On Using Class-Labels in Evaluation of Clusterings

Ines Färber[1], Stephan Günnemann[1], Hans-Peter Kriegel[2], Peer Kröger[2], Emmanuel Müller[1], Erich Schubert[2], Thomas Seidl[1], Arthur Zimek[2]

[1]RWTH Aachen University
Ahornstrasse 55, 52056 Aachen, Germany
http://www.dme.rwth-aachen.de/
{faerber,guennemann,mueller,seidl}@cs.rwth-aachen.de

[2]Ludwig-Maximilians-Universität München
Oettingenstrasse 67, 80538 München, Germany
http://www.dbs.ifi.lmu.de/
{kriegel,kroeger,schube,zimek}@dbs.ifi.lmu.de

ABSTRACT

Although clustering has been studied for several decades, the fundamental problem of a valid evaluation has not yet been solved. The sound evaluation of clustering results in particular on real data is inherently difficult. In the literature, new clustering algorithms and their results are often externally evaluated with respect to an existing class labeling. These class-labels, however, may not be adequate for the structure of the data or the evaluated cluster model. Here, we survey the literature of different related research areas that have observed this problem. We discuss common "defects" that clustering algorithms exhibit w.r.t. this evaluation, and show them on several real world data sets of different domains along with a discussion why the detected clusters do not indicate a bad performance of the algorithm but are valid and useful results. An useful alternative evaluation method requires more extensive data labeling than the commonly used class labels or it needs a combination of information measures to take subgroups, supergroups, and overlapping sets of traditional classes into account. Finally, we discuss an evaluation scenario that regards the possible existence of several complementary sets of labels and hope to stimulate the discussion among different sub-communities — like ensemble-clustering, subspace-clustering, multi-label classification, hierarchical classification or hierarchical clustering, and multiview-clustering or alternative clustering — regarding requirements on enhanced evaluation methods.

1. INTRODUCTION

Evaluating the quality of clustering results is still a challenge in recent research. One kind of evaluation is the use of internal measures as e.g. compactness or density of clusters. Because these measures usually reflect the objective functions of particular clustering models, the evaluation of clustering results based on this technique, however, is problematic. In general, an algorithm specifically designed for the objective function used in the evaluation outperforms its competing approaches in terms of clustering quality. Thus, a fair validation of the results is not achieved by using these measures. Accordingly, the best way for fair evaluations so far is using external evaluation measures. Based on a data set whose true clustering structure is given, the results of an algorithm are compared against this ground truth. A comparable evaluation is ensured because the true clustering structure can be chosen independently of specific models. For this technique, however, the definition of the ground truth is very problematic. The ground truths used so far in the evaluation of clustering results are mostly inadequate, which we will substantiate in the following.

Synthetic data sets can be engineered to match assumptions of the occurrences and properties of meaningful clusters. The underlying assumptions and the thereon based generation of the data allows the deduction of a ground truth for these data sets. The evaluation can then demonstrate to which degree the algorithms actually find clusters that match these assumptions. Thus, this procedure allows the explication of the assumptions of the algorithms themselves. It is important, though, to show the significance of a solution also on real world data to demonstrate the nature of the problem tackled by the specific approach existing beforehand in the real world. For real world data, however, specifying the true clustering is difficult since mostly the knowledge of domain experts is required. This problem does not arise for the supervised mining task of classification. The evaluation procedures in this area are well studied (cf. e.g. [63]) and relatively straightforward. Though the aim of classification usually is to generalize concepts describing previously known knowledge for classifying new, unlabeled data, its success can be assessed quantitatively w.r.t. a relation between the provided set of class-labels and the rediscovered classes. For this kind of evaluation, in the context of classification many different measures are available (see e.g. [50]).

Since using labeled classification benchmark data is rather convenient, the usual approach in evaluation of clustering algorithms is to use such data sets based on the assumption that the natural grouping of the data set (which the clustering algorithm is to find) is reflected by the class labels to a certain degree. Using classification data for the purpose of evaluating clustering results, however, encounters several problems since the class labels do not necessarily correspond to natural clusters. A typical example includes the clustering specific identification of outliers, i.e., of objects that do not belong to any cluster. In classification data, however, usually each object has assigned a certain class label. Thus, a clustering algorithm that detects outliers is actu-

ally punished in an evaluation based on these class labels even though it should be rewarded for identifying outliers as not belonging to a common cluster albeit outliers represent a genuine class of objects. Similar difficulties occur if the labeled classes split up in different sub-clusters or if several classes cannot be distinguished leading to one larger cluster. Consequently, in general, classes cannot be expected to exactly correspond to clusters.

