

Change Profiles

Taneli Mielikäinen
HIIT Basic Research Unit
Department of Computer Science
University of Helsinki, Finland
Taneli.Mielikainen@cs.Helsinki.fi

Abstract

In this paper we introduce a generalization of association rules: change profiles. We analyze their properties, describe their relationship to other structures in pattern discovery and sketch their possible applications. We study how the frequent patterns can be clustered based on their change profiles and propose methods for approximating the frequencies of the patterns from the approximate change profiles and bounding the intervals where the frequencies of the patterns are guaranteed to be. We evaluate empirically the methods for estimating the frequencies and the stability of their frequency estimates under different kinds of noise.

1 Introduction

Pattern discovery is a central task in data mining [19, 27]. Finding frequently occurring patterns from data has been one of the most actively studied problems in the field, and many kinds of frequent patterns can be found efficiently, see e.g. [1, 16, 26, 29, 38]. However, the fundamental problem in pattern discovery is how to benefit from the found patterns.

In this paper we study the global structure of the pattern collection from the local point of view of patterns. This is pursued by a new representation for pattern collections: *change profiles*. A change profile of a pattern describes what are the ratios between the frequency of the pattern and the frequencies of its sub- and superpatterns in the pattern collection, i.e., the patterns comparable with it w.r.t. a given partial order relation. To simplify the considerations, we divide the change profile ch^p of a pattern p into two parts (adapting the terminology of [28, 33]): the specializing change profile ch_s^p describes the changes to the superpatterns of p , and the generalizing change profile ch_g^p describes the changes to the subpatterns of p .

As a concrete example, let us consider the set collection $\{A, B, C, AB, BC, AC, ABC\}$ with frequencies $fr(A) =$

$3/4, fr(B) = 1/2, fr(C) = 3/4, fr(AB) = 1/4, fr(AC) = 1/2, fr(BC) = 1/2$ and $fr(ABC) = 1/4$. For these frequencies, the specializing change profiles of the (singleton) sets A, B and C are mappings $ch_s^A = \{B \mapsto fr(AB) / fr(A) = 1/3, C \mapsto fr(AC) / fr(A) = 2/3, BC \mapsto fr(ABC) / fr(A) = 1/3\}$, $ch_s^B = \{A \mapsto fr(AB) / fr(B) = 1/2, C \mapsto fr(BC) / fr(B) = 1, AC \mapsto fr(ABC) / fr(B) = 1/2\}$ and $ch_s^C = \{A \mapsto fr(AC) / fr(C) = 2/3, B \mapsto fr(BC) / fr(C) = 2/3, AB \mapsto fr(ABC) / fr(C) = 1/3\}$.

The change profiles attempt to reach from the local description of the data, a collection of patterns, to more global view of the relationships between the patterns. As an application to more classical pattern analysis, the change profiles can be used to define (dis)similarity measures between the patterns. Also, the patterns can be scored with help of their change profiles.

In addition of being a potentially useful tool in data analysis, the change profiles can be used to construct condensed representations of pattern collections [5, 6, 7, 8, 9, 25, 30, 32, 34, 35]. Many known condensed representations can be adapted to the change profiles but the change profiles suggest also some novel approaches to frequency estimation.

In this paper we introduce the concept of a change profile, a new representation for pattern collections that attempts to bridge the gap between local and global description of data. We describe several variations of change profiles and examine their properties. As the change profiles have more complex structure than the frequencies of the frequent patterns, they can be used to define dissimilarity measures between the frequent patterns. We consider different approaches to cluster the change profiles and show that many of the clustering problems are NP -hard and inapproximable for a wide variety of natural dissimilarity measures for change profiles. As an application to the condensed representations, we propose methods for approximating frequencies of the frequent patterns from their approximate change profiles and evaluate the methods empirically.

The rest of the paper is organized as follows. The concept of the change profile and its refinements are introduced in Section 2. The clustering of change profiles is studied in Section 3. In Section 4 a few approaches to estimate frequencies of a pattern collection from approximate change profiles are described. In Section 5 the frequency estimation methods are evaluated empirically. Section 6 consists of a short conclusion.

In the remaining sections we describe everything using frequent sets but the discussion generalizes readily to arbitrary pattern collections with a partial order.

2 Association Rules and Change Profiles

The *frequent set mining problem* is, given a sequence $d = d_1 \dots d_n$ of subsets of a finite set R and a threshold value $\sigma \in [0, 1]$, find all subsets X of R such that the frequency of X ,

$$fr(X, d) = \frac{|\{i : X \subseteq d_i, 1 \leq i \leq n\}|}{n},$$

is at least σ . These sets are called σ -*frequent sets* in d and they are denoted by $\mathcal{F}(\sigma, d)$. The frequency of X can be interpreted as the empirical probability $\mathbb{P}(X)$ of the event “the event contains X as its subset”.

A classical set R is the collection of items sold in a supermarket. Then the most popular sequence d of subsets of R is sequence of market baskets of the customers. In addition to the market baskets of the customers, there are several other interesting sequences of subsets of the items: for the logistics division of a company, an interesting sequence d could rather be the sequence of packages arriving from the logistics center.

The frequent set mining is computationally feasible as there are several methods for mining frequent sets (output) efficiently [1, 3, 17, 18, 21, 22, 37]. Although each frequent set contains useful information about the data set, also the relationships between different frequent sets might be valuable. These relationships has been exploited by association rules.

