Subspace outlier mining in large multimedia databases

Ira Assent, Ralph Krieger, Emmanuel Müller, Thomas Seidl

Data management and data exploration group
RWTH Aachen University, Germany
{assent,krieger,mueller,seidl}@cs.rwth-aachen.de

Abstract. Increasingly large multimedia databases in life sciences, e-commerce, or monitoring applications cannot be browsed manually, but require automatic knowledge discovery in databases (KDD) techniques to detect novel and interesting patterns. Clustering, aims at grouping similar objects into clusters, separating dissimilar objects. Density-based clustering has been shown to detect arbitrarily shaped clusters even in noisy data bases. In high-dimensional data bases, meaningful clusters can no longer be detected due to the curse of dimensionality. Consequently, subspace clustering searches for clusters hidden in any subset of the set of dimensions. Clustering information is very useful for applications like fraud detection where outliers, i.e. objects which differ from all clusters, are searched. We propose a density-based subspace clustering model for outlier detection. We define outliers with respect to maximal and non-redundant subspace clusters. We demonstrate the quality of our subspace clustering results in experiments on real world databases and discuss our outlier model as well as future work.

1 Introduction

Many multimedia applications archive huge amounts of data. In life sciences or medicine, e-commerce or sensor networks much information is generated automatically. As data base sizes grow, manual analysis is no longer possible.
Knowledge discovery in databases (KDD) aims at detecting novel and interesting patterns which are useful for the user in that they allow building of knowledge.

1.1 Clustering

Clustering is one of the major KDD tasks. Its goal is grouping of data base objects such that inter-group similarity is minimized, whereas intra-group similarity is maximized. The resulting clustering is a compact representation of the inherent data structure. As class labels are not known apriori, clustering is considered an unsupervised learning approach. Examples of major clustering approaches include partitioning methods which divide the dataset into disjoint groups of objects. Iteratively, initial clusterings are improved, as e.g. in k-means

Dagstuhl Seminar Proceedings 07181
Parallel Universes and Local Patterns
http://drops.dagstuhl.de/opus/volltexte/2007/1257

[1]. EM is a similar approach based on multivariate gaussian mixture models [2]. These approaches require the user to provide a parameter k, the number of clusters, apriori or rely on heuristics to determine the correct number of clusters. Moreover, they assume convex cluster shapes and are sensitive to noise. Hierarchical methods such as BIRCH [3] or CURE [4] construct a decomposition of the data into clusters either top-down or bottom-up. Grid-based technologies, e.g. STING [5], or WaveCluster [6], use multi-resolution data space cells. Cells represent a discretization of the data space which allows for fast detection of clusters, but clusters cut apart by the grid are lost. For the special case of categorical data, **categorical clustering** methods like k-modes [7], CACTUS [8], ROCK [9], COOLCAT [10], and CLICKS [11] have been developed. Density-based algorithms define clusters as dense areas in feature space, separated by sparsely populated ones. Objects are dense if their neighborhood contains at least a minimum number of objects. A connectivity-notion which reflect the transitive closure of dense neighborhoods assigns similar objects to the same cluster (DBSCAN [12], DENCLUE [13]). Density-based clustering has been shown to successfully detect clusters of arbitray shape even in noisy data bases. Unfortunately, all of these clustering methods suffer from the "curse of dimensionality", i.e. with increasing dimensionality, object distances grow more and more similar, making it eventually impossible to find meaningful clusters [14].

