Mining Contrast Subspaces*

Lei Duan[1,6], Guanting Tang[2], Jian Pei[2], James Bailey[3], Guozhu Dong[4], Akiko Campbell[5], and Changjie Tang[1]

[1] School of Computer Science, Sichuan University, China
[2] School of Computing Science, Simon Fraser University, Canada
[3] Dept. of Computing and Information Systems, University of Melbourne, Australia
[4] Dept. of Computer Sci & Engr, Wright State University, USA
[5] Pacific Blue Cross, Canada
[6] State Key Laboratory of Software Engineering, Wuhan University, China
{leiduan, cjtang}@scu.edu.cn, {gta9, jpei}@cs.sfu.ca,
baileyj@unimelb.edu.au, guozhu.dong@wright.edu,
acampbell@pac.bluecross.ca

Abstract. In this paper, we tackle a novel problem of mining contrast subspaces. Given a set of multidimensional objects in two classes C_+ and C_- and a query object o, we want to find top-k subspaces S that maximize the ratio of likelihood of o in C_+ against that in C_-. We demonstrate that this problem has important applications, and at the same time, is very challenging. It even does not allow polynomial time approximation. We present CSMiner, a mining method with various pruning techniques. CSMiner is substantially faster than the baseline method. Our experimental results on real data sets verify the effectiveness and efficiency of our method.

Keywords: contrast subspace, kernel density estimation, likelihood contrast.

1 Introduction

Imagine you are a medical doctor facing a patient having symptoms of being overweight, short of breath, and some others. You want to check the patient on two specific possible diseases: coronary artery disease and adiposity. Please note that clogged arteries are among the top-5 most commonly misdiagnosed diseases. You have a set of reference samples of both diseases. Then, you may naturally ask "In what aspect is this patient most similar to cases of coronary artery disease and, at the same time, dissimilar to adiposity?"

* This work was supported in part by an NSERC Discovery grant, a BCIC NRAS Team Project, NSFC 61103042, SRFDP 20100181120029, and SKLSE2012-09-32. Work by Lei Duan and Guozhu Dong at Simon Fraser University was supported by an Ebco/Eppich visiting professorship. All opinions, findings, conclusions and recommendations in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

The above motivation scenario cannot be addressed well using existing data mining methods, and thus suggests a novel data mining problem. In a multidimensional data set of two classes, given a query object and a target class, we want to find the subspace where the query object is most likely to belong to the target class against the other class. We call such a subspace a *contrast subspace* since it contrasts the likelihood of the query object in the target class against the other class. Mining contrast subspaces is an interesting problem with many important applications. As another example, when an analyst in an insurance company is investigating a suspicious claim, she may want to compare the suspicious case against the samples of frauds and normal claims. A useful question to ask is in what aspects the suspicious case is most similar to fraudulent cases and different from normal claims. In other words, finding the contrast subspace for the suspicious claim is informative for the analyst.

While there are many existing studies on outlier detection and contrast mining, they focus on collective patterns that are shared by many cases of the target class. The contrast subspace mining problem addressed here is different. It focuses on one query object and finds the customized contrast subspace. This critical difference makes the problem formulation, the suitable applications, and the mining methods dramatically different. We will review the related work and explain the differences systematically in Section 2.

To tackle the problem of mining contrast subspaces, we need to address several technical issues. First, we need to have a simple yet informative contrast measure to quantify the similarity between the query object and the target class and the difference between the query object and the other class. In this paper, we use the ratio of the likelihood of the query object in the target class against that in the other class as the measure. This is essentially the Bayes factor on the query object, and comes with a well recognized explanation [1].

Second, the problem of mining contrast subspaces is computational challenging. We show that the problem is MAX SNP-hard, and thus does not allow polynomial time approximation methods unless P=NP. Therefore, the only hope is to develop heuristics that may work well in practice.

Third, one could use a brute-force method to tackle the contrast mining problem, which enumerates every non-empty subspace and computes the contrast measure. This method, however, is very costly on data sets with a non-trivial dimensionality. One major obstacle preventing effective pruning is that the contrast measure does not have any monotonicity with respect to the subspace-superspace relationship. To tackle the problem, we develop pruning techniques based on bounds of likelihood and contrast ratio. Our experimental results on real data sets clearly verify the effectiveness and efficiency of our method.

The rest of the paper is organized as follows. We review the related work in Section 2. In Section 3, we formalize the problem, and analyze it theoretically. We present a heuristic method in Section 4, and evaluate our method empirically using real data sets in Section 5. We conclude the paper in Section 6.

2 Related Work

Our study is related to the existing work on contrast mining, subspace outlier detection and typicality queries. We review the related work briefly here.