An *association rule* $X \Rightarrow Y$ is a rule that associates a set $X \subseteq R$ to a set $Y \subseteq R$. An association rule $X \Rightarrow Y$ is called *simple* if Y is a singleton. The accuracy of the association rule $X \Rightarrow Y$ is

$$acc(X \Rightarrow Y, d) = \frac{fr(X \cup Y, d)}{fr(X, d)}.$$

The frequency of the rule $X \Rightarrow Y$ is defined to be $fr(X \Rightarrow Y, d) = fr(X \cup Y, d)$. The accuracy of the association rule $X \Rightarrow Y$ can be interpreted as a conditional probability $\mathbb{P}(Y|X)$. Thus each association rule $X \Rightarrow Y$ describes one relationship of the set X . (Empirical conditional probabilities with different kinds of events has been studied in data mining under the name of *cubegrades* [23].)

A more global view to relationships between the frequent set X and other frequent sets can be obtained by combining the association rules $X \Rightarrow Y$ into a mapping from the frequent sets to the interval $[0, 1]$. We call this mapping a *specializing change profile*:

Definition 1 (Specializing change profile) A *specializing change profile* of a frequent set X is a mapping

$$ch_s^X : \{Y : X \cup Y \in \mathcal{F}(\sigma, d)\} \rightarrow [0, 1]$$

consisting of the accuracies of frequent rules $X \Rightarrow Y$, i.e.,

$$ch_s^X(Y) = \frac{fr(X \cup Y, d)}{fr(X, d)}$$

where $X \cup Y \in \mathcal{F}(\sigma, d)$.

A specializing change profile ch_s^X can be interpreted as a conditional probability distribution $\mathbb{P}(\mathbf{Y}|X)$ where \mathbf{Y} is a random variable.

Similarly to the specializing change profiles, we can define a *generalizing change profile*:

Definition 2 (Generalizing change profile) A *generalizing change profile* of a frequent set X is a mapping

$$ch_g^X : \mathcal{F}(\sigma, d) \rightarrow [1, 1/\sigma]$$

consisting of the inverse accuracies of the frequent rules $acc(X \setminus Y \Rightarrow X)$, i.e.,

$$ch_g^X(Y) = \frac{fr(X \setminus Y, d)}{fr(X, d)}.$$

We denote change profiles by ch^X when we do not want to fix the type of the change profile. $ch^X(Y)$ is called a *change* of X on Y . The generalizing change profile of X corresponds to $1/\mathbb{P}(X|X \setminus \mathbf{Y})$.

Each specializing and generalizing change profile ch_s^X and ch_g^X describe “upper” and “lower” neighborhoods $N_s(X)$ and $N_g(X)$ of the frequent set X in the collection of frequent sets, respectively. The neighborhood $N(X) = N_s(X) \cup N_g(X)$ of X consists of the frequent sets $Y \in N_s(X)$ that contain X and the frequent sets $Y \in N_g(X)$ that are contained in X , i.e., the frequent sets that are comparable with X .

The change profiles, as we have defined them, are highly redundant due to the following simple properties of sets:

- $X \cup Y = X \cup (Y \setminus X)$ and
- $X \setminus Y = X \setminus (Y \cap X)$.

By exploiting these properties, we can reduce (without loss of information) the number of defined values of the change profile by factor $2^{|X|}$ in the case of a specializing change profile ch_s^X , and by factor $2^{|R \setminus X|}$ in the case of a generalizing change profile ch_g^X . We call this kind of change profiles with reduced redundancy the *concise change profiles*:

Definition 3 (Concise specializing change profile) A concise specializing change profile cch_s^X is a restriction of ch_s^X to elements Y such that $Y \cap X = \emptyset$ and $X \cup Y \in \mathcal{F}(\sigma, d)$.

Definition 4 (Concise generalizing change profile) A concise generalizing change profile cch_g^X is a restriction of ch_g^X to elements Y such that $Y \subseteq X$.

The concise change profiles can be interpreted as affine axis-parallel subspaces of $\mathbb{R}^{|\mathcal{F}(\sigma, d)|}$ that are indexed by Y s.t. $Y \cap X = \emptyset, X \cup Y \in \mathcal{F}(\sigma, d)$ in the specializing case, and Y s.t. $Y \subseteq X$ in the generalizing case. The concise change profiles for $\mathcal{F}(\sigma, d)$ can be computed output-optimally by a careful implementation of the following algorithm:

```

CHANGE-PROFILES( $\mathcal{F}(\sigma, d), fr$ )
1  for each  $X$  in  $\mathcal{F}(\sigma, d)$ 
2    do for each  $Y, Y \subseteq X$ 
3      do  $cch_s^{X \setminus Y}(Y) \leftarrow fr(X) / fr(X \setminus Y)$ 
4      do  $cch_g^X(Y) \leftarrow fr(X \setminus Y) / fr(X)$ 
5  return  $(cch_s, cch_g)$ 

```

The neighborhoods of even the concise change profiles can be too large: For example, if $X \in \mathcal{F}(\sigma, d)$, then $|cch_s^\emptyset| \geq 2^{|X|}$ and $|cch_g^X| \geq 2^{|X|}$. Thus, following the definitions of association rules, we define *simple specializing change profiles* and *simple generalizing change profiles*:

Definition 5 (Simple change profile) A simple specializing (generalizing) change profile sch_s^X (sch_g^X) is a restriction of ch_s^X (of ch_g^X) to singletons Y .

