Delta-Closure Structure for Studying Data Distribution

Aleksey Buzmakov, Sergei O Kuznetsov, Tatiana Makhalova, Amedeo Napoli

Δ-Closure Structure for Studying Data Distribution

Aleksey Buzmakov
National Research University Higher School of Economics
Perm, Russia
avbuzmakov@hse.ru

Sergei O. Kuznetsov
National Research University Higher School of Economics
Moscow, Russia
skuznetsov@hse.ru

Tatiana Makhalova
Université de Lorraine, CNRS, Inria, LORIA
54000 Nancy, France
t.makhalova@gmail.com

Amedeo Napoli
Université de Lorraine, CNRS, Inria, LORIA
54000 Nancy, France
amedeo.napoli@loria.fr

***Abstract*—In this paper, we revisit pattern mining and study the distribution underlying a binary dataset thanks to the closure structure which is based on passkeys, i.e., minimum generators in equivalence classes robust to noise. We introduce Δ-closedness, a generalization of the closure operator, where Δ measures how a closed set differs from its upper neighbors in the partial order induced by closure. A Δ-class of equivalence includes minimum and maximum elements and allows us to characterize the distribution underlying the data. Moreover, the set of Δ-classes of equivalence can be partitioned into the so-called Δ-closure structure. In particular, a Δ-class of equivalence with a high Δ is supported by more observations and thus is more stable. In the experiments, we study the Δ-closure structure of several real-world datasets and show that this structure is very stable for large Δ and does not substantially depend on the data sampling used for the analysis.**

***Index Terms*—pattern mining, closed sets, equivalence class, generators, data distribution**

I. Introduction

In this paper, we are interested in pattern or itemset mining in tabular data. There is a considerable work on this subject, especially regarding algorithms and search for interesting patterns [1]. Here we rather focus on the distribution underlying the dataset under study thanks to closed itemsets, their equivalence classes, and the associated generators.

Many pattern mining approaches produce a particular set of patterns, which provides a certain "view" of the intrinsic structure underlying the data. However, this view, usually, is not directly related to the distribution underlying the data. In this paper we rely on the *closure structure* which was recently introduced [10] and which characterizes the intrinsic structure of a dataset. The closure structure reveals the distribution of the itemsets in the data in terms of frequency and also of stability, i.e., how a closed set depends on its content, and in addition supports an interpretation of the content of a dataset. Here we propose a methodology for computing and understanding the intrinsic structure of a dataset in two main ways: (i) we revisit the *closure structure*, which reveals the distribution of the itemsets in the data in terms of frequency, (ii) we generalize the closure structure to Δ-*closure structure*, which is a more robust closure structure related to the distribution underlying the data. The closure structure and the Δ-closure stracture support an interpretation of the content of a dataset. The closure structure is based on "closed itemsets" and minimum elements of their "equivalence classes" –related to levels–, which are computed independently of any interestingness measure or set of constraints. The closure structure and its levels provide a representation of the complexity of the content of a dataset. We propose a formalization of the closure structure in terms of Formal Concept Analysis [7], which is well adapted to the study of closed itemsets, equivalence classes, and data topology in terms of closed sets.

Then we generalize the notion of closure to Δ-closedness which allows us to work with stability along with frequency. Actually, Δ measures how much a closed set differs from its upper neighbors in the partial order of closed sets. A Δ-class of equivalence allows us to characterize the distribution underlying the data: (i) when Δ is large, there are only a few Δ-classes of equivalence whose elements are very stable, (ii) when Δ is small, the number of Δ-classes increases while the related information is less stable and depends on a smaller number of elements. This allows us to study stable patterns which are robust against noise and also to dynamic changes in the data.

In the experiments, we compute the closure structure of a number of public datasets and show how Δ-closedness is taken into account. The closure structure may roughly be of three types, where the first levels of the structure are the most interesting, stable, and interpretable in standard ("typical") and almost standard ("intermediate") datasets, while in non-standard datasets ("deterministic"), levels are more separated and less easily interpretable. For example, the closure structure allows us to determine the levels where the itemsets are the most diverse. It shows how the frequency and the stability are distributed among the levels, and when a search for interesting itemsets can be stopped without major loss. Moreover, the Δ-closure structure is very stable for large Δ and does not substantially depend on the data sampling used for the analysis. To the best of our knowledge, such kind of study and visualisation

Sergei O. Kuznetsov would like to thank Basic Research Program of the HSE University for its support.

of data is rather unique. The proposed methodology and the related tools are useful for simple navigation in the pattern search space.

