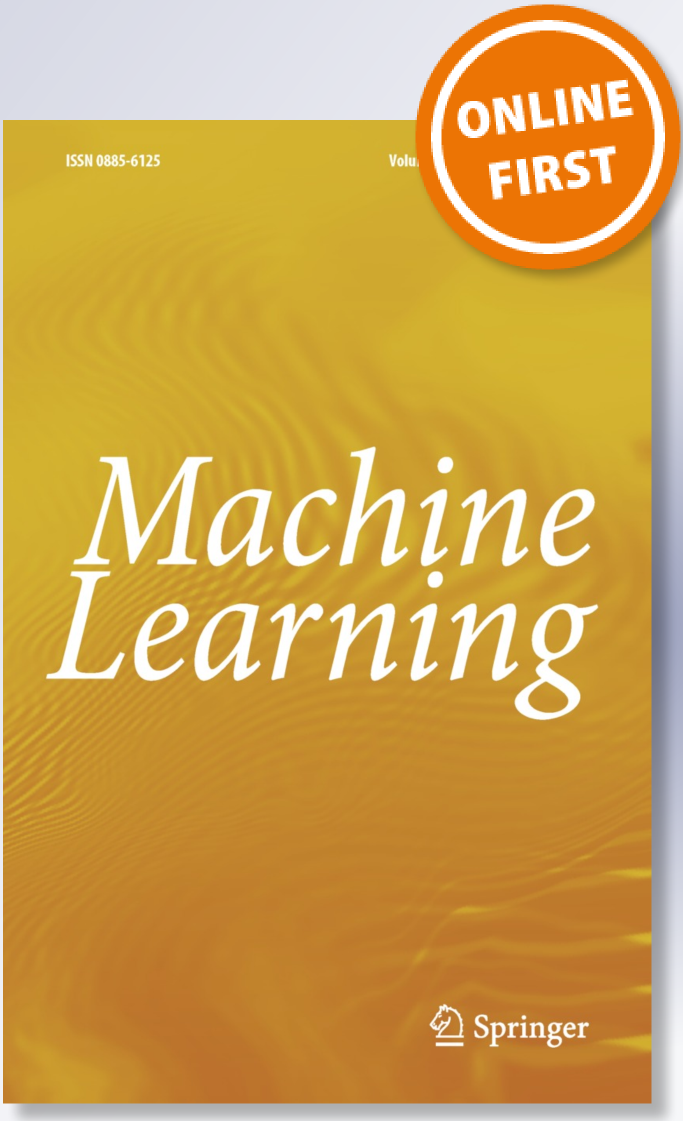
The blind men and the elephant: on meeting the problem of multiple truths in data from clustering and pattern mining perspectives

Arthur Zimek & Jilles Vreeken

Machine Learning

ISSN 0885-6125

Mach Learn
DOI 10.1007/s10994-013-5334-y



Springer

Springer

Mach Learn
DOI 10.1007/s10994-013-5334-y

The blind men and the elephant: on meeting the problem of multiple truths in data from clustering and pattern mining perspectives

Arthur Zimek · Jilles Vreeken

Received: 25 May 2012 / Accepted: 10 February 2013

Abstract In this position paper, we discuss how different branches of research on clustering and pattern mining, while rather different at first glance, in fact have a lot in common and can learn a lot from each other's solutions and approaches. We give brief introductions to the fundamental problems of different sub-fields of clustering, especially focusing on subspace clustering, ensemble clustering, alternative (as a variant of constraint) clustering, and multiview clustering (as a variant of alternative clustering). Second, we relate a representative of these areas, subspace clustering, to pattern mining. We show that, while these areas use different vocabularies and intuitions, they share common roots and they are exposed to essentially the same fundamental problems; in particular, we detail how certain problems currently faced by the one field, have been solved by the other field, and vice versa.

The purpose of our survey is to take first steps towards bridging the linguistic gap between different (sub-) communities and to make researchers from different fields aware of the existence of similar problems (and, partly, of similar solutions or of solutions that could be transferred) in the literature on the other research topic.

Keywords Subspace clustering · Pattern mining · Ensemble clustering · Alternative clustering · Constraint clustering · Multiview clustering

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

A. Zimek (✉)
Department of Computing Science, University of Alberta, Edmonton, AB, Canada
e-mail: zimek@ualberta.ca

J. Vreeken
Advanced Database Research and Modelling, Department of Mathematics and Computer Science, University of Antwerp, Antwerp, Belgium
e-mail: jilles.vreeken@ua.ac.be

Published online: 07 March 2013

1 Introduction

In an old Indian tale, a king has the blind men of his empire examining an elephant. Then he asks about their findings. Their descriptions of what an elephant is, naturally differ widely. The one who felt the ear of the elephant describes it as a hand fan. The others insist that an elephant is like a pot, a ploughshare, a tree trunk, a pillar, a wall, or a brush, depending on whether they felt the elephant's head, tusk, trunk, foot, belly, or the tip of the tail, respectively.

In general for unsupervised data mining, and in particular in its emerging research areas of discovering, summarizing and using *multiple clusterings*, by growing understanding and communication between researchers we see an increase in awareness that each of us can only discover part of the truth a dataset holds. Hence, the better we understand each other's approaches to ultimately the same problem, exploratory analysis of data, the more we can learn from each other; be it from the problems faced, or the (partial) solutions that have been discovered.

In this position paper, surveying different approaches to examine presumably the same elephant, we therefore set out to take first steps towards bridging the linguistic gap between different communities. Our preeminent goal is to make researchers from the different fields aware of the existence of problems in other fields that are very similar to the problems they are facing within their own problem setting, and in particular to raise awareness of the (partial) solutions and theoretical advances made that might be applicable, or in the least learned from and build upon.

Earlier versions of this work appeared at the MultiClust workshops 2010 and 2011 (Kriegel and Zimek 2010; Vreeken and Zimek 2011). Here, we combine and expound these viewpoints, in order to show the bigger picture on how different subfields investigate essentially the same problem. Besides generally extending our discussion, we give more details, include more examples, and have expanded our survey for identifying key references on the considered fields. Furthermore, we follow up discussions presented by Färber et al. (2010) and Kriegel et al. (2011c).

We first, in Sect. 2, set out to discuss similar traits and problems as well as opposing views in specific areas of clustering. As such, we discuss specializations to subspace clustering and ensemble clustering, and we touch on the related aspects of alternative (constraint) clustering and multiview clustering (in the variant aiming at the identification of different clustering solutions). We will also discuss that the fundamental problem, multiple truths, is present in other areas as well but did not find a solution there either. In this wider horizon we touch, e.g., on classification, where a specialization allows for multiple labels or hierarchical classes.

Next, in Sect. 3, we focus on subspace clustering as a representative and detail the many commonalities between clustering and pattern mining. We will show that there is an interesting 'relationship story' (Vreeken and Zimek 2011) between these seemingly different fields, beyond their unsupervised exploratory nature. To this end we will discuss the common roots, similar techniques, and in particular the problems these areas share. We point out problems faced by one area, that in fact have been considered and (partly) solved by the other—albeit with different vocabulary and intuitions.

Finally, in Sect. 4, we round up, summarize the points taken, and suggest further directions to explore.

The research and discussion described in this paper builds upon and extends work published at MultiClust'10 and at MultiClust'11 by Kriegel and Zimek (2010), Vreeken and Zimek (2011).

2 Clustering approaches examining different parts of the elephant

In the following, we first give an overview of the problems and basic solutions of subspace clustering (Sect. 2.1). Afterwards we discuss ensemble clustering approaches (Sect. 2.2) and touch on alternative clustering (Sect. 2.3) and multiview clustering (Sect. 2.4). We summarize similarities and differences between these sub-fields (Sect. 2.5) and we collect questions in these areas where answers from other areas may be helpful in improving the field. Finally (Sect. 2.6), we will also see, however, that the fundamental problem for all these areas, the existence of multiple truths, is acknowledged in other research areas or clustering subtopics as well yet did not find satisfying treatment so far. Let us thus emphasize that we here do neither aim at solving all the open questions, nor aim at unifying different research areas, but instead aim to inspire discussion between different research areas, to put the figure of the elephant together from the different impressions.

2.1 Subspace clustering

Subspace clustering refers to the task of identifying clusters of similar vectors where the similarity is defined w.r.t. a subspace of the data space. The subspace is not necessarily the same (and actually is usually different) for different clusters within one clustering solution. The key-issue in subspace clustering lies in defining similarity over only certain attributes, using possibly different attributes for different (potential) clusters. By how we choose weightings, selections, and/or combinations of attributes, we implicitly choose which subspaces in the data will satisfy our current similarity measure; and hence, we have to choose these properties carefully and appropriately per application domain.

Which subspaces are important for the similarity measure is a task typically left to be learned during the clustering process, since for different clusters within one and the same clustering solution usually different subspaces are relevant. Hence subspace clustering algorithms cannot be thought of as usual clustering algorithms where we simply use a different definition of similarity: the similarity measure and the clustering solution are derived simultaneously and hence strongly depend on each other.

Kröger and Zimek (2009), Kriegel et al. (2012) provide short, merely theoretical overviews on this topic. Kriegel et al. (2009) discussed the problem in depth, also differentiating several partial problems and surveying a number of example algorithms. Latest developments have been surveyed by Sim et al. (2012). Although there is a wide variety of task definitions for clustering in subspaces, the term ‘subspace clustering’ is also often used in a narrower sense to relate to a special category of clustering algorithms in axis-parallel subspaces where the clusters—residing in different subspaces—may overlap. That is, some of the clusters’ objects can be part of several different clusters.

Another family of clustering in axis-parallel subspaces is called ‘projected clustering’, where the dataset is partitioned in a set of disjunct clusters (and, sometimes, an additional set of noise objects). Recent experimental evaluation studies covered some selections of these specific subtypes of subspace clustering (Moise et al. 2009; Müller et al. 2009c) and these subtypes may be the most interesting ones for considering relationships with the other research areas discussed in this survey.

Clustering in axis-parallel subspaces is based on the distinction between relevant and irrelevant attributes. This distinction generally assumes that the variance of attribute values for a relevant attribute over all points of the corresponding cluster is rather small compared to the overall range of attribute values. For irrelevant attributes the variance of values within a given cluster is high or the values are indistinguishable from the rest of the data. For example, one could assume a relevant attribute for a given cluster being normally distributed with a small standard deviation whereas the values of irrelevant attributes are uniformly distributed over the complete data space.

The geometrical intuition of these assumptions relates to the points of a cluster being widely scattered in the direction of irrelevant axes while being densely clustered along relevant attributes. When selecting the relevant attributes only, the cluster would appear as a full-dimensional cluster—within this selected subspace. Whereas in the full dimensional space, also including the irrelevant attributes, the cluster points form a hyperplane parallel to the irrelevant axes as they spread out in these directions. Due to this geometrical appearance, this type of cluster is addressed as 'axis-parallel subspace cluster'.

Since the number of axis-parallel subspaces where clusters can possibly reside scales exponentially with the dimensionality of the data, the main task of research in the field was the development of appropriate subspace search heuristics. Starting from the pioneering approaches to axis-parallel subspace clustering, there have been pursued two opposite basic techniques for searching subspaces, namely (a) *top-down search* (Aggarwal et al. 1999) and (b) *bottom-up search* (Agrawal et al. 1998).

(a) The rationale of top-down approaches is to determine the subspace of a cluster starting from the full-dimensional space. This is usually done by determining a subset of attributes for a given set of points—potential cluster members—such that the points meet the given cluster criterion when projected onto the corresponding subspace.

Obviously a main dilemma is that when determining the subspace of a cluster, we need to know at least some of the cluster members. On the other hand, in order to determine cluster membership, we need to know the subspace of the cluster. To escape from this circular dependency, most of the top-down approaches rely on a rather strict assumption, which has been called the *locality assumption* (Kriegel et al. 2009). It is assumed that the subspace of a cluster can be derived from the local neighborhood of the cluster center or its members. In other words, it is assumed that even in the full-dimensional space, the subspace of each cluster can be learned from the local neighborhood of the cluster. Other top-down approaches that do not rely on the locality assumption use random sampling as a heuristic in order to generate a set of potential cluster members.

(b) For a bottom-up search, we need to traverse the exponential search space of all possible subspaces of the data space. This is strongly related to the search space of the frequent item set mining problem in analysis of market baskets in transaction databases (Agrawal and Srikant 1994).

In our setting, each attribute represents an item and each subspace cluster is a transaction of the items representing the attributes that span the corresponding subspace. Finding item sets with a frequency of at least 1 then relates to finding all combinations of attributes that constitute a subspace containing at least one cluster.[1]

Let us note that there are bottom-up algorithms that do not use an APRIORI-like search, but instead apply other heuristics, such as randomization or greedy-search strategies.

Subspace clustering in a narrower sense pursues the goal to find all clusters in all subspaces of the entire feature space. This goal obviously is defined to correspond to the bottom-up technique used by these approaches, based on some anti-monotonic property of clusters

[1] Later on, we will go more into detail on the relation between subspace clustering and frequent pattern mining; see Sect. 3. Interestingly, the frequent pattern mining principle has been applied to subspace outlier detection far less successfully, for reasons discussed by Zimek et al. (2012).

allowing the application of the APRIORI search strategy.[2] The pioneer approach to finding all clusters in all subspaces coining the term 'subspace clustering' for this specific task has been CLIQUE (Agrawal et al. 1998). Numerous variants have been proposed, e.g., by Cheng et al. (1999), Nagesh et al. (2001), Kailing et al. (2004a), Assent et al. (2007, 2008), Moise and Sander (2008), Müller et al. (2009a), Liu et al. (2009).

Like for frequent itemset mining, one can question the original problem formulation of finding 'all clusters in all subspaces', as by retrieving a huge and highly redundant set of clusters the result will not be very useful or insightful. Subsequent methods therefore often concentrated on possibilities of concisely restricting the resulting set of clusters by somehow assessing and reducing the redundancy of clusters, for example to keep only clusters of highest dimensionality. It also should be noted that the statistical significance of subspace clusters (as defined by Moise and Sander 2008), is not an anti-monotonic property and hence does in general not allow for APRIORI-like bottom-up approaches finding only *meaningful* clusters.

2.2 Ensemble clustering

Ensemble methods are most well-known for boosting performance in classification tasks. In the area of supervised learning, combining several self-contained predicting algorithms to an ensemble to yield a better performance than any of the base predictors, is backed by a sound theoretical background (Hansen and Salamon 1990; Dietterich 2000, 2003; Valentini and Masulli 2002; Brown et al. 2005).

In short, a predictive algorithm can suffer from several limitations such as statistical variance, computational variance, and a strong bias. *Statistical variance* describes the phenomenon that different prediction models result in equally good performance on training data. Choosing arbitrarily one of the models can then result in deteriorated performance on new data. Voting among equally good classifiers can reduce this risk. *Computational variance* refers to the fact, that computing the truly optimal model is usually intractable and hence any classifier tries to overcome computational restrictions by some heuristic. These heuristics, in turn, can lead to local optima in the training phase. Obviously, considering multiple random starting points reduces the risk of ending up in the wrong local optimum.

A restriction of the space of hypotheses a predictive algorithm may create is referred to as *bias* of the algorithm. Usually, the bias allows for learning an abstraction and is, thus, a necessary condition of learning a hypothesis instead of learning by heart the examples of the training data (the latter resulting in random performance on new data). However, a strong bias may also hinder the representation of a good model of the true laws of nature (or other types of characteristics of the data set at hand) one would like to learn. A weighted sum of hypotheses may then expand the space of possible models.

To improve over several self-contained classifiers by building an ensemble of those classifiers requires the base algorithms being accurate (i.e., at least better than random) and diverse (i.e., making different errors on new instances). It is easy to understand why these two conditions are necessary and also sufficient. If several individual classifiers are not diverse, then all of them will be wrong whenever one of them is wrong, as nothing would be gained by voting over wrong predictions.