The number of bits needed for a simple change profile is at most $|R| \log |d|$: Each change profile can be expressed as a vector of length $|R|$ and each value of the change profile can be describe using at most $\log |d|$ bits. Especially for the generalizing change profiles, this upper bound can be quite loose when R is large as the number of sets in $\mathcal{F}(\sigma, d)$ is $\Omega(2^{|X|})$ where X is the largest set in $\mathcal{F}(\sigma, d)$.

3 Clustering the Change Profiles

In order to be able to find groups of similar change profiles, it would be useful to be able to somehow measure the (similarity or) dissimilarity between the change profiles. The dissimilarity between change profiles ch^X and ch^Y can be defined to be their distance in the common domain $Dom(ch^X) \cap Dom(ch^Y)$ w.r.t. some distance function. (We assume the dissimilarity function δ to be such that $\delta(ch^X, ch^Y) = 0 \iff ch^X(Z) = ch^Y(Z)$ holds for all $Z \in Dom(ch^X) \cap Dom(ch^Y)$.) A complementary approach would be to focus on the differences in the structure of the pattern collection, e.g., to measure the difference

between two change profiles by computing the symmetric difference of their domains. This kind of dissimilarity function concentrates wholly on the structure of the pattern collection and thus neglects the frequencies. A sophisticated dissimilarity function should probably consist both points of view. We shall focus on the first one. We further assume that for the distance function δ holds:

There are several ways to define what is a good clustering (and each approach has its own strengths and weaknesses [13, 24]). A simple way to group change profiles based on a dissimilarity function defined in their (pairwise) common domains, is to allow two change profiles ch^X and ch^Y to be in the same group only if $\delta(ch^X, ch^Y) = 0$. The task is, given a collection \mathcal{Ch} of change profiles, to find a partition $\mathcal{Ch}_1, \dots, \mathcal{Ch}_k$ of \mathcal{Ch} such that for each $ch^X, ch^Y \in \mathcal{Ch}_i, 1 \leq i \leq k$, holds: $\delta(ch^X, ch^Y) = 0$. Let us call this problem the *change profile packing problem*. Unfortunately, the problem seems to be very difficult:

Theorem 1 *The change profile packing problem for specializing change profiles is at least as hard as the minimum graph coloring problem*

Proof. The *minimum graph coloring problem* is, given a graph $G = (V, E)$, to find a labeling $l : V \rightarrow \mathbb{N}$ of vertices with the smallest number of different labels $l(u)$ such that if $u, v \in V$ are adjacent then $l(u) \neq l(v)$ [2].

Let v_1, \dots, v_n be an (arbitrary) ordering of the vertices in V and e_1, \dots, e_m an ordering of the edges in E . We first construct an instance (σ, d) of frequent set mining problem and then show that the specializing change profiles \mathcal{Ch} computed from the frequent sets $\mathcal{F}(\sigma, d)$ can be partitioned into k sets $\mathcal{Ch}_1, \dots, \mathcal{Ch}_k$ if and only if the graph G is k -colorable. To simplify the description we shall consider, w.l.o.g., simple change profiles instead of change profiles.

The set R consists of the elements in $V \cup E$. For each v_i there are $3n$ sets $d_{i,1}, \dots, d_{i,3n}$. Each set $d_{i,j}$ contains v_i . We put the edge $\{v_i, v_j\} \in E$ into $d_{i,3(j-1)+1}$ and $d_{i,3(j-1)+2}$ if $i < j$, and otherwise into $d_{i,3j}$. We set $\sigma = 1/(3n^2)$ and compute the collection $\mathcal{F}(\sigma, d)$ of frequent sets.

The collection $\mathcal{F}(\sigma, d)$ consists of the empty set \emptyset , singletons $\{v_1\}, \dots, \{v_n\}, \{e_1\}, \dots, \{e_m\}$ and pairs $\{v_i, e\} \subset V \times E$ where $v_i \in e$. Thus the cardinality of $\mathcal{F}(\sigma, d)$ is polynomial in the size of G .

The simple change profiles of $\mathcal{F}(\sigma, d)$ are the following:

$$\begin{aligned}
sch_s^\emptyset(x) &= \begin{cases} 1/n & \text{if } x \in V \\ 1/n^2 & \text{if } x \in E \end{cases} \\
sch_s^e(v) &= 1 \quad \text{if } v \in e \\
sch_s^{v_i}(\{v_i, v_j\}) &= \begin{cases} 2/(3n) & \text{if } i < j \\ 1/(3n) & \text{if } i > j \end{cases}
\end{aligned}$$

Clearly, $\delta(sch^\emptyset, sch^v) > 0$, $\delta(sch^\emptyset, sch^e) > 0$ and $\delta(sch^v, sch^e) > 0$. Thus no two of sch^v, sch^e and sch^\emptyset

for any $v \in V$ and $e \in E$ can be in same group. On the other hand, all the sch_s^e can be packed into one set Ch_{k+1} and sch_s^0 will always have its own set Ch_{k+2} .

Now we only have show that the simple specializing change profiles $sch_s^v(e)$, can be partitioned into k sets Ch_1, \dots, Ch_k if and only if the graph G is k -colorable: No two simple specializing change profiles $sch_s^{v_i}$ and $sch_s^{v_j}$ can be in the same group if $\{v_i, v_j\} \in E$ since $sch_s^{v_i}(\{v_i, v_j\}) \neq sch_s^{v_j}(\{v_i, v_j\})$. If $\{v_i, v_j\} \notin E$ then $Dom(sch_s^{v_i}) \cap Dom(sch_s^{v_j}) = \emptyset$ they can be in the same group.