Contrast mining discovers patterns and models that manifest drastic differences between datasets. Dong and Bailey [2] presented a comprehensive review. The most renowned contrast patterns include emerging patterns [3], contrast sets [4] and subgroups [5]. Although their definitions vary, the mining methods share heavy similarity [6].

Contrast pattern mining identifies patterns by considering all objects of all classes in the complete pattern space. Orthogonally, contrast subspace mining focuses on one object, and identifies subspaces where a query object demonstrates the strongest overall similarity to one class against the other. These two mining problems are fundamentally different. To the best of our knowledge, the contrast subspace mining problem has not been systematically explored in the data mining literature.

Subspace outlier detection discovers objects that significantly deviate from the majority in some subspaces. It is very different from our study. In contrast subspace mining, the query object may or may not be an outlier. Some recent studies find subspaces that may contain substantial outliers. Böhm *et al.* [7] and Keller *et al.* [8] proposed statistical approaches *CMI* and *HiCS* to select subspaces for a multidimensional database, where there may exist outliers with high deviations. Both *CMI* and *HiCS* are fundamentally different from our method. Technically, they choose subspaces for all outliers in a given database, while our method chooses the most contrasting subspaces for a query object.

Our method uses probability density to estimate the likelihood of a query object belonging to different classes. There are a few density-based outlier detection methods, such as [9–12]. Our method is inherently different from those, since we do not target at outlier objects at all.

Hua *et al.* [13] introduced a novel top-k *typicality query*, which ranks objects according to their typicality in a data set or a class of objects. Although both [13] and our work use density estimation methods to calculate the typicality/likelihood of a query object with respect to a set of data objects, typicality queries [13] do not consider subspaces at all.

3 Problem Formulation and Analysis

In this section, we first formulate the problem. Then, we recall the basics of kernel density estimation, which can estimate the probability density of objects. Last, we investigate the complexity of the problem.

3.1 Problem Definition

Let $D = \{D_1, \ldots, D_d\}$ be a d-dimensional space, where the domain of D_i is $\mathbb{R}$, the set of real numbers. A *subspace* $S \subseteq D$ ($S \neq \emptyset$) is a subset of D. We also call D the *full space*.

Consider an object o in space D. We denote by $o.D_i$ the value of o in dimension D_i ($1 \leq i \leq d$). For a subspace $S = \{D_{i_1}, \ldots, D_{i_l}\} \subseteq D$, the *projection* of o in S is $o^S = (o.D_{i_1}, \ldots, o.D_{i_l})$. For a set of objects $O = \{o_j \mid 1 \leq j \leq n\}$, the *projection* of O in S is $O^S = \{o_j^S \mid o_j \in O, 1 \leq j \leq n\}$.

Given a set of objects O, we assume a latent distribution $\mathcal{Z}$ that generates the objects in O. For a query object q, denote by $L_D(q \mid \mathcal{Z})$ the likelihood of q being generated by $\mathcal{Z}$ in full space D. The posterior probability of q given O, denoted by $L_D(q \mid O)$, can be estimated by $L_D(q \mid \mathcal{Z})$. For a non-empty subspace S ($S \subseteq D$, $S \neq \emptyset$), denote by $\mathcal{Z}^S$ the projection of $\mathcal{Z}$ in S. The *subspace likelihood* of object q with respect to $\mathcal{Z}$ in S, denoted by $L_S(q \mid \mathcal{Z})$, can be estimated by the posterior probability of object q given O in S, denoted by $L_S(q \mid O)$.

In this paper, we assume that the objects in O belong to two classes, C_+ and C_-, exclusively. Thus, $O = O_+ \cup O_-$ and $O_+ \cap O_- = \emptyset$, where O_+ and O_- are the subsets of objects belonging to C_+ and C_-, respectively. Given a query object q, we are interested in how likely q belongs to C_+ and does not belong to C_-. To measure these two factors comprehensively, we define the *likelihood contrast* as $LC(q) = \frac{L(q|O_+)}{L(q|O_-)}$.

Likelihood contrast is essentially the Bayes factor[7] of object q as the observation. In other words, we can regard O_+ and O_- as representing two models, and we need to choose one of them based on query object q. Consequently, the ratio of likelihoods indicates the plausibility of model represented by O_+ against that by O_-. Jeffreys [1] gave a scale for interpretation of Bayes factor. When $LC(q)$ is in the ranges of < 1, 1 to 3, 3 to 10, 10 to 30, 30 to 100, and over 100, respectively, the strength of the evidence is negative, barely worth mentioning, substantial, strong, very strong, and decisive.

We can extend likelihood contrast to subspaces. For a non-empty subspace $S \subseteq D$, we define the likelihood contrast in the subspace as $LC_S(q) = \frac{L_S(q|O_+)}{L_S(q|O_-)}$. To avoid triviality in subspaces where $L_S(q \mid O_+)$ is very small, we introduce a minimum likelihood threshold $\delta > 0$, and consider only the subspaces S where $L_S(q \mid O_+) \geq \delta$.