On the other hand, if the errors made by the classifiers are uncorrelated, more individual classifiers may be correct while some individual classifiers are wrong. Therefore, a majority

[2] There is a certain problematic nature in this circular definition of problem and solution which has been elaborated upon by Kriegel et al. (2012).

vote by an ensemble of these classifiers is more likely to be correct. More formally, suppose an ensemble consisting of k hypotheses, and the error rate of each hypothesis is equal to a certain $p < 0.5$ (assuming a dichotomous problem), though errors occur independently in different hypotheses. The ensemble will be wrong, if more than $k/2$ of the ensemble members are wrong. Thus the overall error rate $\bar{p}$ of the ensemble is given by the area under the binomial distribution, where $k \geq \lceil k/2 \rceil$, that is for at least $\lceil k/2 \rceil$ hypotheses being wrong:

$$\bar{p}\,(k, p) = \sum_{i=\lceil k/2 \rceil}^{k} \binom{k}{i} p^i (1-p)^{k-i}$$

The overall error-rate rapidly decreases for increasingly larger ensembles.

In the unsupervised task of clustering, although a fair amount of approaches has been proposed (see Ghosh and Acharya 2011, for an overview), the theory for building ensembles is still not fully developed.[3] Improvement by application of ensemble techniques has been demonstrated empirically, though. The influence of diversity on ensemble clustering has been acknowledged and studied by Kuncheva and Hadjitodorov (2004), Hadjitodorov et al. (2006), Hadjitodorov and Kuncheva (2007).

Existing approaches have concentrated on creating diverse base clusterings and then combining them somehow to a unified single clustering. The approaches differ in (a) how to create diversity and (b) how to combine different clusterings.

(a) For obtaining diversity, Strehl and Ghosh (2002) discuss (i) non-identical sets of features, (ii) non-identical sets of objects, and (iii) different clustering algorithms.

Clearly, the first of these strategies is related to subspace clustering. It has been pursued in different ensemble clustering approaches (Fern and Brodley 2003; Topchy et al. 2005; Bertoni and Valentini 2005). Usually, however, the projections used here are random projections and not different clusters are sought in different subspaces but true clusters are supposed to be more or less equally apparent in different random projections. It is probably interesting to account for the possibility of different, yet meaningful, clusters in different subspaces. For example, Fern and Brodley (2003) are aware of possibly different clusters existing in different subspaces. Nevertheless, their approach aims at a single unified clustering solution, based on the ensemble framework of Strehl and Ghosh (2002).

(b) For meaningfully combining different clustering solutions, the other dominant question in research on clustering ensembles is how to derive or measure the correspondence between different clustering solutions, see e.g. Topchy et al. (2004, 2005), Long et al. (2005), Fred and Jain (2005), Caruana et al. (2006), Domeniconi and Al-Razgan (2009), Gullo et al. (2009b), Hahmann et al. (2009), Singh et al. (2010). The correspondence between different clusterings is a problem not encountered in classification ensembles where the same class always gets the same label. In clustering, the correspondence between labels assigned by different instances of a clustering learner is to discover by some mapping procedure which makes part of the variety of solutions to ensemble clustering. From the point of view of subspace clustering, this label correspondence problem is interesting as there are no suitable automatic evaluation procedures for possibly highly complex and overlapping clusters over different subspaces.

[3] As a side note, let us remark that the ideas of ensemble learning have begun being applied also in the area of outlier detection, also in an unsupervised manner (Lazarevic and Kumar 2005; Gao and Tan 2006; Nguyen et al. 2010; Kriegel et al. 2011b; Schubert et al. 2012).

It is our impression that these two topics in research on ensemble clustering directly relate to certain questions discussed in subspace clustering:

(a) A lesson research in ensemble clustering may want to learn from subspace clustering could be that diversity of clusterings could be a worthwhile goal in itself. We should differentiate here between significantly differing clusterings and just varying yet similar (i.e., correlated) clusterings (Li and Ding 2008; Azimi and Fern 2009). We believe, however, that it can be meaningful to unify the latter by some sort of consensus while it is in general not meaningful to try to unify substantially different clusterings.

(b) As we have seen above (Sect. 2.1), redundancy is a problem in traditional subspace clustering. Possibly, subspace clustering can benefit from advanced clustering diversity measures in ensemble clustering (Strehl and Ghosh 2002; Fern and Brodley 2003; Hadjitodorov et al. 2006; Gionis et al. 2007b; Fern and Lin 2008). These measures, however, are usually based on some variant of pairwise mutual information where the overlap of clusters (i.e., the simultaneous membership of some subsets of objects in different clusters) is a problem.

Interestingly, there have been recent attempts to combine techniques from both fields, called 'Subspace Clustering Ensembles' (Domeniconi 2012) or 'Projective Clustering Ensembles' (Gullo et al. 2009a, 2010, 2011).

2.3 Alternative (or: constraint) clustering

While in classification (supervised learning) a model is learned based on complete information about the class structure of a data set, and in clustering (cast as unsupervised learning) a model is fit to a data set without using prior information, semi-supervised clustering is using *some* information.

For example, some objects may be labeled, and this class information is used to derive must-link constraints (objects with same labels should end up in the same cluster) or cannot-link constraints (objects with different class labels should be separated into different clusters). In a weaker setting, we could treat these as should-link and should-not-link constraints and only demand a solution to satisfy a certain percentage of these constraints. However, constraint clustering does come in many variants (e.g. Klein et al. 2002; Bade and Nürnberger 2008; Basu et al. 2008; Davidson and Ravi 2009; Lelis and Sander 2009; Zheng and Li 2011).

Interesting in our context is one specific variant of constraint clustering, which imposes a constraint of diversity, or non-redundancy, on the results. In this case, the constraints are based on a given clustering but in an opposite way to what is the common use of class labels. Objects that are clustered together in the given clustering *should not* be clustered together in a new clustering. The clustering process should result in the discovery of different clustering solutions. Hence this variant is known as *alternative clustering*. In the algorithmic variants, these different solutions can be derived sequentially or simultaneously (some variants have been discussed, e.g., by Gondek and Hofmann 2004, 2005; Bae and Bailey 2006; Davidson and Qi 2008; Jain et al. 2008; Davidson et al. 2010; Niu et al. 2010; Dang and Bailey 2010; Dang et al. 2012).

The motivation behind these approaches is that some results may already be known for a specific data set yet other results should be obtained that are new and interesting (cf. the classical definition of knowledge discovery in data as '*the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data*' by Fayyad et al. (1996)). Reproducing known or redundant patterns by clustering algorithms does not qualify as identifying *novel* patterns. A problem could be that the already known facts are dominant in a data set. The idea is, thus, to constrain some clustering algorithm by a set of patterns or links between objects that are *not* to identify. As Gondek and Hofmann (2005) put it: '*users are often unable to* positively *describe what they are looking for, yet may be perfectly capable of expressing what is* not *of interest to them*'.

Some of these approaches again use ensemble techniques. Here, however, we are more interested in the relationship between these approaches and the area of subspace clustering. Hence, for our context, the interesting idea in this area is to use different subspaces as one possibility to find different clustering solutions (Qi and Davidson 2009). As these approaches seek diversity usually constrained by non-redundancy, clearly subspace clustering research tackling the high redundancy level of subspace clusters can learn from these approaches.

Alternatively, allowing a certain degree of redundancy can be meaningful, as in turn can be learned from subspace clustering. In different subspaces, one subset of objects could belong to two different but meaningful clusters and hence increase the redundancy level of these clusters without rendering the report of both overlapping clusters meaningless. Indeed, considerations in this direction can actually be found in the research area of subspace clustering (Günnemann et al. 2009). Also on part of alternative clustering research it has been conceded that it may be desirable to enable the preservation of certain already known properties of known concepts while seeking different clusters (Qi and Davidson 2009).

2.4 Multiview clustering

There are methods of clustering (and other data mining tasks) that treat different subsets of attributes or different data representations separately. The intuition of this separation could be a semantic difference of these subsets such as color features versus texture features of images. Learning distances in the combined color-texture space may easily end up being a completely meaningless effort. Hence, the different representations are treated separately as they are *semantically* different. Examples for this idea are so called 'multi-represented clustering' approaches or distance-learning in multiple views (e.g. Kailing et al. 2004b; Achtert et al. 2005, 2006c; Kriegel et al. 2008; Horta and Campello 2012). Especially in the setting of 'co-learning', i.e., learning one concept simultaneously using different sources of information, the goal is here to find one clustering (or, more general, one similarity measure) guided by the information from different (and separated) subspaces (Kumar and Daumé 2011; Kriegel and Schubert 2012). This work has its roots in the semi-supervised concept of co-training (Blum and Mitchell 1998) and their compatibility-assumption. While the transfer to the unsupervised setting of clustering has often been named 'multi-represented' clustering, the name 'multi-view' clustering was used for such approaches as well (Bickel and Scheffer 2004; Sridharan and Kakade 2008; Chaudhuri et al. 2009). There are even approaches that connect this idea with the idea of having constraints (see Sect. 2.3) that can guide the distance-learning (Yan and Domeniconi 2006). Let us note that, for this general approach of learning one (combined) result based on several representations, strong connections to ensemble clustering (Sect. 2.2) could be pointed out. Essentially, the theory of ensemble learning can explain why 'co-learning' approaches in general work very well.

However, in the context of our study on the blind men and the elephant, we are much more interested in the treatment of different aspects of truth and in deriving different results that are potentially equally valid. This has also been an important motivation for those *multiview clustering* approaches that are technically somewhat similar to multi-represented clustering but do not aim at deriving a combined solution from the different representations

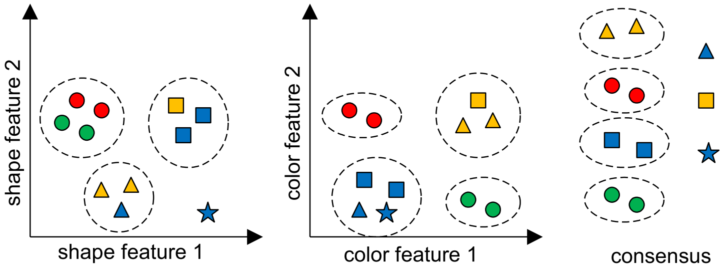


Fig. 1 Different subspaces (views) of a data set can lead to different clustering solutions. The same objects form different meaningful clusters in different views. The consensus alone is less informative and here actually only provides the obvious

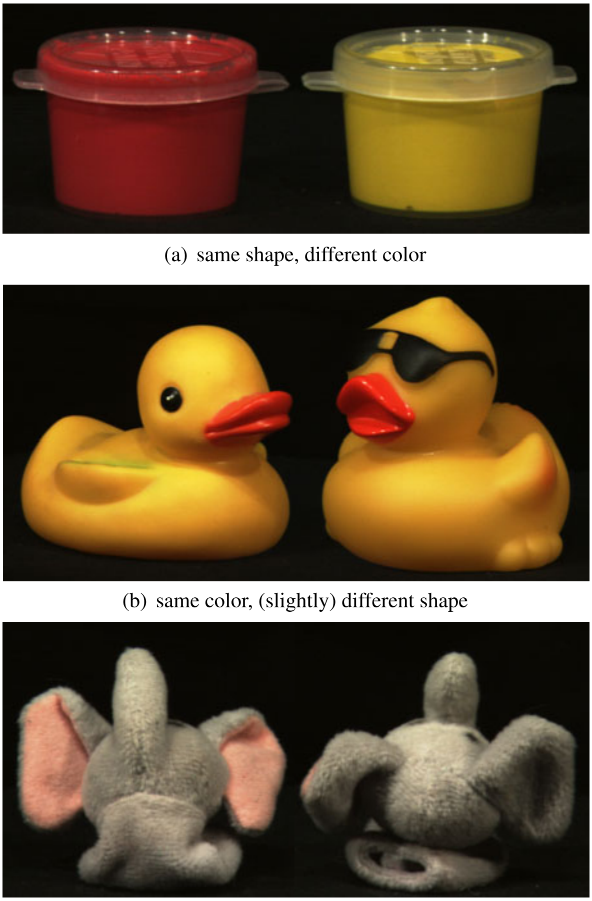
but to find actually different solutions, where the difference could be explained in light of the underlying different representations. This principle is illustrated in Fig. 1: different feature subsets are related to different concepts. As a consequence, it makes perfectly sense to find *different* clustering solutions in (semantically) different subspaces. If we consider the consensus only, we lose information and only see the obvious clusters in this case.

In Fig. 2, we give some examples for this principle, taken from the ALOI data (Geusebroek et al. 2005), where each object is represented by a multitude of different images, taken from different angles, with different illumination conditions and so on. All images from the same object form most naturally a cluster. Aside from this basic ground truth, in image data it can be sensible to cluster objects with similar shape but different color or vice versa. The examples in Figs. 2(a) and 2(b) illustrate these two most straightforward cases. However, sub-clusters, super-clusters, and combinations of parts of the images from different objects can very well also be seen as forming sensible clusters (Kriegel et al. 2011c). Consider, e.g., the possibly meaningful separation of images of the same object seen from different angles as illustrated in Fig. 2(c). Front and back of the same object can be substantially different and in fact relate to meaningful concepts.

Another popular example is the clustering of faces. Clusters of faces of the same person are of course perfectly meaningful. But so are clusters of the same emotional expression on faces of different persons, or clusters of the same orientation of the faces of different persons.

The idea of *multiview clustering* in this notion (e.g. Cui et al. 2007; Jain et al. 2008), i.e., not aiming at a combination and at eventually learning one single concept but in allowing for different concepts in different views, is therefore to enforce clustering results to reside in different but not previously defined subsets of the attributes, i.e., in different 'views' of the data space. In many cases there is no semantically motivated separation of different subsets of attributes available (or, at least, not a single one) but the different representations are found and separated during the clustering process. For example, given a first clustering residing in some subspace, a second clustering result has the additional constraint of being valid in a subspace that is perpendicular to the subspace of the first clustering.

Multiview clustering can be seen as a special case of seeking alternative clusterings, with the additional constraint of the orthogonality of the subspaces (Dang et al. 2012). This is also highlighted by the example of clustering faces, where the cluster of images of the same person could be seen as trivial and uninteresting. The related similarity should therefore not mislead the search for clusters of similar emotional expression. Multiview clustering can also be seen as a special case of subspace clustering allowing maximal overlap yet



(a) same shape, different color

(b) same color, (slightly) different shape

(c) images of the same object (here notably an elephant) seen from different perspectives could be separated, e.g. w.r.t. color

Fig. 2 Examples for different meaningful concepts based on different aspects (subspaces, views) of objects. Concepts could group together different objects or separate on object into different clusters. Examples are taken from the ALOI data set (Geusebroek et al. 2005)

seeking minimally redundant clusters by accommodating different concepts (as proposed by Günnemann et al. 2009).

Because of these close relationships, we do not go into more detail here. Let us note, though, that these approaches shed light on the observation learned in subspace clustering that highly overlapping clusters in different subspaces need not be redundant nor meaningless. That is, certain subsets of objects may meaningfully belong to several clusters simultaneously.

2.5 Summary

In this section on different variants of clustering, we shortly surveyed the research areas of subspace clustering, ensemble clustering, alternative clustering, and multiview clustering. We restricted our survey to the essential parts in order to highlight where similar problems are touched in different areas. As we summarize in Table 1, the different families of clustering algorithms differ in the definition of their goal and, accordingly, face somewhat different problems. However, it is also obvious from this synopsis, that their goals and their problems

Table 1 Goals and problems of the discussed families of clustering algorithms

Approach	Goal	Main problem
Subspace Clustering	Discover clusters in subspaces of the data	Redundancy (very many highly similar clusters are reported)
Ensemble Clustering	Discover different subspaces inducing the same clusters	How do different clusterings correspond? (esp. highly different ones)
Alternative Clustering	Given a clustering, find a different clustering	How much redundancy between the clusterings is admissible?
Multiview Clustering	Discover different cluster concepts in different subspaces	Balancing the overlap of clusters and the difference between concepts

are related. We figure that these families, trying to escape their specific problems, move on towards a common understanding of the problem.