As the minimum graph coloring problem can be mapped to the change profile packing problem for specializing change profiles, the latter is at least as hard as the minimum graph coloring problem. \square

The minimum graph coloring problem is hard to approximate within $|V|^{1-\epsilon}$ for any $\epsilon > 0$ unless $NP = ZPP$ [15]. Assuming that the graph is connected we get from the above mapping from graphs to change profiles the following rough upper bound: $|Ch| = 1 + |V| + 2|E| \leq 1 + |V| + 2 \binom{|V|}{2} = \mathcal{O}(|V|^2)$. Therefore, the change profile packing problem is hard to approximate within $\Omega(|Ch|^{(1/2)-\epsilon})$ for any $\epsilon > 0$.

Although the inapproximability results seem to be devastating, there are efficient heuristics, such as the first-fit and the best-fit heuristics [10], that might be able find sufficiently good partitions efficiently. However, the usefulness of the heuristics should be evaluated carefully for different data sets and pattern collections.

The requirement that two change profiles ch^X and ch^Y can be in the same group Ch_i only if $\delta(ch^X, ch^Y) = 0$, might be too strict. A simple approach to relax this is to discretize the frequencies of frequent sets or the changes in the change profiles. Discretizations minimizing several different loss functions can be found efficiently, see e.g. [31].

Instead of minimizing the number of clusters, one could minimize the error for fixed number of clusters. This kind of clustering is called k -clustering, it is well-studied and good approximation algorithms are known if the dissimilarity function is a metric [11, 12, 14].

Unfortunately, it turns out that distance function defined to be the dissimilarity between change profiles in their common domain cannot be a metric as it cannot satisfy even the triangle inequality:

Theorem 2 *Let d be a function that measures the distance between the change profiles ch^X and ch^Y in their common domain $Dom(ch^X) \cap Dom(ch^Y)$. Then d does not satisfy the triangle inequality.*

Proof. Let ch^X , ch^Y and ch^Z be three change profiles such that $Dom(ch^X) \cap Dom(ch^Y) = Dom(ch^Y) \cap Dom(ch^Z) = \emptyset$ and $\delta(ch^X, ch^Y) > 0$. (Thus

$Dom(ch^X) \cap Dom(ch^Z) \neq \emptyset$.) Clearly, these change profiles do not satisfy the triangle inequality since $\delta(ch^X, ch^Z) > 0 = \delta(ch^X, ch^Y) + \delta(ch^Y, ch^Z)$. \square

It turns out that the k -clustering of specializing change profile is even worse than the change profile packing problem as by combining Theorem 1 and Theorem 2 we get:

Theorem 3 *k -clustering of specializing change profiles cannot be approximated within any ratio.*

Proof. If we could approximate k -clustering of specializing change profiles, then we could, by Theorems 1 and 2, solve the minimum graph coloring problem exactly. \square

A major goal in the clustering of the change profiles is to further understand the relationships between the frequent sets. As the nature of data mining is exploratory, defining a maximum number for clusters or a maximum dissimilarity threshold might be difficult and unnecessary. These parameters can be avoided by searching for hierarchical clustering instead [20]. A hierarchical clustering of Ch is a recursive partition of the elements to $2, 3, \dots, |Ch|$ clusters. It has the enormous benefit from the exploratory data analysis point of view that all the clusterings can be visualized in the same time by a tree.

There are two main categories of hierarchical clustering: agglomerative and divisive. The first begins with $|Ch|$ clusters and recursively merges them and the latter recursively partites the set Ch . Both are optimal in a certain sense: each agglomerative (divisive) hierarchical clustering of Ch into k groups is optimal w.r.t. the clustering into $k+1$ groups (into $k-1$ groups) determined the same agglomerative (divisive) hierarchical clustering.

Divisive strategy seems to be more suitable for clustering the change profiles as the dissimilarity functions we consider are defined to be distances between change profiles in their (pairwise) common domains: The agglomerative clustering groups first change profiles with disjoint domains more or less arbitrary. The choices of groups made in the first few merges can cause huge differences in the clusterings into smaller number of clusters, although the groups of change profiles with disjoint domains are probably quite unimportant for the whole hierarchical clustering. On the other hand, the divisive clustering concentrates first on the nonzero distances and thus the change profiles with disjoint domains do not bias the whole hierarchical clustering.

4 Frequency Estimation from Change Profiles

The change profiles can be used as a basis of condensed representations and several known condensed representations can be adapted to the change profiles. One interesting approach to condense the change profiles is to choose

a small set of representative change profiles (by using e.g. divisive clustering), replace the original change profiles by the chosen representatives, and then estimate the frequencies from the approximate change profiles. This can be seen as a condensed representation of the frequent sets as the approximate change profiles can fit (potentially) into smaller space than the exact change profiles or even the frequent sets. Also, the condensed representations can be applied to further condense the approximate change profiles. In addition to that frequencies can be estimated from the approximate change profiles, the change profiles can benefit from the estimation: the quality of the approximate change profiles can be assessed by evaluating how well the frequencies can be approximated from them.

For the rest of the section we consider only the case where no change profile is missing but they are corrupted. However, the methods can be generalized to handle missing change profiles and missing change profile values.