Given a multidimensional data set O in full space D, a query object q, and a minimum likelihood threshold $\delta > 0$, and a parameter $k > 0$, the *problem of mining contrast subspaces* is to find the top-k subspaces S ordered by the subspace likelihood contrast $LC_S(q)$ subject to $L_S(q \mid O_+) \geq \delta$.

3.2 Kernel Density Estimation

We can use kernel density estimation [14] to estimate likelihood $L_S(q \mid O)$. In this paper, we adopt the Gaussian kernel, which is natural and widely used in density estimation. Given a set of objects O, the density of a query object q in

[7] Generally, given a set of observations Q, the plausibility of two models M_1 and M_2 can be assessed by the Bayes factor $K = \frac{Pr(Q|M_1)}{Pr(Q|M_2)}$.

subspace S, denoted by $\hat{f}_S(q, O)$, can be estimated as

$$\hat{f}_S(q, O) = \hat{f}_S(q^S, O) = \frac{1}{|O|\sqrt{2\pi}h_S} \sum_{o \in O} e^{\frac{-dist_S(q,o)^2}{2h_S^2}}$$

where $dist_S(q, o)^2 = \sum_{D_i \in S} (q.D_i - o.D_i)^2$ and h_S is a bandwidth parameter. Silverman [15] suggested that the optimal bandwidth value for smoothing normally distributed data with unit variance is $h_{S\_opt} = A(K)|O|^{-1/(|S|+4)}$, where $A(K) = \{4/(|S| + 2)\}^{1/(|S|+4)}$.

As the kernel is radially symmetric and the data is not normalized in subspaces, we can use a single scale parameter σ_S in subspace S and set $h_S = \sigma_S \cdot h_{S\_opt}$. As Silverman suggested [15], a possible choice of σ_S is the root of the average marginal variance in S.

Using kernel density estimation, we can estimate $L_S(q \mid O)$ as

$$L_S(q \mid O) = \hat{f}_S(q, O) = \frac{1}{|O|\sqrt{2\pi}h_S} \sum_{o \in O} e^{\frac{-dist_S(q,o)^2}{2h_S^2}} \tag{1}$$

Correspondingly, the likelihood contrast of object q in subspace S is given by

$$LC_S(q, O_+, O_-) = \frac{\hat{f}_S(q, O_+)}{\hat{f}_S(q, O_-)} = \frac{|O_-|h_{S_-}}{|O_+|h_{S_+}} \cdot \frac{\sum_{o \in O_+} e^{\frac{-dist_S(q,o)^2}{2h_{S_+}^2}}}{\sum_{o \in O_-} e^{\frac{-dist_S(q,o)^2}{2h_{S_-}^2}}} \tag{2}$$

We often omit O_+ and O_- and write $LC_S(q)$ if O_+ and O_- are clear from context.

3.3 Complexity Analysis

We have the following theoretical result. It can be proved by a reduction from the emerging pattern mining problem [3], which is MAX SNP-hard [16]. Limited by space, we omit the details here.

Theorem 1 (Complexity). *The problem of mining contrast subspaces is MAX SNP-hard.*

The above theoretical result indicates that the problem of mining contrast subspaces is even hard to approximate – it is impossible to design a good approximation algorithm. In the rest of the paper, we turn to practical heuristic methods.

4 Mining Methods

In this section, we first describe a baseline method that examines every possible non-empty subspace. Then, we present a bounding-pruning-refining method that expedites the search substantially.

4.1 A Baseline Method

A baseline method enumerates all possible non-empty spaces S and calculates the exact values of both $L_S(q \mid O_+)$ and $L_S(q \mid O_-)$. Then, it returns the top-k subspaces S with the largest $LC_S(q)$ values. To ensure the completeness and efficiency of subspace enumeration, the baseline method traverses the set enumeration tree [17] of subspaces in a depth-first manner.

$L_S(q \mid O_+)$ is not monotonic in subspaces. To prune subspaces using the minimum likelihood threshold δ, we develop an upper bound of $L_S(q \mid O_+)$. We sort all the dimensions in their standard deviation descending order. Let $\mathcal{S}$ be the set of children of S in the subspace set enumeration tree using the standard deviation descending order. Define $L^*_S(q \mid O_+) = \frac{1}{|O_+|\sqrt{2\pi}\sigma'_{min}h'_{opt\_min}} \sum_{o \in O_+} e^{\frac{-dist_S(q,o)^2}{2(\sigma_S h'_{opt\_max})^2}}$, where $\sigma'_{min} = min\{\sigma_{S'} \mid S' \in \mathcal{S}\}$, $h'_{opt\_min} = min\{h_{S'\_opt} \mid S' \in \mathcal{S}\}$, and $h'_{opt\_max} = max\{h_{S'\_opt} \mid S' \in \mathcal{S}\}$. We have the following result.

