pi)^{kd}|\boldsymbol{\Sigma}|^k}}\exp\left(-\frac{1}{2}\operatorname{trace}\left[\mathbf{P}_\mathbf{E}\cdot\left(\mathbf{X}-\mathbf{e}\boldsymbol{\mu}'\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{X}-\mathbf{e}\boldsymbol{\mu}'\right)'\right]\right).$$

Thus the self-information of the set of patterns defined by the columns of $\mathbf{E}$*, defined as its negative log probability and denoted as* $SelfInformation_\mathcal{I}$*, is equal to*:

$$SelfInformation_\mathcal{I} = \frac{k}{2}\log\left((2\pi)^d|\boldsymbol{\Sigma}|\right) + \frac{1}{2}Q_\mathcal{I},$$
$$where\ Q_\mathcal{I} = \operatorname{trace}\left[\mathbf{P}_\mathbf{E}\cdot\left(\mathbf{X}-\mathbf{e}\boldsymbol{\mu}'\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{X}-\mathbf{e}\boldsymbol{\mu}'\right)'\right].$$

Again, since the self-information depends on I only through Q_I, we choose to use Q_I as a quality measure for a cluster further below.

Proof A constraint $\mathbf{X}'\mathbf{E} = \mathbf{M}$ constrains the data $\mathbf{X}$ to an $(n-k)\times d$ dimensional affine subspace in the following way. Let us write the singular value decomposition for $\mathbf{E}$ as:

$$\mathbf{E} = \begin{pmatrix} \mathbf{U} & \mathbf{U}_0 \end{pmatrix} \begin{pmatrix} \boldsymbol{\Lambda} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} \begin{pmatrix} \mathbf{V} & \mathbf{V}_0 \end{pmatrix}'.$$

Then, this constraint can be written in the following form:

$$\mathbf{X} = \mathbf{U}\mathbf{Z} + \mathbf{U}_0\mathbf{Z}_0,$$

where $\mathbf{Z} = \boldsymbol{\Lambda}^{-1}\mathbf{V}'\mathbf{M}'$ is a constant fixed by $\mathbf{E}$ and $\mathbf{M}$, and $\mathbf{Z}_0 \in \mathbb{R}^{(n-k)\times d}$ is a variable. In general, writing $\mathbf{X} = \mathbf{U}\mathbf{Z} + \mathbf{U}_0\mathbf{Z}_0$, we can write the probability density for $\mathbf{X}$ as:

$$\begin{aligned}
P(\mathbf{X}) &= P(\mathbf{Z}, \mathbf{Z}_0), \\
&= \frac{1}{\sqrt{(2\pi)^{nd}|\boldsymbol{\Sigma}|^n}} \exp\left(-\frac{1}{2}\operatorname{trace}\left[\left(\mathbf{U}\mathbf{Z} + \mathbf{U}_0\mathbf{Z}_0 - \mathbf{e}\boldsymbol{\mu}'\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{U}\mathbf{Z} + \mathbf{U}_0\mathbf{Z}_0 - \mathbf{e}\boldsymbol{\mu}'\right)'\right]\right), \\
&= \frac{1}{\sqrt{(2\pi)^{nd}|\boldsymbol{\Sigma}|^n}} \exp\left(-\frac{1}{2}\operatorname{trace}\left[\left(\mathbf{U}_0\mathbf{Z}_0 - \mathbf{U}_0\mathbf{U}_0'\mathbf{e}\boldsymbol{\mu}'\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{U}_0\mathbf{Z}_0 - \mathbf{U}_0\mathbf{U}_0'\mathbf{e}\boldsymbol{\mu}'\right)'\right]\right) \\
&\quad \cdot \exp\left(-\frac{1}{2}\operatorname{trace}\left[\left(\mathbf{U}\mathbf{Z} - \mathbf{U}\mathbf{U}'\mathbf{e}\boldsymbol{\mu}'\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{U}\mathbf{Z} - \mathbf{U}\mathbf{U}'\mathbf{e}\boldsymbol{\mu}'\right)'\right]\right), \\
&= \frac{1}{\sqrt{(2\pi)^{nd}|\boldsymbol{\Sigma}|^n}} \exp\left(-\frac{1}{2}\operatorname{trace}\left[\left(\mathbf{Z}_0 - \mathbf{U}_0'\mathbf{e}\boldsymbol{\mu}'\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{Z}_0 - \mathbf{U}_0'\mathbf{e}\boldsymbol{\mu}'\right)'\right]\right) \\
&\quad \cdot \exp\left(-\frac{1}{2}\operatorname{trace}\left[\left(\mathbf{Z} - \mathbf{U}'\mathbf{e}\boldsymbol{\mu}'\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{Z} - \mathbf{U}'\mathbf{e}\boldsymbol{\mu}'\right)'\right]\right), \\
&= \frac{1}{\sqrt{(2\pi)^{(n-k)d}|\boldsymbol{\Sigma}|^{n-k}}} \exp\left(-\frac{1}{2}\operatorname{trace}\left[\left(\mathbf{Z}_0 - \mathbf{U}_0'\mathbf{e}\boldsymbol{\mu}'\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{Z}_0 - \mathbf{U}_0'\mathbf{e}\boldsymbol{\mu}'\right)'\right]\right) \\
&\quad \cdot \frac{1}{\sqrt{(2\pi)^{(k)d}|\boldsymbol{\Sigma}|^{k}}} \exp\left(-\frac{1}{2}\operatorname{trace}\left[\left(\mathbf{Z} - \mathbf{U}'\mathbf{e}\boldsymbol{\mu}'\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{Z} - \mathbf{U}'\mathbf{e}\boldsymbol{\mu}'\right)'\right]\right)
\end{aligned}$$

We can now compute the marginal probability density for $\mathbf{Z}$ by integrating over $\mathbf{Z}_0$, yielding:

$$P(\mathbf{Z}) = \frac{1}{\sqrt{(2\pi)^{kd}|\boldsymbol{\Sigma}|^k}} \exp\left(-\frac{1}{2}\operatorname{trace}\left[\left(\mathbf{Z} - \mathbf{U}'\mathbf{e}\boldsymbol{\mu}'\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{Z} - \mathbf{U}'\mathbf{e}\boldsymbol{\mu}'\right)'\right]\right).$$

The probability density value for the pattern's presence, i.e. for $\mathbf{X}'\mathbf{E} = \mathbf{M}$ or equivalently $\mathbf{Z} = \boldsymbol{\Lambda}^{-1}\mathbf{V}'\mathbf{M}'$, is thus:

$$\begin{aligned}
&P\left(\mathbf{Z} = \boldsymbol{\Lambda}^{-1}\mathbf{V}'\mathbf{M}'\right) \\
&\quad = \frac{1}{\sqrt{(2\pi)^{kd}|\boldsymbol{\Sigma}|^k}} \exp\left(-\frac{1}{2}\operatorname{trace}\left[\left(\boldsymbol{\Lambda}^{-1}\mathbf{V}'\mathbf{M}' - \mathbf{U}'\mathbf{e}\boldsymbol{\mu}'\right)\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{\Lambda}^{-1}\mathbf{V}'\mathbf{M}' - \mathbf{U}'\mathbf{e}\boldsymbol{\mu}'\right)'\right]\right), \\
&\quad = \frac{1}{\sqrt{(2\pi)^{kd}|\boldsymbol{\Sigma}|^k}} \exp\left(-\frac{1}{2}\operatorname{trace}\left[\mathbf{P}_{\mathbf{E}} \cdot \left(\mathbf{X} - \mathbf{e}\boldsymbol{\mu}'\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{X} - \mathbf{e}\boldsymbol{\mu}'\right)'\right]\right),
\end{aligned}$$

where $\mathbf{P}_{\mathbf{E}} = \mathbf{E}(\mathbf{E}'\mathbf{E})^{\dagger}\mathbf{E}' = \mathbf{U}\mathbf{U}'$ is a projection matrix projecting onto the k-dimensional column space of $\mathbf{E}$.

The required expression is obtained immediately by taking the negative logarithm of the last equation. □

Note that Theorem 1 is indeed a special case of Theorem 2 as can be seen by substituting $\mathbf{E} = \mathbf{e}_I$ and $\mathbf{P}_{\mathbf{E}} = \frac{\mathbf{e}_I\mathbf{e}_I'}{|I|}$.