Let us consider the problem from a different perspective. The already annotated classes usually do not completely exploit the knowledge hidden in the data. The whole point in performing unsupervised methods in data mining is to find previously unknown knowledge. Or to put it another way, additionally to the (approximately) given object groupings based on the class labels, several further views or concepts can be hidden in the data that the data miner would like to detect. If a clustering algorithm detects structures that deviate considerably from annotated classes, this could actually be a good and desirable result. Consider for example the case of subspace clustering algorithms [39], where not only groups of data objects define a cluster but also certain subsets (or combinations) of the attributes of the data. The general task definition of "subspace clustering" is to look for "all clusters in all subspaces" which allows for overlap of clusters, i.e., one object can belong to different clusters simultaneously in different subspaces. Therefore, determining the true clustering based on class labels cannot be accurate since different groupings for each individual object are possible. This observation goes beyond subspace clustering and is also true for other approaches. Alternative clustering and multi-view clustering methods are interested in finding clusterings that differ from a previously known clustering. Again, given the class labels, an evaluation may not be correct. Thus, for a meaningful evaluation of clustering algorithms, multiple labeled data should be used where for each object a set of possible clusters is specified — and such annotations of evaluation benchmark data sets should be updated if new, meaningful concepts can be identified in the data. Even the recently conducted experimental evaluations of subspace clustering algorithms [48, 46] suffer from these problems. In both studies, for real data sets, class labels have been assumed to correspond also to true clusters. A recent subspace clustering method for detecting orthogonal views in data, resorts to a different evaluation approach [28]. Multiple groupings are obtained by concatenating multiple times the object vectors of usual classification data (permuting in each step the order of objects). Thereby, for each object several clusters are generated and more meaningful conclusions are possible. This approach, however, is more like synthesizing data based on real data than solving the problems sketched here. Actually, the original problem is not solved since already the unmodified data comprises several groupings that are not incorporated into the evaluation. At last, the approaches in [19, 52] perform only a subjective evaluation by manually inspecting their results avoiding the problematic objective quantitative external evaluation.

Overall, the true hindrance of scientific progress in the area of clustering hence becomes apparent as being the lack of data sets carefully studied regarding their clustering structure and suitably prepared for a more meaningful evaluation of clustering approaches. Here, we discuss and demonstrate the problems in using classification data for cluster evaluation. We aim at studying the characteristics of natural groupings in real world data and we describe different meaningful views or clusters within these data sets. We try to derive general problems and challenges for evaluating clusterings in order to attract the attention of the research community to this problem and to inspire discussions among researchers about how enhanced evaluation procedures could and should look like. This would be beneficial because we envision to annotate real world data with the information on alternative clusters in order to obtain reasonable groupings for the task of clustering and to eventually substitute the naive class label approach. It is our hope that if these more carefully annotated data sets can be provided along with enhanced evaluation procedures, this may initiate research for a more meaningful evaluation of clustering algorithms in the future.

Although the problem we are sketching here has not found much attention so far, we find some related observations in the literature which we lay out in Section 2. The formalization of the introduced problems and case studies on exemplary data sets that contain more than one meaningful set of concepts are provided in Section 3. Our observations lead to a new procedure for the evaluation of clustering results. In Section 4, we discuss possible evaluation scenarios. Section 5 concludes the paper.

2. RELATED WORK

2.1 The Problem in Different Research Areas

It has been observed now and then that classes and clusters need not be identical but, for example, one class can comprise several clusters. Such observations on the difference between clusters and classes have been occasionally reported in different research areas. However, in none of these research areas this difference has been truly taken into account for the evaluation of approaches, as we survey below.

An example concerned with *traditional clustering* is [15], where the difference between clusters and classes is noted though not taken into account for the evaluation. The clustering research community did not pay much attention, however, on this rather obvious possibility. For example, although the observation of [15] is quoted and used for motivating *ensemble clustering* in [57], the evaluation in the latter study uses uncritically classification benchmark data sets (including the pendigits data set that we survey below in Section 3.4). The problematic variants of approaches to ensemble clustering cannot be discussed here. It is, however, our conviction that the results we report in this study are of utmost importance especially for the research on ensemble clustering since focussing on one single clustering result in the presence of different, possibly equally important clustering solutions, is an inherent flaw of many approaches to ensemble clustering.

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 is provided in [60]. 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. There are, of course, many other possibilities of creating a flat set of class labels (see e.g. [53, 29, 12, 60]). 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., in [17, 59]). 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. Again, our present study envisions to facilitate development of clustering approaches more apt to treat such complex data. This is, however, also partly motivating subspace clustering and biclustering (see below) but even there, no suitable evaluation technique has been developed until now.

A special case of multi-label classification is *hierarchical classification* (see e.g. [38, 44, 14, 11, 13]). Here, each class is either a top-level class or a subclass of any other class. 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) or classes not sharing a common superclass. Evaluation of approaches to hierarchical classification is usually performed by choosing one specific level corresponding to a certain granularity of the classification task. Hierarchical problems have also been studied in clustering and actually represent the majority of older work in this area [55, 62, 54, 33]. Recent work includes [7, 58, 1, 2, 4]. Though there are approaches to evaluate several or all levels of the hierarchy [23], there has never been a systematic, numerical methodology of evaluating such hierarchical clusterings *as hierarchies*. 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.

In the areas of *bi-clustering* or *co-clustering* [43, 39] 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 (starting with [32]), there has not been described a general method of evaluation of clustering results in this field either. Some approaches are commonly accepted for the biological domain of protein data, though, which we will shortly describe below.

Subspace clustering pursues the goal to find all clusters in all subspaces of the entire feature space [39]. This goal obviously is defined to correspond to the bottom-up technique used by these approaches, based on some anti-monotonic property of clusters allowing the application of the APRIORI [56] search heuristic. Examples for subspace clustering include [5, 16, 49, 37, 9, 47, 10, 45, 41]. Since the initial problem formulation of finding all clusters in all subspaces is rather questionable (the information gained by retrieving such a huge set of clusters with high redundancy is not very useful), subsequent methods often concentrated on possibilities of restricting the resulting set of clusters by somehow assessing and reducing the redundancy of clusters, for example to keep only clusters of highest dimensionality. However, none of these methods were designed for detection of alternative subspace clusters. Only recently, for subspace clustering the notion of orthogonal concepts has been introduced [28] constituting a direct connection between the areas of subspace clustering and alternative clustering.