1.2 Subspace clustering

Subspace clustering aims at finding clusters in any subspace of high-dimensional feature spaces. As opposed to projective clustering, as in ORCLUS [15] or Monte Carlo projective clustering [16], overlapping clusters in different subspaces are detected. Grid-based subspace clustering such as CLIQUE discretize the search space [17]. Monotonicity on the density of cells is used to prune the search space in a bottom-up algorithm. Grids greatly reduce the computational complexity, yet clusters which spread across several cells might be missed as mentioned above. Density-based subspace clustering as in SUBCLU extends the DBSCAN approach to subspace clustering [18]. The algorithm uses an apriori like scheme (discussed first in association rule mining [19]) to detect subspace cluster in a bottom-up fashion. Recent approaches like SCHISM adopt a more complex density definiton for subspace clusters. As complexity of computation grows, heuristics and a grid-based discretization for pruning are used. Subspace search algorithms like RIS search for subspaces which might contain subspace clusters and are thus considered "interesting" [20]. The actual clusters are then mined using any traditional clustering algorithm. As no concrete subspace clusters are mined, the interestingness value of subspaces does not always reflect the actual number of clusters contained. As the clustering step is not included, overall runtimes are infeasible for high-dimensional data sets.

1.3 Outlier detection

Outlier mining is used for fraud detection in a variety of applications such as credit card fraud detection, data consistency checks, abnormal reactions in pharmaceutical studies, etc. [21]. Statistical outlier mining measures the deviation from an assumed distribution model using discordancy tests. A number of input parameters has to be specified by the user such as the number of outliers [22]. Deviation-based outlier mining assumes implicit redundancy in the data. By computing series of subsets, deviations from the structure of the previous subset are detected [23]. However, the order of subsets may influence which outliers are actually detected. Distance based outlier mining labels those objects as outliers whose neighborhood does not contain enough objects or where nearest-neighbor distances are large [24, 25]. In these approaches, choosing the size of the neighborhood and the number of objects required is often difficult.

Data mining output from algorithms in rule mining or clustering is also used to detect outliers as those objects which do not fit in with the predominant patterns. Algorithms like k-means have been adapted to this end, and new approaches like compact micro-clusters or local outlier factors (ranking approach) have been proposed [26–30].

2 Subspace outlier mining

2.1 Density-based subspace clusters

Our subspace clustering model extends density-based approaches. In density-based clustering, clusters are sets of density-connected objects. As illustrated in Figure 1, an ε-neighborhood around each object is defined. If the number of objects within this neighborhood exceeds a certain threshold *minpts*, a new cluster is started and iteratively the objects within successive neighborhoods are picked up. This is formalized in the notions of *density* and $density - connectedness$.

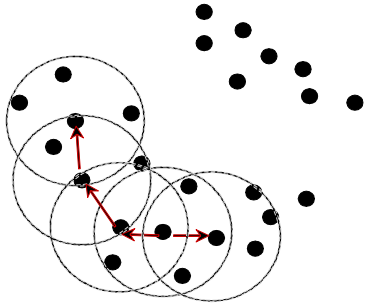

Fig. 1. Density-connected clustering

Objects are defined as dense, if the number of objects in the neighborhood, that is within a distance of ε exceeds a threshold *minpts*:

Definition 1. ***Density***

An object o in a subspace S is dense with respect to a neighborhood range parameter ε and a minimum points parameter minpts:

$$dense(o^S) \Leftrightarrow |N_\varepsilon(o^S)| \geq minpts \text{ with } N_\varepsilon(o^S) = \{p \in DB | dist(o^S, p^S) \leq \varepsilon\}$$

Dense objects are connected to one another via transitive inclusion in their ε- neighborhoods.

Definition 2. ***Density-connectivity.***

Two objects o_1, o_2 are density-connected if there is a chain of objects

$$q_1 \ldots q_n \in DB : q_1 = o_1, q_n = o_2 \text{ with } \forall i \in \{1, \ldots, n-1\}\ q_i \in N_\varepsilon(q_{i+1}),\ dense(q_i)$$

In subspace clustering, these definitions have to be restricted to the respective dimensions of the subspace. Density and density-connectedness are defined for o^S instead of o, where o^S denotes the projection of o to the subspace S. Figure 2 denotes two different subspace projections of the same data objects.