Thus, given approximations of change profiles for a frequent set collection $\mathcal{F}(\sigma, d)$, it is possible to estimate (approximations of) the frequencies of the sets in $\mathcal{F}(\sigma, d)$ from the approximative change profiles. This estimation can be done in many ways and which approach is the best depends on how the change profiles are approximated. Here we describe an approach that is based on the estimates given by different paths from the empty set \emptyset to the set X of which frequency is being estimated. Especially, we concentrate on computing the average of the frequencies given by the paths from \emptyset to X . (We describe the methods using simple specializing change profiles but the adaptation to simple generalizing change profiles is straightforward.)

The number of paths equals to the number of permutations of elements in X which is $|X|!$. If $\mathcal{F}(\sigma, d)$ consists of X and all of its subsets, then $|X| = \log |\mathcal{F}(\sigma, d)|$ and the number of paths is superpolynomial in $|\mathcal{F}(\sigma, d)|$. (This example is the worst case.) However, the average frequency estimate over all paths can be computed much faster by observing that the average frequency of X is the average of the average frequencies of $Y \subseteq X, |Y| = |X| - 1$, scaled by $ch_s^Y(X \setminus Y)$'s, i.e.,

$$fr(X) = \frac{1}{|X|} \sum_{Y \subseteq X, |Y|=|X|-1} fr(Y) sch_s^Y(X \setminus Y).$$

From this observation we can derive a dynamic programming solution for the problem:

DP-FROM-SCHS(X, sch_s)

```

1   $fr(\emptyset) \leftarrow 1$ 
2  for  $i = 1$  to  $|X|$ 
3      do for each  $Y \subseteq X, |Y| = i$ 
4          do  $fr(Y) \leftarrow 0$ 
5              for each  $Z \subset Y, |Z| = |Y| - 1$ 
6                  do  $fr(Y) \leftarrow fr(Y) + fr(Z) sch_s^Z(Y \setminus Z)$ 
7           $fr(Y) \leftarrow fr(Y) / |Y|$ 

```

8 **return** $fr(X)$

The time complexity of the algorithm is $\mathcal{O}(|X|2^{|X|}) = \mathcal{O}(|\mathcal{F}(\sigma, d)| \log |\mathcal{F}(\sigma, d)|)$ where k is the number of paths sampled. Even this can be too much. We can further speed up the estimation by simply sampling uniformly from the paths from \emptyset to X :

SAMPLER-FROM-SCHS(X, sch_s, k)

```

1   $fr(\emptyset) \leftarrow 1$ 
2   $fr(X) \leftarrow 0$ 
3  for  $j = 1$  to  $k$ 
4      do  $Y \leftarrow \emptyset$ 
5          for  $i = 1$  to  $|X| - 1$ 
6              do  $A \leftarrow \text{RANDOM-ELEMENT}(X \setminus Y)$ 
7                   $fr(Y \cup \{A\}) \leftarrow fr(Y) sch_s^Y(\{A\})$ 
8                   $Y \leftarrow Y \cup \{A\}$ 
9                   $fr(X) \leftarrow fr(X) + fr(Y) sch_s^Y(X \setminus Y)$ 
10  $fr(X) \leftarrow fr(X) / k$ 
11 return  $fr(X)$ 

```

The time complexity of the algorithm is $\mathcal{O}(k|X|)$. The algorithm can be easily modified to an any-time algorithm. This is very useful for interactive data mining (and for resource bounded computation in general).

We can apply the same algorithms to other kinds of estimates, too. Especially, if we have upper and lower bounds for $sch_s^Y(Z)$ for all $Y \subseteq X, Z \in X \setminus Y$, then we can compute upper and lower bounds for $fr(X)$: the frequency of $fr(X)$ is at most (at least) the maximum of the minimum estimates (the minimum of the maximum estimates) given by the paths from \emptyset to X .

5 Empirical Evaluation of the Frequency Estimation

In this section we experimentally evaluate the stability of the (noisified) simple (specializing) change profiles in frequency approximation. More specifically, we want to find out how the frequency estimation from change profiles tolerates different kinds of noise and how the number of sampled paths affects the quality of approximation compared to the dynamic programming solution of the problem.

Our primary data were two data sets from UCI KDD Repository:¹ Internet Usage data set consisting of 10104 rows and 10674 attributes, and IPUMS Census data set consisting of 88443 rows and 39954 attributes.

The noisified simple change profiles were constructed as follows: We computed the frequent sets with different minimum frequency thresholds from the above mentioned data sets using Christian Borgelt's implementation of the Apriori algorithm [4]. From the frequent sets, we computed the

¹<http://kdd.ics.uci.edu>

simple change profiles.

In order to study how estimation methods tolerate different kinds of noise, the change profiles were noisified in three different ways:

- randomly perturbing the values of the change profiles by $\pm\epsilon$,
- adding uniform noise from the interval $[-\epsilon, \epsilon]$ to the values of the change profiles, and
- adding zero-mean Gaussian noise with standard deviation ϵ to the values of the change profiles.

We truncated the noisified changes to the interval $[0, 1]$ as, by the definition of specializing change profile, the changes in a generalizing change profile must be in that interval.

We tested the dependency of the approximation on the number of sampled paths by evaluating the absolute difference between the correct and the estimated frequencies for the dynamic programming solution corresponding to the average over all paths, and the sampling solution corresponding to the average over different number of paths. We experimented with different number of paths, minimum frequency thresholds σ and noise levels ϵ . Also multiplicative variants of the noising schemes were tested. The results were similar to the representative results shown in Figure 1 and Figure 2. Clearly, quite small number of paths suffices to give approximations close to the dynamic programming solution.