Recently, the problem description of *multiview clustering* [19, 36] or *finding alternative clusterings* [21, 52] has been brought forward. The results stated in these studies concur with our observations that different (alternative) clustering structures in one data set may be possible, meaningful, and a good result for any clustering algorithm. However, they implicitly also support our claim that new evaluation techniques are necessary. Since they cannot rely on the flat and simple set of class labels for evaluation, they evaluate their alternative clustering results mainly by manual inspection and interpretation of the clusters found. While the notion of multiview clustering [19] is closely related to the problems tackled in subspace clustering, the notion of alternative clusterings discussed in [21, 52] provide a different perspective not relying on the notion of subspaces or different views of a data set but constraining clustering solutions to differ from each other. In [30], the problem setting is explicitly described as seeking a clustering different from the already known classification. Nevertheless, regarding a quantitative evaluation of results, all these approaches concur with the other research areas sketched above in motivating the considerations we present in this study.

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. This observation motivates our detailed discussion of the problems resulting for an appropriate evaluation of clustering algorithms in order to trigger efforts for developing enhanced evaluation techniques.

2.2 Observations in Different Data Domains

Similarly to these problem observations in different research areas we observe also the usage of several data domains showing multiple hidden clusters per object. In general all of the multiview and alternative clustering evaluations are based on such databases. Furthermore, there are similar observations in the more application oriented domain of gene expression analysis.

Data used in the experiments of recent techniques [19, 21, 52], range from the pendigits database provided by UCI ML repository [24] up to image databases containing Escher images which are known to have multiple interpretations to the human eye [52]. In general, the observation for all of these databases is that data are known to contain multiple interpretations and thus also multiple hidden clusters per object. However, all of these databases provide only single class labels for evaluation.

For gene expression data, one observes multiple functional relationships for each gene to be detected by clustering algorithms. A couple of methods has been proposed in order to evaluate clusters retrieved by arbitrary clustering methods [8, 64, 6]. 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) [8]. For example, FatiGO[6] tries to judge whether GO terms are over- or under-represented in a set of genes w.r.t. a reference set. More theoretically, the cluster validity is measured by means

of how good the obtained clusters match the class system where the class system exists of several directed graphs, i.e., there are hierarchical elements and elements of overlap or multiple labels. However, examples of such measures include precision/recall values, or the measures reviewed in [31]. 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 [20, 25, 27, 40, 51].

Previous evaluations do in fact report many found clusters to not obviously reflect known structure, possibly, however, due to the fact that the used biological knowledge bases are very incomplete [18]. 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 [61] or a relationship between sequence similarity and semantic similarity [42].

In summary, although the research community concerned with gene expression data is pretty aware of the fact that there is not a one-and-only flat set of class-labels and that, if there were one, to rediscover it by using clustering is in itself not a meaningful goal of scientific progress, there is also no readily generalizable methodology for evaluating clustering algorithms in presence of complex, possibly overlapping layers of different classes.

3. ALTERNATIVE CLUSTERINGS

As motivated in the previous sections there are multiple meaningful reasons to allow alternative clusterings. In such cases there is not one ground truth to be detected by a clustering task but there are multiple alternative valid solutions. Let us formalize this before providing some case studies on real world data.

3.1 Classic Evaluation of Clusterings

A typical clustering algorithm aims at detecting a grouping of objects in disjoint clusters $\{G_1, \ldots, G_k\}$ such that similar objects are grouped in one cluster while dissimilar objects are separated in different clusters. For evaluation usually a given ground truth, such as the class labels in classification data sets is used. In such databases, each object is assigned to exactly one class out of $\{C_1, \ldots, C_l\}$. The basic assumption is that these class labels constitute an ideal clustering solution which should be detected by any useful clustering algorithm. In typical clustering evaluations (e.g. [46, 48]), one uses purity and coverage measures for comparison of a detected cluster G_i with the given classes C_j. Such evaluations result in perfect quality of a clustering solution if its clusters represent exactly the given classes. The quality defined by these measures decreases if a detected cluster does not represent one of the given class labels, if class labels are split in multiple clusters, or if different class labels are mixed in one cluster. Different possibilities for comparing partitions exist (see e.g. [35, 50]) but they suffer from the same problem.

We doubt that such an evaluation based only on one given class label per object is a fair and meaningful criterion for high quality clusterings. In fact it can be seen as an "overfitting" of the evaluation method to the class labels. Clustering is usually used to discover new structures in the data, instead of reproducing known structure.

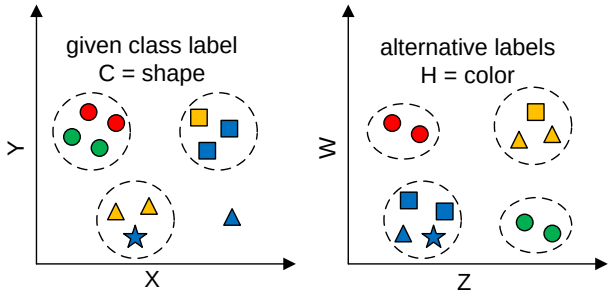

Figure 1: Toy example on alternative clusters

3.2 Clustering "Defects" Analyzed

When analyzing clustering results, one often encounters clusters that are meaningful, but do not necessarily correspond exactly to the class labels used in evaluation. In the following, we want to discuss some of the common "defects" and their causes theoretically. Below (Section 3.4), we will discuss exemplary real-data occurrences of these defects.