The paper has the following structure. Section II presents related work and motivation. In Section III we firstly introduce Δ-closure with the related equivalence classes and passkeys. We also provide a concrete example and we make precise an algorithm for discovering the Δ-closure and the related elements. Then in Section IV we present and discuss experiments about the computing of the Δ-closure structure and the stability of passkeys. We also show how this structure can be visualized and interpreted. Then we conclude the paper and give directions for future work.

II. Related work and Motivation

Following the research directions discussed in [2] and [10], we are interested in discovering closed and stable patterns together with their related equivalence classes and minimum generators (passkeys). As in [2], we introduce a Δ parameter for studying Δ-closure and the robustness of the discovered patterns against noise, and as well the stability of the associated equivalence classes and their characteristic elements such as passkeys. Moreover, as in [10] we are also interested in investigating the data distribution through the closure structure. The objective is to build a kind of "internal picture" of the datasets under study and to check at which level –related to closure and Δ-closure– one can discover the most interesting patterns, and as well interesting implications or association rules. In addition, in the present paper, we introduce Δ-closure structure based on Δ-closure operator which subsumes ordinary closure.

The interest in data distribution is frequently appearing in the literature. Indeed, even the standard frequent itemset mining indirectly focuses on properties of the distribution. Such interest in the distribution can be stated more formally if the itemset frequency is verified as a statistical hypothesis [14]. Some papers study how to take noise into account in pattern mining, which can also be considered as indirect characterization of the distribution underlying the data. In particular, δ-free sets [3] are special classes of patterns that encode association rules with a high confidence –only few counter-examples are accepted– for allowing identification of noise-resistant patterns [12]. Other studies were also performed to find out classes of patterns resistant to noise [8], [18].

Still about data distribution, in [13] the authors study distribution of sizes of frequent and maximal frequent itemsets in a database. In [6] the authors are mostly interested in the distributions related to the border between frequent and infrequent itemsets together with the distribution of three concise representations: frequent closed, frequent free, and frequent essential itemsets. In [16] the authors propose an efficient approach to estimation of the number of frequent patterns for arbitrary minsup thresholds. In our study we assume that "important" patterns enabling better interpretation are not necessarily frequent, but are certainly stable to noise in data. So, here we study empirical distribution of noise tolerance in terms of Δ-stability introduced below.

Computational complexity remains an important challenge in itemset mining. The number of itemsets can be exponential w.r.t. the dataset size. Focusing on closed itemsets allows for a substantial reduction of this number by replacing an equivalence class, i.e., a whole class of itemsets having the same support, by the largest one which is the closed itemset [11]. For dealing with large amounts of closed itemsets, it is possible to evaluate itemsets thanks to quality metrics [9], [17]. For example Δ-closedness evaluates the robustness of patterns and the corresponding closure operation [2]. Roughly speaking, a Δ-closed set cannot be augmented by any further item without decreasing its support by at least Δ. Actually, Δ-closed sets allow to capture interesting itemsets that are stable, i.e., robust to noise and changes in the data. A very similar notion is δ-tolerance closed itemsets which are defined w.r.t. itemset support rather than an absolute change in support for Δ-closure [4]. An alternative to exhaustive enumeration of itemsets is based on "sampling" [5] and on a gradual search for itemsets according to an interestingness measure or a set of constraints [15]. Such algorithms usually output a rather small set of itemsets while they may provide only an approximate solution.

The papers discussed above do not address the problem of discovering an intrinsic structure underlying the set of closed itemsets, and –more generally– the dataset itself. By contrast, in this paper, we define such a level-wise structure for representing this intrinsic structure of the datasets. Moreover, we build this structure by means of Δ-closedness operator and we study the stability of the related Δ-closure levels w.r.t. the size of Δ. The Δ-closure reveals the internal structure of a dataset and as well relates this structure to the distribution underlying the dataset.