Based on this overview, we identify as possible topics for the discussion between different areas:

1. How to treat diversity of clustering solutions? Should diverse clusterings always be unified? Allegedly, they should not—but under which conditions is a unification of divergent clusterings meaningful and when is it not?
2. Contrariwise, can we learn also from diversity itself? If in an ensemble of several clusterings in several arbitrary random subspaces, one clustering is exceptionally different from the others, it will be outnumbered in most voting procedures and lost. Could it not be especially interesting to report? Speaking in the figure of the elephant discovery: the tip of the elephant's tail, or its tusks, though seemingly insignificant when compared to the belly, may be well worth being described and appreciated.
3. How to treat redundancy of clusters, especially in the presence of overlap between clusters? When does a cluster qualify as redundant w.r.t. another cluster, and when does it represent a different concept although many objects are part of both concepts? We have seen research on subspace clustering more and more trying to get rid of too much redundancy in clustering results while research on alternative clustering recently tends to allow some degree of redundancy. May there be a point where both research directions meet?
4. How to assess similarity between clustering solutions or overlapping clusters within a single clustering? Again, the presence of overlap between clusters increases the complexity of a mapping of clusters where the label correspondence problem makes it already non-trivial.

2.6 Discussion—clustering evaluation considering multiple ground truths?

While we pointed out several questions for discussion, here we set out to discuss one central question: if we admittedly can face several valid solutions to the clustering problem, such as in different subspaces, how do we evaluate and compare these results? This problem, in our opinion, is somehow forming the basis of all dispute between these areas, and also of evaluation of clusterings in general: the possibility that more than one observation or rather clustering solution could be true and worthwhile, as the story of the blind men and the elephant is setting out to teach us. This is truly a fundamental problem not only in clustering. The presence of multiple ground truths has been touched by various data mining areas, as surveyed by Färber et al. (2010). Consequently, it is a problem for clustering evaluation in general that classification data with a single layer of class labels is not necessarily the best choice for testing a new clustering algorithm.

We conjecture that it is even an inherent flaw in design of clustering algorithms if the researcher designing the algorithm evaluates it only w.r.t. the class labels of classification data sets. It is an important difference between classifiers and clustering algorithms that most classification algorithms aim at learning borders of separation of different classes while clustering algorithms aim at grouping similar objects together. Hence the design of clustering algorithms oriented towards learning a class structure may be strongly biased in the wrong direction. The clustering algorithm may be designed to overfit on the class-structure of some data sets.

That is, in classification we are explicitly looking for structure in the data that aids us in predicting the class labels. In unsupervised, exploratory data mining on the other hand, we are after learning about the structure of the data in general—regardless of whether this helps us to predict. As such, evaluating unsupervised methods on how well their results predict class labels only makes sense when the data and class labels at hand correlate with the type of structure the method can discover.

In fact, in general it can be highly desirable if a clustering algorithm detects structures that deviate considerably from annotated classes. If it is a good and interesting result, the clustering algorithm should not be punished for deviating from the class labels. The judgement on new clustering results, however, requires difficult and time-consuming validation based on external domain-knowledge beyond the existing class-labels. Reconsider Fig. 2: all the pairs of images from the ALOI data could sensibly be part of the same cluster as well as being separated in meaningful different clusters. The MultiClust workshop series has seen some first steps in these directions, theoretically (Färber et al. 2010) as well as practically (Kriegel et al. 2011c).

While there is no satisfying solution to this problem around, let us point out with a broader survey, that this problem of evaluation when there exist multiple 'ground truths' has been acknowledged in other research areas as well (but did not find a solution in these other areas either):

- As an example concerned with *traditional clustering*, Chakravarthy and Ghosh (1996) note the difference between clusters and classes though they do not take it into account for the evaluation. The clustering research community overall did not pay much attention, however, on this rather obvious possibility. For example, although the observation of Chakravarthy and Ghosh (1996) is quoted and used for motivating *ensemble clustering* in Strehl and Ghosh (2002), the evaluation in the latter study uses uncritically classification benchmark data sets—including the Pen Digits data set (Frank and Asuncion 2010) which is known to contain several meaningful clustering solutions, as elaborated by Färber et al. (2010).
- In the research on classification, the topic of *multi-label classification* is highly related. Research on this topic is concerned with data where each object can be multiply labeled, i.e., belong to different classes simultaneously. An overview on this topic has been provided by Tsoumakas and Katakis (2007). In this area, the problem of different simultaneously valid ground truths is usually tackled by transforming the complex, nested or intersecting class labels to flat label sets. One possibility is to treat each occurring combination of class labels as an artificial class in its own for training purposes. Votes for this new class are eventually mapped to the corresponding original classes at classification time, i.e., if some objects could belong to either class A, class B, or both, in fact three classes are considered, namely A', B', and (A', B'). (A', B') could actually be defined (in this simple case) as $A \cap B$, while $A' = A \setminus B$ and $B' = B \setminus A$ which shows that the typical treatment of the class labels in this area is quite heuristically deviating from the actual ground truth. There are, of course, many other possibilities of creating a flat set of

class labels, see, e.g., work of Schapire and Singer (2000), Godbole and Sarawagi (2004), Boutell et al. (2004) or Tsoumakas and Katakis (2007). It is, however, remarkable that none of these methods treat the multi-label data sets in their full complexity. This is only achieved when algorithms are adapted to the problem, as, e.g., by Clare and King (2001), Thabtah et al. (2004). But even if training of classifiers takes the complex nature of a data set into account, the other side of the coin, evaluation, remains traditional. For clustering, no comparable approaches exist yet.

- As a special case of multi-label classification let us consider *hierarchical classification* (see e.g. Koller and Sahami 1997; McCallum et al. 1998; Chakrabarti et al. 1998; Barutcuoglu et al. 2006; Cai and Hofmann 2004; Zimek et al. 2010; Fürnkranz and Sima 2010; Silla and Freitas 2011). Here, each class is either a top-level class or a subclass of some other class or several other classes. Overlap among different classes is present only between superclass and its subclasses (i.e., comparing different classes vertically), but not between classes on the same level of the hierarchy (horizontally), i.e., if we have classes A, B, and (A, B), it holds that $A \subseteq (A, B)$ and $B \subseteq (A, B)$ and $A \cap B = \emptyset$. Evaluation of approaches for hierarchical classification is usually based on one specific level of choice, corresponding to a certain granularity of the classification task.
- Hierarchical problems have also been studied in clustering and actually represent the majority of classical work in this area (Sneath 1957; Wishart 1969; Sibson 1973; Hartigan 1975)—see also the overview by Kriegel et al. (2011a). Recent work includes papers by Ankerst et al. (1999), Stuetzle (2003), Achtert et al. (2006b, 2007a, 2007b). There are only some early approaches to evaluate several or even all levels of the hierarchy (Fowlkes and Mallows 1983), or some comparison measures to compare a clustering with the so called cophenetic matrix (see Jain and Dubes 1988, pp. 165–172). The cluster hierarchy can be used to retrieve a flat clustering if a certain level of the hierarchy is selected, for example at a certain density level (Kriegel et al. 2011a).
- In the areas of *bi-clustering* or *co-clustering* (Madeira and Oliveira 2004; Kriegel et al. 2009) it is also a common assumption in certain problem settings that one object can belong to different clusters simultaneously. Surprisingly, although methods in this field have been developed for four decades (initiated by Hartigan 1972), there has not been described a general method of evaluation of clustering results in this field either.

  What is usually done is a (manual) inspection of clusters, reviewing them for prevalent representation of some meaningful concept. We could figure this as 'evaluation by example'. There are attempts to formalize this as 'enrichment' w.r.t. some known concept.
- This technique is automated to a certain extent in *biological analysis* of gene or protein data.
  - A couple of methods has been proposed in order to evaluate clusters retrieved by arbitrary clustering methods (Ashburner et al. 2000; Zeeberg et al. 2003; Al-Shahrour et al. 2004). These methods assume that a class label is assigned to each object of the data set (in the case of gene expression data to each gene/ORF), i.e., a class system is provided. In most cases, the accuracy and usefulness of a method is proven by identifying sample clusters containing 'some' genes/ORFs with functional relationships, e.g. according to Gene Ontology (GO).[4] For example, FatiGO (Al-Shahrour et al. 2004) tries to judge whether GO terms are over- or under-represented in a set of genes w.r.t. a reference set.
  - More theoretically, cluster validity is here measured by means of how good they match the class system where the class system exists of several directed graphs, i.e., there are

[4] http://www.geneontology.org/ (Ashburner et al. 2000).

hierarchical elements and elements of overlap or multiple labels. However, examples of such measures include precision/recall values, or the measures reviewed by Halkidi et al. (2001). This makes it methodically necessary to concentrate the efforts of evaluation at one set of classes at a time. In recent years, multiple class-driven approaches to validate clusters on gene expression data have been proposed (Datta and Datta 2006; Gat-Viks et al. 2003; Gibbons and Roth 2002; Lee et al. 2004; Prelić et al. 2006).

- Previous evaluations do in fact report many found clusters to not obviously reflect known structure. This is possibly due to the fact that the used biological knowledge bases are very incomplete (Clare and King 2002). Others, however, report a clear relationship between strong expression correlation values and high similarity and short distance values w.r.t. distances in the GO-graph (Wang et al. 2004) or a relationship between sequence similarity and semantic similarity (Lord et al. 2003).

In the context of multiple clusterings and overlapping clusters—as are expected in the specialized clustering areas we surveyed above, like subspace clustering, ensemble clustering, alternative or multiview clustering—it becomes even more important to find methods of evaluating clusterings w.r.t. each other, w.r.t. existing knowledge, and w.r.t. their usefulness as interpreted by a human researcher.

Though the problem of overlapping ground truths, and hence the impossibility of using a flat set of class labels directly, is pre-eminent in such research areas as subspace clustering (Kriegel et al. 2009, 2012; Sim et al. 2012; Assent 2012), alternative clustering (Qi and Davidson 2009), or multiview clustering (Cui et al. 2007), it is, in our opinion, actually relevant for all non-naïve approaches that set out to learn something new and interesting about the world—where 'naïve' approaches would require the world to be simple and the truth to be one single flat set of propositions only. Also, instead of sticking to one evaluation measure only, visual approaches and approaches involving several measures of similarities between different partitions of the data (e.g., Achtert et al. 2012) seem more promising, although a lot of work needs to be done in the area of clustering evaluation measures (even when given a ground truth—let alone several truths). For some pointers here see, e.g., Halkidi et al. (2001), Campello (2010), Vendramin et al. (2010).

In summary, we conclude that (i) the difference between clusters and classes, and (ii) the existence of multiple truths in data (i.e., overlapping or alternative—labeled or unlabeled—natural groups of data objects) are important problems in a range of different research areas. These problems have been observed partly since decades yet they have not found appropriate treatment. Especially for clustering and the specialized areas we have surveyed, this is the core problem. The point we are making in this position paper is, however, that this is a *common* problem where researchers working in different areas should probably not only focus on their own version of the problem but learn from approaches taken in other areas, since we eventually want to describe the one elephant. While there probably is no satisfactory general solution for describing elephants—there is no free lunch, after all—we argue for increasing, within these highly related fields, the awareness that we are investigating highly similar problems.

3 Pattern mining—considering the same elephant?

Let us next consider how clustering relates to that other branch of exploratory data mining, namely pattern mining. Whereas in clustering research we are after identifying sets of objects that are highly similar, in pattern mining we are typically after discovering recurring structures. While at first glance these goals and the accompanying problem definitions seem quite different, in the following we will argue that they in fact have quite a lot in common and that researchers in the two fields can learn much from each other. As sketched in Sect. 2.1, 'subspace clustering' and pattern mining share common roots, as well as a number of technical approaches. It seems, however, important to us to note that clustering, esp. subspace clustering, and pattern mining, are currently all considering problems similar to those that have been studied (or, sometimes, even solved) by one of the others.

As such, it seems that the blind men of pattern mining and those of subspace clustering are not only examining the same elephant but also the same parts. Alas, as both fields do not share a common language, much of their work appears unrelated to each other—or when it actually does seem related, it is not easily understood from the other's perspective. Therefore, we will here try to bridge the linguistic gap. However, let us note that we do not argue that clustering, subspace clustering, or pattern mining are one and the same; there are of course genuine differences between the fields. Here, however, we emphasize the similarities, as well as those points in research from which researchers in the different fields might learn from each other's insights and mutually benefit from a common language.

Let us first coarsely sketch the two fields. Pattern mining, to start with, is concerned with developing theory and algorithms for identifying groups of attributes and some selection criteria on those; such that the most 'interesting' attribute-value combinations in the data are returned. That is, by that selection we identify objects in the data that somehow stand out. The prototypical example is supermarket basket analysis, in which by frequent itemset mining we identify items that are frequently sold together—with the infamous *beer* and *nappies* as an example of an 'interesting' pattern.

In contrast to *models*, patterns describe only part of the data. While there are many different formulations for pattern mining, we are here particularly interested in patterns that describe interesting subsets of the database. That is, we are concerned with *theory mining*. Formally, this task has been described by Mannila and Toivonen (1997) as follows. Given a database $\mathcal{D}$, a language $\mathcal{L}$ defining subsets of the data, and a selection predicate q that determines whether an element $\phi \in \mathcal{L}$ describes an interesting subset of $\mathcal{D}$, the task is to find

$$\mathcal{T}(\mathcal{L}, \mathcal{D}, q) = \{\phi \in \mathcal{L} \mid q(\mathcal{D}, \phi) \text{ is true}\}$$

or in other words, the task is to find *all* interesting subsets.

Clearly, this formulation allows us a lot of freedom; the choice of $\mathcal{L}$ and q together determine what we are after, and what *can* be discovered in the data. For the traditional frequent itemset mining setting, the pattern language $\mathcal{L}$ consists of all possible subsets of all items $\mathcal{I}$ in the data, and the selection predicate q corresponds to checking whether this combination of values occurs sufficiently frequently, or has a sufficient support. Typically, and different from clustering approaches, the selection predicate q does not specify much about how the data is distributed *within* the selected sub-part of the data.

Clustering, on the other hand, in general aims at finding groups of similar objects in a database. Aside from algorithmic variations in the process of identifying these groups, the major differences between various clustering approaches are in the actual meaning of 'similar'. If we can agree that, in an abstract sense, both pattern mining and (subspace) clustering identify sub-parts of the data, the key aspect of clustering is that, by requiring similarity, the 'q' here specifies a requirement on the distribution within this part of the data. Further, we note that a key difference between the two is how we (typically) define what sub-parts of the data we want to find; whether these are required to be tall enough (i.e., frequency), or sufficiently similar. Defining similarity, however, is not easy—especially in high-dimensional data the notion of similarity is not a trivial one.

One obvious difference between clustering and pattern mining we need to mention is that, in traditional pattern mining one typically ‘simply’ wants to find ‘all’ interesting subparts of the data, while in clustering one often aims at finding a partitioning of the data such that within each part similarity is high. In clustering, groups of objects, or rows, of the data are the first class citizens, whereas in pattern mining any sub-matrix of the data is what we consider interesting—not necessarily full rows. This more general view on sub-matrices is what it shares with subspace clustering.

We argue that research in subspace clustering, while having a common origin with pattern mining, and even sharing some early ideas, has seen significantly different developments than those made in pattern mining. Interestingly, both fields now face problems already studied by the other. In this section we will survey what main developments both fields have seen, and in particular aim to point out those recent developments in pattern mining from which research on subspace clustering can possibly benefit, and vice versa. For example, the explosion in numbers of results, and reducing their redundancy, are currently open problems in subspace clustering (although some papers already addressed this issue, we will discuss that below) while this has recently has been studied in detail by the pattern mining community. On the other hand, the notion of alternative results, as well as the generalization beyond binary data, are topics where pattern miners may draw much inspiration from recent work in (subspace) clustering or the other specialized clustering approaches we have discussed above (Sect. 2).

The goal of this section is to identify a number of developments in these fields that should not go unnoticed; we are convinced that solutions for pattern mining problems are applicable in subspace clustering, and vice versa.