6 Conclusions

In this paper we have introduced the concept of change profiles, a new representation for a pattern collection that attempts to reach for a more global view to data without losing the benefits of the local view. We have studied their basic properties and applications. Approaches to cluster change profiles has been suggested and several change profile clustering problems has been shown to be computationally hard for wide variety of natural dissimilarity measures. Also, we have shown how the collection of frequent sets can be approximated from a collection of their approximate change profiles.

There are several interesting open problems for further research:

- The connection between change profiles and local search should be explored systematically. For example, the change profiles could be interpreted through the research on combinatorial landscapes [36].
- The applicability of clustering methods to cluster change profiles should be evaluated. Although the theoretical results about the change profile clustering

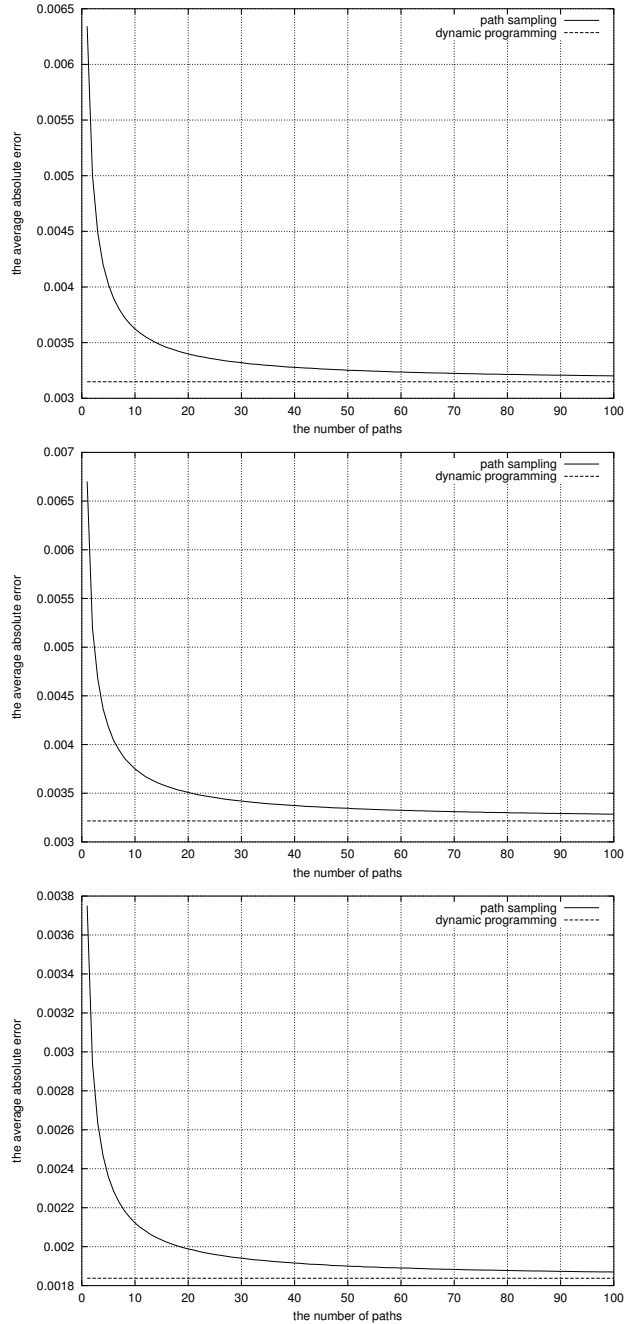


Figure 1. The error of the frequency estimation of the 0.20-frequent sets in Internet Usage data from the change profiles noisified by additive Gaussian zero-mean noise with standard deviation 0.01 (up), perturbed by ± 0.01 (middle), and noisified by additive uniform noise from the interval $[-0.01, 0.01]$ (down). The curves are averages of 1000 random experiments.