Remark 1 (Description Length of a set of clusters) The framework from De Bie (2011) suggests to take into account not only the self-information of a pattern, but also the cost to communicate a pattern, i.e. its description length. This depends on the coding scheme used, which should reflect the perceived complexity of a pattern as perceived by the data miner. Choosing this coding scheme can also be done so as to bias the results toward specific patterns.

In the current context, describing a pattern amounts to describing the subset I and the mean vector $\boldsymbol{\mu}_I$ for every cluster in the clustering. For simplicity, we assume the description length is constant for all patterns, independent of I and $\boldsymbol{\mu}_I$. However, note that different costs could be used, which can bias the interestingness of the clusterings towards specific goals.

4 Subjective interestingness of an alternative clustering

Since a single clustering as well as a set of clusterings can be described by means of a set of cluster patterns, we can rely on Theorem 2 for verifying the optimality of a clustering or of a set of clusterings according to the *self-information* measure.

Unfortunately, though by no means unusually for clustering formulations, optimizing $Q_{\mathcal{I}}$ over the set of all possible clusterings is a difficult combinatorial problem. A fortiori, searching for the globally optimal set of clusterings is infeasible in practice. In fact in De Bie (2011), it was shown that in general this amounts to an NP-hard set coverage problem, where each possible pattern (the clustering patterns, in the present paper) corresponds to a set in the set coverage problem.

Fortunately, a set coverage optimization problem can be approximated provably well by a greedy iterative approach. Applied to our problem setting, this strategy would iteratively select the clustering that maximizes the increase of the quality measure $Q_{\mathcal{I}}$. In this section we quantify this increase as ΔQ_k. We show that ΔQ_k is the *self-information* of a clustering on data in the k'th iteration that is projected onto the space orthogonal to the clustering assignments already discovered by the algorithm in previous $k-1$ iterations. Although also maximizing this increase is hard as well, we present an approximate solution to this problem in Sect. 5, relying on a spectral relaxation of the problem inspired by the spectral clustering literature.

Before we proceed, however, we want to point out another beneficial aspect of the iterative approach.

Usually it is not a priori clear how many clusterings are required for the data miner to be sufficiently satisfied with his new understanding of the data. The idea of alternative clustering, as we view it, is to provide the user the opportunity to request new clusterings as long as more are desired. Optimizing the quality measure over a growing set of clusterings by iteratively optimizing over newly added columns of $\mathbf{E}$ is thus a type of alternative clustering. Hence, the iterative approach can be regarded as an approximation, but one with usability benefits over a global optimizing approach.

Let us say that we have already found $k-1 \geq 1$ clusterings, and the matrices $\mathbf{E}$ and $\mathbf{M}$ respectively contain the indicator vectors and scaled cluster means as their columns. We are interested in finding the k'th clustering so as to optimize the quality measure from Theorem 2 but keeping the first $k-1$ clustering patterns as they are.

To do this, it is convenient to write the quality measure as a function of the k'th clustering with indicator matrix $\mathbf{E}_k$ (which we can safely assume to be of full column rank). Let us denote the augmented indicator matrix for the k-th iteration as $\mathbf{E}^* = (\mathbf{E}\ \mathbf{E}_k)$. The projection matrix of $\mathbf{E}^*$ can be expressed as

$$\begin{aligned}\mathbf{P}_{\mathbf{E}^*} &= \mathbf{E}^*\left(\mathbf{E}^{*\prime}\mathbf{E}^*\right)^{\dagger}\mathbf{E}^{*\prime} = \begin{pmatrix}\mathbf{E} & \mathbf{E}_k\end{pmatrix}\left[\begin{pmatrix}\mathbf{E} & \mathbf{E}_k\end{pmatrix}'\begin{pmatrix}\mathbf{E} & \mathbf{E}_k\end{pmatrix}\right]^{\dagger}\begin{pmatrix}\mathbf{E} & \mathbf{E}_k\end{pmatrix}',\\ &= \mathbf{P}_{\mathbf{E}} + \mathbf{Q}_{\mathbf{E}}\mathbf{E}_k\left(\mathbf{E}_k'\mathbf{Q}_{\mathbf{E}}\mathbf{E}_k\right)^{-1}\mathbf{E}_k'\mathbf{Q}_{\mathbf{E}},\end{aligned}$$

where $\mathbf{Q}_{\mathbf{E}} = \mathbf{I} - \mathbf{P}_{\mathbf{E}}$, the projection matrix on the null column space of $\mathbf{E}$.

Using the definition quality measure Q from Theorem 2 and the expression for the projection matrix derived above we obtain:

$$\begin{aligned}Q_{\bigcup_{i=1}^{k}\mathcal{I}_i} &= \operatorname{trace}\left[\mathbf{P}_{\mathbf{E}^*}\cdot\left(\mathbf{X}-\mathbf{e}\boldsymbol{\mu}'\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{X}-\mathbf{e}\boldsymbol{\mu}'\right)'\right],\\ &= \operatorname{trace}\left[\mathbf{P}_{\mathbf{E}}\cdot\left(\mathbf{X}-\mathbf{e}\boldsymbol{\mu}'\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{X}-\mathbf{e}\boldsymbol{\mu}'\right)'\right]\\ &\quad + \operatorname{trace}\left[\mathbf{Q}_{\mathbf{E}}\mathbf{E}_k\left(\mathbf{E}_k'\mathbf{Q}_{\mathbf{E}}\mathbf{E}_k\right)^{-1}\mathbf{E}_k'\mathbf{Q}_{\mathbf{E}}\left(\mathbf{X}-\mathbf{e}\boldsymbol{\mu}'\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{X}-\mathbf{e}\boldsymbol{\mu}'\right)'\right],\\ &= Q_{\bigcup_{i=1}^{k-1}\mathcal{I}_i}\\ &\quad + \operatorname{trace}\left\{\left(\mathbf{E}_k'\mathbf{Q}_{\mathbf{E}}\mathbf{E}_k\right)^{-1}\mathbf{E}_k'\left[\mathbf{Q}_{\mathbf{E}}\left(\mathbf{X}-\mathbf{e}\boldsymbol{\mu}'\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{X}-\mathbf{e}\boldsymbol{\mu}'\right)'\mathbf{Q}_{\mathbf{E}}\right]\mathbf{E}_k\right\},\end{aligned}$$

where $\mathcal{I}_i$ is the partition corresponding to the i-th clustering, $\mathbf{E}_k$ is the indicator matrix corresponding to the partition $\mathcal{I}_k$ and $\mathbf{E}$ contains the indicator vectors for all clusters in $\bigcup_{i=1}^{k-1}\mathcal{I}_i$.

Each of the iterative steps thus reduces to the maximization of the following increase of the quality measure:

$$\Delta Q_k = \operatorname{trace}\left\{\left(\mathbf{E}_k'\mathbf{Q}_{\mathbf{E}}\mathbf{E}_k\right)^{-1}\mathbf{E}_k'\left[\mathbf{Q}_{\mathbf{E}}\left(\mathbf{X}-\mathbf{e}\boldsymbol{\mu}'\right)\boldsymbol{\Sigma}^{-1}\left(\mathbf{X}-\mathbf{e}\boldsymbol{\mu}'\right)'\mathbf{Q}_{\mathbf{E}}\right]\mathbf{E}_k\right\}.$$

The initial iteration, where no clustering is present in the prior knowledge, is a special case of the quality measure increment ΔQ_k due to an alternative clustering with settings $Q_{\emptyset} = 0$ and $\mathbf{Q}_{\mathbf{E}} = \mathbf{I}$. Using these settings the expression of Theorem 2 for the *self-information* of a clustering is derived.

We can interpret the above reformulation of the quality measure for the k'th clustering conditioned on the first $k-1$ clusterings as being the quality measure for a first clustering on data that is projected onto the space orthogonal to the $k-1$ sets of columns of $\mathbf{E}$, i.e. the $k-1$ previously selected sets of indicator vectors. It is as if the data was deflated to take account of the knowledge of the previously found clustering patterns, thus automatically accounting for redundancy. In other words, the value ΔQ_k can be seen as the *self-information* of a clustering pattern after considering a set of other clustering patterns.