3.1 Patterns—and the curse of dimensionality

The so-called ‘curse of dimensionality’ is often credited for causing problems in similarity computations in high-dimensional data, and, hence, is given as motivation for specialized approaches such as ‘subspace clustering’ (Kriegel et al. 2009). Let us consider two aspects of the ‘curse’ that are often confused in the literature: (i) the concentration effect of L_p-norms and (ii) the presence of irrelevant attributes.

Regarding the concentration effect (i), the key result of Beyer et al. (1999) states that, if the ratio of the variance of the length of any point vector $\vec{x} \in \mathbb{R}^d$ (denoted by $\|\vec{x}\|$) with the length of the mean point vector (denoted by $E[\|\vec{x}\|]$) converges to zero with increasing data dimensionality, then the proportional difference between the farthest-point distance D_{max} and the closest-point distance D_{min} (the *relative contrast*) vanishes, i.e., all distances concentrate around a mean, and look alike.

$$\lim_{d\to\infty} \operatorname{var}\left(\frac{\|\vec{x}\|}{E[\|\vec{x}\|]}\right) = 0 \quad\Longrightarrow\quad \frac{D_{max} - D_{min}}{D_{min}} \to 0 \tag{1}$$

This observation is often quoted for motivating subspace clustering as a specialized procedure. It should be noted, though, that the problem is neither well enough understood (see e.g. François et al. 2007) nor actually relevant when the data follows different, well separated distributions (Bennett et al. 1999; Houle et al. 2010; Bernecker et al. 2011).

Regarding the separation of clusters, the second problem (ii) is far more important for subspace clustering: In order to find structures describing phenomena of the world some researcher or businessman is interested in, abundances of highly detailed data are collected. Among the features of a high-dimensional data set, for any given query object, many attributes can be expected to be irrelevant to describing that object. Irrelevant attributes can easily obscure clusters that are clearly visible when we consider only the relevant 'subspace' of the dataset. Hence, they interfere with the performance of similarity measures in general, but in a far more fundamental way for clustering. The relevance of certain attributes may differ for different groups of objects within the same data set. Thus, many subsets of objects are defined only by some of the available attributes, and the irrelevant attributes ('noise') will interfere with the efforts to find these subsets. This second problem is actually the true motivation for designing specialized methods to look for clusters in subspaces of a dataset.

Pattern mining, on the other hand, typically does not suffer at all from this aspect of the curse of dimensionality. This is because most interestingness measures do not consider *similarity*, but rather compute a statistic over the selected objects; frequency for instance, is a linear statistic over objects, and when calculating the frequency of a pattern we simply have to count which objects satisfy the constraints (e.g., which transactions contain all the items identified by an itemset). Alternatively, many recent interestingness measures consider patterns that are more surprising given a probabilistic background model (Wang and Parthasarathy 2006; Webb 2007; Tatti 2008; Kontonasios and De Bie 2010; Mampaey et al. 2011); a simple example is *lift* as defined by Brin et al. (1997), which considers the observed frequency of a pattern, and compares it to the expected frequency under the independence model.

As long as the employed interestingness measure does not define a comparison *within* the selected sub-part of the data, which is where the curse stems from, there is no fundamental problem in pattern mining w.r.t. high-dimensional data. However, this is not to say that high-dimensional data does not come with any problems. Computationally, things naturally become much more complex; if anything by the simple fact that the number of possible subsets grows exponentially—which is actually the problem that Bellman (1961), more generally and in a more abstract manner, originally described coining the figurative term 'curse of dimensionality'.

Which leads us to the key problem in traditional pattern mining, the so-called *pattern explosion*. In short, whenever we set a high threshold on interestingness (e.g., a high frequency) we will get returned only few patterns. These patterns are typically not very interesting, as they convey common knowledge; the super-marked manager long since knew that many people that buy bread also buy butter. However, whenever we lower the threshold, we very quickly will find ourselves swamped with extremely many patterns. All of these patterns have passed our interestingness test, and are hence potentially interesting. However, many of the returned patterns will be variations of the same theme, essentially convey the same message, but including or excluding an extra clause (i.e., item). It has proven to be rather difficult to predict at which frequency threshold the pattern explosion hits (Geerts et al. 2005; Boley and Grosskreutz 2008).

When, and how badly, the pattern explosion hits is determined by the characteristics of the data. For dense data, it is often already impossible to mine frequent itemsets below trivially high support thresholds without ending up with unreasonably many results; making it inherently difficult to analyze such data with traditional frequent itemset mining techniques. The link to high-dimensional data here being that for higher number of dimensions, there are more possibilities to create variations of a pattern. And hence, for higher dimensional data, we are likely to discover even more (redundant variations of) patterns.

In short, in pattern mining, controlling the number of results, and the redundancy within has been a major research focus. We will discuss the here relevant advances in Sect. 3.4.

3.2 Common roots of pattern mining and subspace clustering

In general, in subspace clustering similarity is defined in some relation to subsets or combinations of attributes (dimensions) of database objects. Hence, a clustering with n clusters for a database $\mathcal{D} \times \mathcal{A}$, with the set of objects $\mathcal{D}$ and with the full set of attributes $\mathcal{A}$, can be seen as a set $\mathcal{C} = \{(\mathcal{C}_1, \mathcal{A}_1), \ldots, (\mathcal{C}_n, \mathcal{A}_n)\}$, where $\mathcal{C}_i \subseteq \mathcal{D}$ and $\mathcal{A}_i \subseteq \mathcal{A}$, i.e., a cluster is defined w.r.t. a set of objects and w.r.t. a set of attributes (a.k.a. a subspace). Which set of attributes is most appropriate to describe which cluster is dependent on the cluster members. Membership to a cluster is dependent on the similarity-measure which, in turn, depends on the selected attributes (or their combination). This circular dependency is the ultimate reason responsible for subspace clustering being an even harder problem than just clustering.

Subspace clustering algorithms are typically categorized into two groups; in 'projected clustering' objects belong to at most one cluster, while 'subspace clustering' (in a more narrow sense) seeks to find all possible clusters in all available subspaces, allowing overlap (Kriegel et al. 2009). The distinction (and terminology) originates from the two pioneering papers in the field, namely CLIQUE (Agrawal et al. 1998) for 'subspace clustering' and PROCLUS (Aggarwal et al. 1999) for projected clustering; and the two definitions allow a broad field of hybrids.

Since we are interested in the relationship between pattern mining and subspace clustering, we will let aside projected clustering and hybrid approaches and concentrate on subspace clustering in the narrower sense as defined by Agrawal et al. (1998). In this setting, subspace clustering is usually related to a bottom-up traversal of the search space of all possible subspaces, i.e., starting with all one-dimensional subspaces, two-dimensional combinations of these subspaces, three-dimensional combinations of the two-dimensional subspaces and so on, all (relevant) subspaces are searched for clusters residing therein.

Considering CLIQUE, we find the intuition of subspace clustering promoted there closely related to pattern mining. To this end, we consider frequent itemset mining (Agrawal and Srikant 1994), in which we consider binary transaction data, where transactions are sets of items A, B, C, etc. The key idea of APRIORI (Agrawal and Srikant 1994) is to start with itemsets of size 1 that are frequent, and exclude all itemsets from the search that cannot be frequent anymore, given the knowledge which smaller itemsets are frequent. For example, if we count a 1-itemset containing A less than the given minimum support threshold, all 2-itemsets, 3-itemsets, etc. containing A (e.g., $\{A, B\}$, $\{A, C\}$, $\{A, B, C\}$) cannot be frequent either and need not be considered. While theoretically the search space remains exponential, in practice searching becomes feasible even for very large datasets.

Transferring this problem to subspace clustering, each attribute represents an item, and each subspace cluster is then an itemset containing the items representing the attributes of the subspace. This way, finding itemsets with support 1 relates to finding all combinations of attributes constituting a subspace of at least one cluster. This observation is the rationale of most bottom-up subspace clustering approaches: subspaces containing clusters are determined starting from all one-dimensional subspaces accommodating at least one cluster, employing a search strategy similar to that of itemset mining algorithms. To apply any efficient algorithm, the cluster criterion must implement a downward closure property (i.e., (anti-)monotonicity): *If subspace $\mathcal{A}_i$ contains a cluster, then any subspace $\mathcal{A}_j \subseteq \mathcal{A}_i$ must also contain a cluster.*

The anti-monotonic reverse implication, *if a subspace $\mathcal{A}_j$ does not contain a cluster, then any superspace $\mathcal{A}_i \supseteq \mathcal{A}_j$ also cannot contain a cluster*, can subsequently be used for pruning.

Clearly, this is a rather naïve use of the concept of frequent itemsets in subspace clustering. What constitutes a good subspace clustering result is defined here apparently in close relationship to the design of the algorithm, i.e., the desired result appears to be defined according to the expected result (as opposed to: in accordance to what makes sense). The tool 'frequent itemset mining' was first, and the problem of 'finding *all* clusters in *all* subspaces' has apparently been defined in order to have some new problem where the tool can be applied straightforwardly—as has been discussed by Kriegel et al. (2009, 2012), the resulting clusters are usually highly redundant and, hence, mostly useless.

This issue is strongly related to the pattern explosion we discussed above. Recently, pattern miners have started to acknowledge they have been asking the wrong question: instead of asking for *all* patterns that satisfy some constraints, we should ask for small, non-redundant, and high quality *sets of patterns*—where by high-quality we mean that each of the patterns in the set satisfies the thresholds we set on interestingness or similarity, and that the set is optimal with regard to some criterion, e.g. mutual information (Knobbe and Ho 2006a, 2006b), data description (or compression) (Vreeken et al. 2011; Kontonasios and De Bie 2010; De Bie 2011), or covered area (Geerts et al. 2004; Xiang et al. 2011).

Research on subspace clustering inherited all the deficiencies from this originally ill-posed problem. However, early research on subspace clustering as follow-ups of CLIQUE apparently also tried to transfer improvements from pattern mining. As an example, ENCLUS (Cheng et al. 1999) uses several quality criteria for subspaces, not only implementing the downward closure property, but also an upward closure (i.e., allowing search for interesting subspaces as specializations as well as generalizations). This most probably relates to the concept of positive and negative borders known from closed frequent itemsets (Pasquier et al. 1999a). Both can be seen as implementations of the classical concept of version spaces (Mitchell 1977).

However, while pattern mining and subspace clustering share common roots, research in both fields has subsequently diverged in different directions.

3.3 Patterns, subspace clusters—and the elephant

Methods aside, there are two important notions we have to discuss before we can continue; notions that may, or may not make the two fields fundamentally different. First and foremost, what is a result? And, second, can we relate interestingness and similarity? To start with the former, in subspace clustering, a single result is defined by the Cartesian product of objects $C \subseteq \mathcal{D}$ and attributes $A \subseteq \mathcal{A}$. In order to be considered as a result, each of the objects in the selection should be similar to the others, over the selected attributes, according to the employed similarity function.

In order to make a natural connection to pattern mining, we adopt a visual approach; if we are allowed to re-order both attributes and objects freely, we can reorder $\mathcal{D}$ and $\mathcal{A}$ such that C and A define a rectangle in the data, or in pattern mining vocabulary, a *tile*. We give a simple overview in Fig. 3. The left-most plot shows, for some toy data, a traditional pattern *bcd* for which some minimal interestingness measure holds, such as minimal frequency of a value instantiation over these attributes. Note that we do not specify *which* rows support this pattern. Opposing, the right-most plot of the figure shows an example cluster. Here, we are reported that given some similarity measure, rows 2, 3, and 4 are particularly similar regarded over all attributes.

In the middle plot, we show both modern pattern mining and subspace clustering. Both these branches of exploratory data mining research focus on identifying informative submatrices of the data. In pattern mining we are generally interested in discovering all submatrices, or *tiles*, that satisfy some statistic that captures our notion of interestingness. For

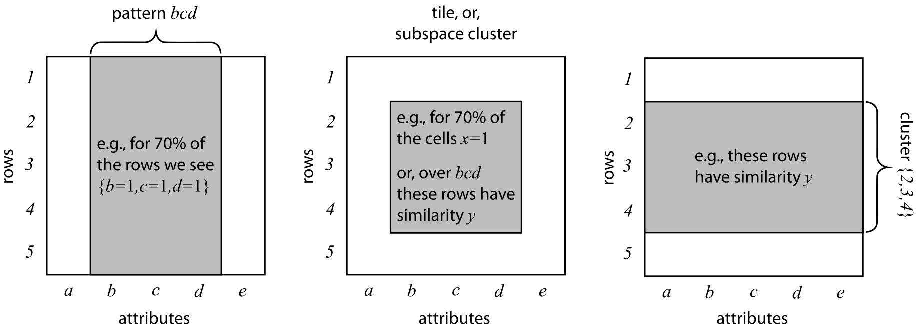

pattern *bcd*
e.g., for 70% of the rows we see {b=1,c=1,d=1}
rows
1
2
3
4
5
a
b
c
d
e
attributes
tile, or, subspace cluster
e.g., for 70% of the cells x=1
or, over *bcd* these rows have similarity y
e.g., these rows have similarity y
cluster {2,3,4}

Fig. 3 (*left*) Traditional pattern mining returns patterns, sets of attributes, that satisfy some statistic, such as minimal frequency, over all rows of the data (e.g., here *bcd*). That is, we are told a particular value-instantiation over attributes *bcd* occurs (at least) σ times—but not *which* rows support this pattern. (*right*) Traditional clustering returns object groups, or clusters, (e.g., here {2, 3, 4}), that satisfy some similarity constraint, over all attributes of the data. That is, we are told the group has similarity y calculated over all attributes, possibly including irrelevant ones. (*center*) Modern pattern mining aims at discovering *tiles*, or sub-matrices, that satisfy some statistical property, such as a minimal density of 1s. Likewise, subspace clustering aims at finding sub-matrices, those with a minimal similarity over the rows within the sub-matrix

example, a sub-matrix with very many 1s, or with very skewed value-instantiations. In subspace clustering this ‘statistic’ is made concrete through a similarity measure, and we are after finding all sub-matrices for which the rows show at least a certain similarity.

In recent years the notion of a tile has become very important in pattern mining (Geerts et al. 2004; Kontonasios and De Bie 2010; De Bie 2011; Gionis et al. 2004). Originally the definition of a pattern was very much along the lines of an SQL query, posing selection criteria on which objects in the data are considered to *support* the pattern or not. As such, beyond whether they contribute to such a global statistic, the selected objects were not really taken into account. However, we can trivially create a tile by simply selecting those objects that support the pattern, and only consider the attributes the pattern identifies. In many recent papers, however, the supporting objects are explicitly taken into account, and by doing so, patterns also naturally define *tiles*.

As such, both fields identify *tiles*, sub-parts of the data. However, both have different ways of arriving at these tiles. In pattern mining, results are typically selected by some measure of ‘interestingness’—of which support, the number of selected objects, is the most well-known example. In subspace clustering, on the other hand, we measure results by how similar the selected objects are over the considered attributes. Clearly, while this may lead to discovering rather different tiles, it is important to realize that both approaches do find tiles, and provide some statistics for the contents of each tile.

We observe that in pattern mining the selection of the objects ‘belonging’ to the pattern is very strict—and that as such those objects will exhibit high similarity over the subspace of attributes the pattern identifies. For example, in standard frequent itemset mining, transactions (i.e., objects) are only selected if they are a strict superset of the pattern at hand—and in fault-tolerant itemset mining (see, e.g., Hébert and Crémilleux 2005; Poernomo and Gopalkrishnan 2009) typically only very few attribute-value combinations are allowed to deviate from the template the pattern identifies. Linking this to similarity, in this strict selection setting, it is easy to see that for the attributes identified by the pattern, the selected objects are highly similar.

The same also holds for subgroup discovery (Wrobel 1997). Subgroup discovery is a supervised branch of pattern mining, aimed at mining patterns by which a target attribute can be predicted well. It is also known, amongst others, as contrast set mining, as correlated pattern mining, or as emerging pattern mining (see Novak et al. 2009, for an overview of the differences and similarities). The main differences between these sub-areas are geographically, as well as in what measure is employed for calculating the correlation to the target attribute. Recently, Leman et al. (2008) introduced a generalized version of subgroup discovery, where multiple target attributes are allowed.