Theorem 2 (Monotonic density bound). *For a query object q, a set of objects O, and subspaces S_1, S_2 such that S_1 is an ancestor of S_2 in the subspace set enumeration tree using the standard deviation descending order in O_+, $L^*_{S_1}(q \mid O_+) \geq L_{S_2}(q \mid O_+)$.*

Using Theorem 2, in addition to $L_S(q \mid O_+)$ and $L_S(q \mid O_-)$, we also compute $L^*_S(q \mid O_+)$ for each subspace S. Once $L^*_S(q \mid O_+) < \delta$ in a subspace S, all superspaces of S can be pruned.

Using Equations 1 and 2, the baseline algorithm computes the likelihood contrast for every subspace where $L_S(q \mid O_+) \geq \delta$, and returns the top-$k$ subspaces. The time complexity is $O(2^{|D|} \cdot (|O_+| + |O_-|))$.

4.2 A Bounding-Pruning-Refining Method

For a query object q and a set of objects O, the *ϵ-neighborhood* ($\epsilon > 0$) of q in subspace S is $N^\epsilon_S(q) = \{o \in O \mid dist_S(q,o) \leq \epsilon\}$. We can divide $L_S(q \mid O)$ into two parts, that is, $L_S(q \mid O) = L_{N^\epsilon_S}(q \mid O) + L^{rest}_S(q \mid O)$. The first part is contributed by the objects in the ϵ-neighborhood, that is, $L_{N^\epsilon_S}(q \mid O) = \frac{1}{|O|\sqrt{2\pi}h_S} \sum_{o \in N^\epsilon_S(q)} e^{\frac{-dist_S(q,o)^2}{2h_S^2}}$, and the second part is by the objects outside the ϵ-neighborhood, that is, $L^{rest}_S(q \mid O) = \frac{1}{|O|\sqrt{2\pi}h_S} \sum_{o \in O \setminus N^\epsilon_S(q)} e^{\frac{-dist_S(q,o)^2}{2h_S^2}}$.

Let $\overline{dist}_S(q \mid O)$ be the maximum distance between q and all objects in O in subspace S. We have,

$$\frac{|O| - |N^\epsilon_S(q)|}{|O|\sqrt{2\pi}h_S} \cdot e^{-\frac{\overline{dist}_S(q,O)^2}{2h_S^2}} \leq L^{rest}_S(q \mid O) \leq \frac{|O| - |N^\epsilon_S(q)|}{|O|\sqrt{2\pi}h_S} \cdot e^{-\frac{\epsilon^2}{2h_S^2}}$$

Using the above, we have the following upper and lower bounds of $L_S(q \mid O)$ using ϵ-neighborhood.

Theorem 3 (Bounds). *For a query object q, a set of objects O and $\epsilon \geq 0$,*

$$LL_S^\epsilon(q \mid O) \leq L_S(q \mid O) \leq UL_S^\epsilon(q \mid O)$$

where

$$LL_S^\epsilon(q \mid O) = \frac{1}{|O|\sqrt{2\pi}h_S}\left(\sum_{o \in N_S^\epsilon(q)} e^{\frac{-dist_S^\epsilon(q,o)^2}{2h_S^2}} + (|O| - |N_S^\epsilon(q)|)e^{-\frac{\overline{dist}_S(q,O)^2}{2h_S^2}}\right)$$

and

$$UL_S^\epsilon(q \mid O) = \frac{1}{|O|\sqrt{2\pi}h_S}\left(\sum_{o \in N_S^\epsilon(q)} e^{\frac{-dist_S^\epsilon(q,o)^2}{2h_S^2}} + (|O| - |N_S^\epsilon(q)|)e^{-\frac{\epsilon^2}{2h_S^2}}\right)$$

We obtain an upper bound of $LC_S(q)$ based on Theorem 3 and Equation 2.

Corollary 1 (Likelihood Contrast Upper Bound). *For a query object q, a set of objects O_+, a set of objects O_-, and $\epsilon \geq 0$, $LC_S(q) \leq \frac{UL_S^\epsilon(q|O_+)}{LL_S^\epsilon(q|O_-)}$.*

Using Corollary 1, for a subspace S, if there are at least k subspaces whose likelihood contrast are greater than $\frac{UL_\epsilon^S(q|O_+)}{LL_\epsilon^S(q|O_-)}$, then S cannot be a top-k subspaces of the largest likelihood contrast.