Remark 2 (Kernel variant) Note that for $\boldsymbol{\Sigma} = \mathbf{I}$ and $\boldsymbol{\mu} = \mathbf{0}$, the quality measures depend on $\mathbf{X}$ only through the inner product matrix $\mathbf{X}\mathbf{X}'$. This means that a kernel-variant is readily derived, by substituting this inner product matrix with any suitable kernel matrix. In this way non-linearly shaped clusters can be obtained, similar to spectral clustering methods and kernel K-Means.

5 A spectral algorithm for optimizing subjective interestingness

Computing the binary matrix $\mathbf{E}_k$ which maximizes ΔQ_k is a combinatorial problem. However, if we relax the vector $\mathbf{E}_k$ to be real-valued instead of containing only 0's and 1's we obtain the *quotient trace* optimization problem. The *quotient trace* problem is a well-known problem encountered often in fields such as classification and dimensionality reduction (Wang et al. 2007).

This trace is maximized by the matrix $\mathbf{E}_k$ with columns the k dominant eigenvectors of the generalized eigenvalue problem $\mathbf{AE_k} = \boldsymbol{\Lambda}\mathbf{BE_k}$, where $\mathbf{A} = \mathbf{Q_E}(\mathbf{X} - \mathbf{e}\boldsymbol{\mu}')\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \mathbf{e}\boldsymbol{\mu}')'\mathbf{Q_E}$ and $\mathbf{B} = \mathbf{Q_E}$ (Hersch 1961).

In order to obtain a crisp $0/1$ matrix $\mathbf{E}_k$, a binary approximation to the real solution must be employed. This is a hard task and various methods have been suggested in the literature to derive a crisp solution from a relaxed one. Here we use the method proposed in Ng et al. (2001), Shi and Malik (2000). The method consists of the following steps:

- Create the matrix $\mathbf{E}_k^n$ by normalizing the rows of the real solution matrix $\mathbf{E}_k$.
- Apply K-means algorithm for clustering the rows of $\mathbf{E}_k^n$ in k clusters.
- Create each row in the binary $\mathbf{E}_k$ by assigning the value 1 to the column indicating the cluster in which the row was contained and 0 everywhere else.

According to Ng et al. (2001), Shi and Malik (2000), this procedure produces a binary matrix $\mathbf{E}_k$ with quotient trace close to the one of the optimal real-valued solution obtained using the generalized eigenvalue problem. Thus we can use the obtained $\mathbf{E}_k$ as the kth clustering returned by our alternative clustering scheme, representing an alternative to the $k-1$ previously considered clusterings.

Note that this spectral relaxation approach is highly efficient, requiring little more than the solution of a generalized eigenvalue problem (with dimensionality equal to n), and a subsequent low-dimensional K-means problem. This makes our approach highly scalable and easy to implement using widely available numerical matrix analysis packages.[1] Remarkably, the complexity also does not increase in consecutive iterations of alternative clustering.

The computation of the iterative step concludes the presentation of our algorithm. In the next section (Sect. 6) we demonstrate the use of our algorithm and compare with existing solutions in artificial and real-world experiments.

6 Experimental evaluation

6.1 Evaluation scheme

Assessing alternative clustering results is an inherently difficult task and many different approaches have been proposed in the literature. A general scheme, common in most of them, is that a good alternative clustering has to contain clusters of high quality and to be as dissimilar to the previous clusterings as possible. The use of various quality/dissimilarity measures for the instantiations of this general scheme leads to different overall evaluations.

In this paper we employ several different techniques for evaluating the discovered clusterings. In order to provide means for subjective evaluation, we visualize the discovered patterns. Although visualization is easy for low dimensional artificial datasets, it is increasingly difficult for high dimensional real-world datasets. In order to deal with this issue, we opted to use real datasets consisting of images for our experiments.

In addition, we compare our results with an intuitive ground truth initial clustering. For this reason we used labelled datasets for our more complex real-world experiments. Labelled datasets directly motivate an external evaluation strategy using confusions matrices and measures such as the adjusted Rand index.

[1] Though this is beyond the scope of the current paper, note that one can also rely on the considerable body of literature on approximation techniques for generalized eigenvalue problems should the data set be particularly large.

External evaluation A first type of evaluation we will consider in this paper relies on the availability of an external ground truth clustering of the data. We refer to this type of evaluation as the external evaluation, and it relies on the adjusted Rand index.

The adjusted Rand index (ARI) (Rand 1971) measures the similarity between two clusterings in a conveniently calibrated manner: its value lies between -1 (absolute dissimilarity) and 1 (absolute similarity), and it is 0 in expectation for statistically independent clusterings making its values easy to interpret.

The ARI is used in this paper to measure both the quality of the clustering at hand and its redundancy with the previous clusterings. We assess the quality of the given clustering as the similarity, measured by ARI, with the given ground truth clustering. We denote this quantity with ARI_g. Higher ARI_g values indicate higher quality clusterings. For measuring the redundancy, we calculate ARIs between the current clustering and all the clusterings previously discovered (denoted ARI_b). Lower values in this setting are better as they indicate more dissimilar clusterings.

A more detailed view of the information carried by the discovered patterns is given by confusion matrices between ground truth labels and labels assigned by our clustering method for each iteration of the algorithm.

Internal evaluation Internal evaluation approaches for clustering attempt to quantify the quality of a clustering without referring to any external ground truth labeling of the data. Following common practice in the alternative clustering literature, we do this here by means of the Dunn and Jaccard indices for measuring the quality and redundancy of clusterings respectively.

More in detail, according to the Dunn index (DI) (Dunn 1971), a high quality clustering consists of both individual clusters which present low variance between the points they contain (compact clusters) and large distances between the cluster centers. Higher values for the DI are better.

The Jaccard index (JI) (Tan et al. 2005) is a similarity measure between two different sets and can be generalized to measure the similarity of two partitions in order to be used for measuring the similarity between clusterings. Its value lies between 0 (totally dissimilar sets) and 1 (identical sets). We calculate Jaccard indices between the current clustering and all previously discovered ones in the same fashion as the calculation of ARI_b. Thus, lower values for the JI are better as they indicate lower redundancy between the alternative clustering and the previously discovered clusterings.

The F-measure: aggregating cluster quality and non-redundancy In order to summarize the quality of an alternative clustering using a single score, one commonly relies on the so-called F-measure to trade-off cluster quality with redundancy. The overall score of a clustering C is then defined as:

$$F(C) = 2 \cdot \frac{Quality \cdot (1 - Similarity)}{1 + Quality - Similarity},$$

where *Quality* can be instantiated by the ARI_g measure (in the external evaluation) or the Dunn index (in the internal evaluation) and *Similarity* with the ARI_b (in the external evaluation) or the Jaccard index (in the internal evaluation) respectively.

At this point, we should note that our method is not designed to optimize any of these 'objective' clustering measures. However, it is nonetheless interesting to observe that our approach manages to perform well on these independently proposed notions of quality and dissimilarity of clustering results (as will be shown below), arguably extending the credibility of our approach.

Algorithms used for comparison We compare the performance of our methods with three well-known alternative clustering algorithms, namely the COALA (Bae and Bailey 2006), ADFT (Qi and Davidson 2009) and minCEntropy$^+$ (Vinh and Epps 2010) algorithms.

The COALA algorithm is based on an agglomerative hierarchical clustering algorithm. A set of 'cannot-link' constraints encode the prior knowledge provided by the initial clustering. Merging of pairs that satisfy these constraints are performed according to quality and dissimilarity criteria calculated using as a similarity function, the 'average linkage' distance, i.e. the Euclidean distance of all pairwise objects between clusters. A manually tuned parameter ω is used to establish the trade-off between quality and dissimilarity.

On the other hand, the ADFT algorithm transforms the original data using suitable distance metrics in a way that points clustered together initially are less likely to be clustered together in the alternative clustering. Then clustering on the transformed data using any traditional clustering algorithm is performed.

The minCEntropy$^+$ algorithm is an information-theoretic approach. It searches for an alternative clustering that minimizes the conditional entropy of the data (quality objective) while maximizing its entropy conditioned on previous discovered clusterings (dissimilarity objective). A manually tuned parameter λ is used to control the trade-off between these two objectives.

