Subjectively Interesting Alternative Clusters

Tijl De Bie

Intelligent Systems Laboratory, University of Bristol, UK
tijl.debie@gmail.com
http://www.tijldebie.net

Abstract. We deploy a recently proposed framework for mining subjectively interesting patterns from data [3] to the problem of clustering, where patterns are clusters in the data. This framework outlines how subjective interestingness of patterns (here, clusters) can be quantified using sound information theoretic concepts. We demonstrate how it motivates a new objective function quantifying the interestingness of (a set of) clusters, automatically accounting for a user's prior beliefs and for redundancies between the clusters.
Directly searching for the optimal set of clusters defined in this way is hard. However, the optimization problem can be solved to a provably good approximation if clusters are generated iteratively, paralleling the iterative data mining setting discussed in [3]. In this iterative scheme, each subsequent cluster is maximally interesting given the previously generated ones, automatically trading off interestingness with non-redundancy. Thus, this implementation of the clustering approach can be regarded as a method for alternative clustering. Although generating each cluster in an iterative fashion is computationally hard as well, we develop an approximation technique similar to spectral clustering algorithms.
We end with a few visual demonstrations of the alternative clustering approach to artificial datasets.

Keywords: Subjective interestingness, alternative clustering

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).

Despite this variety in approaches, the goal of clustering is almost always to provide a user with insight 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 take the perspective that a clustering is more useful if it conveys more novel information. We make a specific choice for a cluster syntax, and we deploy ideas from [3] to quantify the interestingness of a cluster as the amount of information conveyed to the user when told about the cluster's presence.

Our approach attempts to quantify *subjective* interestingness of clusters, in that it takes prior beliefs held by the user into account. As a result, different clusters might be deemed interesting to different users. One particular example is the situation where a user has already been informed about previously discovered clusters in the data, which is the *alternative clustering* setting. In that case, clusters that are individually informative while non-redundant will be the most interesting ones. Our approach naturally deals with alternative clustering, by regarding already communicated clusters as prior beliefs.

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

2 A framework for data mining: a brief overview

For completeness, we here provide a short overview of a framework for data mining that was introduced in [3], and readers familiar with this paper can skip this section. Earlier and more limited versions of this framework, as well as its application to frequent pattern mining, can be found in [4, 2, 5]. 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, and to this end we use 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 one of 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:

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

Such constraints represent the fact that the expected value of certain statistics (the functions f_i) are equal to a given number. The 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.)

We argued in [3] that among all distributions $P \in \mathcal{P}$, the 'best' choice for P is the one of maximum entropy 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 [3].

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 $\mathcal{X}'$. We 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 $-\log\left(P(\mathbf{X} \in \mathcal{X}')\right)$. 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 [3] we 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 [3] we showed 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, and the iterative data mining approach is equivalent to such a greedy approximation.

Thus, the alternative clustering method we will detail below generates clusters in an iterative manner, in such a way that at any time the clusters generated so far are approximately the most informative set of clusters of that size.

3 Subjective interestingness of clusters

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 well known that for data with unbounded domain, this distribution is the multivariate Gaussian distribution with mean and covariance matrix equal to the prescribed cumulants:

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

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 [3] 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 $\mathbf{e}_I \in \mathbb{R}^d$ 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:

$$\frac{1}{n_I}\sum_{i\in I}\mathbf{x}_i = \boldsymbol{\mu}_I,$$
$$\Leftrightarrow \mathbf{X}'\frac{\mathbf{e}_I}{\mathbf{e}_I'\mathbf{e}_I} = \boldsymbol{\mu}_I.$$

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 The self-information of a cluster pattern

The following theorem shows how to assess the self-information of a pattern.

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

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

Thus the self-information of the pattern for a cluster specified by the set I, defined as the negative log probability of the cluster pattern and denoted as $SelfInformation_I$, is equal to:

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

Note that the self-information depends on I only through Q_I, so we may choose to use Q_I as a quality metric 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 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 Sec. 3.4, 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 Sec. 3.4.