In subgroup discovery the patterns typically strongly resemble SQL range-based selection queries, where the goal is to identify those patterns (intention) that select objects (extension) that correlate strongly to some target attribute(s). In other words, the goal is to find a pattern by which we can select those rows for which the distribution of values over the target is significantly different from the global distribution. Intuitively, the more strict the selection criteria are per attribute, the more alike the selected objects will be on those attributes.

Consequently, in the traditional sense, patterns identified as interesting by a measure using support, are highly likely to correspond to highly-similar subspace clusters, the more strict conditions the pattern defines on its attributes. The other way around, we can say that the higher the similarity of a subspace cluster, the easier it will be to define a pattern that covers the same area of the database. And, the larger this highly-similar subspace cluster is, the more likely it is that it will be discovered by pattern mining using any support-based interestingness measure. In particular for the branch of pattern mining research concerned with mining dense areas in binary matrices (Seppanen and Mannila 2004; Gionis et al. 2004; Miettinen et al. 2008), there exists a clear link to subspace clustering; sub-parts of the data with very many, or very few ones, are likely to exhibit high similarity.

Besides this link, it is interesting to consider what the main differences are. In our view, it is a matter of perspective; whether to take a truly local stance at the objects, and from *within* a tile, like in subspace clustering, or, whether to take a slightly more global stance and look at how we can select those objects by defining conditions on the attributes.

Let us further remark that both interestingness and similarity are very vague concepts, for which many proposals exist. A unifying theory, likely from a statistical point of view, would be very welcome. Although not yet unifying, recently Tatti and Vreeken (2011) indirectly gave an attempt in this direction. They discuss how results from possibly different methods can be compared by taking an information theoretic perspective. By transforming data mining results into sets of tiles, where the tiles essentially specify some statistics on a sub-part of the data; and subsequently constructing probabilistic maximum entropy models (Jaynes 1982) of the data based on these sets of tiles (De Bie 2011), we can use information theoretic measures to see how much information is shared between these distributions (Cover and Thomas 2006). The paper gives a proof of concept for noisy tiles in binary data. The recent paper by Kontonasios et al. (2011) gives theory for modelling real-valued data by maximum entropy, and may hence serve as a stepping stone for comparing results on real-valued data.

3.4 Advances of interest in pattern mining

In this section we discuss some recent advances in pattern mining research that may likewise be applicable for issues in subspace clustering.

3.4.1 Summarizing sets of patterns

Reducing redundancy, which we have identified as a major issue in subspace clustering (Sect. 2), has been studied for a long period of time in pattern mining. Very roughly speaking, two main approaches can be distinguished: finding those patterns that best summarize a large collection of patterns, and mining that set of patterns that together best summarize the data.

In this subsection we discuss the former, in which we find well-known examples. The main idea of this approach is that we have a set of results $\mathcal{F}$, consisting of results that have passed the constraints that we have set, e.g. they all pass the interestingness threshold. Now, with the goal of reducing redundancy in $\mathcal{F}$, we want to select a subset $\mathcal{S} \subseteq \mathcal{F}$ such that $\mathcal{S}$ contains as much information on the whole of $\mathcal{F}$ while being minimal in size.

Perhaps the most well-known examples of this approach are *closed* (Pasquier et al. 1999b) and *maximal* (Bayardo 1998) frequent itemsets, by which we only allow elements $X \in \mathcal{F}$ into $\mathcal{S}$ for which no superset exists that has the same support, or no superset exists that does not meet the mining constraints, respectively. As such, for closed sets, given $\mathcal{S}$ we can reconstruct $\mathcal{F}$ without loss of information—and for maximal sets we can reconstruct only the itemsets, not their frequencies. Non-derivable itemsets (Calders and Goethals 2007) follow a slightly different approach, and only provide those itemsets for which their frequency cannot be derived from the rest. While the concepts of closed and maximal have been applied in subspace clustering, non-derivability has not been explored yet, to the best of our knowledge.

Reduction by closure only works well when the data are highly structured. Its usefulness deteriorates rapidly with noise. A recent improvement to this end is margin-closedness (Fradkin and Mörchen 2010; Mörchen et al. 2011), where we consider elements into the closure for which our measurement falls within a given margin. This provides strong reduction in redundancy, and higher noise resistance; we expect it to be applicable for subspace clusters. Tatti and Mörchen (2011) discuss how to determine by sampling how *robustly* a pattern has a particular property, such as closedness. The idea is that if the itemset remains, e.g., closed, in many subsamples of the data, it is more informative than itemsets that lose the property more easily. As a side note, this seems closely related to the concept of lifetime of a cluster within a hierarchy that has been used sometimes in cluster validation (see, e.g., Ling 1972, 1973; Jain and Dubes 1988; Fred and Jain 2005).

Perhaps not trivially translatable to subspace clusters, another branch of summarization is that of picking or creating a number of representative results. Yan et al. (2005) choose $\mathcal{S}$ such that the error of predicting the frequencies in $\mathcal{F}$ is minimized. Here, it may well be reasonable to replace frequency with similarity. There are some attempts in this direction, e.g. in biclustering (Segal et al. 2001). Mampaey et al. (2011) give an information theoretic approach to identifying that $\mathcal{S}$ by which the frequencies in either $\mathcal{F}$, or the data in general, can best be approximated. To this end, they define a maximum entropy model for data objects, given knowledge about itemset frequencies. The resulting models do capture the general structure of the data very well, without redundancy.

3.4.2 Pattern set mining

While the above-mentioned techniques do reduce redundancy, they typically still result in large collections of patterns that still do contain many variations of the same theme. A recent major insight in pattern mining is that we were asking the wrong question. Instead of asking for all patterns that satisfy some constraints, we should be asking for a small non-redundant group of high-quality patterns. With this insight, the attention shifted from attempting to summarize the full result $\mathcal{F}$, to provide a good summarization of the *data*. In terms of subspace clustering, this means that we would select a certain group of subspace clusters such that we can approximate (explain, describe, etc.) the data optimally. This comes close to the general idea of seeing data mining in terms of compression of data (Faloutsos and Megalooikonomou 2007). Here we discuss a few examples of such *pattern set mining* techniques that we think are applicable to subspace clustering in varying degrees.

The most straightforward technique we discuss is tiling (Geerts et al. 2004). It proposes to not just consider itemsets, but also the transactions they occur in—the same notion we adopted in Sect. 3.3. The main idea here is that patterns covering a large area are more interesting than patterns covering a small area. 'Area' here is defined as the product of the number of items and number of transactions that support the itemset. Most importantly, the authors give an algorithm for approximating the optimal *tiling* of the data—those k tiles that together cover as much of the data as possible. As the paper only considers exact tiles, for which exactly what the data values are, namely 1s, the returned tilings are good summarizations of the data. It is not trivially translated to subspace clustering. One could extract a cluster profile, e.g. a centroid, and take the deviation between the current summary and the real data into account—something that one could minimize.

In this direction, other promising approaches take cues from Information Theory, the Minimum Description Length (MDL) principle (Rissanen 1978; Grünwald 2007) in particular. That is, they approach the pattern set selection problem from the perspective of lossless compression; the best set of patterns is that set of patterns that together compress the data best. Gionis et al. (2004) propose a hierarchical model for discovering informative regions (patterns, subspaces) in the data by employing MDL. It does not consider a candidate set $\mathcal{F}$, but looks for interesting regions directly, assuming a given fixed order of the attributes and objects. The hierarchical nature potentially does link strongly to subspace clustering, where nested clusters are sometimes considered, e.g. by Achtert et al. (2006a, 2007a). An MDL-related method for clusters was proposed by Böhm et al. (2010).

Vreeken et al. (2011) proposed the KRIMP algorithm to approximate the set of itemsets that together optimally compress the data from a candidate collection $\mathcal{F}$. The resulting code tables have been shown to be of very high quality, while reducing the number of patterns up to 7 orders of magnitude. Subsequent work showed that natural clusterings of binary data can be discovered by partitioning the data such that the combined cost of the code tables per part are minimized (van Leeuwen et al. 2009). Extensions of KRIMP to sequences and multi-table settings have been proposed by Tatti and Vreeken (2012) and Koopman and Siebes (2008, 2009), respectively. Recently, Smets and Vreeken (2012) gave the SLIM algorithm to induce high-quality code tables directly from (binary) data. The SQS algorithm (Tatti and Vreeken 2012) builds upon this idea for event sequence data.

In turn, Kontonasios and De Bie (2010) combine the ideas of MDL and Tiling, although they do not simply accumulate tiles to maximize the covered area of the database. Instead, they measure how informative candidate results are with regard to a static probabilistic background model, while also taking their complexity into account. In other words, how many bits does adding result X save us when explaining the data, and how many does it cost to understand X.

Each of the above methods has as of yet only been defined for (single- and multi-table) binary data. However, MDL theory does exist for richer data types, and we would like to point out the strong potential for reducing redundancy in subspace clustering by aiming at that set of subspace clusters that together describe the data best. That is, those clusters by which we can encode the data and the model most succinctly. Note that this approach naturally allows for overlapping clusters, as well as refinements (a big general cluster, and a smaller sub-region of it)—results will be selected if they provide sufficient extra information by which the data can be compressed better than without, while not costing too much to be described themselves. The concept of compression is however difficult to transfer to all density-based (subspace) clustering approaches since density-based clustering does not readily provide a compressed representation of the data (see, e.g., Kriegel et al. 2011a).[5]

As a side note, consider the cross-over field of bi-clustering (Madeira and Oliveira 2004; Kriegel et al. 2009) as seen, e.g., by Chakrabarti et al. (2004), Pensa et al. (2005): especially when applied to binary data, it has very strong links to pattern set mining, as well as subspace clustering. If anything, bi-clustering clearly identifies tiles, and typically, provides a *tiling* of the data; a full coverage of the data in terms of these tiles (bi-clusters).

3.4.3 Significance and randomization

Perhaps the largest problem in exploratory data mining is validation. Unlike in settings where there is a clear formal goal, as there typically is in many supervised machine learning tasks, our goal is as ill-defined as to find 'interesting things'. Like in clustering a plethora of different similarity measures have been considered, all of which may identify some interesting interplay between objects, also in pattern mining a broad spectrum of interestingness measures have been discussed, yet there is no gold standard by which we can compare results.

One approach that recently received ample attention in pattern mining is that of statistical significance. If a result can be easily explained by background knowledge, it will most likely not be interesting to the end user, never mind how large its support or similarity. Webb (2007) proposes to rank patterns depending on their individual statistical significance. As a more general framework, Gionis et al. (2007a) propose to investigate significance of results *in general* through randomization. To this end, they introduce swap randomization as a means to sample random binary datasets of the same row and column margins as the original data, and consequently calculate empirical p-values. Ojala et al. (2008, 2009), Ojala (2010) gave variants for numerical data, easing the use of the model for subspace clustering. Hanhijärvi et al. (2009) extended the framework such that more complex background information, as, e.g., cluster densities and itemset frequencies, can be entered into the model—making the approach applicable for iterative data mining.

De Bie (2011) proposes to model probability distributions over datasets *analytically*, by employing the Maximum Entropy principle (Jaynes 1982; Csiszár 1975). A major advantage of this approach is that it allows for the calculation of exact p-values, and that randomized data can be sampled much faster than with swap-randomization. Kontonasios et al. (2011) recently extended this approach, and gave theory and algorithms to infer MaxEnt models for real-valued data, using background knowledge beyond simple row and column means.

Both, the swap-randomization as well as the MaxEnt approaches, allow us to calculate, empirically and analytically, the probability of seeing a particular structure (e.g., a subspace cluster) in the database. When employed in an iterative approach, such as advocated by Hanhijärvi et al. (2009) and Mampaey et al. (2011), we can iteratively identify that result that is most surprising with regard to what we have seen so far—and hence effectively reduce redundancy to a minimum.

These approaches seem highly promising with regard to reducing redundancy in subspace clustering. Although current theory does not include notions of 'similarity' as background knowledge, theory for including means, variances and histograms of the distribution of subparts of the data are available. Hence, although possibly not trivial, we expect these methods to be relatively easily applicable for assessing whether a subspace cluster, subspace

[5] Let us note that this is the main reason why adaptations of density-based (subspace) clustering to a data stream scenario are particularly difficult, see, e.g., the work of Ntoutsi et al. (2012).

clustering or multiple clustering is significant—whether in light of some basic properties of the data, or with regard to more involved known structure.

3.5 Interesting advances in subspace clustering

Let us now, vice versa, discuss advances in subspace clustering that may be of particular worth in progressing pattern mining research.

In early approaches to subspace clustering, the fixed grid, that allows an easy translation to frequent itemsets, introduced a strong bias towards certain cluster properties, depending on the granularity and positioning of the grid. Thus, it has been of major interest in research on subspace clustering to find ways to reduce this bias. The MAFIA (Nagesh et al. 2001) method uses an adaptive grid, while its generation of subspace clusters is similar to CLIQUE. Another variant, nCluster (Liu et al. 2007), allows overlapping windows of length δ as 1-dimensional units of the grid. Obviously, the accuracy and the efficiency of all these variants still depends on the grid, primarily its granularity and positioning. A higher grid granularity results in higher runtime-requirements but will most likely produce more accurate results. Letting go of the grid-based approach led to the adaptation of density-based clustering concepts (Kriegel et al. 2011a). The first of these adaptations, SUBCLU (Kailing et al. 2004a), is directly using the DBSCAN cluster model of density-connected sets (Ester et al. 1996) in all (interesting) subspaces, getting rid of the grid-approach completely. Nevertheless, density-connected sets satisfy the downward closure property. This enables SUBCLU to search for density-based clusters in subspaces also in an APRIORI-like style. The resulting clusters may exhibit an arbitrary shape and size in the corresponding subspaces. In fact, for each subspace, SUBCLU computes all clusters that would have been found by DBSCAN if applied to that subspace only.

A global density threshold, as used by SUBCLU and the grid-based approaches, leads to a bias towards a certain dimensionality: a tighter threshold separates clusters from noise well in low-dimensional subspaces, but tends to lose clusters in higher dimensions. A more relaxed threshold detects high-dimensional clusters but produces an excessive amount of low-dimensional clusters. This problem has been of major interest in the research on subspace clustering in the recent years. See e.g. the work of Assent et al. (2007), Müller et al. (2009b), where the density-threshold is adapted in turn to the dimensionality of the subspace currently being scrutinized during the run of the algorithm. This is also an aspect closely related to the pattern explosion problem in pattern mining and pattern miners as well as subspace clusterers may probably inspire each other in this specific problem.

Closely related to choosing the appropriate density level, but also a problem in itself, the redundancy issue has found much interest recently as well, e.g., in the work of Assent et al. (2010), Günnemann et al. (2010), Müller et al. (2009a). These approaches aim at reporting only the most representatives of a couple of redundant subspace clusters. A broader overview on these issues in subspace clustering has been recently presented by Sim et al. (2012).

While technically the approaches differ, in concept, adaptive density-thresholds show high similarity with selection of patterns based on statistical significance (Webb 2007; Tatti 2008; Kontonasios and De Bie 2010; Mampaey et al. 2011). Significance of subspace clusters, though, has only been addressed once so far, by Moise and Sander (2008). Hence, we regard it as highly likely that both approaches can learn from each other's endeavours.

Another recent development in both fields, subspace clustering and pattern mining, is finding alternatives to results. The techniques we employ in exploratory data mining can only rarely be shown to provide optimal results. Instead, they typically return (hopefully?) good results heuristically. However, while one good result might shed light on one aspect of the data, it might ignore other parts of the data—for which other results will be informative. As we stated in the previous section (Sect. 2.6), a clustering result that can be judged valid by known structures may even be completely uninteresting. Hence a proposal to improve cluster evaluation relies on the *deviation* of a clustering from known structures instead of judging the coverage of known structures (Kriegel et al. 2011c). Algorithmically, this clearly shows the connection of these questions to the other sub-topics of clustering, alternative clustering and multiview clustering (cf. Sects. 2.3 and 2.4, respectively).