Unlike our methods, the COALA and ADFT algorithms take account of only one clustering as prior information and they require an existing clustering as an initial seed. This restricts our experimental setting to a single iteration. The required initial clustering was computed by the MaxEntLinear method and it was common for all methods.

Parameter settings For all experiments we consider $\boldsymbol{\Sigma} = \mathbf{I}$ and $\boldsymbol{\mu} = \mathbf{0}$. The K-Means algorithm was restarted 100 times. When using the RBF kernel version of our algorithm, the kernel parameter σ was set as the median value of the Euclidean distances between all pairs of data points. When using the COALA algorithm we set the parameter ω to its default value. The default values for the trade-off parameter λ and the kernel parameter σ were used for the minCEntropy$^+$ algorithm. For convenience we refer to the linear variant of our method as MaxEntLinear and the RBF-kernel variant as MaxEntRBF.

6.1.1 Datasets

We used two well known real-world UCI datasets (Blake and Merz 1998), namely the CMUFace and the Digits datasets, for our experiments. The CMUFace dataset contains 640 grayscale images of 20 individuals, i.e. 32 images for each individual. The resolution of each image is 32×30, which provides 960 features of integer values between 1 and 256. This dataset is well suited for alternative clustering methods since it contains 4 different labellings with varying number of labels in each of them. More in detail, each image presents an individual in different poses (straight, up, right, left), different expressions (angry, happy, sad, neutral) and wearing sunglasses or not.

The Digits dataset consists of 5620 images of handwritten digits. Each image contains one digit between 0 and 9. The value of each pixel in the image varies between 1 and 16 and the image resolution is 8×8, providing 64 features.

We note that the same datasets were used for the experimental evaluation of Cui et al. (2007). The CMUFace dataset was also used in Niu et al. (2010) and Vinh and Epps (2010). Here we follow closely and expand the settings of their experiments. In the next subsections we refer to the results presented in these papers and compare them with the ones obtained by our methods.

We also created two artificial datasets. The first one consists of four clusters of 25 two-dimensional points each. The second one contains a central cluster of 20 two-dimensional points and two half-moon shaped clusters around this central cluster containing 40 points each.

6.2 Experiments with artificial data

6.2.1 Visualization

As our first experiment we generate three clusterings for the datasets at hand using both versions of our algorithm. Figures 1 and 2 depict the results for the first and the second dataset respectively. In both cases the results show that the algorithm produces non-redundant results not only between consecutive iterations but also between all pre-discovered clusterings. Moreover the clusters in each clustering make intuitive sense as they consist mostly of mergings of the initial clusters in the dataset. A small number of erratic points appear most probably due to the fact that we are only approximating the ideal solution by solving the relaxed eigenvalue problem.

The ΔQ value for each cluster is displayed on the top of every plot in Figs. 1(a, b) and 2(a, b). We observe that for all cases this value is relatively high for the first iteration and drops rapidly in the next ones. This is natural as the algorithm is totally unaware of the presence of any clusters in the first iteration. From iteration two onwards successively more information is encoded typically resulting in smaller values for ΔQ.

6.2.2 Comparison with existing methods

Next we compare our methods with the ADFT, COALA and minCEntropy$^+$ algorithms. The discovered clusterings are presented in the bottom plots of Figs. 1 and 2 for the two datasets.

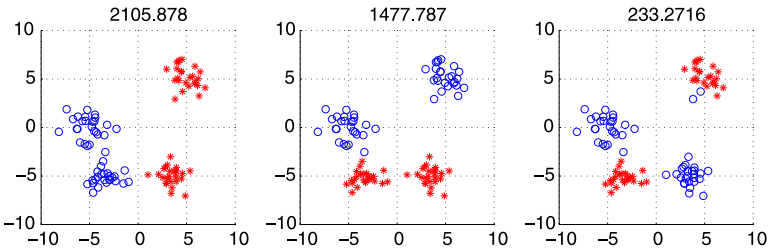
The first plot in the top row for each set of plots is the initial clustering for all methods (computed by MaxEntLinear). For the first dataset we observe that our methods, ADFT and minCEntropy$^+$ produce nearly identical results. Essentially they all perform a linear separation of the data orthogonal to the one performed in the initial clustering. On the other hand, COALA produces a different clustering. One can argue that this solution is more redundant with the initial clustering than the solution computed by the other methods.

Regarding the second dataset (bottom plot in Fig. 2) again MaxEntLinear, ADFT and minCEntropy$^+$ produce almost the same results. It again consists of a linear separation of the data almost orthogonal to the separation performed in the first clustering. The MaxEntRBF separates the central cluster from the rest which is the solution with the highest quality and dissimilarity. It is made prominent that the MaxEntRBF method can easily identify non-linearly shaped clusters. COALA generates a clustering which overlaps largely with the initial one.

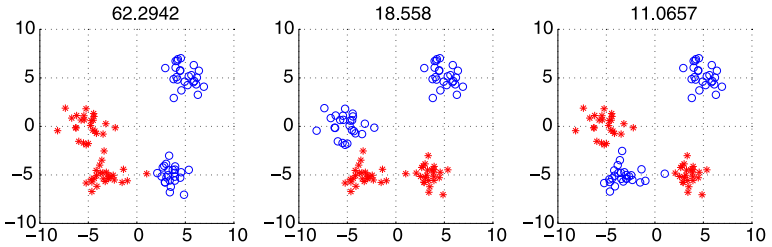
6.3 Digits dataset

6.3.1 Visualization and external evaluation

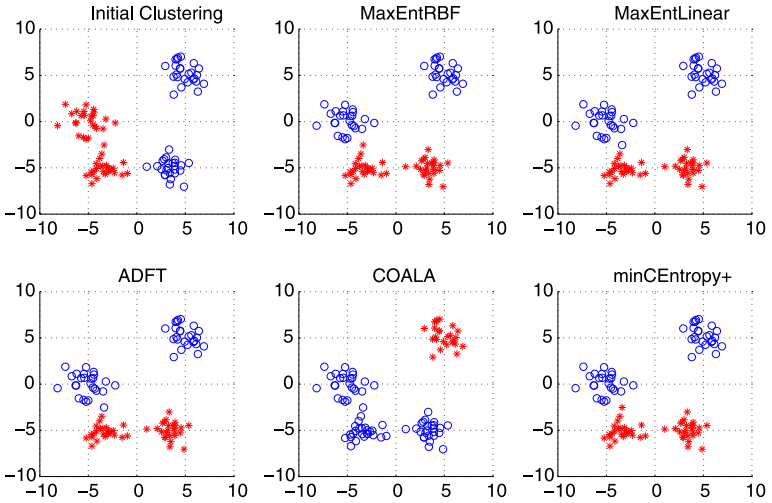
The Digits dataset provides a single ground truth labelling for ten clusters. However, we do not require a number of clusters in each clustering to be equal to the number of clusters in the ground truth labelling since we want to avoid overfitting the model already in the



(**a**) Three clusterings for the *first artificial dataset* using the MaxEnt method with a *linear* kernel. The value on the top of each plot is the ΔQ measure

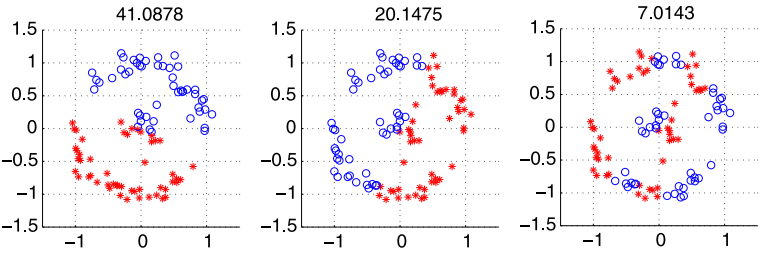


(**b**) Three clusterings for the *first artificial dataset* using the MaxEnt method with a *RBF* kernel. The value on the top of each plot is the ΔQ measure