3.4 The self-information of a set of cluster patterns

Let us discuss the more general case, where the patterns are constraints of the form $\mathbf{X}'\mathbf{E} = \mathbf{M}$ with $\mathbf{E} \in \mathbb{R}^{n\times k}$ and $\mathbf{M} \in \mathbb{R}^{d\times k}$. Clearly, the type of patterns we wish to consider are a special case of this, with $\mathbf{E} = \frac{\mathbf{e}_I}{\mathbf{e}_I'\mathbf{e}_I}$ and $\mathbf{M} = \boldsymbol{\mu}_I$. Furthermore, it allows us to consider a *composite pattern*, a pattern defined as the union of a set of k patterns. Indeed, if we have k different cluster patterns specified by the sets from $\mathcal{I} = \{I_i\}$, we can write down this set of constraints concisely as $\mathbf{X}'\mathbf{E} = \mathbf{M}$ where $\mathbf{E}$ and $\mathbf{M}$ contain $\frac{\mathbf{e}_{I_i}}{\mathbf{e}_{I_i}'\mathbf{e}_{I_i}}$ and the mean vector $\boldsymbol{\mu}_i$ of the i'th cluster as their i'th columns.

Theorem 2. *Let the columns of the matrix* $\mathbf{E}$ *be the indicator vectors of the sets in* $\mathcal{I} = \{I_i\}$*, and let* $\mathbf{P}_\mathbf{E} = \mathbf{E}(\mathbf{E}'\mathbf{E})^{-1}\mathbf{E}'$*, 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:*

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

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:*

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

Again, since the self-information depends on I only through Q_I, we choose to use Q_I as a quality metric 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:

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

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

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

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:

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

where $\mathbf{P}_\mathbf{E} = \mathbf{E}(\mathbf{E}'\mathbf{E})^{-1}\mathbf{E}'$ is a projection matrix projecting onto the k-dimensional column space of $\mathbf{E}$. □

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

According to the framework, a (composite) pattern specified by matrices $\mathbf{E}$ and $\mathbf{M}$ in this way is thus more informative if trace $\left[\mathbf{P}_\mathbf{E}\cdot(\mathbf{X}-\mathbf{e}\boldsymbol{\mu}')\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\mathbf{e}\boldsymbol{\mu}')'\right]$ is larger. Thus, we could search for the most informative *set* of clusters by maximizing this quality measure with respect to a set $\mathcal{I} = \{I_i\}$ of clusters. This is a hard problem though, even in the case where only one cluster is sought. Thus, we developed an approximation algorithm.

Our approach is approximate in two ways. First, the search for a set of clusters is approximated using a greedy algorithm, searching clusters one by one, thus operating like an alternative clustering algoritm. From [3], it can be seen that this greedy approximation is guaranteed to approximate the true optimum provably well. Second, the search for each cluster is relaxed to an eigenvalue problem. These issues are discussed in greater detail in Sec. 4.

3.5 The cost of describing a cluster

The framework from [3] 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 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 if patterns with smaller sets I are more easy to understand (i.e. have a smaller cost), or vice versa.

4 Alternative clustering: finding the next most informative cluster

Here we discuss an iterative approach to optimizing the quality measure $Q_{\mathcal{I}}$ from Theorem 2. There are two reasons for choosing an iterative approach.

Firstly, directly optimizing the quality measure is equivalent to a set covering type problem (see [3] for more background on why this is the case). While NP-hard, this problem can be approximated well by optimizing over the different clusters (and thus the columns of $\mathbf{E}$) in a greedy iterative manner.

Secondly, usually it is not a priori clear how many clusters 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 clusters (or clusterings) as long as more are desired. Optimizing the quality measure over a growing set of clusters by iteratively optimizing over a newly added column 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.

4.1 The iterative scheme: alternative clustering

To search for the first cluster, we attempt to optimize the quality function from Theorem 1. This is itself a hard problem, but we explain how we approximately solve it in Sec. 4.2.