Using the ϵ-neighborhood, $L_S^*(q \mid O_+)$ is computed by

$$L_S^*(q \mid O_+) = \frac{\sum_{o \in N_S^\epsilon(q)} e^{\frac{-dist_S^\epsilon(q,o)^2}{2(\sigma_S h'_{opt\_max})^2}} + (|O_+| - |N_S^\epsilon(q)|)e^{-\frac{\epsilon^2}{2(\sigma_S h'_{opt\_max})^2}}}{|O_+|\sqrt{2\pi}\sigma'_{min}h'_{opt\_min}}$$

Our bounding-pruning-refining method, CSMiner (for Contrast Subspace Miner), conducts a depth-first search on the subspace set enumeration tree. For a candidate subspace S, CSMiner calculates $UL_S^\epsilon(q \mid O_+)$ and $LL_S(q \mid O_-)$ using the ϵ-neighborhood. If $UL_S^\epsilon(q \mid O_+)$ is less than the minimum likelihood threshold, S cannot be a contrast subspace. Otherwise, CSMiner checks whether the likelihood contrasts of the current top-k subspaces are larger than the upper bound of $LC_S(q)$. If not, CSMiner refines $L_S(q \mid O_+)$ and $L_S(q \mid O_-)$ by involving objects that are out of the ϵ-neighborhood. S will be added into the current top-k list if its likelihood contrast is larger than one of the current top-k ones. Algorithm 1 gives the pseudo-code of CSMiner. Due to the hardness of the problem shown in Theorem 1 and the heuristic nature of this method, the time complexity of CSMiner is $O(2^{|D|} \cdot (|O_+| + |O_-|))$, the same as the exhaustive baseline method. However, as shown by our empirical study, CSMiner is substantially faster than the baseline method.

Computing ϵ-neighborhood is critical in CSMiner. The distance between objects increases when dimensionality increases. Thus, the value of ϵ should not be

Algorithm 1 $CSMiner(q, O_+, O_-, \delta, k)$

Input: q: a query object, O_+: the set of objects belonging to C_+, O_-: the set of objects belonging to C_-, δ: a likelihood threshold, k: positive integer
Output: k subspaces with the highest likelihood contrast

let Ans be the current top-k list of subspaces, initialize Ans as k null subspaces associated with likelihood contrast 0
for each subspace S in the subspace set enumeration tree, searched in the depth-first manner **do**
  if $UL^{\epsilon}_S(q \mid O_+) \geq \delta$ and $\exists S' \in Ans$ s.t. $\frac{UL^{\epsilon}_S(q|O_+)}{LL^{\epsilon}_S(q|O_-)} > LC_{S'}(q)$ **then**
    calculate $L_S(q \mid O_+)$, $L_S(q \mid O_-)$ and $LC_S(q)$; // refining
    if $L_S(q \mid O_+) \geq \delta$ and $\exists S' \in Ans$ s.t. $LC_S(q) > LC_{S'}(q)$ **then**
      insert S into the top-k list
    end if
  end if
  if $L^*_S(q \mid O_+) < \delta$ **then**
    prune all super-spaces of S;
  end if
end for
return Ans;

Table 1. Data set characteristics

Data set	# objects	# attributes
Breast Cancer Wisconsin (BCW)	683	9
Climate Model Simulation Crashes (CMSC)	540	18
Glass Identification (Glass)	214	9
Pima Indians Diabetes (PID)	768	8
Waveform	5000	21
Wine	178	13

fixed. The standard deviation expresses the variability of a set of data. For subspace S, we set $\epsilon = \sqrt{r \cdot \sum_{D_i \in S} (\sigma^2_{D_i^+} + \sigma^2_{D_i^-})}$ $(r \geq 0)$, where $\sigma^2_{D_i^+}$ and $\sigma^2_{D_i^-}$ are the marginal variances of O_+ and O_-, respectively, on dimension D_i $(D_i \in S)$, and r is a system defined parameter. Our experiments show that r can be set in the range of $0.3 \sim 0.6$, and is not sensitive.

5 Empirical Evaluation

In this section, we report a systematic empirical study using real data sets to verify the effectiveness and efficiency of our method. All experiments were conducted on a PC computer with an Intel Core i7-3770 3.40 GHz CPU, and 8 GB main memory, running Windows 7 operating system. All algorithms were implemented in Java and compiled by JDK 7.

5.1 Effectiveness

We use 6 real data sets from the UCI machine learning repository [18]. We remove non-numerical attributes and all instances containing missing values. Table 1 shows the data characteristics.

For each data set, we take each record as a query object q, and all records except q belonging to the same class as q forming the set O_1, and records belonging to the other classes forming the set O_2. Using CSMiner, we compute for each record (1) the *inlying contrast subspace* taking O_1 as O_+ and O_2 as O_-, and (2) the *outlying contrast subspace* taking O_2 as O_+ and O_1 as O_-. In this experiment, we only compute the top-1 subspace. For clarity, we denote the likelihood contrasts of inlying contrast subspace by $LC_S^{in}(q)$ and those of outlying contrast subspace by $LC_S^{out}(q)$. The minimum likelihood threshold is set to 0.001.