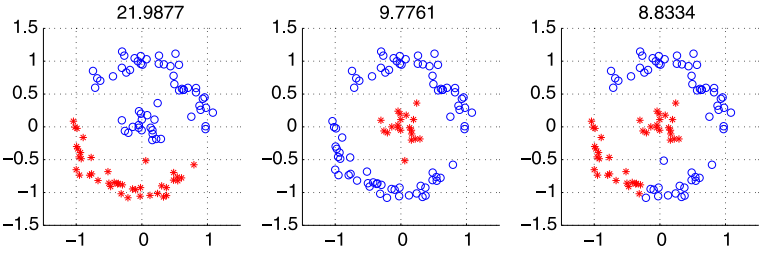


(**c**) Comparison with existing methods for one iteration of each algorithm for *the first artificial dataset*. The first plot on the first row is the initial clustering, the second in the top row is an alternative clustering generated by the MaxEntRBF method, the third in the top row by the MaxEntLinear method, the first in the bottom row by the ADFT algorithm, the second in the bottom row by the COALA algorithm and the third in the bottom row by the minCEntropy$^+$ algorithm

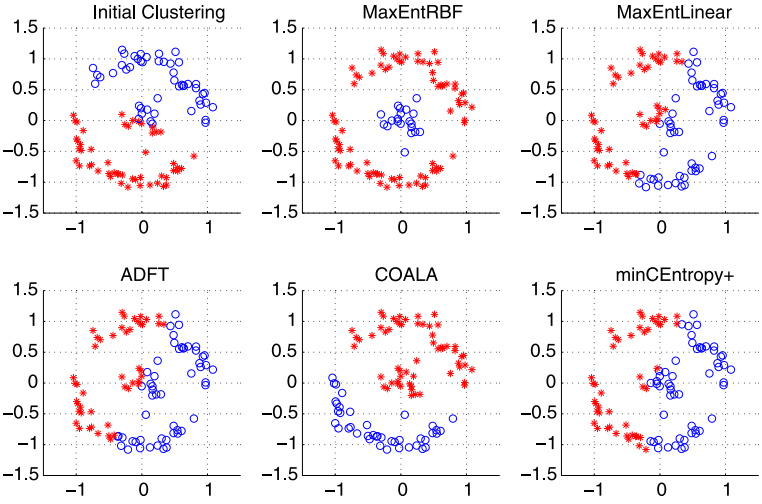
Fig. 1 Clusterings created in three consecutive iterations by the MaxEnt method using a linear kernel (*top image*) and a RBF kernel (*middle image*) for the *first artificial dataset*. A comparison with existing measures is presented in the *bottom image*



(**a**) Three clusterings for the *second artificial dataset* using the MaxEnt method with a *linear* kernel. The value on the top of each plot is the ΔQ measure



(**b**) Three clusterings for *the second artificial dataset* using the MaxEnt method with a *RBF* kernel. The value on the top of each plot is the ΔQ measure



(**c**) Comparison with existing methods for one iteration of each algorithm for the *second artificial dataset*. The first plot on the first row is the initial clustering, the second in the top row is an alternative clustering generated by the MaxEntRBF method, the third in the top row by the MaxEntLinear method, the first in the bottom row by the ADFT algorithm, the second in the bottom row by the COALA algorithm and the third in the bottom row by the minCEntropy$^+$ algorithm

Fig. 2 Clusterings created in three consecutive iterations by the MaxEnt method using a linear kernel (*top image*) and a RBF kernel (*middle image*) for the *second artificial dataset*. A comparison with existing measures is presented in the *bottom image*

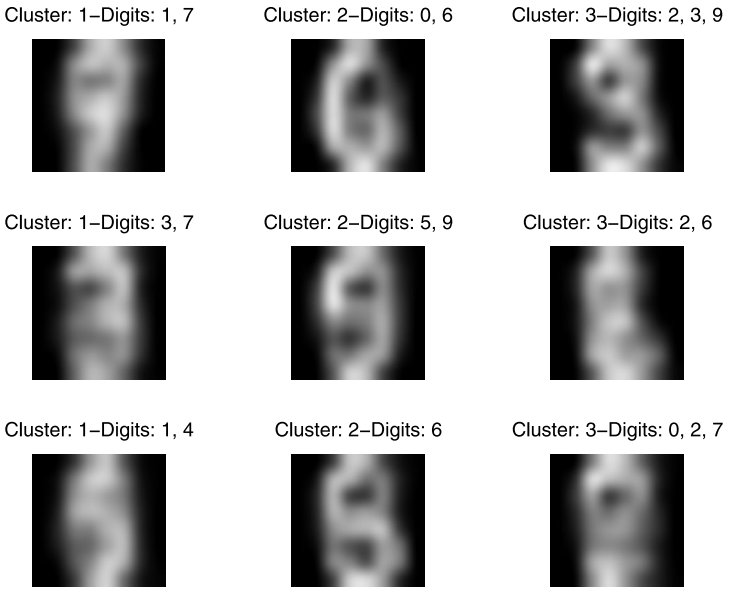


Fig. 3 Mean images for 3 clusterings of 3 clusters for the *Digits* dataset using a *RBF* kernel. The values on the *top* of the images present the recognized digits in each cluster. Adigit is recognized if more than 70 % of its images are contained in one cluster

first iteration. Instead, we generate five clusterings each one containing three clusters. The same experiment was performed in Cui et al. (2007). We compare the performance of our algorithm with the algorithm in Cui et al. (2007) without presenting the results from this paper here. The reader can refer to Cui et al. (2007) for the complete results.

Figure 3 depicts the mean image for each cluster for three iterations performed by our algorithm using the RBF kernel variant. Following the settings of Cui et al. (2007), we consider a digit recognized if more than 70 % of its instances are clustered in the same cluster. The Digits labels above each plot contains the digits recognized by the corresponding cluster. Table 1 presents the confusion matrix for the whole set of five iterations performed.

Regarding the discovered clusterings, we observe that all digits, except '8', are recognized and clustered in the course of 5 iterations. In the first iteration we recognize 7 out of 10 digits. The number of recognized digits remains high for the next two iterations (six in both of them) and drops in the final two. This is something we expect as an increasing amount of information is encoded into our model with each iteration. However, in iteration 5 digit '8', the only unrecognised digit, gets its higher value over all iterations and almost reaches the recognition threshold with 69 % of their instances clustered in the same cluster.

With respect to the redundancy, we note that no pair of recognized digits are recognized together in more than one cluster for all iterations which suggests that intuitively a highly non-redundant set of clusterings is generated.

Table 2 presents the ARI_g values for the five clusterings. We observe a decrease in the ARI_g value with the number of iterations. However, this decrease is not rapid, indicating that the method is able to produce high quality clusterings for a relatively high number of iterations.

Table 1 Confusion Matrix for the *Digits* dataset using the *RBF* kernel for the inner product calculation. *Bold* faced numbers correspond to recognized digits

Digit	Iteration 1			Iteration 2		
	Cluster 1	Cluster 2	Cluster 3	Cluster 1	Cluster 2	Cluster 3
Zero	1	**505**	48	200	353	1
One	**470**	73	28	157	37	377
Two	20	22	**515**	117	7	**433**
Three	48	4	**520**	**458**	25	89
Four	282	286	0	181	310	77
Five	199	69	290	6	**519**	33
Six	1	**557**	0	12	38	**508**
Seven	**562**	0	4	**520**	25	21
Eight	280	60	214	95	182	277
Nine	156	2	**404**	127	**426**	9
Digit	Iteration 3			Iteration 4		
	Cluster 1	Cluster 2	Cluster 3	Cluster 1	Cluster 2	Cluster 3
Zero	50	21	**483**	**528**	20	6
One	**502**	1	68	290	140	141
Two	66	1	**490**	38	233	286
Three	299	193	80	185	59	328
Four	**433**	90	45	14	**501**	53
Five	175	105	278	61	51	**446**
Six	107	**405**	46	56	159	343
Seven	5	92	**469**	10	299	257
Eight	110	197	247	287	235	32
Nine	239	267	56	**425**	51	86
Digit		Iteration 5				
		Cluster 1		Cluster 2		Cluster 3
Zero		74		142		338
One		281		240		50
Two		153		60		344
Three		195		318		59
Four		152		71		345
Five		287		259		12
Six		244		214		100
Seven		5		328		233
Eight		381		140		33
Nine		77		251		234

Figure 4 and Tables 2 and 3 present corresponding results using the linear kernel for the computation of inner products. The results obtained are qualitatively similar with these obtained using the RBF variant. Again all symbols, except '8', are recognized, the first iteration produces identical results and the ARI_g values are of the same scale.