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

To do this, it is convenient to write the quality measure as a function of the k'th cluster with indicator vector $\mathbf{e}_k$. Let $\mathbf{Q}_{\mathbf{E}} = \mathbf{I} - \mathbf{P}_{\mathbf{E}}$, the projection matrix on the null column space of $\mathbf{E}$. Furthermore, let us denote $\mathbf{E}^* = \left(\mathbf{E}\ \mathbf{e}_k\right)$. Then, using the definition of a projection matrix and the matrix inversion lemma:

$$\begin{aligned}
Q_{\{I_i|i=1:k\}} &= \text{trace}\left[\mathbf{P}_{\mathbf{E}^*} \cdot (\mathbf{X} - \mathbf{e}\boldsymbol{\mu}')\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \mathbf{e}\boldsymbol{\mu}')'\right], \\
&= \text{trace}\left[\mathbf{P}_{\mathbf{E}} \cdot (\mathbf{X} - \mathbf{e}\boldsymbol{\mu}')\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \mathbf{e}\boldsymbol{\mu}')'\right] \\
&\quad + \text{trace}\left[\frac{\mathbf{Q}_{\mathbf{E}}\mathbf{e}_k\mathbf{e}_k'\mathbf{Q}_{\mathbf{E}}}{\mathbf{e}_k'\mathbf{Q}_{\mathbf{E}}\mathbf{e}_k} \cdot (\mathbf{X} - \mathbf{e}\boldsymbol{\mu}')\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \mathbf{e}\boldsymbol{\mu}')'\right], \\
&= Q_{\{I_i|i=1:k-1\}} + \frac{\mathbf{e}_k'\left[\mathbf{Q}_{\mathbf{E}} \cdot (\mathbf{X} - \mathbf{e}\boldsymbol{\mu}')\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \mathbf{e}\boldsymbol{\mu}')'\mathbf{Q}_{\mathbf{E}}\right]\mathbf{e}_k}{\mathbf{e}_k'\mathbf{Q}_{\mathbf{E}}\mathbf{e}_k}.
\end{aligned}$$

Note that if we define $Q_{\emptyset} = 0$ and with $\mathbf{Q}_{\mathbf{E}} = \mathbf{I}$ the quality measure from Theorem 1 is retrieved. We can thus interpret the above reformulation of the quality metric for the k'th cluster conditioned on the first $k-1$ clusters as being the quality metric for a first cluster on data that is projected onto the space

orthogonal to the $k-1$ columns of $\mathbf{E}$, i.e. the $k-1$ previously selected indicator vectors. It is as if the data was deflated to take account of the knowledge of the previously found cluster patterns, thus automatically accounting for redundancy.

4.2 A spectral relaxation of the iterations

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

$$\Delta Q_k = \frac{\mathbf{e}_k' \left[\mathbf{Q_E} \cdot (\mathbf{X} - \mathbf{e}\boldsymbol{\mu}')\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \mathbf{e}\boldsymbol{\mu}')'\mathbf{Q_E}\right] \mathbf{e}_k}{\mathbf{e}_k'\mathbf{Q_E}\mathbf{e}_k}. \tag{2}$$

If we relax the vector $\mathbf{e}_k$ to be real-valued instead of containing only 0's and 1's, this Rayleigh quotient is maximized by the dominant eigenvector of the matrix $\mathbf{Q_E} \cdot (\mathbf{X} - \mathbf{e}\boldsymbol{\mu}')\boldsymbol{\Sigma}^{-1}(\mathbf{X} - \mathbf{e}\boldsymbol{\mu}')'\mathbf{Q_E}$. Thus, as an approximation technique we will use this dominant eigenvector, and threshold it to obtain a crisp 0/1 vector $\mathbf{e}_k$. To determine a suitable threshold, we simply do an exhaustive search over $n+1$ threshold values that generate a different set I of indices $i \in I$ for which $\mathbf{e}_k(i) = 1$, selecting the threshold that maximizes the quantity in Eq. (2).

4.3 A kernel-based version

Note that for $\boldsymbol{\Sigma} = \mathbf{I}$, the quality metrics 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 nonlinearly shaped clusters can be obtained, similar to spectral clustering methods and kernel K-Means.