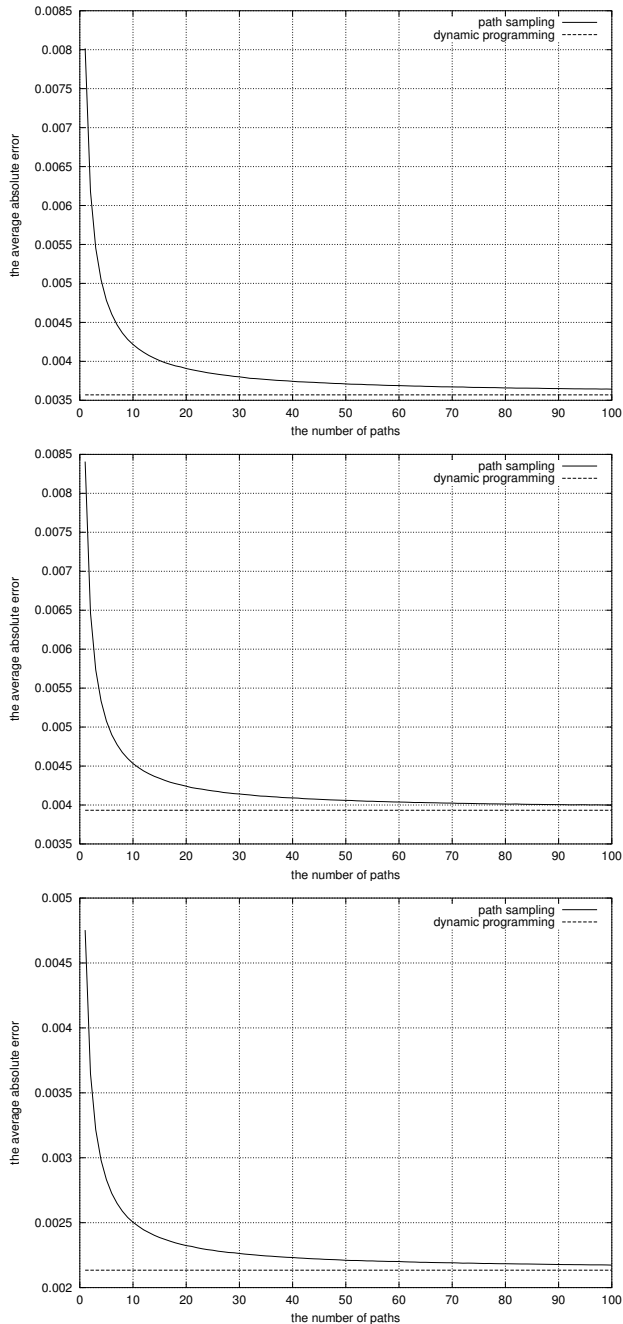


Figure 2. The error of the frequency estimation of the 0.30-frequent sets in IPUMS Census data from the change profiles noisified by additive Gaussian zero-mean noise with standard deviation 0.01 (up), perturbed by ± 0.01 (middle), and noisified by additive uniform noise from the interval $[-0.01, 0.01]$ (down). The curves are averages of 200 random experiments.

are somewhat negative, some clustering methods still seem to be useful for the change profile clustering. Also, distance functions suitable for comparing change profiles should be examined rigorously.

- The condensed representations based on change profiles should be studied further. For example, exploring the possible rule systems (see e.g. [8]) to find ones that are well-suited for condensing collections of change profiles could be useful. The frequency estimation from change profiles seems to generalize nicely to outside the frequent sets. Thus the possibilities to use change profiles to estimate the joint probability distribution from the change profiles of frequent sets should be investigated.

Acknowledgments. I wish to thank Floris Geerts and Heikki Mannila for constructive comments and encouragement.

References

- [1] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. I. Verkamo. Fast discovery of association rules. In U. M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in Knowledge Discovery and Data Mining*, chapter 12, pages 307–328. AAAI/MIT Press, 1996.
- [2] G. Ausiello, P. Crescenzi, V. Kann, A. Marchetti-Spaccamela, and M. Protasi. *Complexity and Approximation: Combinatorial Optimization Problems and Their Approximability Properties*. Springer-Verlag, 1999.
- [3] Y. Bastide, R. Taouil, N. Pasquier, G. Stumme, and L. Lakhai. Mining frequent patterns with counting inference. *SIGKDD Explorations*, 2(2):66–75, 2000.
- [4] C. Borgelt and R. Kruse. Induction of association rules: Apriori implementation. In W. Härdle and B. Rönz, editors, *15th Conference on Computational Statistics (Compstat 2002)*, pages 395–400. Physika Verlag, 2002.
- [5] J.-F. Boulicaut and A. Bykowski. Frequent closures as a concise representation for binary data mining. In T. Terano, H. Liu, and A. L. P. Chen, editors, *Knowledge Discovery and Data Mining*, volume 1805 of *Lecture Notes in Artificial Intelligence*, pages 62–73. Springer-Verlag, 2000.
- [6] J.-F. Boulicaut, A. Bykowski, and C. Rigotti. Free-sets: a condensed representation of Boolean data for the approximation of frequency queries. *Data Mining and Knowledge Discovery*, 7(1):5–22, 2003.
- [7] A. Bykowski and C. Rigotti. A condensed representation to find frequent patterns. In *Proceedings of the Twentieth ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems*. ACM, 2001.
- [8] T. Calders and B. Goethals. Mining all non-derivable frequent itemsets. In T. Elomaa, H. Mannila, and H. Toivonen, editors, *Principles of Data Mining and Knowledge Discovery*, volume 2431 of *Lecture Notes in Artificial Intelligence*, pages 74–865. Springer-Verlag, 2002.