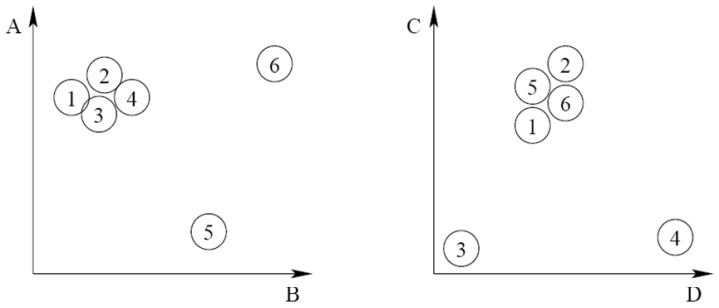

Fig. 2. Subspace projections of an example data set

Using these definitions, we define subspace clusters as density-connected maximal sets in any subspace.

Definition 3. ***Subspace Cluster.***

A set of data base objects $C \subseteq DB$ in subspace $S \subseteq d$ with $|S| > 1$ and $|C| > minSize$ is a subspace cluster if

- *C* ***density-connected****: $\forall o_1, o_2 \in C$: o_1^S, o_2^S density-connected.*
- *C* ***maximal****: $\forall o_1, o_2 \in DB$: $o_1 \in C \ \wedge\ o_1^S, o_2^S$ density-connected $\Leftrightarrow o_2 \in C$.*
- *C* ***non-redundant****: there is no higher-dimensional cluster containing points in C.*

For reasonable output sizes, subspace clusters are restricted to sets of a certain minimum size *minsize* which can be set by the user, and redundant repetitions of subspace clusters in lower dimensional projections are removed.

2.2 Outlier mining

In high-dimensional spaces, meaningful separation between outliers and clusters is typically not possible [14]. We therefore propose to study subspace outlier mining based on density-based subspace clustering approaches. In this sense, an outlier is an object which cannot be explained by existing subspace cluster patterns.

As opposed to full space clustering, in subspace clustering a set of local patterns is mined. An outlier is an object for which there is **no** density-connected subspace cluster of *minsize* which does not overfit in the sense that neither too few nor too many dimensions are covered. If one or two dimensions are covered by subspace clusters, hardly any subspace outliers will be detected as any object is bound to be similar to some cluster in one or two attributes. Likewise, very high dimensional subspace clusters may indicate overfitting.

Definition 4. ***Subspace Outlier.***

An object o is an outlier with respect to a subspace clustering as in Definition 3, if it is ***not*** *density-connected to a relevant, non-overfitting (according to parameters min and max) subspace cluster:*

- o, C ***density-connected****:* $\exists\ o_1 \in C$*:* o_1^S, o^S *density-connected.*
- C ***relevant****:* $|\{\hat{o} \mid \hat{o} \in C\}| \geq minsize$*.*
- C ***not overfitting****:* $min \leq |S| \leq max$*.*

2.3 Parallel universes in subspace outlier mining

In the presence of heterogeneous attributes sometimes no meaningful distance function can be found. For example, in an application from the financial domain, where the focus was on detecting potential money laundering, the challenge is to detect deviating behavior in transaction data. Deviations have to be compared to local patterns. Obviously, financial transactions in students and businesspeople is very different. Consequently, meaningful outlier detection should group customers according to the information available before searching for deviations.

The information on customers available contains two very heterogeneous types of attributes. One, for each customer, address information, customer segment, etc. are recorded. And second, the actual transactions are stored. There is no meaningful distance function which could model deviations in terms of attributes like profession and deviations in terms of money transfers simultaneously. These types of information constitute two very distinct models of the same customer. We consider them two universes in which outliers may be detected. These universes are not independent; transaction data can only be analyzed for outliers once the meta data has been group. For example, once a subspace cluster of bakers in San Francisco has been identified, conspicuous transactions within in this group may be identified. We thus propose a two-step model:

Definition 5. ***Parallel Universe Subspace Outlier.***

An object o is an outlier with respect to a meta data universe and a specialized data universe if:

– *$o \in C$ where C is a subspace cluster in the meta data universe.*
– *o is a subspace outlier according to Definition 4 in the specialized data universe restricted to the objects in C.*

Thus, a parallel universe subspace outlier is an object which deviates from the behavior found in its local pattern.