5 Relations to existing work

There appear to be strong relations between spectral clustering and our spectral relaxation of the problem [7]. Additionally, the quality measure is strongly related to the K-Means cost function [8, 6]. Finally, there seem to be interesting connections to (0-1) SDP problems used for solving combinatorial optimization problems such as clustering (e.g. K-Means and graph cut clustering) [6, 1].

6 Experiments

We conducted 3 experiments, each time reporting the result of 6 iterations of the alternative clustering scheme. Initial prior beliefs are always $\boldsymbol{\mu} = \mathbf{0}$ and $\boldsymbol{\Sigma} = \mathbf{I}$.

- A plain application to a synthetic dataset of 100 points and 2 dimensions in 4 clusters. See Fig. 1.
- An application to the same data but now with a Radial Basis Function (RBF) kernel used for the inner products. See Fig. 2.
- An application with an RBF kernel to a different synthetic dataset, with one central cluster and two half-moon shaped clusters around this central cluster. See Fig. 3.

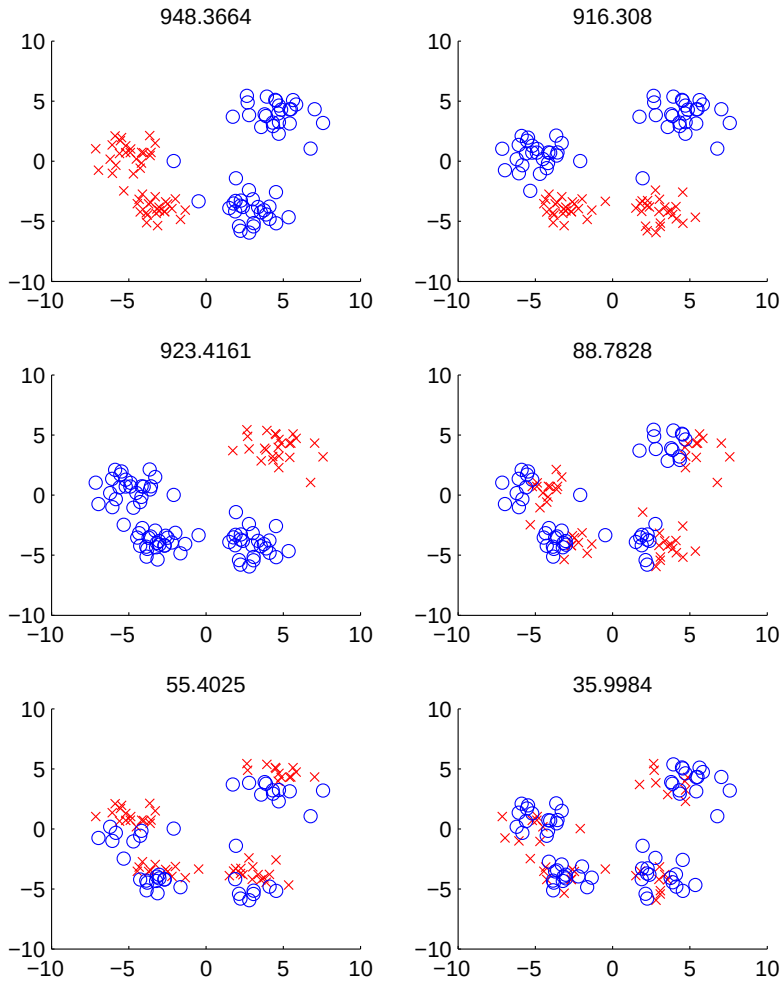

Fig. 1. A synthetic dataset with 25 data points sampled from each of 4 2-dimensional Gaussian distributions with identity covariance matrix and different means. From left to right and top to bottom, the plots show the first 6 consecutive alternative clusters found by our method when a standard inner product is used (data points belonging to the cluster are plotted using crosses). The numbers above the plots are the values of ΔQ_k from Eq. (2) for the cluster shown. Note that it is high for the first 3 clusters, which reveal the enforced cluster structure, before dropping to a much lower level.