- [9] T. Calders and B. Goethals. Minimal k -free representations of frequent sets. In N. Lavrac, D. Gamberger, L. Todorovski, and H. Blockeel, editors, *Knowledge Discovery in Databases: PKDD 2003*, volume 2838 of *Lecture Notes in Artificial Intelligence*. Springer-Verlag, 2003.
- [10] E. G. Coffman Jr., C. Courcoubetis, M. R. Garey, D. S. Johnson, P. W. Shor, R. R. Weber, and M. Yannakakis. Perfect packing theorems and the average-case behaviour of optimal and online bin packing. *SIAM Review*, 44(1):95–108, 2002.
- [11] S. Dasgupta. Performance guarantees for hierarchical clustering. In J. Kivinen and R. H. Sloan, editors, *Computational Learning Theory*, volume 2375 of *Lecture Notes in Artificial Intelligence*, pages 351–363. Springer-Verlag, 2002.
- [12] W. F. de la Vega, M. Karpinski, C. Kenyon, and Y. Rabani. Approximation schemes for clustering problems. In *Proceedings on 35th Annual ACM Symposium on Theory of Computing*. ACM, 2003.
- [13] V. Estivill-Castro. Why so many clustering algorithms – a position paper. *SIGKDD Explorations*, 4(1):65–75, 2002.
- [14] T. Feder and D. H. Greene. Optimal algorithms for approximate clustering. In *Proceedings of the twentieth annual ACM Symposium on Theory of Computing, Chicago, Illinois, May 2–4, 1988*, pages 434–444. ACM, 1988.
- [15] U. Feige and J. Kilian. Zero knowledge and the chromatic number. *Journal of Computer and Systems Science*, 57(2):187–199, 1998.
- [16] M. Garofalakis, R. Rastogi, and K. Shim. Mining sequential patterns with regular expression constraints. *IEEE Transactions on Knowledge and Data Engineering*, 14(3):530–552, 2002.
- [17] F. Geerts, B. Goethals, and J. Van den Bussche. A tight upper bound on the number of candidate patterns. In N. Cercone, T. Y. Lin, and X. Wu, editors, *Proceedings of the 2001 IEEE International Conference on Data Mining (ICDM 2001)*, pages 155–162. IEEE Computer Society, 2001.
- [18] J. Han, J. Pei, and Y. Yin. Mining frequent patterns without candidate generation. In W. Chen, J. F. Naughton, and P. A. Bernstein, editors, *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, pages 1–12. ACM, 2000.
- [19] D. J. Hand. Pattern detection and discovery. In D. Hand, N. Adams, and R. Bolton, editors, *Pattern Detection and Discovery*, volume 2447 of *Lecture Notes in Artificial Intelligence*, pages 1–12. Springer-Verlag, 2002.
- [20] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Series in Statistics. Springer-Verlag, 2001.
- [21] J. Hipp, U. Güntzer, and G. Nakhaeizadeh. Algorithms for association rule mining – a general survey and comparison. *SIGKDD Explorations*, 1(2):58–64, 2000.
- [22] J. D. Holt and S. M. Chung. Mining association rules using inverted hashing and pruning. *Information Processing Letters*, 82:211–220, 2002.
- [23] T. Imieliński, L. Khachiyan, and A. Abdulghani. Cube-grades: Generalizing association rules. *Data Mining and Knowledge Discovery*, 6(3):219–257, 2002.
- [24] J. Kleinberg. An impossibility theorem for clustering. In *Advances in Neural Information Processing Systems (NIPS)*, volume 15, 2002.
- [25] M. Kryszkiewicz. Concise representation of frequent patterns based on disjunction-free generators. In N. Cercone, T. Y. Lin, and X. Wu, editors, *Proceedings of the 2001 IEEE International Conference on Data Mining*, pages 305–312. IEEE Computer Society, 2001.
- [26] M. Kurakochi and G. Karypis. Discovering frequent geometric subgraphs. In *Proceedings of the 2002 IEEE International Conference on Data Mining*. IEEE Computer Society, 2002.
- [27] H. Mannila. Local and global methods in data mining: Basic techniques and open problems. In P. Widmayer, F. Triguero, R. Morales, M. Hennessy, S. Eidenbenz, and R. Conejo, editors, *Automata, Languages and Programming*, volume 2380 of *Lecture Notes in Computer Science*, pages 57–68. Springer-Verlag, 2002.
- [28] H. Mannila and H. Toivonen. Levelwise search and borders of theories in knowledge discovery. *Data Mining and Knowledge Discovery*, 1(3):241–258, 1997.
- [29] H. Mannila, H. Toivonen, and A. I. Verkamo. Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery*, 1(3):259–289, 1997.
- [30] T. Mielikäinen. Chaining patterns. In G. Grieser, Y. Tanaka, and A. Yamamoto, editors, *Discovery Science*, volume 2843 of *Lecture Notes in Artificial Intelligence*, pages 232–243. Springer-Verlag, 2003.
- [31] T. Mielikäinen. Frequency-based views to pattern collections. In *IFIP/SIAM Workshop on Discrete Mathematics and Data Mining*, 2003.
- [32] T. Mielikäinen and H. Mannila. The pattern ordering problem. In N. Lavrac, D. Gamberger, L. Todorovski, and H. Blockeel, editors, *Knowledge Discovery in Databases: PKDD 2003*, volume 2838 of *Lecture Notes in Artificial Intelligence*. Springer-Verlag, 2003.
- [33] T. M. Mitchell. Generalization as search. *Artificial Intelligence*, 18(2):203–226, 1982.
- [34] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal. Discovering frequent closed itemsets for association rules. In C. Beeri and P. Buneman, editors, *Database Theory - ICDT'99*, volume 1540 of *Lecture Notes in Computer Science*, pages 398–416. Springer-Verlag, 1999.
- [35] J. Pei, G. Dong, W. Zou, and J. Han. On computing condensed pattern bases. In *Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM 2002)*, pages 378–385. IEEE Computer Society, 2002.
- [36] C. M. Reidys and P. F. Stadler. Combinatorial landscapes. *SIAM Review*, 44(1):3–54, 2002.
- [37] M. J. Zaki. Scalable algorithms for association mining. *IEEE Transactions on Knowledge and Data Engineering*, 12(3):372–390, 2000.
- [38] M. J. Zaki. Efficiently mining frequent trees in a forest. In D. Hand, D. Keim, and R. Ng, editors, *Proceedings of the Eight International Conference on Knowledge Discovery and Data Mining (KDD-2002)*. ACM, 2002.