III. Formalism

A dataset is modeled as a *formal context* [7], which is a triple $\mathbb{K} = (G, M, I)$, where G is a set of objects, M is a set of attributes, and $I \subseteq G \times M$ is an incidence relation such that $(g, m) \in I$ if object g has attribute m. Two derivation operators, both denoted by $(\cdot)'$, are defined for $A \subseteq G$ and $B \subseteq M$ as follows:

$$A' = \{m \in M \mid \forall g \in A : gIm\} \quad (1)$$

$$B' = \{g \in G \mid \forall m \in B : gIm\} \quad (2)$$

Intuitively, A' is the set of attributes common to objects in A, while B' is the set of objects which have all attributes in B. Sets $A \subseteq G$, $B \subseteq M$, such that $A = A''$ and $B = B''$, are closed sets. Moreover, $(\cdot)''$ is a closure operator equivalent to instance counting closure in data mining as introduced in [11]. For $A \subseteq G$, $B \subseteq M$, a pair (A, B) such that $A' = B$ and $B' = A$, is called a *formal concept*, then A and B are closed sets and called *extent* and *intent*, respectively.

	m_1	m_2	m_3	m_4	m_5	m_6	m_7	m_8	m_9
	a	b	c	d	e	f	g	h	i
g_1	×	×	×	×	×				
g_2	×	×	×	×	×				
g_3	×	×	×	×					
g_4	×	×	×		×				
g_5	×	×	×			×			
g_6	×	×	×				×		
g_7	×	×		×	×			×	
g_8	×	×		×	×				×
g_9	×								
g_{10}		×							

Fig. 1: A toy binary dataset where 'x' in cell (i, j) indicates that object g_i is related to attribute m_j.

A. Δ-classes of equivalence

Definition III.1. *An itemset B is called Δ-closed if for any attribute $m \in M$:*

$$|B'| - |(B \cup \{m\})'| \geq \Delta \geq 1. \tag{3}$$

If $\Delta = 1$, then the itemset is just *closed* w.r.t. object counting [2], [11]. In [2] it was also shown that Δ-closedness is associated with a closure operator that we call Δ-closure and that works as follows. Given a non Δ-closed itemset B, i.e., $\exists m \in M(|B'| - |(B \cup \{m\})'| < \Delta)$, B can be closed by iteratively changing B to $B \cup \{m\}$, for any m violating (3) until no such attribute is found. The result is the Δ-closure of B. The corresponding closure operator is denoted by ϕ_Δ. Moreover any closure operator is associated with a class of equivalence.

Definition III.2. *Given an itemset X, the equivalence class $Equiv_\Delta(X)$ of X is the set of all itemsets whose Δ-closure is equal to the Δ-closure of X, i.e.,*

$$Equiv_\Delta(X) = \{Y \subseteq M \mid \phi_\Delta(Y) = \phi_\Delta(X)\}. \tag{4}$$

Some elements of a class of equivalence can be highlighted since they have special properties [10].

Definition III.3. *A Δ-*key $X \in Equiv_\Delta(B)$ *is any minimal (w.r.t. subset relation) itemset in $Equiv_\Delta(B)$. We denote the set of Δ-keys of B, or Δ-key set, by $Key_\Delta(B)$.*

Definition III.4. *An itemset $X \in Key_\Delta(B)$ is called a Δ-*passkey *if it has the smallest size among all keys in $Key_\Delta(B)$.*

*For a Δ-closed itemset B, the Δ-*passkey set*, denoted by $pKey_\Delta(B)$, is given by $pKey_\Delta(B) = \{X \mid X \in Key_\Delta(B), |X| = min_{Y \in Key_\Delta(B)}|Y|\}$, and it verifies $pKey_\Delta(B) \subseteq Key_\Delta(B)$.*

A concrete example about Δ-closure, Δ-keys, and Δ-passkeys is given in Section III-B. It can be checked that Δ-classes of equivalence with higher Δ make coarser partitions of the set of concept intents (i.e., Δ-closed itemsets for $\Delta = 1$)) than Δ-classes of equivalence with smaller Δ, i.e., $(\forall \Delta \geq 2)(\forall X \subseteq M)(\exists X_1, \ldots, \exists X_k) Equiv_\Delta(X) = \bigcup Equiv_{\Delta-1}(X_i)$. Moreover, we can associate every Δ-class of equivalence with the size of its passkeys. This size corresponds to a *level* in the class of equivalence. Then for any Δ the complexity of the dataset can be captured by means of the distribution of the classes of equivalence within the levels, called here the *level structure*.