Figure 1 shows two example clusterings of the same data set. The first clustering extracted the feature projections (also: view, subspace) X and Y and identified the groups on the left side, which correspond closely to the class labeling, visualized as shape of the objects. A typical evaluation scenario would assign a high score to this result. The clustering on the right hand side, using W and Z projections, scores rather low in traditional evaluation: none of the clusters corresponds closely to a class label, even the circle objects are broken apart into two separate clusters. However, when you take the color of the objects into account, this partitioning of the data is perfect. So assuming that the color is a hidden structure in the data, this is a very good clustering result (albeit it is not as useful for re-identification of the original classes). In particular, the second clustering discovered a hidden structure in the circular shaped objects. As such, the second result actually provides more insight into the data: the circular class can be split into two meaningful groups, while the other classes have a grouping that is orthogonal to their classes. When taking this result into a classification context, this allows for improved training of classifiers, by training two classifiers for the circular objects and by removing the W and Z features for separating the other classes.

Commonly observed differences between a real clustering result and a class labeling include:

Splitting of Classes into Multiple Clusters: A clustering algorithm might have detected subgroups of one class as multiple clusters. Thus, they split the original class structure in more than one cluster. In our example we observe this phenomenon for the green and red circles. Based on the clustering, the elements of such subgroups might be quite dissimilar to each other although they are labeled with the same class. This is in general true for classes representing a multi-modal distribution. However, using class labels (e.g. the shape in our example) in the evaluation often results in allegedly bad clustering quality.

Merging of Classes into a Single Cluster: Then again, multiple class labels might be detected together in one cluster. Objects from different classes might share common properties for some attributes, such that they are grouped

together. In our example, both the orange and blue cluster in the W and Z projection are mixing-up different shapes. There is no completely merged cluster in the toy example, but we will observe many examples in the real world experiments.

Missing Class Outliers: Clustering algorithms often determine only obvious groupings as clusters and do not try to assign outlying objects. However in a class label context, also "class outliers" (that is unusual members of the class) are assigned the same label. In a classification context, these objects are important. For example, a Support Vector Machine relies on such objects as support vectors. In a clustering context, models are derived and the focus is on typical objects. When learning correlation models [3], such untypical objects can significantly reduce model quality. An example of treating class outliers in an evaluation different than class inliers is [34].

Multiple (Overlapping) Hidden Structures: As a common observation, one class label per object is not sufficient to represent multiple hidden structures. Recent clustering applications aim at such multiple and alternative cluster detection [30, 19, 36, 21, 52]. As a consequence, also the methodology of evaluation should be reconsidered in these cases. As depicted in our example, the hidden structure is not only the object shape but also its color. Using this alternative hidden grouping $\{H_1, \ldots, H_m\}$ is essential for a fair evaluation of alternative clusterings. Each alternative hidden group H_i might represent a split of the original class labels or a totally different grouping by mixing-up multiple classes or parts thereof. Thus, classes C_i and alternative hidden structures H_j together provide a more enhanced ground truth for cluster evaluation.

3.3 Evaluation Challenge

Evaluating a clustering against a data set containing alternative hidden groupings requires more sophisticated techniques. For example, when a clustering algorithm merges two classes that belong to the same hidden group, or splits a class into two clusters according to two hidden groups, this does not indicate an entirely bad performance of the algorithm. Similarly, a clustering algorithm identifying some class outliers as noise objects should not be heavily penalized in an evaluation of its results. But of course the clustering should still convey some meaning and not just contain arbitrary groups. In general, the evaluation of clusterings can happen at two levels: The first level deals with the question whether a sensible structure has been found or not while the second level evaluates the significance of this structure, i.e., whether it is trivial or not.

For the first level of evaluating clustering results typically the idea of partitionings of data is considered. Class labels usually define a *disjoint* and *complete* partitioning of the data. Many clustering algorithms return a similar partitioning, often with a special partition N of "noise" objects, that could also be seen as partitions that only contain a single object each. Frequently, this partition is treated the same way as a cluster, and the same happens for class labels that identify outliers. Hierarchical class labels and clusterings represent a more advanced structure. In a strict hierarchy, every cluster has at most one parent cluster. Some algorithms such as ERiC [4] can also produce arbitrarily (multi-) nested clusters. For many data sets there exists more than one meaningful grouping. Sometimes these groupings are

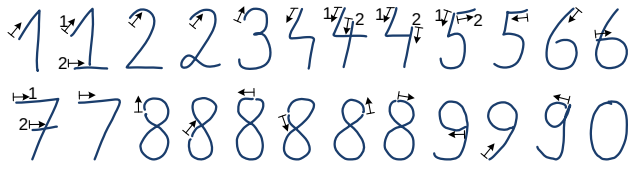

Figure 2: Different ways of digit notation

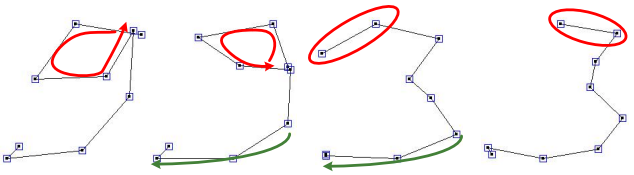

Figure 3: Different types of digits 9 and 3

orthogonal in the sense that there is no or only little correlation between the clusters of each grouping. Algorithms such as OSCLU [28] are able to detect these orthogonal clusters that often occur when the data is comprised of multiple views.