Table 2 ARI_g values between different clusterings and the ground truth labels for the *Digits* dataset

ARI_g	Clustering 1	Clustering 2	Clustering 3	Clustering 4	Clustering 5
MaxEntRBF	0.2029	0.1638	0.1241	0.1236	0.0638
MaxEntLinear	0.2031	0.1400	0.1212	0.0927	0.0804

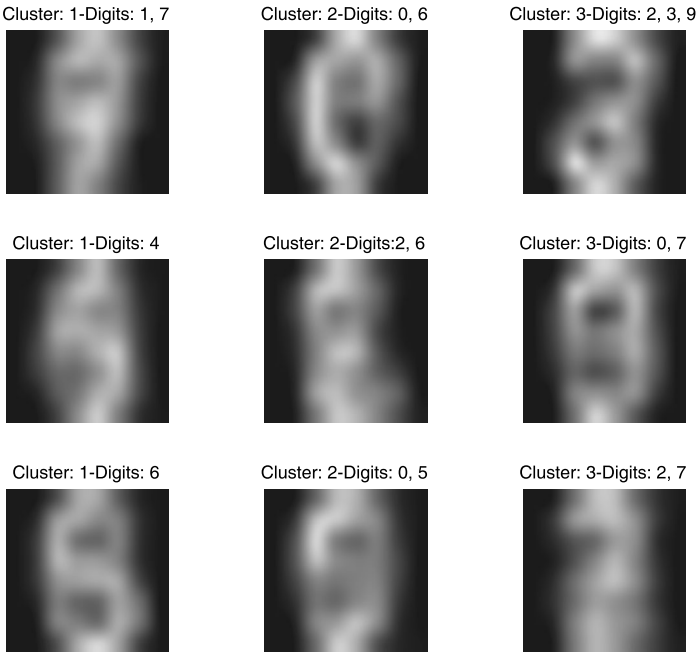


Fig. 4 Mean images for 3 clusterings of 3 clusters for the *Digits* dataset using a *linear* kernel. The values on the *top* of the images present the recognized digits in each cluster. A digit is recognized if more than 70 % of its images are contained in one cluster

6.3.2 Comparison with existing methods

Table 4 presents a comparison between our methods and the ADFT, COALA and minCEntropy$^+$ algorithms for the Digits dataset. We perform two iterations for the RBF variant of the MaxEnt methods (the first iteration provides the initial clustering for the other algorithms and we assess our method in the second clustering) and a single iteration for the rest of them. Each clustering in this setting has five clusters.

The RBF variant of the MaxEnt method and the minCEntropy$^+$ algorithm produce the clusterings of the highest quality measured using the ARI_g and the Dunn Index respectively. The most diverse solution was generated by the minCEntropy$^+$ algorithm and the Linear MaxEnt variant for the ARI_b and Jaccard index respectively. Measuring simultaneously quality and diversity, however, the MaxEnt RBF method provides the highest rated solution for both F-measures used in this paper (i.e. using both the external and the internal evaluation measures). Values for the COALA algorithm are missing because the provided implementation failed to execute due to memory overflow issues.

Table 3 Confusion matrix for 5 iterations for the *Digits* dataset using the *standard inner product* for the computation of inner products. *Bold* faced numbers correspond to recognized digits

Digit	Iteration 1			Iteration 2		
	Cluster 1	Cluster 2	Cluster 3	Cluster 1	Cluster 2	Cluster 3
Zero	1	**505**	48	11	1	**542**
One	**470**	74	27	285	284	2
Two	20	22	**515**	22	**498**	37
Three	48	4	**520**	231	90	251
Four	282	286	0	**462**	50	56
Five	198	69	291	125	200	233
Six	1	**557**	0	79	**470**	9
Seven	**563**	0	3	7	136	**423**
Eight	281	60	213	119	346	89
Nine	157	2	**403**	187	9	366
Digit	Iteration 3			Iteration 4		
	Cluster 1	Cluster 2	Cluster 3	Cluster 1	Cluster 2	Cluster 3
Zero	62	**449**	43	254	15	285
One	158	176	237	168	356	47
Two	5	141	411	**403**	42	112
Three	180	44	348	78	386	108
Four	143	230	195	**457**	92	19
Five	115	**436**	7	**468**	72	18
Six	**478**	37	43	186	224	148
Seven	28	21	517	180	81	305
Eight	292	197	65	84	97	373
Nine	326	214	22	314	111	137

Digit	Iteration 5		
	Cluster 1	Cluster 2	Cluster 3
Zero	31	177	346
One	2	276	293
Two	49	357	151
Three	62	228	282
Four	**394**	70	104
Five	388	137	33
Six	248	228	82
Seven	**413**	120	33
Eight	98	82	374
Nine	101	208	253

6.4 CMUFace dataset

Our algorithm deals naturally with different number of clusters for each iteration. Since we are already aware of different underlying structures in the CMUFace dataset, we develop an iterative scheme different from the one used in the Digits dataset. We develop a four-

Table 4 Comparison of different methods using two sets of measures for the *Digits* dataset. ARI_g computes the overlap of the discovered clustering with the ground truth labels. ARI_g and DI are measures of quality and higher values are better. ARI_b computes the overlap between the discovered clusterings. ARI_b and J are measures of dissimilarity between clusterings. Lower values are better. The F-measure indicates a trade-off between quality and dissimilarity measures

Methods	Set 1			Set 2		
	ARIg	ARIb	F	DI	J	F
MaxEntRBF	**0.2134**	0.0467	**0.3480**	0.7335	0.0918	**0.8116**
MaxEntLinear	0.2031	0.0382	0.3354	0.6904	**0.0873**	0.7861
ADFT	0.0692	0.0251	0.1292	0.6071	0.1037	0.7239
COALA	NA	NA	NA	NA	NA	NA
minCEntropy$^+$	0.2092	**0.0188**	0.3449	**0.7535**	0.1545	0.7968

Fig. 5 Average images for 20 clusters generated by the first iteration of the algorithm using the *standard inner product* version (linear kernel) for the *CMUFace* dataset. Each image correspond to a different individual. Values above each image indicate the percentage of the individual's images inside the cluster (L value) and the percentage of the individual's images inside the specified cluster

iterations setting with 20, 4, 4, and 2 clusters respectively in each clustering. We hope that each iteration will reveal a clustering that correlates mostly with the corresponding ground truth label.

6.4.1 Visualization and external evaluation

Figure 5 shows the average image for the 20 clusters generated in the first iteration using the standard inner product version of our algorithm. The L value above each image is the

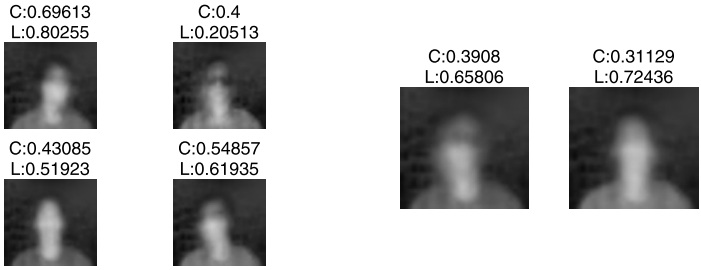


Fig. 6 Average images for 4 and 2 clusters generated by the second and fourth iteration of the algorithm using the *standard inner product* version (linear kernel) for the *CMUFace* dataset. The leftmost set of images corresponds to the pose labelling of the dataset and the rightmost set the sunglasses labelling

ratio of the number of individual's images inside the corresponding cluster over the total number of the individual's images in the dataset. The C value is the ratio of the number of individual's images over the total number of images in the cluster.

From visual inspection it is prominent that the vast majority of the images correspond to different individuals. Furthermore, L and C values indicate that seven of the individuals were clustered in clusters that contain all their images and none else. In total 10 clusters present values above 0.85 for both L and C. Comparing with the results in Cui et al. (2007), we obtain a perfect clustering of a person, i.e. $L = 1$, for 10 clusters, significantly larger than 7 and 8 which were obtained by the algorithms in Cui et al. (2007) and minCEntropy$^+$ algorithm respectively (C values are not reported in either Cui et al. 2007 and Vinh and Epps 2010).