Now, given an itemset X, let us define the following measures:

$$\Delta(X) = \max\{0 \leq \Delta \leq |G| \mid X \text{ is } \Delta\text{-closed}\} \tag{5}$$

$$\Delta_{key}(X) = \max\{0 \leq \Delta \leq |G| \mid X \text{ is a } \Delta\text{-key}\} \tag{6}$$

$$\Delta_{passkey}(X) = \max\{0 \leq \Delta \leq |G| \mid X \text{ is a } \Delta\text{-passkey}\} \tag{7}$$

All these measures capture the maximal Δ for which the itemset X preserves a certain property, i.e., being closed, being a key, or being a passkey. We will call these values Δ-values of closed itemsets, keys, and passkeys, respectively. Since any passkey is a key, $\Delta_{passkey}(X) \leq \Delta_{key}(X)$. It is clear from the definitions that for all $1 \leq \Delta \leq \Delta_{key}(X)$, X is a Δ-key, and for all $1 \leq \Delta \leq \Delta_{passkey}(X)$, X is a Δ-passkey.

Let us now relate these measures to the variations existing in a dataset. More precisely, is it possible to discover the itemsets that are likely to be preserved when a dataset is changed?

Proposition III.1. *Let $\Delta_{passkey}(X) = \delta$. Then at least δ objects should be removed from the dataset $\mathbb{K}$ to obtain a subdataset $\mathbb{K}_s \subseteq \mathbb{K}$ such that X is not a passkey in $\mathbb{K}_s$.*

Proposition III.2. *Let $\Delta_{key}(X) = \delta$. Then at least δ objects should be removed from the dataset $\mathbb{K}$ to obtain a subdataset $\mathbb{K}_s \subseteq \mathbb{K}$ such that X is not a key in $\mathbb{K}_s$.*

Proposition III.3. *Let $\Delta(X) = \delta$. Then at least δ objects should be removed from the dataset $\mathbb{K}$ to obtain a subdataset $\mathbb{K}_s \subseteq \mathbb{K}$ such that X is not closed in $\mathbb{K}_s$.*

The proofs of these propositions can be found in a technical report at http://arxiv.org/abs/2210.06926. These propositions show that, the higher is Δ, the deeper should the dataset be modified for removing a certain property of the dataset. Thus, if the structure of the dataset is captured with itemsets of high Δ, then the discovered itemsets are stable w.r.t. object removal. This fact is related to the distribution underlying the dataset the objects are taken from.

As a final remark, let us notice that an itemset X is a Δ-key only if $\forall Y \subset X\ (|Y'| - |X'| \geq \Delta)$. This is only a necessary condition. For example, in the following context with 3 objects $G = \{g_1, g_2, g_3\}$, $\{g_1\}' = \emptyset$, $\{g_2\}' = \{m_1\}$, and $\{g_3\}' = \{m_1, m_2\}$, the aforementioned condition is satisfied for $\{m_2\}$ and $\Delta = 2$. However, $\{m_2\}$ is not a Δ-key, but a δ-free set [3]. This shows that not every δ-free set is a Δ-key.

B. Example

Figure 1 displays a toy formal context. Every object except objects 1, 2, 3, and 4, has its own attribute set and thus the size of their passkeys is 1. The closed itemsets (intents) along with their supports (extents) are shown in Fig. 2. For brevity sake, we write 123 instead of $\{1, 2, 3\}$ and so on. The concepts

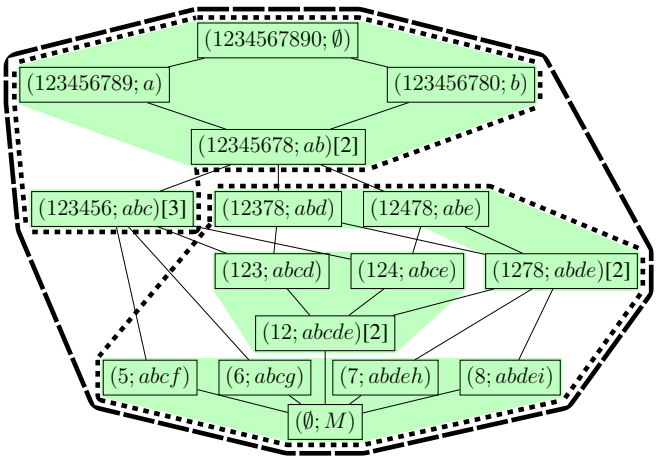


Fig. 2: The hierarchical structure of Δ-equivalence classes.

Algorithm 1: A general algorithm for finding Δ-values for all concept intents and for their passkeys.