The most basic formulation however is that a data set can contain any number of non-disjoint "concepts" C_i. A data object can belong to any number of concepts. In a classification context, one can discern "interesting" and "uninteresting" concepts based on their correlation with the classification task. In a clustering context, concepts can be rated based on their information value (concepts having a reasonable size) but more importantly based on their discoverability using naive methods. This is more related to the second level of evaluating clusterings mentioned above. For example in an image data context, the concept of images with "dominant color red" can be considered of low value since this concept can be defined using a static, data-independent classifier. A clustering algorithm that merely groups images by color indeed detects a proper grouping of objects — but a rather trivial grouping.

3.4 Case Studies

The general phenomena of multiple alternative clusterings have been observed in multiple applications as surveyed above (Section 2). Here, we highlight and illustrate our key observations using two well-known real world data sets.

Pendigits Data Set: First, we consider the hidden clustering structures in the Pendigits database from the UCI repository [24]. In the Pendigits data set, hand written pen digits (objects) are described by 8 measurements of (x, y) positions. These are concatenated in a digit trajectory described by a 16 dimensional feature vector. For distinction of pen digits, clearly not all of the pen positions are important. Some digits show very similar values in the first positions as they all start in the upper left area. Thus, using a certain projection of the database one can distinguish certain subgroups of digits. As we will see, these subgroups do not necessarily correspond to the digit values given as class labels.

A careful manual analysis of this data set reveals that there exist different ways of notation for equal digits, as can be seen in Figure 2. Most of the common clustering methods, especially in full space, will therefore not detect

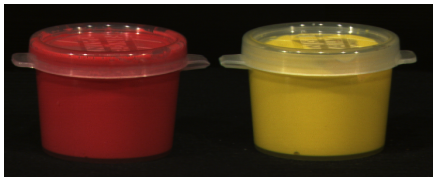

Figure 4: Two objects with similar shape in ALOI. The left can contains red, the right can contains yellow play dough.

all instances of one digit in one cluster but will split classes into multiple clusters. In the following we will take a closer look at differences in digit writing exemplarily for digits 3 and 9 in Figure 3. We consider the two basic observations of splitting classes and mixing-up classes in this real world example. For the class of digit 9 we observe a clear split into two detected clusters. While some of the digit 9 objects start in the upper right corner, a second subgroup of digit 9 objects start also at the rightmost position but with an offset downwards. Considering the first pen position stored in the first attributes, a subspace clustering algorithm should clearly separate these two types. On the other side, there are digits exhibiting highly similar trajectories although they represent different digit values. For the depicted digit 3 and digit 9 we observe a common ending of the trajectory. Both show a clear round curve stored as similar values in their last attributes. Thus, considering these attributes only as a subspace will lead to mixing-up of these two classes.

These are only two examples of alternative clusters hidden in this real world data set. We have found many more valid groupings of digits, representing almost 30 different groups of digits in contrast to the 10 given classes provided by the original classification database from the UCI archive. Thus, using only the given class labels might result in an unfair comparison especially for recent clustering tasks like subspace and alternative clustering.

ALOI Data Set: The real world data set "ALOI" comes from the ALOI image database [26]. The ALOI database consists of 110,250 images (instances) of 1,000 objects taken from different orientations and in different lighting conditions, each object being treated as a class. We produced color histograms based on the HSB color model using $7\times3\times3$ bins. Running DBSCAN with Histogram Intersection distance with $\varepsilon = 0.20$ and minPts $= 20$ the data set clusters into 370 clusters and noise, where noise contains around half of the data, three large clusters and a variety of small clusters often around 50 images. Many of these small clusters are pure, with all images coming from the same object. But there are some clusters with more interesting contents. Figure 4 shows two objects from a larger cluster found containing images from exactly 2 objects: The rather coarse binning into 7 hue values made the two objects next to identical with respect to this representation. Note that we used a rather traditional clustering algorithm and that the mixing up in this case truly happens already at the feature extraction process but this is not the point here. While the objects were distinguishable using larger color histograms, they do have a clear semantic relationship as the objects differ only in their color. Even the best shape based features (and clustering algorithms on top of that) would (and *should*) not be able to separate them.

Figure 5: Three objects in two views from ALOI. DBSCAN on color histograms generated one cluster containing front views (top row), another cluster containing side and back views (bottom row).

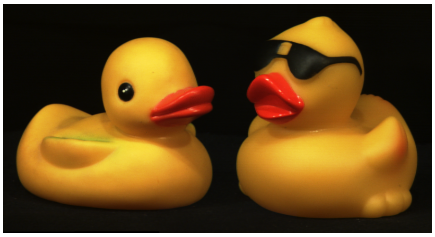

Figure 6: Different rubber duckies in ALOI, not separated by DBSCAN.

An even more interesting sample is formed by two clusters containing images from 3 different objects. Figure 5 contains images from these clusters. But instead of separating objects into different clusters, the algorithm separated different views on the objects, with one cluster containing the front views of all three objects (46 images) and the other cluster containing back and side views (66 images), again of all three objects. Obviously, this is due to the three objects being very similar baking mixes and having next to identical colors on the front, with the back and side views having different characteristic colors. Again, adding shape features would not yet help the algorithm to discern the objects. It can be claimed that "it is a feature, not a bug" that the extracted features, distance function and clustering algorithm produce these results. The detected groups have a valid semantic meaning that can be easily described in natural language as "baking mix front views" and "baking mix side and back views".