Figure 6 presents average images for iterations 2 and 4 respectively. In this case we observe that the second clustering correlates with the labelling of the images according to the pose of the individual and the fourth clustering with the 'sunglasses' label. Similar results are obtained in Cui et al. (2007) for iterations two and four.

For the second clustering we get the lowest value of L for clusters 2 (second on top in Fig. 6). This image depicts a person mostly in two very similar posses, front and up, clearly wearing sunglasses. It can be argued that the algorithm discriminates based on a combination of the different labellings rather than based on single ones. In particular for the second clustering the logical expression '(up OR front) AND sunglasses' is depicted. Consequently since we are computing L and C values in single labellings their values are small. The third cluster, which depicts the expression '(up OR front) AND NOT sunglasses' is also affected. The labelling according to facial expressions was not retrieved by either our method or the algorithms in Cui et al. (2007) and Vinh and Epps (2010) due to the same effect. Nevertheless the resulting clusterings are of high quality and make intuitive sense.

The correlation of each clustering with the labellings mentioned above is backed up by the ARI_g values presented in the first row of Table 5. The table presents maximum ARI_g values between the discovered solution and all ground truth labellings for iterations 1 to 4. These values were indeed observed for labellings 1 (individuals), 2 (poses), 2 (poses) and 4 (sunglasses) respectively.

Figures 7 and 8 present the clusterings discovered using our algorithm with a RBF kernel for the computation of inner products.

Qualitatively, the discovered clusterings are similar with the ones obtained by the linear methods. Table 5 contains the ARI_g values for each iteration. They are higher than the ones

Table 5 ARI_g values between different clusterings and the ground truth labels generated for the *CMUFace* dataset. Results displayed refer to the labelling with the maximum ARI_g. These labellings are 1, 2, 2, 4 for clusterings 1, 2, 3, 4 respectively for both the standard inner product and the RBF kernel

ARI_g	Clustering 1	Clustering 2	Clustering 3	Clustering 4
MaxEntLinear	0.7179	0.2127	0.0863	0.0434
MaxEntRBF	0.7534	0.2772	0.1152	0.0162

Fig. 7 Average images for 20 clusters generated by the first iteration of the algorithm using an *RBF* kernel for the *CMUFace* dataset. Each image correspond to a different individual. Values above each image indicate the percentage of the individual's images inside the cluster (L value) and the percentage of the individual's images inside the specified cluster

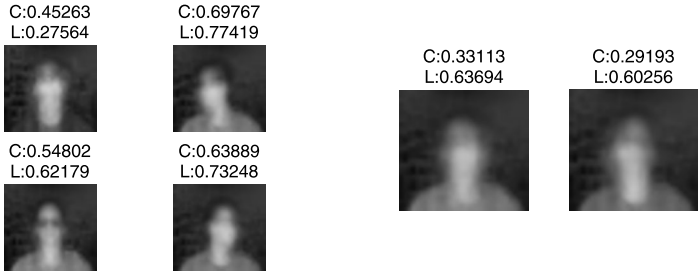


Fig. 8 Average images for 4 and 2 clusters generated by the second and fourth iteration of the algorithm using an *RBF* kernel for the *CMUFace* dataset. The *leftmost* set of images corresponds to the pose labelling of the dataset and the *rightmost* set the sunglasses labelling

Table 6 Comparison of different methods using two sets of measures for the *CMUFaces* dataset. ARI_g computes the overlap of the discovered clustering with the ground truth labels. ARI_g and DI are measures of quality and higher values are better. ARI_b computes the overlap between the discovered clusterings. ARI_b and J are measures of dissimilarity between clusterings. Lower values are better. The F-measure indicates a trade-off between quality and dissimilarity measures

Methods	Set 1			Set 2		
	ARI_g	ARI_b	F	DI	J	F
MaxEntRBF	**0.1341**	0.0254	**0.2358**	**0.6580**	0.0546	**0.7759**
MaxEntLinear	0.0712	0.0171	0.1328	0.6226	0.0633	0.7480
ADFT	0.0013	**0.0026**	0.0026	0.6413	**0.0333**	0.7711
COALA	0.0630	0.0675	0.1180	0.5895	0.0521	0.7269
minCEntropy$^+$	0.1143	0.0212	0.2046	0.6540	0.1252	0.7485

obtained by the linear method, which indicates clusters more correlated with the ground truth labels.

6.4.2 Comparison with existing methods

Next we compare our methods with the ADFT, COALA and minCEntropy$^+$ algorithms using two different sets of quality measures. As before, the initial solution is a clustering with five clusters, it was provided by the MaxEntRBF method and it was common for all methods. For the ARI_g and ARI_b calculation we use the labelling according to individual/person and the other labellings are dropped. The results are displayed in Table 6.

The MaxEntRBF method produces the highest quality results using both the ARI_g and the Dunn index. In addition it presents the highest score for the F-measures indicating that the best trade-of between quality and diversity is achieved by this method (using both the external and the internal evaluation measures). The most diverse solutions were obtained by the ADFT algorithm but with significantly lower quality values as measured by the ARI_g measure.

7 Related work

A number of information-theoretic approaches for finding alternative clusterings have been proposed. In Gondek and Hofmann (2003) (and later refined in Gondek and Hofmann 2004) the Conditional Information Bottleneck (CIB) method was adjusted in the alternative clustering setting. More in particular, the authors associate the quality of a clustering with compression and define constrained optimization problems which compute a clustering that simultaneously presents a good compression and preserves the knowledge of the given clustering.

The NACI algorithm (Dang and Bailey 2010) searches for non-linearly shaped alternative clusterings. It quantifies the quality of a clustering using the quadratic mutual information between the data and the clustering labels. The minimization of mutual information is also employed to measure the dissimilarity between the clustering at hand and a given one. The quality and dissimilarity objectives are combined in a constrained optimization problem.

A similar rationale was implemented in the minCEntropy$^+$ method in Vinh and Epps (2010). The mutual information—or equivalently the conditional entropy—is used again to quantify quality and distinctiveness of clusterings. A hill climbing strategy is employed to

iteratively optimize the conditional entropy objective. One of the most promising features of the method is that it can handle more than one previously discovered clusterings as prior knowledge.

In Dasgupta and Ng (2010) alternative clusterings, 'suboptimal' with respect to a reference—'optimal'—clustering, are discovered using distinct dimensions of the optimal clustering. The 'optimal' clustering is computed as the solution of a relaxed eigenvalue problem motivated by the standard spectral clustering objective of finding the optimal normalized cut in a directed graph. For learning 'suboptimal' clusterings, the same constraint optimization problem is employed but with an increasing number of constraints in every iteration.

The CIB and NACI methods detect only a single alternative clustering. In contrast, our setting naturally deals with the multiple alternative clustering setting.

While being very attractive in the way they handle non-linearity in data, information theoretic approaches proposed so far present some technical difficulties. For example, modelling joint distributions between features and cluster labels in the CIB can be computationally difficult. In order to bypass this obstacle without making any assumptions on the distribution the NACI and minCEntropy$^+$ methods employ complex techniques, such as KL-divergence variants, Havrda-Charvat structural α entropy and Parzen-window density estimation. Our method however, does not require a probabilistic model for the data at all. It rather computes a very simple model for the data miner's prior knowledge, expressed using only simple statistics (mean, variance) on the dataset.

The NACI and minCEntropy$^+$ algorithms involve a manually-tuned parameter to control the trade-off between quality and distinctiveness of clustering results. Although one can argue that this is a desired feature, tuning the parameter would require some user expertise. Our method is essentially parameter-free, with the only parameter involved being the σ value on the kernel setting which is easier to tune.

Furthermore, all methods proposed so far use information theory to quantify how similar cluster indexes are. On the contrary, our method considers the data points directly in the data space for defining both the probabilistic model and the quality measure to be optimized.

Finally, we consider it a strength that the objective function to be optimized in our method is readily derived by instantiating a general data mining framework for the specific case of alternative clustering. This framework has been proven useful in the past for other machine learning and data mining applications, such as principal component analysis (De Bie 2011), frequent itemset mining (De Bie 2010; Kontonasios and De Bie 2010, 2012), multi-relational pattern mining (Spyropoulou and De Bie 2011) and statistical assessment of data mining patterns (Kontonasios et al. 2011).