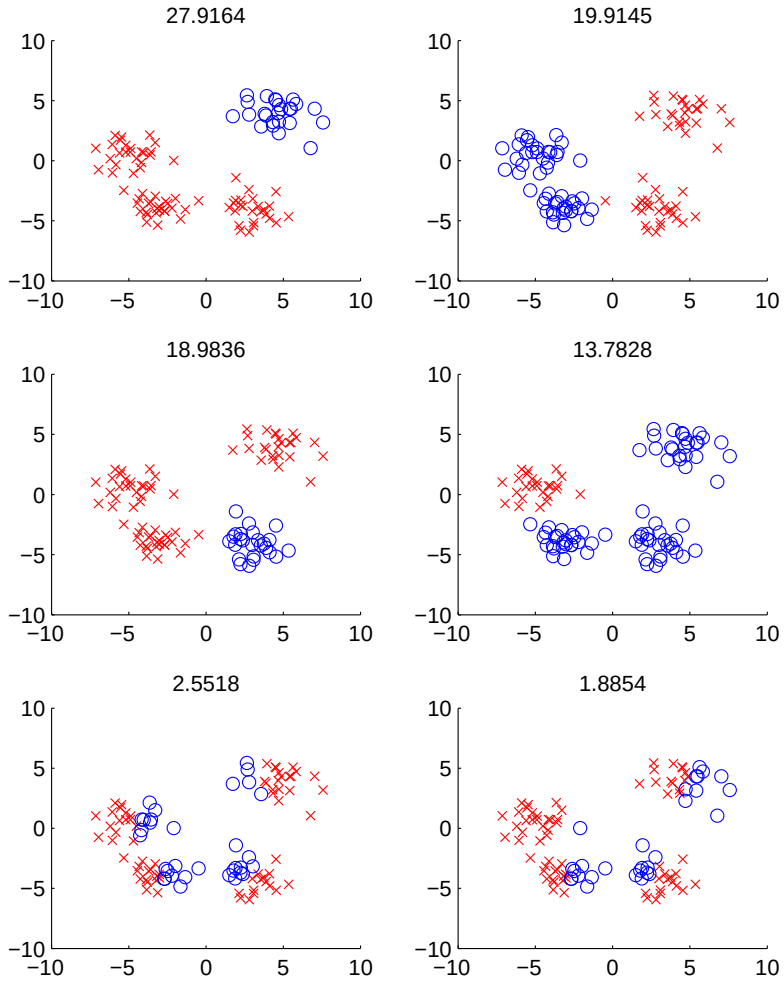

Fig. 2. A synthetic dataset with 25 data points sampled from each of 4 2-dimensional Gaussian distributions with identity covariance matrix and different means. From left to right and top to bottom, the plots show the first 6 consecutive alternative clusters when an RBF kernel with kernel width 3 is used. The numbers above the plots are the values of ΔQ_k from Eq. (2) for the cluster shown.