Figure 6 is yet another example from the same DBSCAN run, where it failed to separate two classes. This cluster contains images of two different rubber duckies contained in the ALOI set. While they are separate objects (one rubber duck is wearing sunglasses), it is debatable whether or not a clustering algorithm should cluster these objects into two separate clusters or merge them into a single one. There are plenty examples of this kind: two similar red molding forms, two painted easter eggs in similar colors, five yellow plastic objects of a larger size, two next to identical metal cans.

Table 1 gives some example concepts on ALOI. Some of these can be seen as hierarchical concepts (e.g. exact object

Concept	Available	Example
Exact Object	Filename	123
Lighting condition	Filename	l8c3 *(incomplete)*
Viewing angle	Filename	r35 *(not comparable)*
Object type	no	bell pepper, fruit, ...
Dominant color	no	yellow
Size	no	small, large, ...
Basic shape	no	rectangular, ...

Table 1: Example concepts on ALOI

being a subdivision of object type), others are clearly orthogonal: there are red bell peppers (dominant color red, object type bell pepper) as well as yellow bell peppers, red play dough and yellow play dough (dominant color yellow, object type play dough). Specialized features can be useful to identify some of these concepts (e.g. color bias for lighting conditions), but in particular the object type concepts are a human concept that does not necessarily map to object properties (e.g. "fruit" as informal human concept that often disagrees with the biological notion of a fruit).

4. POSSIBLE EVALUATION SCENARIOS

So far, we surveyed different research areas and different data domains backing up the conjecture that a single gold-standard provided in terms of class labels is not sufficient information to judge about different merits of different clustering algorithms. Annotated class labels are insufficient to function as a ground truth for clustering evaluation for several reasons. First, class labels represent a theoretical aggregation of data objects. This categorization may not become spatially manifested in the sense of cluster definitions. Class structures therefore do not necessarily represent an underlying clustering structure. Second, for many databases more than one view with meaningful and interesting clusters can be determined. With multiple views, several, potentially overlapping clusters are possible such that one object can be assigned to more than one cluster.

As a consequence of these insights, an evaluation solely based on traditional class labels as ground truth does not allow to draw any conclusions about the quality of the clustering approach under consideration. A more careful annotation of data sets, that tries to account for the inherent clustering structure of the data, will lead to more objective and meaningful conclusions and quality statements for the evaluation results (cf. Fig. 7). Previous approaches compared their clustering results solely against the class labels. The idea is now to compare the results with additional sets H_i of concept annotations. While the former evaluation procedures more or less tested the applicability of a particular cluster approach as classifier, an evaluation enhanced by considering also hidden structures allows insights into the ability of the cluster approach itself to capture particular concepts, also in dependance on the views involved. For example, a hierarchical clustering may rely on one view to detect a larger concept, then use another view to split this into subconcepts.

The provision of such cluster oriented ground truths is a major step towards enhanced, meaningful and objective clustering evaluation. However, a such structured ground truth rises new challenges and questions. Solving these challenges is beyond the scope of this paper but one aim of this

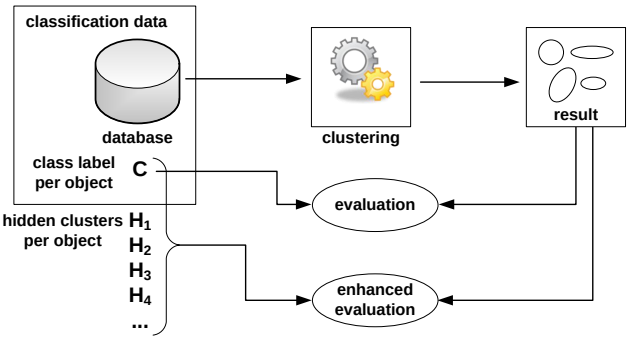

Figure 7: Traditional vs. enhanced evaluation.

contribution is to attract the attention of the research community to these problems and inspire discussions on how such solutions should look like.

If multiple hidden clusters are annotated to the data, commonly used evaluation measures are not appropriate anymore. A new, more meaningful evaluation measure has essentially to cope with several of the following scenarios and the question is whether to allow or to punish them:

- A result cluster covers exactly one concept but only contains part of the objects (e.g. missing outlier members).
- A result cluster covers the join of multiple concepts either completely or also only partially.
- The most challenging problem will probably be the case of newly detected clusters, not yet covered by a concept in the ground truth. To punish or reward the mentioned newly detected clusters presumes, that one has fully understood the clustering structure.
- Since the ground truth may represent different views like e.g. *rotation*, *color*, or *shape*, it is also reasonable to discuss the intersection of ground truth concepts from different views. These deviated clusters describe multiple views simultaneously, like *color* $\wedge$ *shape*.

In some cases, the experienced user might be able to select a feature space and a clustering algorithm that is biased towards the desired concept level. However, here we are concerned with a fair evaluation of clustering algorithms without focussing on certain application areas. By this survey and our observations, we finally hope to stimulate discussion about the problems between the different research communities and to enhance the mutual understanding of scientists concerned with different, but related problem formulations.