Tables 2 $\sim$ 7 list the joint distributions of $LC_S^{in}(q)$ and $LC_S^{out}(q)$ in each data set. As expected, for most objects $LC_S^{in}(q)$ are larger than $LC_S^{out}(q)$. However, interestingly a good portion of objects have strong outlying contrast subspaces. For example, in CMSC, more than 50% of the objects have outlying contrast subspaces satisfying $LC_S^{out}(q) \geq 10^3$. Moreover, we can see that, except PID, a non-trivial number of objects in each data set have both strong inlying and outlying contrast subspaces (e.g., $LC_S^{in}(q) \geq 10^4$ and $LC_S^{out}(q) \geq 10^2$).

Figures 1, 2 show the distributions of dimensionality of inlying and outlying contrast subspaces, respectively. The dimensionality distribution is an interesting feature characterizing a data set. For example, in most cases the dimensionality of contrast subspaces follows a two-side bell-shape distribution. However, in BCW and PID, the outlying contrast subspaces tend to have low dimensionality.

5.2 Efficiency

To the best of our knowledge, there is no previous method tackling the exact same mining problem. Therefore, we evaluate the efficiency of only CSMiner and the baseline method. Limited by space, we report the results on the Waveform data set only, since it is the largest one with the highest dimensionality. We randomly select 100 records from Waveform as query objects, and report the average runtime. The results on the other data sets follow similar trends.

Figure 3(a) shows the runtime (in logarithmic scale) with respect to the minimum likelihood threshold δ. As δ decreases, the runtime increases exponentially. However, the heuristic pruning techniques in CSMiner expedites the search substantially in practice. Figures 3(b) and 3(c) show the scalability on data set size and dimensionality. CSMiner is substantially faster than the baseline method.

CSMiner uses a user defined parameter r to define ϵ-neighborhood. Figure 4 shows the relative runtime with respect to r. The runtime of CSMiner is not very sensitive to r in general. Experimentally, the shortest runtime of CSMiner happens when r is in $[0.3, 0.6]$. Figure 5 illustrates the relative runtime of CSMiner with respect to k, showing that CSMiner is linearly scalable with respect to k.

6 Conclusions

In this paper, we studied a novel and interesting problem of mining contrast subspaces to discover the aspects that a query object most similar to a class and dissimilar to the other class. We showed theoretically that the problem is very

Table 2. Distribution of $LC_S(q)$ in BCW

		$LC_S^{out}(q)$					
		< 1	[1,3)	[3,10)	$[10, 10^2)$	$\geq 10^2$	Total
$LC_S^{in}(q)$	$< 10^2$	0	0	0	2	21	23
	$[10^2, 10^3)$	6	7	5	8	11	37
	$[10^3, 10^4)$	176	37	18	15	18	264
	$[10^4, 10^5)$	99	7	6	4	5	121
	$\geq 10^5$	38	25	87	82	6	238
Total		319	76	116	111	61	683

Table 3. Distribution of $LC_S(q)$ in Glass

		$LC_S^{out}(q)$					
		< 1	[1,3)	[3,10)	$[10, 10^2)$	$\geq 10^2$	Total
$LC_S^{in}(q)$	< 10	0	4	0	2	4	10
	$[10, 10^2)$	11	70	26	6	4	117
	$[10^2, 10^3)$	2	24	5	3	2	36
	$[10^3, 10^4)$	0	0	4	0	1	5
	$\geq 10^4$	0	23	14	6	3	46
Total		13	121	49	17	14	214

Table 4. Distribution of $LC_S(q)$ in PID

		$LC_S^{out}(q)$					
		< 1	[1,3)	[3,10)	$[10, 10^2)$	$\geq 10^2$	Total
$LC_S^{in}(q)$	< 1	0	0	1	1	0	2
	$[1, 3)$	0	124	99	19	2	244
	$[3, 10)$	17	241	54	4	0	316
	$[10, 10^2)$	28	146	19	4	0	197
	$\geq 10^2$	1	8	0	0	0	9
Total		46	519	173	28	2	768

Table 5. Distribution of $LC_S(q)$ in Wine

		$LC_S^{out}(q)$					
		< 1	[1,3)	[3,10)	$[10, 10^2)$	$\geq 10^2$	Total
$LC_S^{in}(q)$	$< 10^3$	2	22	10	13	9	56
	$[10^3, 10^4)$	0	17	11	6	2	36
	$[10^4, 10^5)$	0	10	4	2	2	18
	$[10^5, 10^6)$	0	5	5	2	0	12
	$\geq 10^6$	4	21	15	12	4	56
Total		6	75	45	35	17	178