3 Experiments

We have evaluated our subspace clustering model on several real world data sets, measuring both quality and coverage. Corresponding roughly to the notions of precision and recall in classification, these notions describe the purity of clusters with respect to a class label and the percentage of objects assigned to some cluster. Coverage thus indicates the number of objects which could be termed outliers according to the above model. This requires further experimental evaluation. Quality is determined using the entropy, i.e. $H(C) = -\sum_{i=1}^{k} p_i(C) \cdot \log(p_i(C))$

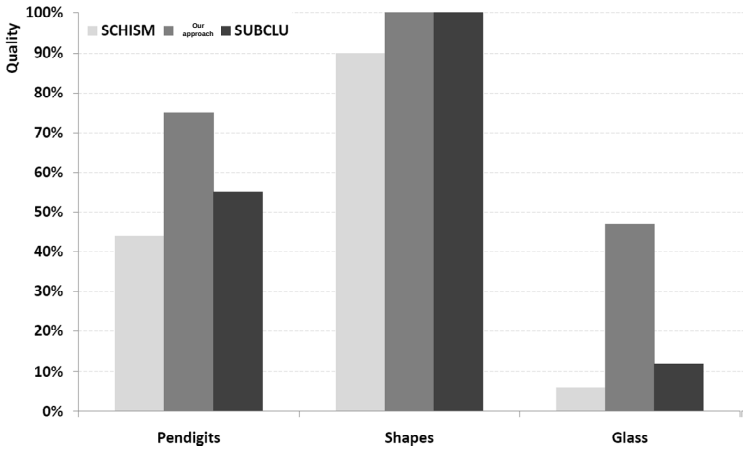

Fig. 3. Quality on real world data sets

for k class labels in cluster C. Entropy is an information theoretic indicator for the homogeneity of the data. For readability, we take the inverse entropy and normalize it to a range of 0% to 100% by dividing by the maximum entropy $log(k)$. More precisely,

$$Quality(C) = 1 - \frac{\text{entropy}(C_1, \ldots, C_n)}{\log(k)}$$

Coverage is the percentage of objects in any subspace cluster. Coverage was found to be around 80% to 90% in these experiments. As we can see in Figure 3, the quality of our approach is superior to competing approaches. As discussed

in the related work section, SUBCLU is an extension of DBSCAN to subspace clustering, whereas SCHISM is a grid-based approach [18, 31]. This experiment demonstrates that our approach indeed detects pure clusters. These preliminary experiments indicate that our subspace clustering algorithm is capable of detecting pure subspace clusters. Quite interestingly, the coverage ratios indicate that outliers exist. As about 10% to 20% of the data is not assigned to clusters, ranking of outliers seems a crucial requirement. We plan to investigate this further; especially with respect to our two-step approach for heterogeneous data.

4 Conclusion

In this work, we present a density-based subspace clustering model for outlier detecting in heterogeneous data. Density-based subspace clustering detects local patterns in arbitrary projections of the feature space. Incorporating information on heterogeneous data is helpful in a number of applications, where distinct features cannot be compared in a meaningful way. Our preliminary experiments are very promising in that our approach outperforms existing subspace clustering algorithms. Moreover, they indicate that outliers may indeed be common in the data. In future work, we plan to validate this hypothesis on these data sets and additionally on financial data. Moreover, ranking of outliers may be helpful to allow users to analyze the most urgent cases first. We also plan to incorporate expert knowledge on typical cases of money laundering.