5. CONCLUSION

In this study, we surveyed different research areas where the observation that different clustering solutions may be equally meaningful has been reported. The obvious conclusion is that the evaluation of clustering results w.r.t. some one-and-only gold standard does not seem to be the method of choice. It is not only somewhat questionable to evaluate unsupervised methods as clustering in the same way as one evaluates supervised methods where the concept to be learned is known beforehand. The already annotated classes are not even interesting in terms of finding new, previously

unknown knowledge. And this is, after all, the whole point in performing unsupervised methods in data mining [22].

We conjecture that it is 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.

It could actually be a good and desirable result 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 judgment on new clustering results, however, requires difficult and time-consuming validation based on external domain-knowledge beyond the existing class-labels. Here, we are interested in the discussion of requirements for an evaluation tool allowing for enhancement of annotated concepts and hence allowing for an evaluation adapted to new insights on well-studied data sets. Evaluation of clustering algorithms should then assess how well a clustering algorithm can rediscover clustering (not class!) structures that are already known or, if they are unknown, comprehensible and validated by insight. Such new structures should then be annotated in benchmark data sets in order to include the new knowledge in future evaluations of new algorithms. Our vision is hence to provide a repository for clustering benchmark data sets that are studied and annotated — and that are continued being studied and annotated — in order to facilitate enhanced possibilities of evaluation truly considering the observations reported but not yet fully taken into account in different research areas: classes and clusters are not the same.

Acknowledgment

This work has been supported in part by the UMIC Research Centre, RWTH Aachen University, Germany.

6. REFERENCES

[1] E. Achtert, C. Böhm, P. Kröger, and A. Zimek. Mining hierarchies of correlation clusters. In *Proc. SSDBM*, 2006.

[2] E. Achtert, C. Böhm, H.-P. Kriegel, P. Kröger, I. Müller-Gorman, and A. Zimek. Detection and visualization of subspace cluster hierarchies. In *Proc. DASFAA*, 2007.

[3] E. Achtert, C. Böhm, H.-P. Kriegel, P. Kröger, and A. Zimek. Deriving quantitative models for correlation clusters. In *Proc. KDD*, 2006.

[4] E. Achtert, C. Böhm, H.-P. Kriegel, P. Kröger, and A. Zimek. On exploring complex relationships of correlation clusters. In *Proc. SSDBM*, 2007.

[5] R. Agrawal, J. Gehrke, D. Gunopulos, and P. Raghavan. Automatic subspace clustering of high dimensional data for data mining applications. In *Proc. SIGMOD*, 1998.

[6] F. Al-Shahrour, R. Diaz-Uriarte, and J. Dopazo. FatiGO: a web tool for finding significant associations of Gene Ontology terms with groups of genes. *Bioinformatics*, 20(4):578–580, 2004.

[7] M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander. OPTICS: Ordering points to identify the clustering structure. In *Proc. SIGMOD*, 1999.

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

[9] I. Assent, R. Krieger, E. Müller, and T. Seidl. DUSC: dimensionality unbiased subspace clustering. In *Proc. ICDM*, 2007.

[10] I. Assent, R. Krieger, E. Müller, and T. Seidl. INSCY: indexing subspace clusters with in-process-removal of redundancy. In *Proc. ICDM*, 2008.

[11] Z. Barutcuoglu, R. E. Schapire, and O. G. Troyanskaya. Hierarchical multi-label prediction of gene function. *Bioinformatics*, 22(7):830–836, 2006.

[12] M. R. Boutell, J. Luo, X. Shen, and C. M. Brown. Learning multi-label scene classification. *Pattern Recognition*, 37(9):1757–1771, 2004.

[13] L. Cai and T. Hofmann. Hierarchical document categorization with support vector machines. In *Proc. CIKM*, 2004.

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

[15] S. V. Chakravarthy and J. Ghosh. Scale-based clustering using the radial basis function network. *IEEE TNN*, 7(5):1250–1261, 1996.

[16] C. H. Cheng, A. W.-C. Fu, and Y. Zhang. Entropy-based subspace clustering for mining numerical data. In *Proc. KDD*, pages 84–93, 1999.

[17] A. Clare and R. King. Knowledge discovery in multi-label phenotype data. In *Proc. PKDD*, 2001.

[18] A. Clare and R. King. How well do we understand the clusters found in microarray data? *In Silico Biol.*, 2(4):511–522, 2002.

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

[20] S. Datta and S. Datta. Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes. *BMC Bioinformatics*, 7(397), 2006.

[21] I. Davidson and Z. Qi. Finding alternative clusterings using constraints. In *Proc. ICDM*, 2008.

[22] U. Fayyad, G. Piatetsky-Shapiro, and P. Smyth. Knowledge discovery and data mining: Towards a unifying framework. In *Proc. KDD*, 1996.

[23] E. B. Fowlkes and C. L. Mallows. A method for comparing two hierarchical clusterings. *JASA*, 78(383):553–569, 1983.

[24] A. Frank and A. Asuncion. UCI machine learning repository, 2010. `http://archive.ics.uci.edu/ml`.

[25] I. Gat-Viks, R. Sharan, and R. Shamir. Scoring clustering solutions by their biological relevance. *Bioinformatics*, 19(18):2381–2389, 2003.

[26] J. M. Geusebroek, G. J. Burghouts, and A. Smeulders. The Amsterdam Library of Object Images. *Int. J. Computer Vision*, 61(1):103–112, 2005.

[27] F. D. Gibbons and F. P. Roth. Judging the quality of gene expression-based clustering methods using gene annotation. *Genome Res.*, 12:1574–1581, 2002.