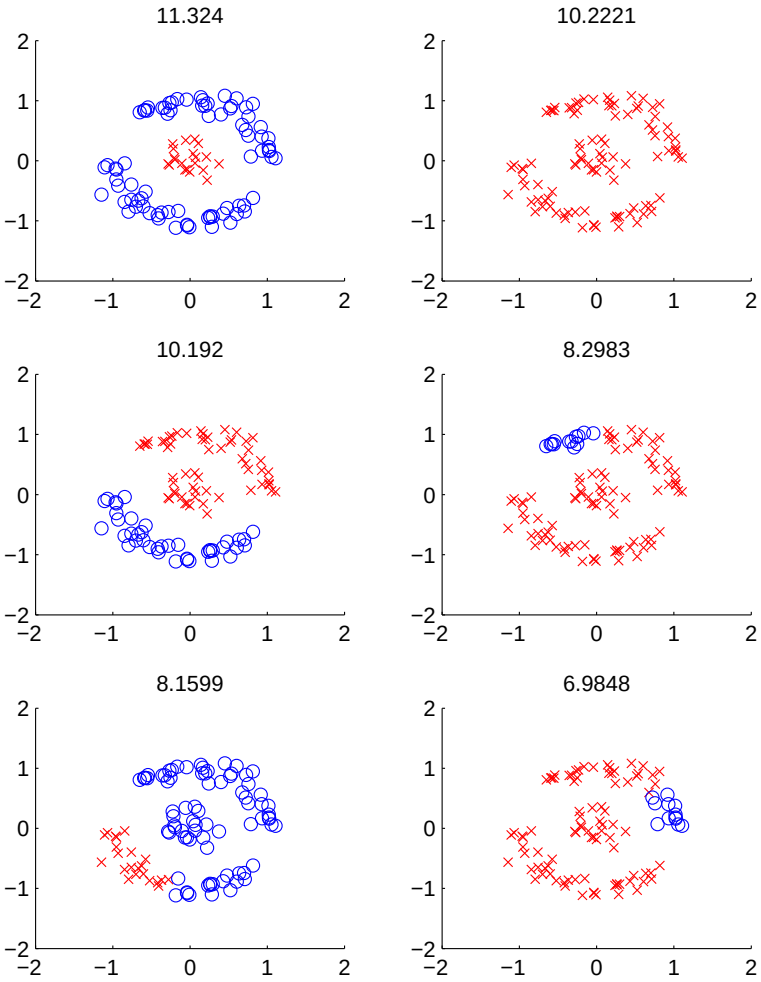

Fig. 3. A synthetic dataset with a central set of 20 data points surrounded by two half moons of 40 data points each. The plots show the first 6 consecutive alternative clusters when an RBF kernel with kernel width 0.3 is used. The numbers above the plots are the values of ΔQ_k from Eq. (2) for the cluster shown. Note that the second cluster generated contains all data points. This is possible and sensible from the perspective of our approach if the mean of the entire data set is significantly different from the expected mean in the initial background model. This may well be the case when working in a Hilbert space induced by the RBF kernel, where all data points lie in one orthant such that their mean cannot be in the origin.

7 Conclusions

In [3] we introduced a framework for data mining, aiming to quantify the subjective interestingness of patterns. We showed 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. In earlier work we also showed the potential of the framework in quantifying subjective interestingness for frequent itemset mining [2, 4, 5]. Now, 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 clusters in data in an iterative data mining scheme.

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. We will also investigate the use of other pattern syntaxes for cluster(ing)s, and the use of more complex types of prior beliefs. Lastly, we plan to demonstrate the power of the framework by also applying these ideas to other types of data and corresponding types of prior beliefs, such as positive real-valued, integer, binary, and more structured types of data.

Due to space constraints, in this workshop paper we could not situate the contributions within the wider literature on alternative clustering. We will rectify this important shortcoming in a later version of this workshop paper.

Acknowledgements
This work is supported by the EPSRC grant EP/G056447/1.

References

1. T. De Bie and N. Cristianini. Fast sdp relaxations of graph cut clustering, transduction, and other combinatorial problems. *Journal of Machine Learning Research*, 7:1409–1436, 2006.
2. T. De Bie. Maximum entropy models and subjective interestingness: an application to tiles in binary databases. Data Mining and Knowledge Discovery, 2010.
3. T. De Bie. An information-theoretic framework for data mining. In *Proc. of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD11)*, 2011.
4. T. De Bie, K.-N. Kontonasios, and E. Spyropoulou. A framework for mining interesting pattern sets. *SIGKDD Explorations*, 2010.
5. K.-N. Kontonasios and T. De Bie. An information-theoretic approach to finding informative noisy tiles in binary databases. In *Proceedings of the 2010 SIAM International Conference on Data Mining*, 2010.
6. Jiming Peng and Yu Wei. Approximating k-means-type clustering via semidefinite programming. *SIAM Journal on Optimization*, 18(1), 2007.
7. J. Shi and J. Malik. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):888–905, 2000.
8. H. Zha, C. Ding, M. Gu, X. He, and H. Simon. Spectral relaxation for k-means clustering. In *Advances in Neural Information Processing Systems 14 (NIPS01)*, pages 1057–1064, 2002.