In particular in light of iterative data mining (Hanhijärvi et al. 2009; Mampaey et al. 2011), where we take into account what the user already knows to determine what might be informative, considering multiple different models is very much an open problem. Most approaches in pattern mining to this end, however, simply iteratively optimize a given score heuristically in a greedy fashion (Hanhijärvi et al. 2009; Mampaey et al. 2011; Smets and Vreeken 2012). Alternatively, we find the approaches that rank results according to interestingness (Kontonasios and De Bie 2010; Kontonasios et al. 2011; De Bie 2011), and assume the user to determine which of the (hopefully top-ranked) results they want to investigate—after which these should be added to the scoring model, and the results are re-ranked. While this approach allows the user to steer the process, it is still far away from the ideal of being presented a number of alternative results.

Another approach in pattern mining related to alternatives, is redescription mining (Ramakrishnan et al. 2004; Gallo et al. 2008). In redescription mining, we are given a target pattern, and are after finding patterns that are syntactically different (i.e., redescriptions), yet select as well as possible the same set of rows (objects). One way to view this problem is to regard patterns as tiles in the data, and having access to a large collection of these tiles, find those tiles that have a large overlap in the identified transactions, but as little overlap as possible in the attributes. While most work on redescription mining has been focused on binary data, recently Galbrun and Miettinen (2011) discuss how to find redescriptions in the form of simple conjunctive queries for data that includes real-valued attributes.

4 Conclusions

Under the theme of the figurative tale of the blind men examining the elephant and stating different, allegedly opposing findings, we took position here in mainly two aspects of the big topic of the ‘MultiClust’ community:

First, the fields of subspace clustering, ensemble clustering, alternative clustering (as a specialization of constraint clustering), and multiview clustering, seemingly so different, are closely tied, partly by applying similar strategies, but mainly by one fundamental common problem: what to make out of apparently opposing findings, i.e., multiple ‘ground truths’.

Second, to be more outward-looking, we elaborated on the strong links between subspace clustering and pattern mining. Although the topics of research within the two fields have diverged over time, we argue the case that both fields are not as different as one might think seeing the researchers from both fields barely exchanging ideas. Moreover, researchers from both areas can probably learn much from the experience gained by the other.

To make both points, that are related by our desire to stimulate the learning from each other’s progress in research, and in order to improve the mutual understanding despite a sometimes different vocabulary, we surveyed a plethora of literature—many of these references probably for the first time find each other in the same paper. It is our hope that the metaphorical ‘blind men’ (including ourselves) will learn from each others insight, and so will be able, eventually, to see the entire elephant.

Acknowledgements Jilles Vreeken is supported by a Post-Doctoral Fellowship of the Research Foundation—Flanders (FWO). Arthur Zimek is partially supported by NSERC. The elephant was satisfied with a handful of peanuts.

References

Achtert, E., Kriegel, H. P., Pryakhin, A., & Schubert, M. (2005). Hierarchical density-based clustering for multi-represented objects. In *Workshop on mining complex data (MCD) on the 5th IEEE international conference on data mining (ICDM)*, Houston, TX (p. 9).

Achtert, E., Böhm, C., Kriegel, H. P., Kröger, P., Müller-Gorman, I., & Zimek, A. (2006a). Finding hierarchies of subspace clusters. In *Proceedings of the 10th European conference on principles and practice of knowledge discovery in databases (PKDD)*, Berlin, Germany (pp. 446–453). doi:10.1007/11871637_42.

Achtert, E., Böhm, C., Kröger, P., & Zimek, A. (2006b). Mining hierarchies of correlation clusters. In *Proceedings of the 18th international conference on scientific and statistical database management (SSDBM)*, Vienna, Austria (pp. 119–128). doi:10.1109/SSDBM.2006.35.

Achtert, E., Kriegel, H. P., Pryakhin, A., & Schubert, M. (2006c). Clustering multi-represented objects using combination trees. In *Proceedings of the 10th Pacific-Asia conference on knowledge discovery and data mining (PAKDD)*, Singapore (pp. 174–178). doi:10.1007/11731139_21.

Achtert, E., Böhm, C., Kriegel, H. P., Kröger, P., Müller-Gorman, I., & Zimek, A. (2007a). Detection and visualization of subspace cluster hierarchies. In *Proceedings of the 12th international conference on database systems for advanced applications (DASFAA)*, Bangkok, Thailand (pp. 152–163). doi:10.1007/978-3-540-71703-4_15.

Achtert, E., Böhm, C., Kriegel, H. P., Kröger, P., & Zimek, A. (2007b). On exploring complex relationships of correlation clusters. In *Proceedings of the 19th international conference on scientific and statistical database management (SSDBM)*, Banff, Canada (pp. 7–16). doi:10.1109/SSDBM.2007.21.

Achtert, E., Goldhofer, S., Kriegel, H. P., Schubert, E., & Zimek, A. (2012). Evaluation of clusterings—metrics and visual support. In *Proceedings of the 28th international conference on data engineering (ICDE)*, Washington, DC (pp. 1285–1288). doi:10.1109/ICDE.2012.128.

Aggarwal, C. C., Procopiuc, C. M., Wolf, J. L., Yu, P. S., & Park, J. S. (1999). Fast algorithms for projected clustering. In *Proceedings of the ACM international conference on management of data (SIGMOD)*, Philadelphia, PA (pp. 61–72).

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. In *Proceedings of the 20th international conference on very large data bases (VLDB)*, Santiago de Chile, Chile (pp. 487–499).

Agrawal, R., Gehrke, J., Gunopulos, D., & Raghavan, P. (1998). Automatic subspace clustering of high dimensional data for data mining applications. In *Proceedings of the ACM international conference on management of data (SIGMOD)*, Seattle, WA (pp. 94–105).

Al-Shahrour, F., Diaz-Uriarte, R., & Dopazo, J. (2004). FatiGO: a web tool for finding significant associations of Gene ontology terms with groups of genes. *Bioinformatics*, *20*(4), 578–580. doi:10.1093/bioinformatics/btg455.

Ankerst, M., Breunig, M. M., Kriegel, H. P., & Sander, J. (1999). OPTICS: ordering points to identify the clustering structure. In *Proceedings of the ACM international conference on management of data (SIGMOD)*, Philadelphia, PA (pp. 49–60).

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. *The Gene Ontology Consortium. Nature Genetics*, *25*(1), 25–29.

Assent, I. (2012). Clustering high dimensional data. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *2*(4), 340–350.

Assent, I., Krieger, R., Müller, E., & Seidl, T. (2007). DUSC: dimensionality unbiased subspace clustering. In *Proceedings of the 7th IEEE international conference on data mining (ICDM)*, Omaha, NE (pp. 409–414). doi:10.1109/ICDM.2007.49.

Assent, I., Krieger, R., Müller, E., & Seidl, T. (2008). INSCY: indexing subspace clusters with in-process-removal of redundancy. In *Proceedings of the 8th IEEE international conference on data mining (ICDM)*, Pisa, Italy (pp. 719–724). doi:10.1109/ICDM.2008.46.

Assent, I., Müller, E., Günnemann, S., Krieger, R., & Seidl, T. (2010). Less is more: non-redundant subspace clustering. In *MultiClust: 1st international workshop on discovering, summarizing and using multiple clusterings held in conjunction with KDD 2010*, Washington, DC.

Azimi, J., & Fern, X. (2009). Adaptive cluster ensemble selection. In *Proceedings of the 21st international joint conference on artificial intelligence (IJCAI)*, Pasadena, CA (pp. 992–997).

Bade, K., & Nürnberger, A. (2008). Creating a cluster hierarchy under constraints of a partially known hierarchy. In *Proceedings of the 8th SIAM international conference on data mining (SDM)*, Atlanta, GA (pp. 13–23).

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 IEEE international conference on data mining (ICDM)*, Hong Kong, China (pp. 53–62). doi:10.1109/ICDM.2006.37.

Barutcuoglu, Z., Schapire, R. E., & Troyanskaya, O. G. (2006). Hierarchical multi-label prediction of gene function. *Bioinformatics*, *22*(7), 830–836. doi:10.1093/bioinformatics/btk048.

Basu, S., Davidson, I., & Wagstaff, K. (Eds.) (2008). *Constraint clustering: advances in algorithms, applications and theory*. Boca Raton, London, New York: CRC Press.

Bayardo, R. (1998). Efficiently mining long patterns from databases. In *Proceedings of the ACM international conference on management of data (SIGMOD)*, Seattle, WA (pp. 85–93).

Bellman, R. (1961). *Adaptive control processes. a guided tour*. Princeton: Princeton University Press.

Bennett, K. P., Fayyad, U., & Geiger, D. (1999). Density-based indexing for approximate nearest-neighbor queries. In *Proceedings of the 5th ACM international conference on knowledge discovery and data mining (SIGKDD)*, San Diego, CA (pp. 233–243). doi:10.1145/312129.312236.

Bernecker, T., Houle, M. E., Kriegel, H. P., Kröger, P., Renz, M., Schubert, E., & Zimek, A. (2011). Quality of similarity rankings in time series. In *Proceedings of the 12th international symposium on spatial and temporal databases (SSTD)*, Minneapolis, MN (pp. 422–440). doi:10.1007/978-3-642-22922-0_25.

Bertoni, A., & Valentini, G. (2005). Ensembles based on random projections to improve the accuracy of clustering algorithms. In *16th Italian workshop on neural nets (WIRN), and international workshop on natural and artificial immune systems (NAIS)*, Vietri sul Mare, Italy (pp. 31–37). doi:10.1007/11731177_5.

Beyer, K., Goldstein, J., Ramakrishnan, R., & Shaft, U. (1999). When is "nearest neighbor" meaningful? In *Proceedings of the 7th international conference on database theory (ICDT)*, Jerusalem, Israel (pp. 217–235). doi:10.1007/3-540-49257-7_15.

Bickel, S., & Scheffer, T. (2004). Multi-view clustering. In *Proceedings of the 4th IEEE international conference on data mining (ICDM)*, Brighton, UK (pp. 19–26). doi:10.1109/ICDM.2004.10095.

Blum, A., & Mitchell, T. (1998). Combining labeled and unlabeled data with Co-training. In *Proceedings of the 11th annual conference on computational learning theory (COLT)*, Madison, WI (pp. 92–100). doi:10.1145/279943.279962.

Böhm, C., Fiedler, F., Oswald, A., Plant, C., Wackersreuther, B., & Wackersreuther, P. (2010). ITCH: information-theoretic cluster hierarchies. In *Proceedings of the European conference on machine learning and knowledge discovery in databases (ECML PKDD)*, Barcelona, Spain.

Boley, M., & Grosskreutz, H. (2008). A randomized approach for approximating the number of frequent sets. In *Proceedings of the 8th IEEE international conference on data mining (ICDM)*, Pisa, Italy (pp. 43–52). New York: IEEE Press.

Boutell, M. R., Luo, J., Shen, X., & Brown, C. M. (2004). Learning multi-label scene classification. *Pattern Recognition*, *37*(9), 1757–1771. doi:10.1016/j.patcog.2004.03.009.

Brin, S., Motwani, R., & Silverstein, C. (1997). Beyond market baskets: generalizing association rules to correlations. In *Proceedings of the ACM international conference on management of data (SIGMOD)*, Tucson, AZ (pp. 265–276). New York: ACM Press.

Brown, G., Wyatt, J., Harris, R., & Yao, X. (2005). Diversity creation methods: a survey and categorisation. *Information Fusion*, *6*, 5–20. doi:10.1016/j.inffus.2004.04.004.

Cai, L., & Hofmann, T. (2004). Hierarchical document categorization with support vector machines. In *Proceedings of the 13th ACM conference on information and knowledge management (CIKM)*, Washington, DC (pp. 78–87). doi:10.1145/1031171.1031186.

Calders, T., & Goethals, B. (2007). Non-derivable itemset mining. *Data Mining and Knowledge Discovery*, *14*(1), 171–206.

Campello, R. J. G. B. (2010). Generalized external indexes for comparing data partitions with overlapping categories. *Pattern Recognition Letters*, *31*(9), 966–975. doi:10.1016/j.patrec.2010.01.002.

Caruana, R., Elhawary, M., Nguyen, N., & Smith, C. (2006). Meta clustering. In *Proceedings of the 6th IEEE international conference on data mining (ICDM)*, Hong Kong, China (pp. 107–118). doi:10.1109/ICDM.2006.103.

Chakrabarti, D., Papadimitriou, S., Modha, D. S., & Faloutsos, C. (2004). Fully automatic cross-associations. In *Proceedings of the 10th ACM international conference on knowledge discovery and data mining (SIGKDD)*, Seattle, WA (pp. 79–88).

Chakrabarti, S., Dom, B., Agrawal, R., & Raghavan, P. (1998). Scalable feature selection, classification and signature generation for organizing large text databases into hierarchical topic taxonomies. *The VLDB Journal*, *7*(3), 163–178.

Chakravarthy, S. V., & Ghosh, J. (1996). Scale-based clustering using the radial basis function network. *IEEE Transactions on Neural Networks*, *7*(5), 1250–1261.

Chaudhuri, K., Kakade, S. M., Livescu, K., & Sridharan, K. (2009). Multi-view clustering via canonical correlation analysis. In *Proceedings of the 26th international conference on machine learning (ICML)*, Montreal, QC, Canada (pp. 129–136).

Cheng, C. H., Fu, A. W. C., & Zhang, Y. (1999). Entropy-based subspace clustering for mining numerical data. In *Proceedings of the 5th ACM international conference on knowledge discovery and data mining (SIGKDD)*, San Diego, CA (pp. 84–93). doi:10.1145/312129.312199.

Clare, A., & King, R. (2001). Knowledge discovery in multi-label phenotype data. In *Proceedings of the 5th European conference on principles of data mining and knowledge discoverys (PKDD)*, Freiburg, Germany (pp. 42–53). doi:10.1007/3-540-44794-6_4.

Clare, A., & King, R. (2002). How well do we understand the clusters found in microarray data? *In Silico Biology*, *2*(4), 511–522.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory*. New York: Wiley-Interscience.

Csiszár, I. (1975). I-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, *3*(1), 146–158.

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)*, Omaha, NE (pp. 133–142). doi:10.1109/ICDM.2007.94.

Dang, X. H., & Bailey, J. (2010). Generation of alternative clusterings using the CAMI approach. In *Proceedings of the 10th SIAM international conference on data mining (SDM)*, Columbus, OH (pp. 118–129).

Dang, X. H., Assent, I., & Bailey, J. (2012). Multiple clustering views via constrained projections. In *3rd MultiClust workshop: discovering, summarizing and using multiple clusterings held in conjunction with SIAM data mining 2012*, Anaheim, CA.

Datta, S., & Datta, S. (2006). Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes. *BMC Bioinformatics*, *7*, 397. doi:10.1186/1471-2105-7-397.

Davidson, I., & Qi, Z. (2008). Finding alternative clusterings using constraints. In *Proceedings of the 8th IEEE international conference on data mining (ICDM)*, Pisa, Italy (pp. 773–778). doi:10.1109/ICDM.2008.141.

Davidson, I., & Ravi, S. (2009). Using instance-level constraints in agglomerative hierarchical clustering: theoretical and empirical results. *Data Mining and Knowledge Discovery*, *18*, 257–282.

Davidson, I., Ravi, S. S., & Shamis, L. (2010). A SAT-based framework for efficient constrained clustering. In *Proceedings of the 10th SIAM international conference on data mining (SDM)*, Columbus, OH (pp. 94–105).

De Bie, T. (2011). Maximum entropy models and subjective interestingness: an application to tiles in binary databases. *Data Mining and Knowledge Discovery*, *23*(3), 1–40.

Dietterich, T. G. (2000). Ensemble methods in machine learning. In *First international workshop on multiple classifier systems (MCS)*, Cagliari, Italy (pp. 1–15). doi:10.1007/3-540-45014-9_1.

Dietterich, T. G. (2003). Ensemble learning. In M. A. Arbib (Ed.), *The handbook of brain theory and neural networks* (2nd edn., pp. 405–408). Cambridge: MIT Press.