[28] S. Günnemann, E. Müller, I. Färber, and T. Seidl. Detection of orthogonal concepts in subspaces of high dimensional data. In *Proc. CIKM*, 2009.

[29] S. Godbole and S. Sarawagi. Discriminative methods for multi-labeled classification. In *Proc. PAKDD*, 2004.

[30] D. Gondek and T. Hofmann. Non-redundant clustering with conditional ensembles. In *Proc. KDD*, 2005.

[31] M. Halkidi, Y. Batistakis, and M. Vazirgiannis. On clustering validation techniques. *JIIS*, 17(2-3):107–145, 2001.

[32] J. A. Hartigan. Direct clustering of a data matrix. *JASA*, 67(337):123–129, 1972.

[33] J. A. Hartigan. *Clustering Algorithms*. John Wiley&Sons, 1975.

[34] M. E. Houle, H.-P. Kriegel, P. Kröger, E. Schubert, and A. Zimek. Can shared-neighbor distances defeat the curse of dimensionality? In *Proc. SSDBM*, 2010.

[35] L. Hubert and P. Arabie. Comparing partitions. *J. Classif.*, 2(1):193–218, 1985.

[36] P. Jain, R. Meka, and I. S. Dhillon. Simultaneous unsupervised learning of disparate clusterings. *Stat. Anal. Data Min.*, 1(3):195–210, 2008.

[37] K. Kailing, H.-P. Kriegel, and P. Kröger. Density-connected subspace clustering for high-dimensional data. In *Proc. SDM*, 2004.

[38] D. Koller and M. Sahami. Hierarchically classifying documents using very few words. In *Proc. ICML*, 1997.

[39] H.-P. Kriegel, P. Kröger, and A. Zimek. Clustering high dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM TKDD*, 3(1):1–58, 2009.

[40] S. G. Lee, J. U. Hur, and Y. S. Kim. A graph-theoretic modeling on go space for biological interpretation of gene clusters. *Bioinformatics*, 20(3):381–388, 2004.

[41] G. Liu, K. Sim, J. Li, and L. Wong. Efficient mining of distance-based subspace clusters. *Stat. Anal. Data Min.*, 2(5-6):427–444, 2009.

[42] P. W. Lord, R. D. Stevens, A. Brass, and C. A. Goble. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. *Bioinformatics*, 19(10):1275–1283, 2003.

[43] S. C. Madeira and A. L. Oliveira. Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM TCBB*, 1(1):24–45, 2004.

[44] A. McCallum, R. Rosenfeld, T. M. Mitchell, and A. Y. Ng. Improving text classification by shrinkage in a hierarchy of classes. In *Proc. ICML*, 1998.

[45] E. Müller, I. Assent, S. Günnemann, R. Krieger, and T. Seidl. Relevant subspace clustering: Mining the most interesting non-redundant concepts in high dimensional data. In *Proc. ICDM*, 2009.

[46] E. Müller, S. Günnemann, I. Assent, and T. Seidl. Evaluating clustering in subspace projections of high dimensional data. *PVLDB*, 2(1):1270–1281, 2009.

[47] G. Moise and J. Sander. Finding non-redundant, statistically significant regions in high dimensional data: a novel approach to projected and subspace clustering. In *Proc. KDD*, 2008.

[48] G. Moise, A. Zimek, P. Kröger, H.-P. Kriegel, and J. Sander. Subspace and projected clustering: Experimental evaluation and analysis. *KAIS*, 21(3):299–326, 2009.

[49] H. Nagesh, S. Goil, and A. Choudhary. Adaptive grids for clustering massive data sets. In *Proc. SDM*, 2001.

[50] D. Pfitzner, R. Leibbrandt, and D. Powers. Characterization and evaluation of similarity measures for pairs of clusterings. *KAIS*, 19(3):361–394, 2009.

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

[52] Z. J. Qi and I. Davidson. A principled and flexible framework for finding alternative clusterings. In *Proc. KDD*, 2009.

[53] R. E. Schapire and Y. Singer. BoosTexter: a boosting-based system for text categorization. *Mach. Learn.*, 39(2-3):135–168, 2000.

[54] R. Sibson. SLINK: An optimally efficient algorithm for the single-link cluster method. *The Computer Journal*, 16(1):30–34, 1973.

[55] P. H. A. Sneath. The application of computers to taxonomy. *J. gen. Microbiol.*, 17:201–226, 1957.

[56] R. Srikant and R. Agrawal. Mining quantitative association rules in large relational tables. In *Proc. SIGMOD*, 1996.

[57] A. Strehl and J. Ghosh. Cluster ensembles – a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.*, 3:583–617, 2002.

[58] W. Stuetzle. Estimating the cluster tree of a density by analyzing the minimal spanning tree of a sample. *J. Classif.*, 20(1):25–47, 2003.

[59] F. A. Thabtah, P. Cowling, and Y. Peng. MMAC: a new multi-class, multi-label associative classification approach. In *Proc. ICDM*, 2004.

[60] G. Tsoumakas and I. Katakis. Multi-label classification: An overview. *IJDWM*, 3(3):1–13, 2007.

[61] H. Wang, F. Azuaje, O. Bodenreider, and J. Dopazo. Gene expression correlation and gene ontology-based similarity: An assessment of quantitative relationships. In *Proc. CIBCB*, 2004.

[62] D. Wishart. Mode analysis: A generalization of nearest neighbor which reduces chaining effects. In *Numerical Taxonomy*, 1969.

[63] I. H. Witten and E. Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2nd edition, 2005.

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