Table 6. Distribution of $LC_S(q)$ in CMSC

		$LC_S^{out}(q)$					
		$[10, 10^2)$	$[10^2, 10^3)$	$[10^3, 10^4)$	$[10^4, 10^5)$	$\geq 10^5$	Total
$LC_S^{in}(q)$	$< 10^3$	2	6	41	15	0	64
	$[10^3, 10^4)$	4	28	47	17	4	100
	$[10^4, 10^5)$	7	38	44	17	7	113
	$[10^5, 10^6)$	1	30	36	10	3	80
	$\geq 10^6$	4	82	75	16	6	183
Total		18	184	243	75	20	540

Table 7. Distribution of $LC_S(q)$ in Waveform

		$LC_S^{out}(q)$					
		[1, 3)	[3,10)	$[10, 10^2)$	$[10^2, 10^3)$	$\geq 10^3$	Total
$LC_S^{in}(q)$	< 10	0	8	24	10	7	49
	$[10, 10^2)$	88	462	695	222	98	1565
	$[10^2, 10^3)$	235	686	956	299	104	2280
	$[10^3, 10^4)$	151	346	383	71	23	974
	$\geq 10^4$	36	46	45	5	0	132
Total		510	1548	2103	607	232	5000

challenging, and cannot even be approximated in polynomial time. We presented a heuristic method based on upper and lower bounds of likelihood and likelihood contrast. Our experiments on real data sets show that our method expedites contrast subspace mining substantially comparing to the baseline method.

References

1. Jeffreys, H.: The Theory of Probability. 3rd edn. Oxford (1961)
2. Dong, G., Bailey, J., eds.: Contrast Data Mining: Concepts, Algorithms, and Applications. CRC Press (2013)
3. Dong, G., Li, J.: Efficient mining of emerging patterns: discovering trends and differences. In KDD (1999) 43–52
4. Bay, S.D., Pazzani, M.J.: Detecting group differences: Mining contrast sets. Data Mining and Knowledge Discovery **5**(3) (2001) 213–246
5. Wrobel, S.: An algorithm for multi-relational discovery of subgroups. In PKDD (1997) 78–87