Domeniconi, C. (2012). Subspace clustering ensembles (invited talk). In *3rd MultiClust workshop: discovering, summarizing and using multiple clusterings held in conjunction with SIAM data mining 2012*, Anaheim, CA.

Domeniconi, C., & Al-Razgan, M. (2009). Weighted cluster ensembles: methods and analysis. *ACM Transactions on Knowledge Discovery from Data*, *2*(4), 1–40. doi:10.1145/1460797.1460800.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the 2nd ACM international conference on knowledge discovery and data mining (KDD)*, Portland, OR (pp. 226–231).

Faloutsos, C., & Megalooikonomou, V. (2007). On data mining, compression and Kolmogorov complexity. In *Data mining and knowledge discovery* (Vol. 15, pp. 3–20). Berlin: Springer.

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). Knowledge discovery and data mining: towards a unifying framework. In *Proceedings of the 2nd ACM international conference on knowledge discovery and data mining (KDD)*, Portland, OR (pp. 82–88).

Fern, X. Z., & Brodley, C. E. (2003). Random projection for high dimensional data clustering: a cluster ensemble approach. In *Proceedings of the 20th international conference on machine learning (ICML)*, Washington, DC (pp. 186–193).

Fern, X. Z., & Lin, W. (2008). Cluster ensemble selection. *Statistical Analysis and Data Mining*, *1*(3), 128–141. doi:10.1002/sam.10008.

Fowlkes, E. B., & Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, *78*(383), 553–569.

Fradkin, D., & Mörchen, F. (2010). Margin-closed frequent sequential pattern mining. In *Proc. ACM SIGKDD workshop on useful patterns (UP'10)*.

François, D., Wertz, V., & Verleysen, M. (2007). The concentration of fractional distances. *IEEE Transactions on Knowledge and Data Engineering*, *19*(7), 873–886. doi:10.1109/TKDE.2007.1037.

Frank, A., & Asuncion, A. (2010). UCI machine learning repository. http://archive.ics.uci.edu/ml, http://archive.ics.uci.edu/ml.

Fred, A. L. N., & Jain, A. K. (2005). Combining multiple clusterings using evidence accumulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(6), 835–850.

Fürnkranz, J., & Sima, J. F. (2010). On exploiting hierarchical label structure with pairwise classifiers. *ACM SIGKDD Explorations*, *12*(2), 21–25. doi:10.1145/1964897.1964903.

Färber, I., Günnemann, S., Kriegel, H. P., Kröger, P., Müller, E., Schubert, E., Seidl, T., & Zimek, A. (2010). On using class-labels in evaluation of clusterings. In *MultiClust: 1st international workshop on discovering, summarizing and using multiple clusterings held in conjunction with KDD 2010*, Washington, DC.

Galbrun, E., & Miettinen, P. (2011). From black and white to full colour: extending redescription mining outside the boolean world. In *Proceedings of the 11th SIAM international conference on data mining (SDM)*, Mesa, AZ (pp. 546–557).

Gallo, A., Miettinen, P., & Mannila, H. (2008). Finding subgroups having several descriptions: algorithms for redescription mining. In *Proceedings of the 8th SIAM international conference on data mining (SDM)*, Atlanta, GA.

Gao, J., & Tan, P. N. (2006). Converting output scores from outlier detection algorithms into probability estimates. In *Proceedings of the 6th IEEE international conference on data mining (ICDM)*, Hong Kong, China (pp. 212–221). doi:10.1109/ICDM.2006.43.

Gat-Viks, I., Sharan, R., & Shamir, R. (2003). Scoring clustering solutions by their biological relevance. *Bioinformatics*, *19*(18), 2381–2389. doi:10.1093/bioinformatics/btg330.

Geerts, F., Goethals, B., & Mielikäinen, T. (2004). Tiling databases. In *Proceedings of the 7th international conference on discovery science*, Padova, Italy (pp. 278–289).

Geerts, F., Goethals, B., & Van den Bussche, J. (2005). Tight upper bounds on the number of candidate patterns. *ACM Transactions on Database Systems*, *30*(2), 333–363.

Geusebroek, J. M., Burghouts, G. J., & Smeulders, A. (2005). The Amsterdam library of object images. *International Journal of Computer Vision*, *61*(1), 103–112. doi:10.1023/B:VISI.0000042993.50813.60.

Ghosh, J., & Acharya, A. (2011). Cluster ensembles. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(4), 305–315. doi:10.1002/widm.32.

Gibbons, F. D., & Roth, F. P. (2002). Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Research*, *12*, 1574–1581.

Gionis, A., Mannila, H., & Seppänen, J. K. (2004). Geometric and combinatorial tiles in 0-1 data. In *Proceedings of the 8th European conference on principles and practice of knowledge discovery in databases (PKDD)*, Pisa, Italy (pp. 173–184).

Gionis, A., Mannila, H., Mielikäinen, T., & Tsaparas, P. (2007a). Assessing data mining results via swap randomization. *ACM Transactions on Knowledge Discovery from Data*, *1*(3), 167–176.

Gionis, A., Mannila, H., & Tsaparas, P. (2007b). Clustering aggregation. *ACM Transactions on Knowledge Discovery from Data*. doi:10.1145/1217299.1217303.

Godbole, S., & Sarawagi, S. (2004). Discriminative methods for multi-labeled classification. In *Proceedings of the 8th Pacific-Asia conference on knowledge discovery and data mining (PAKDD)*, Sydney, Australia (pp. 22–30). doi:10.1007/978-3-540-24775-3_5.

Gondek, D., & Hofmann, T. (2004). Non-redundant data clustering. In *Proceedings of the 4th IEEE international conference on data mining (ICDM)*, Brighton, UK (pp. 75–82). doi:10.1109/ICDM.2004.10104.

Gondek, D., & Hofmann, T. (2005). Non-redundant clustering with conditional ensembles. In *Proceedings of the 11th ACM international conference on knowledge discovery and data mining (SIGKDD)*, Chicago, IL (pp. 70–77). doi:10.1145/1081870.1081882.

Grünwald, P. (2007). *The minimum description length principle*. Cambridge: MIT Press.

Gullo, F., Domeniconi, C., & Tagarelli, A. (2009a). Projective clustering ensembles. In *Proceedings of the 9th IEEE international conference on data mining (ICDM)*, Miami, FL.

Gullo, F., Tagarelli, A., & Greco, S. (2009b). Diversity-based weighting schemes for clustering ensembles. In *Proceedings of the 9th SIAM international conference on data mining (SDM)*, Sparks, NV (pp. 437–448).

Gullo, F., Domeniconi, C., & Tagarelli, A. (2010). Enhancing single-objective projective clustering ensembles. In *Proceedings of the 10th IEEE international conference on data mining (ICDM)*, Sydney, Australia.

Gullo, F., Domeniconi, C., & Tagarelli, A. (2011). Advancing data clustering via projective clustering ensembles. In *Proceedings of the 17th ACM international conference on knowledge discovery and data mining (SIGKDD)*, San Diego, CA.

Günnemann, S., Müller, E., Färber, I., & Seidl, T. (2009). Detection of orthogonal concepts in subspaces of high dimensional data. In *Proceedings of the 18th ACM conference on information and knowledge management (CIKM)*, Hong Kong, China (pp. 1317–1326). doi:10.1145/1645953.1646120.

Günnemann, S., Färber, I., Müller, E., & Seidl, T. (2010). ASCLU: alternative subspace clustering. In *MultiClust: 1st international workshop on discovering, summarizing and using multiple clusterings held in conjunction with KDD 2010*, Washington, DC.

Hadjitodorov, S. T., & Kuncheva, L. I. (2007). Selecting diversifying heuristics for cluster ensembles. In *7th international workshop on multiple classifier systems (MCS)*, Prague, Czech Republic (pp. 200–209).

Hadjitodorov, S. T., Kuncheva, L. I., & Todorova, L. P. (2006). Moderate diversity for better cluster ensembles. *Information Fusion*, *7*(3), 264–275. doi:10.1016/j.inffus.2005.01.008.

Hahmann, M., Volk, P. B., Rosenthal, F., Habich, D., & Lehner, W. (2009). How to control clustering results? Flexible clustering aggregation. In *Proceedings of the 8th international symposium on intelligent data analysis (IDA)*, Lyon, France (pp. 59–70). doi:10.1007/978-3-642-03915-7_6.

Halkidi, M., Batistakis, Y., & Vazirgiannis, M. (2001). On clustering validation techniques. *Journal of Intelligent Information Systems*, *17*(2–3), 107–145. doi:10.1023/A:1012801612483.

Hanhijärvi, S., Ojala, M., Vuokko, N., Puolamäki, K., Tatti, N., & Mannila, H. (2009). Tell me something I don't know: randomization strategies for iterative data mining. In *Proceedings of the 15th ACM international conference on knowledge discovery and data mining (SIGKDD)*, Paris, France (pp. 379–388). New York: ACM Press.

Hansen, L. K., & Salamon, P. (1990). Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *12*(10), 993–1001. doi:10.1109/34.58871.

Hartigan, J. A. (1972). Direct clustering of a data matrix. *Journal of the American Statistical Association*, *67*(337), 123–129.

Hartigan, J. A. (1975). *Clustering algorithms*. New York, London, Sydney, Toronto: Wiley.

Hébert, C., & Crémilleux, B. (2005). Mining frequent *delta*-free patterns in large databases. In *Proceedings of the 8th international conference discovery science*, Singapore (pp. 124–136).

Horta, D., & Campello, R. J. G. B. (2012). Automatic aspect discrimination in data clustering. *Pattern Recognition*, *45*(12), 4370–4388.

Houle, M. E., Kriegel, H. P., Kröger, P., Schubert, E., & Zimek, A. (2010). Can shared-neighbor distances defeat the curse of dimensionality? In *Proceedings of the 22nd international conference on scientific and statistical database management (SSDBM)*, Heidelberg, Germany (pp. 482–500). doi:10.1007/978-3-642-13818-8_34.

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data*. Englewood Cliffs: Prentice Hall.

Jain, P., Meka, R., & Dhillon, I. S. (2008). Simultaneous unsupervised learning of disparate clusterings. *Statistical Analysis and Data Mining*, *1*(3), 195–210. doi:10.1002/sam.10007.

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

Kailing, K., Kriegel, H. P., & Kröger, P. (2004a). Density-connected subspace clustering for high-dimensional data. In *Proceedings of the 4th SIAM international conference on data mining (SDM)*, Lake Buena Vista, FL (pp. 246–257).

Kailing, K., Kriegel, H. P., Pryakhin, A., & Schubert, M. (2004b). Clustering multi-represented objects with noise. In *Proceedings of the 8th Pacific-Asia conference on knowledge discovery and data mining (PAKDD)*, Sydney, Australia (pp. 394–403). doi:10.1007/978-3-540-24775-3_48.

Klein, D., Kamvar, S. D., & Manning, C. D. (2002). From instance-level constraints to space-level constraints: making the most of prior knowledge in data clustering. In *Proceedings of the 19th international conference on machine learning (ICML)*, Sydney, Australia (pp. 307–314).

Knobbe, A., & Ho, E. (2006a). Pattern teams. In *Proceedings of the 10th European conference on principles and practice of knowledge discovery in databases (PKDD)* (Vol. 4213, pp. 577–584). Berlin: Springer.

Knobbe, A. J., & Ho, E. K. Y. (2006b). Maximally informative k-itemsets and their efficient discovery. In *Proceedings of the 12th ACM international conference on knowledge discovery and data mining (SIGKDD)*, Philadelphia, PA (pp. 237–244).

Koller, D., & Sahami, M. (1997). Hierarchically classifying documents using very few words. In *Proceedings of the 14th international conference on machine learning (ICML)*, Nashville, TN (pp. 170–178).

Kontonasios, K. N., & De Bie, T. (2010). An information-theoretic approach to finding noisy tiles in binary databases. In *Proceedings of the 10th SIAM international conference on data mining (SDM)*, Columbus, OH, SIAM (pp. 153–164).

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)*, Vancouver, BC, ICDM.

Koopman, A., & Siebes, A. (2008). Discovering relational items sets efficiently. In *Proceedings of the 8th SIAM international conference on data mining (SDM)*, Atlanta, GA (pp. 108–119).

Koopman, A., & Siebes, A. (2009). Characteristic relational patterns. In *Proceedings of the 15th ACM international conference on knowledge discovery and data mining (SIGKDD)*, Paris, France (pp. 437–446).

Kriegel, H. P., & Schubert, M. (2012). Co-RCA: unsupervised distance-learning for multi-view clustering. In *3rd MultiClust workshop: discovering, summarizing and using multiple clusterings held in conjunction with SIAM data mining 2012*, Anaheim, CA (pp. 11–18).

Kriegel, H. P., & Zimek, A. (2010). Subspace clustering, ensemble clustering, alternative clustering, multiview clustering: what can we learn from each other? In *MultiClust: 1st international workshop on discovering, summarizing and using multiple clusterings held in conjunction with KDD 2010*, Washington, DC.

Kriegel, H. P., Kunath, P., Pryakhin, A., & Schubert, M. (2008). Distribution-based similarity for multi-represented multimedia objects. In *Proceedings of the 14th IEEE international MultiMedia modeling conference (MMM)*, Kyoto, Japan (pp. 155–164). doi:10.1007/978-3-540-77409-9_15.

Kriegel, H. P., Kröger, P., & Zimek, A. (2009). Clustering high dimensional data: a survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Transactions on Knowledge Discovery from Data*, *3*(1), 1–58. doi:10.1145/1497577.1497578.

Kriegel, H. P., Kröger, P., Sander, J., & Zimek, A. (2011a). Density-based clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *1*(3), 231–240. doi:10.1002/widm.30.

Kriegel, H. P., Kröger, P., Schubert, E., & Zimek, A. (2011b). Interpreting and unifying outlier scores. In *Proceedings of the 11th SIAM international conference on data mining (SDM)*, Mesa, AZ (pp. 13–24).

Kriegel, H. P., Schubert, E., & Zimek, A. (2011c). Evaluation of multiple clustering solutions. In *2nd MultiClust workshop: discovering, summarizing and using multiple clusterings held in conjunction with ECML PKDD 2011*, Athens, Greece (pp. 55–66).

Kriegel, H. P., Kröger, P., & Zimek, A. (2012). Subspace clustering. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *2*(4), 351–364.

Kröger, P., & Zimek, A. (2009). Subspace clustering techniques. In L. Liu & M. T. Özsu (Eds.), *Encyclopedia of database systems* (pp. 2873–2875). Berlin: Springer. doi:10.1007/978-0-387-39940-9_607.

Kumar, A., & Daumé, H. (2011). A co-training approach for multi-view spectral clustering. In *Proceedings of the 28th international conference on machine learning (ICML)*, Bellevue, Washington, DC, USA (pp. 393–400).

Kuncheva, L. I., & Hadjitodorov, S. T. (2004). Using diversity in cluster ensembles. In *Proceedings of the 2004 IEEE international conference on systems, man, and cybernetics (ICSMC)*, The Hague, Netherlands (pp. 1214–1219).

Lazarevic, A., & Kumar, V. (2005). Feature bagging for outlier detection. In *Proceedings of the 11th ACM international conference on knowledge discovery and data mining (SIGKDD)*, Chicago, IL (pp. 157–166). doi:10.1145/1081870.1081891.

Lee, S. G., Hur, J. U., & Kim, Y. S. (2004). A graph-theoretic modeling on GO space for biological interpretation of gene clusters. *Bioinformatics*, *20*(3), 381–388. doi:10.1093/bioinformatics/btg420.

Lelis, L., & Sander, J. (2009). Semi-supervised density-based clustering. In *Proceedings of the 9th IEEE international conference on data mining (ICDM)*, Miami, FL (pp. 842–847). doi:10.1109/ICDM.2009.143.

Leman, D., Feelders, A., & Knobbe, A. J. (2008). Exceptional model mining. In *Proceedings of the European conference on machine learning and knowledge discovery in databases (ECML/PKDD)*, Antwerp, Belgium (pp. 1–16).