References

1. MacQueen, J.: Some methods for classification and analysis of multivariate observations. In: Berkeley Symp. Math. stat. & prob. (1967) 281–297
2. Lauritzen, S.: The em algorithm for graphical association models with missing data. Comp. Statistics & Data Analysis **19** (1995) 191–201
3. Zhang, T., Ramakrishnan, R., Livny, M.: Birch : an efficient data clustering method for very large databases. In: SIGMOD. (1996) 103–114
4. Guha, S., Rastogi, R., Shim, K.: Cure: An efficient clustering algorithm for large databases. In: SIGMOD. (1998) 73–84
5. Wang, W., Yang, J., Muntz, R.: Sting: A statistical information grid approach to spatial data mining. In: VLDB. (1997) 517–528
6. Sheikholeslami, G., Chatterjee, S., Zhang, A.: Wavecluster: A multi-resolution clustering approach for very large spatial databases. In: VLDB. (1998) 428–439
7. Huang, Z.: A fast clustering algorithm to cluster very large categorical data sets in data mining. In: DKDM. (1997) 1–8
8. Ganti, V., Gehrke, J., Ramakrishnan, R.: Cactus: Clustering categorical data using summaries. In: KDD. (1999) 73–83
9. Guha, S., Rastogi, R., Shim, K.: A robust clustering algorithm for categorical attributes. In: ICDE. (1999) 512–521
10. Barbara, D., Li, Y., Couto, J.: Coolcat: an entropy-based algorithm for categorical clustering. In: CIKM. (2002) 582–589
11. Zaki, M., Peters, M., Assent, I., Seidl, T.: Clicks: An effective algorithm for mining subspace clusters in categorical datasets. DKE **60** (January 2007) 51–70

12. Ester, M. et al.: A density-based algorithm for discovering clusters in large spatial databases. In: KDD. (1996) 226–231
13. Hinneburg, A., Keim, D.: An efficient approach to clustering in large multimedia databases with noise. In: KDD. (1998) 58–65
14. Beyer, K. et al.: When is nearest neighbors meaningful. In: IDBT. (1999) 217–235
15. Aggarwal, C., Yu, P.: Finding generalized projected clusters in high dimensional spaces. In: SIGMOD. (2000) 70–81
16. Procopiuc, C. et al.: A monte carlo algorithm for fast projective clustering. In: SIGMOD. (2002) 418–427
17. Agrawal, R. et al.: Automatic subspace clustering of high dimensional data for data mining applications. In: SIGMOD. (1998) 94–105
18. Kailing, K., Kriegel, H.-P., Kroeger, P.: Density-connected subspace clustering for high-dimensional data. In: ICDM. (2004) 246–257
19. Agrawal, R., Srikant, R.: Fast algorithms for mining association rules. In: VLDB. (1994) 487–499
20. Kailing, K. et al.: Ranking interesting subspaces for clustering high dimensional data. In: PKDD. (2003) 241–252
21. Han, J., Kamber, M.: Data Mining: Concepts and Techniques. Morgan Kaufmann (2001)
22. Barnett, V., Lewis, T.: Outliers in statistical data. Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics, Chichester: Wiley, 1984, 2nd ed. (1984)
23. Arning, A., Agrawal, R., Raghavan, P.: A linear method for deviation detection in large databases. In: Knowledge Discovery and Data Mining. (1996) 164–169
24. Knorr, E.M., Ng, R.T.: Algorithms for mining distance-based outliers in large datasets. In: VLDB '98: Proceedings of the 24rd International Conference on Very Large Data Bases, San Francisco, CA, USA, Morgan Kaufmann Publishers Inc. (1998) 392–403
25. Ramaswamy, S., Rastogi, R., Shim, K.: Efficient algorithms for mining outliers from large data sets. SIGMOD Rec. **29**(2) (2000) 427–438
26. Jiang M.F.; Tseng S.S.1; Su C.M.: Two-phase clustering process for outliers detection. Pattern Recognition Letters **22**(6) (2001) 691–700
27. Jin, W., Tung, A.K.H., Han, J.: Mining top-n local outliers in large databases. In: KDD '01: Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining, New York, NY, USA, ACM Press (2001) 293–298
28. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: Lof: identifying density-based local outliers. SIGMOD Rec. **29**(2) (2000) 93–104
29. He, Z., Xu, X., Deng, S.: Discovering cluster-based local outliers. Pattern Recogn. Lett. **24**(9-10) (2003) 1641–1650
30. Chen, Z., Fu, A.W.C., Tang, J.: On complementarity of cluster and outlier detection schemes. In: 5th International Conference, DaWaK. (2003) 234–243
31. Sequeira, K., Zaki, M.: Schism: A new approach for interesting subspace mining. In: ICDM. (2004) 186–193