8 Conclusions

In De Bie (2011) a framework for data mining was introduced, aiming to quantify the subjective interestingness of patterns. There it was shown that Principal Component Analysis can be seen as implementing this framework for a particular pattern type and prior beliefs, thus providing an alternative justification for this method. More importantly, also the potential of the framework in quantifying subjective interestingness for other machine learning and data mining applications was introduced (De Bie 2010, 2011; Kontonasios and De Bie 2010; Spyropoulou and De Bie 2011; Kontonasios et al. 2011).

In the present paper, we showed in detail how the framework can also be applied successfully to the case of clustering, leading to a new approach for alternative clustering that

presents subjectively interesting clusterings in data in an iterative data mining scheme. We showed that our method deals naturally with some of the major challenges in mining multiple clustering solutions and compares favourably with already proposed algorithms.

The core of our alternative clustering approach is based on a generalized eigenvalue problem, such that it can be implemented easily and efficiently using widely-available and highly optimized numerical matrix analysis packages. A further feature of our approach is that it is able to identify nonlinearly shaped clusters, similar to spectral clustering methods.

In further work, we will investigate the quality of the spectral relaxation, and consider the development of tighter relaxations (e.g. to semi-definite programs). We will also further develop links with spectral clustering and other existing clustering approaches, to provide alternative justifications and insights or to improve on these approaches. In addition a more refined approach for the *Description Length* of a clustering will be examined.

Acknowledgements This work is supported by the EPSRC grant EP/G056447/1 and a Centenary Scholarship from the University of Bristol.

References

Bae, E., & Bailey, J. (2006). Coala: a novel approach for the extraction of an alternate clustering of high quality and high dissimilarity. In *Proceedings of the 6th international conference on data mining*, ICDM'06 (pp. 53–62).

Blake, C., & Merz, C. (1998). In *UCI machine learning repository*.

Cui, Y., Fern, X. Z., & Dy, J. G. (2007). Non-redundant multi-view clustering via orthogonalization. In *Proceedings of the 7th IEEE international conference on data mining*, ICDM'07 (pp. 133–142). Los Alamitos: IEEE Comput. Soc.

Dang, X.-H., & Bailey, J. (2010). A hierarchical information theoretic technique for the discovery of non linear alternative clusterings. In *Proceedings of the 16th ACM SIGKDD international conference on knowledge discovery and data mining*, KDD'10 (pp. 573–582). New York: ACM.

Dasgupta, S., & Ng, V. (2010). Mining clustering dimensions. In *Proceedings of the 27th international conference on machine learning*, ICML'10 (pp. 263–270).

Davidson, I., & Qi, Z. (2008). Finding alternative clusterings using constraints. In *Proceedings of the 8th IEEE international conference on data mining*, ICDM'08 (pp. 773–778). Los Alamitos: IEEE Comput. Soc.

De Bie, T. (2010). Maximum entropy models and subjective interestingness: an application to tiles in binary databases. *Data Mining and Knowledge Discovery*, *23*(3), 407–446.

De Bie, T. (2011). An information-theoretic framework for data mining. In *Proceedings of the 17th ACM SIGKDD international conference on knowledge discovery and data mining*, KDD'11 (pp. 564–572).

De Bie, T. (2011). Subjectively interesting alternative clusters. In *Proceedings of the 2nd MultiClust workshop: discovering, summarizing and using multiple clusterings*, September 2011.

Dunn, J. C. (1971). A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *Journal of Cybernetics*, *3*(3), 32–57.

Geng, L., & Hamilton, H. J. (2006). Interestingness measures for data mining: a survey. *ACM Computing Surveys*, *September*, 38.

Gondek, D., & Hofmann, T. (2003). Conditional information bottleneck clustering. In *Proceedings of the 3rd IEEE international conference on data mining, workshop on clustering large datasets*, ICDM'03 (pp. 36–42).

Gondek, D., & Hofmann, T. (2004). Non-redundant data clustering. In *Proceedings of the 4th IEEE international conference on data mining*, ICDM'04 (pp. 75–82). Los Alamitos: IEEE Comput. Soc.

Hersch, J. (1961). Caractérisation variationelle d'une somme de valeurs propres consécutives; généralisation d'inégalités de pólya-schiffer et de weyl. *Comptes Rendus de l'Académie des Sciences, Paris*, *252*, 1714–1716.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. *ACM Computing Surveys*, *31*(3), 264–323.

Jain, P., Meka, R., & Dhillon, I. S. (2008). Simultaneous unsupervised learning of disparate clusterings. *Statistical Analysis and Data Mining*, *1*(3), 195–210.

Jaynes, E. T. (1982). On the rationale of maximum-entropy methods. *Proceedings of the IEEE*, *70*(9), 939–952.

Kontonasios, K.-N., & De Bie, T. (2010). An information-theoretic approach to finding informative noisy tiles in binary databases. In *Proceedings of the 2010 SIAM international conference on data mining*, SDM'10.

Kontonasios, K.-N., & De Bie, T. (2012). Formalizing complex prior information to quantify subjective interestingness of frequent pattern sets. In *11th international symposium on intelligent data analysis*, IDA 2012 (pp. 161–171).

Kontonasios, K.-N., Spyropoulou, E., & De Bie, T. (2012). Knowledge discovery interestingness measures based on unexpectedness. *WIREs Data Mining and Knowledge Discovery*, *2*, 386–399.

Kontonasios, K.-N., Vreeken, J., & De Bie, T. (2011). Maximum entropy modelling for assessing results on real-valued data. In *Proceedings of the 11th IEEE international conference on data mining*, ICDM'11, Vancouver, BC, Canada, December 11–14, 2011 (pp. 350–359). New York: IEEE Press.

McGarry, K. (2005). A survey of interestingness measures for knowledge discovery. *Knowledge Engineering Review*, *20*, 39–61.

Müller, E., Assent, I., Günnemann, S., Krieger, R., & Seidl, T. (2009). Relevant subspace clustering: mining the most interesting non-redundant concepts in high dimensional data. In *Proceedings of the 9th IEEE international conference on data mining*, ICDM'09, Washington, DC, USA (pp. 377–386). Los Alamitos: IEEE Comput. Soc.

Müller, E., Günnemann, S., Farber, I., & Seidl, T. (2010). Discovering multiple clustering solutions: grouping objects in different views of the data. In *Proceedings of the 2010 IEEE international conference on data mining*, ICDM'10, Washington, DC, USA (pp. 1220–1223). Los Alamitos: IEEE Comput. Soc.

Ng, A. Y., Jordan, M. I., & Weiss, Y. (2001). On spectral clustering: analysis and an algorithm. In *Advances in neural information processing systems* (pp. 849–856). Cambridge: MIT Press.

Niu, D., Dy, J. G., & Jordan, M. I. (2010). Multiple non-redundant spectral clustering views. In *Proceedings of the 27th international conference on machine learning 2010*, ICML'10 (pp. 831–838).

Qi, Z.J., & Davidson, I. (2009). A principled and flexible framework for finding alternative clusterings. In *Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining*, KDD'09 (pp. 717–726). New York: ACM.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical Association*, *66*(336), 846–850.

Sequeira, K., & Zaki, M. (2004). Schism: a new approach for interesting subspace mining. In *Proceedings of the 4th IEEE international conference on data mining* (pp. 186–193). Los Alamitos: IEEE Comput. Soc.

Shi, J., & Malik, J. (2000). Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(8), 888–905.

Spyropoulou, E., & De Bie, T. (2011). Interesting multi-relational patterns. In *Proceedings of the 11th IEEE international conference on data mining*, ICDM'11, Vancouver, BC, Canada, December 11–14, 2011 (pp. 675–684). New York: IEEE Press.

Tan, P.-N., Steinbach, M., & Kumar, V. (2005). *Introduction to data mining*. Boston: Addison-Wesley.

Vinh, N. X., & Epps, J. (2010). minCEntropy: a novel information theoretic approach for the generation of alternative clusterings. In *Proceedings of the 10th IEEE international conference on data mining*, ICDM'10 (pp. 521–530).

Wang, H., Yan, S., Xu, D., Tang, X., & Huang, T. (2007). Trace ratio vs. ratio trace for dimensionality reduction. In *IEEE conference on computer vision and pattern recognition*, CVPR'07, June 2007 (pp. 1–8). New York: IEEE Press.