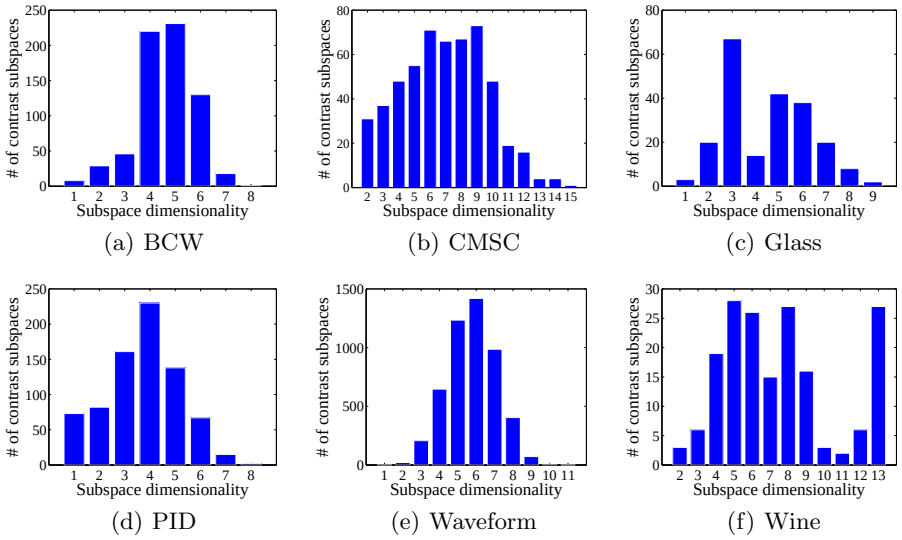


Fig. 1. Dimensionality distributions of inlying contrast subspaces

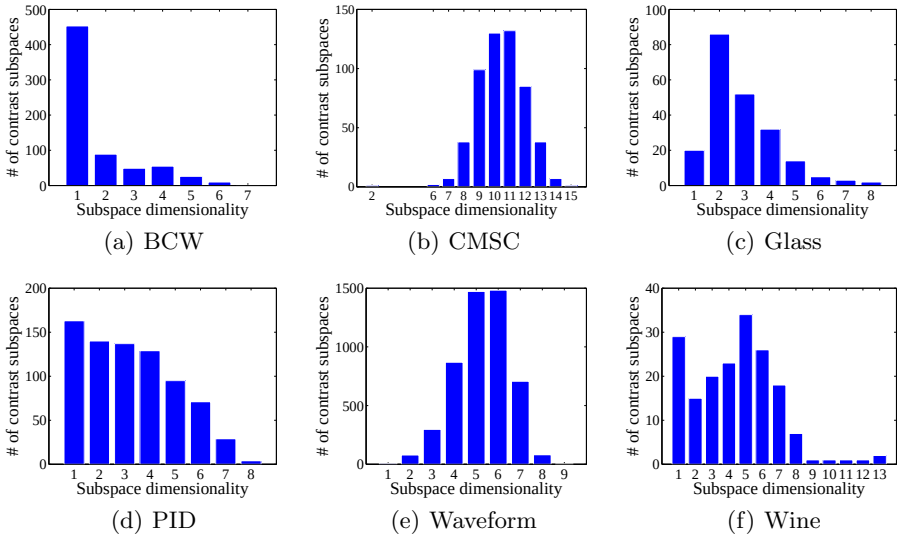
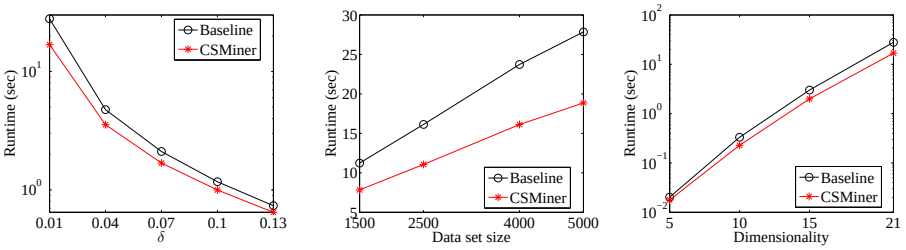


Fig. 2. Dimensionality distributions of outlying contrast subspaces

6. Novak, P.K., Lavrac, N., Webb, G.I.: Supervised descriptive rule discovery: A unifying survey of contrast set, emerging pattern and subgroup mining. Journal of Machine Learning Research **10** (2009) 377–403
7. Böhm, K., Keller, F., Müller, E., Nguyen, H.V., Vreeken, J.: CMI: An information-theoretic contrast measure for enhancing subspace cluster and outlier detection. In SDM (2013) 198–206



(a) w.r.t δ ($k = 10, r = 0.4$) (b) w.r.t data set size ($k = 10, \delta = 0.01, r = 0.4$) (c) w.r.t dimensionality ($k = 10, \delta = 0.01, r = 0.4$)

Fig. 3. Scalability test

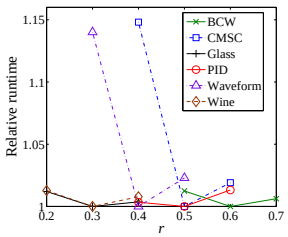


Fig. 4. Relative runtime w.r.t r ($k = 10, \delta = 0.01$)

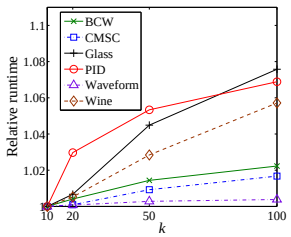


Fig. 5. Relative runtime w.r.t k ($\delta = 0.01$)

8. Keller, F., Müller, E., Böhm, K.: HiCS: High contrast subspaces for density-based outlier ranking. In ICDE (2012) 1037–1048
9. Breunig, M.M., Kriegel, H.P., Ng, R.T., Sander, J.: LOF: Identifying density-based local outliers. In SIGMOD (2000) 93–104
10. Kriegel, H.P., Schubert, M., Zimek, A.: Angle-based outlier detection in high-dimensional data. In KDD (2008) 444–452
11. He, Z., Xu, X., Huang, Z.J., Deng, S.: FP-outlier: Frequent pattern based outlier detection. Computer Science and Information Systems **2**(1) (2005) 103–118
12. Aggarwal, C.C., Yu, P.S.: Outlier detection for high dimensional data. ACM Sigmod Record. Volume 30. (2001) 37–46
13. Hua, M., Pei, J., Fu, A.W., Lin, X., Leung, H.F.: Top-k typicality queries and efficient query answering methods on large databases. The VLDB Journal **18**(3) (2009) 809–835
14. Breiman, L., Meisel, W., Purcell, E.: Variable kernel estimates of multivariate densities. Technometrics **19**(2) (1977) 135–144
15. Silverman, B.W.: Density Estimation for Statistics and Data Analysis. Chapman and Hall/CRC, London (1986)
16. Wang, L., Zhao, H., Dong, G., Li, J.: On the complexity of finding emerging patterns. Theor. Comput. Sci. **335**(1) (2005) 15–27
17. Rymon, R.: Search through systematic set enumeration. In Proc. of the 3rd Int'l Conf. on Principles of Knowledge Representation and Reasoning. (1992) 539–550
18. Bache, K., Lichman, M.: UCI machine learning repository (2013)