Li, T., & Ding, C. (2008). Weighted consensus clustering. In *Proceedings of the 8th SIAM international conference on data mining (SDM)*, Atlanta, GA (pp. 798–809).

Ling, R. F. (1972). On the theory and construction of k-clusters. *Computer Journal*, *15*(4), 326–332.

Ling, R. F. (1973). A probability theory of cluster analysis. *Journal of the American Statistical Association*, *68*(341), 159–164.

Liu, G., Li, J., Sim, K., & Wong, L. (2007). Distance based subspace clustering with flexible dimension partitioning. In *Proceedings of the 23rd international conference on data engineering (ICDE)*, Istanbul, Turkey (pp. 1250–1254). doi:10.1109/ICDE.2007.368985.

Liu, G., Sim, K., Li, J., & Wong, L. (2009). Efficient mining of distance-based subspace clusters. *Statistical Analysis and Data Mining*, *2*(5–6), 427–444. doi:10.1002/sam.10062.

Long, B., Zhang, Z., & Yu, P. S. (2005). Combining multiple clustering by soft correspondence. In *Proceedings of the 5th IEEE international conference on data mining (ICDM)*, Houston, TX (pp. 282–289). doi:10.1109/ICDM.2005.45.

Lord, P. W., Stevens, R. D., Brass, A., & Goble, C. A. (2003). Investigating semantic similarity measures across the Gene ontology: the relationship between sequence and annotation. *Bioinformatics*, *19*(10), 1275–1283. doi:10.1093/bioinformatics/btg153.

Madeira, S. C., & Oliveira, A. L. (2004). Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *1*(1), 24–45. doi:10.1109/TCBB.2004.2.

Mampaey, M., Tatti, N., & Vreeken, J. (2011). Tell me what I need to know: succinctly summarizing data with itemsets. In *Proceedings of the 17th ACM international conference on knowledge discovery and data mining (SIGKDD)*, San Diego, CA. New York: ACM Press.

Mannila, H., & Toivonen, H. (1997). Levelwise search and borders of theories in knowledge discovery. *Data Mining and Knowledge Discovery*, *1*(3), 241–258.

McCallum, A., Rosenfeld, R., Mitchell, T. M., & Ng, A. Y. (1998). Improving text classification by shrinkage in a hierarchy of classes. In *Proceedings of the 15th international conference on machine learning (ICML)*, Madison, WI (pp. 359–367).

Miettinen, P., Mielikäinen, T., Gionis, A., Das, G., & Mannila, H. (2008). The discrete basis problem. *IEEE Transactions on Knowledge and Data Engineering*, *20*(10), 1348–1362.

Mitchell, T. M. (1977). Version spaces: a candidate elimination approach to rule learning. In *Proceedings of the 5th international joint conference on artificial intelligence (IJCAI)*, Cambridge, MA (pp. 305–310).

Moise, G., & Sander, J. (2008). Finding non-redundant, statistically significant regions in high dimensional data: a novel approach to projected and subspace clustering. In *Proceedings of the 14th ACM international conference on knowledge discovery and data mining (SIGKDD)*, Las Vegas, NV (pp. 533–541). doi:10.1145/1401890.1401956.

Moise, G., Zimek, A., Kröger, P., Kriegel, H. P., & Sander, J. (2009). Subspace and projected clustering: experimental evaluation and analysis. *Knowledge and Information Systems*, *21*(3), 299–326. doi:10.1007/s10115-009-0226-y.

Mörchen, F., Thies, M., & Ultsch, A. (2011). Efficient mining of all margin-closed itemsets with applications in temporal knowledge discovery and classification by compression. *Knowledge and Information Systems*, *29*(1), 55–80.

Müller, E., Assent, I., Günnemann, S., Krieger, R., & Seidl, T. (2009a). 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)*, Miami, FL (pp. 377–386). doi:10.1109/ICDM.2009.10.

Müller, E., Assent, I., Krieger, R., Günnemann, S., & Seidl, T. (2009b). DensEst: density estimation for data mining in high dimensional spaces. In *Proceedings of the 9th SIAM international conference on data mining (SDM)*, Sparks, NV (pp. 173–184).

Müller, E., Günnemann, S., Assent, I., & Seidl, T. (2009c). Evaluating clustering in subspace projections of high dimensional data. In *Proceedings of the 35th international conference on very large data bases (VLDB)*, Lyon, France (pp. 1270–1281).

Nagesh, H. S., Goil, S., & Choudhary, A. (2001). Adaptive grids for clustering massive data sets. In *Proceedings of the 1st SIAM international conference on data mining (SDM)*, Chicago, IL.

Nguyen, H. V., Ang, H. H., & Gopalkrishnan, V. (2010). Mining outliers with ensemble of heterogeneous detectors on random subspaces. In *Proceedings of the 15th international conference on database systems for advanced applications (DASFAA)*, Tsukuba, Japan (pp. 368–383). doi:10.1007/978-3-642-12026-8_29.

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 (ICML)*, Haifa, Israel (pp. 831–838).

Novak, P. K., Lavrac, N., & Webb, G. I. (2009). Supervised descriptive rule discovery: a unifying survey of contrast set, emerging pattern and subgroup mining. *Journal of Machine Learning Research*, *10*, 377–403.

Ntoutsi, E., Zimek, A., Palpanas, T., Kröger, P., & Kriegel, H. P. (2012). Density-based projected clustering over high dimensional data streams. In *Proceedings of the 12th SIAM international conference on data mining (SDM)*, Anaheim, CA (pp. 987–998).

Ojala, M. (2010). Assessing data mining results on matrices with randomization. In *Proceedings of the 10th IEEE international conference on data mining (ICDM)*, Sydney, Australia (pp. 959–964).

Ojala, M., Vuokko, N., Kallio, A., Haiminen, N., & Mannila, H. (2008). Randomization of real-valued matrices for assessing the significance of data mining results. In *Proceedings of the 8th SIAM international conference on data mining (SDM)*, Atlanta, GA (pp. 494–505).

Ojala, M., Vuokko, N., Kallio, A., Haiminen, N., & Mannila, H. (2009). Randomization methods for assessing data analysis results on real-valued matrices. *Statistical Analysis and Data Mining*, *2*(4), 209–230.

Pasquier, N., Bastide, Y., Taouil, R., & Lakhal, L. (1999a). Discovering frequent closed itemsets for association rules. In *Proceedings of the 7th international conference on database theory (ICDT)*, Jerusalem, Israel.

Pasquier, N., Bastide, Y., Taouil, R., & Lakhal, L. (1999b). Discovering frequent closed itemsets for association rules. In *Proceedings of the 7th international conference on database theory (ICDT)*, Jerusalem, Israel (pp. 398–416). New York: ACM Press.

Pensa, R. G., Robardet, C., & Boulicaut, J. F. (2005). A bi-clustering framework for categorical data. In *Proceedings of the 9th European conference on principles and practice of knowledge discovery in databases (PKDD)*, Porto, Portugal (pp. 643–650).

Poernomo, A. K., & Gopalkrishnan, V. (2009). Towards efficient mining of proportional fault-tolerant frequent itemsets. In *Proceedings of the 15th ACM international conference on knowledge discovery and data mining (SIGKDD)*, Paris, France (pp. 697–706).

Prelić, A., Bleuler, S., Zimmermann, P., Wille, A., Bühlmann, P., Guissem, W., Hennig, L., Thiele, L., & Zitzler, E. (2006). A systematic comparison and evaluation of biclustering methods for gene expression data. *Bioinformatics*, *22*(9), 1122–1129. doi:10.1093/bioinformatics/btl060.

Qi, Z. J., & Davidson, I. (2009). A principled and flexible framework for finding alternative clusterings. In *Proceedings of the 15th ACM international conference on knowledge discovery and data mining (SIGKDD)*, Paris, France (pp. 717–726). doi:10.1145/1557019.1557099.

Ramakrishnan, N., Kumar, D., Mishra, B., Potts, M., & Helm, R. F. (2004). Turning cartwheels: an alternating algorithm for mining redescriptions. In *Proceedings of the 10th ACM international conference on knowledge discovery and data mining (SIGKDD)*, Seattle, WA (pp. 266–275).

Rissanen, J. (1978). Modeling by shortest data description. *Automatica*, *14*(1), 465–471.

Schapire, R. E., & Singer, Y. (2000). BoosTexter: a boosting-based system for text categorization. *Machine Learning*, *39*(2–3), 135–168. doi:10.1023/A:1007649029923.

Schubert, E., Wojdanowski, R., Zimek, A., & Kriegel, H. P. (2012). On evaluation of outlier rankings and outlier scores. In *Proceedings of the 12th SIAM international conference on data mining (SDM)*, Anaheim, CA (pp. 1047–1058).

Segal, E., Taskar, B., Gasch, A., Friedman, N., & Koller, D. (2001). Rich probabilistic models for gene expression. *Bioinformatics*, *17(Suppl*(1), S243–S252.

Seppanen, J. K., & Mannila, H. (2004). Dense itemsets. In *Proceedings of the 10th ACM international conference on knowledge discovery and data mining (SIGKDD)*, Seattle, WA (pp. 683–688).

Sibson, R. (1973). SLINK: an optimally efficient algorithm for the single-link cluster method. *Computer Journal*, *16*(1), 30–34. doi:10.1093/comjnl/16.1.30.

Silla, C. N., & Freitas, A. A. (2011). A survey of hierarchical classification across different application domains. *Data Mining and Knowledge Discovery*, *22*(1–2), 31–72. doi:10.1007/s10618-010-0175-9.

Sim, K., Gopalkrishnan, V., Zimek, A., & Cong, G. (2012). A survey on enhanced subspace clustering. *Data Mining and Knowledge Discovery*. doi:10.1007/s10618-012-0258-x.

Singh, V., Mukherjee, L., Peng, J., & Xu, J. (2010). Ensemble clustering using semidefinite programming with applications. *Machine Learning*, *79*(1–2), 177–200.

Smets, K., & Vreeken, J. (2012). Slim: directly mining descriptive patterns. In *Proceedings of the 12th SIAM international conference on data mining (SDM)*, Anaheim, CA (pp. 1–12). Philadelphia: Society for Industrial and Applied Mathematics (SIAM).

Sneath, P. H. A. (1957). The application of computers to taxonomy. *Journal of General Microbiology*, *17*, 201–226.

Sridharan, K., & Kakade, S. M. (2008). An information theoretic framework for multiview learning. In *Proceedings of the 21st annual conference on learning theory (COLT)*, Helsinki, Finland (pp. 403–414).

Strehl, A., & Ghosh, J. (2002). Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, *3*, 583–617.

Stuetzle, W. (2003). Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample. *Journal of Classification*, *20*(1), 25–47. doi:10.1007/s00357-003-0004-6.

Tatti, N. (2008). Maximum entropy based significance of itemsets. *Knowledge and Information Systems*, *17*(1), 57–77.

Tatti, N., & Mörchen, F. (2011). Finding robust itemsets under subsampling. In *Proceedings of the 11th IEEE international conference on data mining (ICDM)*, Vancouver, BC (pp. 705–714).

Tatti, N., & Vreeken, J. (2011). Comparing apples and oranges: measuring differences between data mining results. In *Proceedings of the European conference on machine learning and knowledge discovery in databases (ECML PKDD)*, Athens, Greece (pp. 398–413). Berlin: Springer.

Tatti, N., & Vreeken, J. (2012). The long and the short of it: summarizing event sequences with serial episodes. In *Proceedings of the 18th ACM international conference on knowledge discovery and data mining (SIGKDD)*, Beijing, China.

Thabtah, F. A., Cowling, P., & Peng, Y. (2004). MMAC: a new multi-class, multi-label associative classification approach. In *Proceedings of the 4th IEEE international conference on data mining (ICDM)*, Brighton, UK (pp. 217–224). doi:10.1109/ICDM.2004.10117.

Topchy, A., Jain, A., & Punch, W. (2005). Clustering ensembles: models of consensus and weak partitions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *27*(12), 1866–1881. doi:10.1109/TPAMI.2005.237.

Topchy, A. P., Law, M. H. C., Jain, A. K., & Fred, A. L. (2004). Analysis of consensus partition in cluster ensemble. In *Proceedings of the 4th IEEE international conference on data mining (ICDM)*, Brighton, UK (pp. 225–232). doi:10.1109/ICDM.2004.10100.

Tsoumakas, G., & Katakis, I. (2007). Multi-label classification: an overview. *International Journal of Data Warehousing and Mining*, *3*(3), 1–13.

Valentini, G., & Masulli, F. (2002). Ensembles of learning machines. In *Proceedings of the 13th Italian workshop on neural nets*, Vietri, Italy (pp. 3–22). doi:10.1007/3-540-45808-5_1.

van Leeuwen, M., Vreeken, J., & Siebes, A. (2009). Identifying the components. *Data Mining and Knowledge Discovery*, *19*(2), 173–292.

Vendramin, L., Campello, R. J. G. B., & Hruschka, E. R. (2010). Relative clustering validity criteria: a comparative overview. *Statistical Analysis and Data Mining*, *3*(4), 209–235. doi:10.1002/sam.10080.

Vreeken, J., & Zimek, A. (2011). When pattern met subspace cluster—a relationship story. In *2nd MultiClust workshop: discovering, summarizing and using multiple clusterings held in conjunction with ECML PKDD 2011*, Athens, Greece (pp. 7–18).

Vreeken, J., van Leeuwen, M., & Siebes, A. (2011). Krimp: mining itemsets that compress. *Data Mining and Knowledge Discovery*, *23*(1), 169–214.

Wang, C., & Parthasarathy, S. (2006). Summarizing itemset patterns using probabilistic models. In *Proceedings of the 12th ACM international conference on knowledge discovery and data mining (SIGKDD)*, Philadelphia, PA (pp. 730–735).

Wang, H., Azuaje, F., Bodenreider, O., & Dopazo, J. (2004). Gene expression correlation and gene ontology-based similarity: an assessment of quantitative relationships. In *Proceedings of the 2004 IEEE symposium on computational intelligence in bioinformatics and computational biology (CIBCB)*, La Jolla, CA.

Webb, G. I. (2007). Discovering significant patterns. *Machine Learning*, *68*(1), 1–33.

Wishart, D. (1969). Mode analysis: a generalization of nearest neighbor which reduces chaining effects. In A. J. Cole (Ed.), *Numerical taxonomy* (pp. 282–311).

Wrobel, S. (1997). An algorithm for multi-relational discovery of subgroups. In *Proceedings of the 1st European symposium on principles of data mining and knowledge discovery (PKDD)*, Trondheim, Norway (pp. 78–87).

Xiang, Y., Jin, R., Fuhry, D., & Dragan, F. (2011). Summarizing transactional databases with overlapped hyperrectangles. *Data Mining and Knowledge Discovery*, *23*(2), 215–251.

Yan, B., & Domeniconi, C. (2006). Subspace metric ensembles for semi-supervised clustering of high dimensional data. In *Proceedings of the 17th European conference on machine learning (ECML)*, Berlin, Germany (pp. 509–520).

Yan, X., Cheng, H., Han, J., & Xin, D. (2005). Summarizing itemset patterns: a profile-based approach. In *Proceedings of the 11th ACM international conference on knowledge discovery and data mining (SIGKDD)*, Chicago, IL (pp. 314–323).

Zeeberg, B. R., Feng, W., Wang, G., Wang, M. D., Fojo, A. T., Sunshine, M., Narasimhan, S., Kane, D. W., Reinhold, W. C., Lababidi, S., Bussey, K. J., Riss, J., Barrett, J. C., & Weinstein, J. N. (2003). GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biology*, *4*(4), R28.

Zheng, L., & Li, T. (2011). Semi-supervised hierarchical clustering. In *Proceedings of the 11th IEEE international conference on data mining (ICDM)*, Vancouver, BC (pp. 982–991).

Zimek, A., Buchwald, F., Frank, E., & Kramer, S. (2010). A study of hierarchical and flat classification of proteins. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *7*(3), 563–571. doi:10.1109/TCBB.2008.104.

Zimek, A., Schubert, E., & Kriegel, H. P. (2012). A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining*, *5*(5), 363–387. doi:10.1002/sam.11161.