Data: $\mathbb{K}$ is the dataset
Result: A set of concepts associated with their levels and the values of Δ for the concept intent and for the concept passkey

1 $\{c_i \Rightarrow level_i\} =$ `GDPM(`$\mathbb{K}$`)`;
2 $\{c_i \Rightarrow \langle \Delta_{cls}(c_i), Ref(c_i)\rangle\} =$ `ComputeDeltas(`$\{c_i\}$`)`;
3 $\{c_i \Rightarrow \Delta_{passkey}(c_i)\} =$ `ComputePKDeltas(`$\{c_i \Rightarrow \langle level_i, \Delta_{cls}(c_i), Ref(c_i)\rangle\}$`)`;

with extents 123456 and 12478 have passkeys equal to c and e, respectively. The passkey for concept with extent 12345678 is ab and the passkey for the top concept is $\emptyset$. The concept with extent 12 has a passkey larger than 2, namely cde.

In Fig. 2, the Δ of the intent or Δ-measure is indicated in the square brackets for every concept when it is different from 1. The concept intents are the maximal elements in the Δ-classes of equivalence for Δ not larger than their Δ-measure.

Now consider the Δ-classes of equivalence for $\Delta = 2$. There are only 4 concepts with a Δ-measure not smaller than 2, thus, there are only 4 classes of equivalence for $\Delta = 2$ plus the "technical" class of equivalence at the bottom. These classes of equivalence with $\Delta = 2$ are formed by joining smaller classes of equivalence, i.e., the classes of equivalence with $\Delta = 1$ which are related to single concepts. Then in Fig. 2, the classes of equivalence with $\Delta = 2$ are shown in green areas, while the simple concepts, i.e., closed sets, correspond to classes of equivalence with $\Delta = 1$.

Now, if we consider Δ-classes of equivalence for $\Delta = 3$, there are just two closures abc and M. These equivalence classes are surrounded by the dotted line in the diagram. The Δ-class containing the closed set abc contains the concepts with the intents abc, ab, a, b, $\emptyset$. The other concepts belong to the Δ-class of equivalence whose maximal element is M. If we further increase Δ, the whole lattice will collapse into a single class of equivalence.

C. Computational considerations

The algorithm for computing Δ-equivalence classes is sketched in Algorithm 1 while details are provided in Algorithm 2. Firstly, we compute the concepts and the associated concept levels using GDPM algorithm [10], which iterates over keys searching for passkeys. Then, for every concept (A, B) we compute the value of $\Delta(B)$ as given by equation (5). Finally, for every concept (A, B) we compute the value of $\Delta_{passkey}(B)$ as given by equation (7). This procedure is made precise in Algorithm 2.

At the beginning the value of $\Delta_{passkey}(B)$ is initialized to 1 for every concept (A, B). Then, since any $(\Delta + 1)$-class of equivalence is a union of Δ-classes of equivalence, we close every Δ-closed element with $(\Delta + 1)$-closure (lines 10–19). Finally, if two Δ-classes of equivalence have the same $(\Delta+1)$-closure, we select one passkey among the possible $(\Delta + 1)$-passkeys (lines 20–29).

The computational time complexity of Algorithm 2 is $O(|G| \cdot \log(|G|) \cdot |\{c_i\}|)$, since we need to iterate over all possible thresholds of Δ and for any given threshold we need to find the Δ-closure of every concept, which can be found in at most $\log(|G|)$ time, since on every new iteration of loop 12–17 the distance to the closure from the concept is either doubled or the closure is found.

IV. Experiments

The experiments are carried out using a machine with Intel Core i5 CPU, 16GB of RAM and Nvidea GeForce RTX 3080 video card operated under Ubuntu 20.04 operating system. Table I shows datasets with total computation time larger than 10 seconds. It also shows the number of objects, attributes, and closed itemsets. The computation time is divided into the computation time of the level structure with GDPM and the time for computing Δ-classes of equivalence. We can observe that the most important factor in the computation time is the number of closed itemsets in the dataset, the second factor being the size of the dataset.

A. Stability of passkeys

The goal of the experiments is to verify that Δ-closure structures are becoming more robust when Δ is increasing. In particular, Δ-closure structures when Δ is greater than 1 are more robust than closure structures or 1-closure structures.

For any Δ-equivalence class we distinguish its minimum elements called the Δ-passkeys and the maximum one called the Δ-closed itemset. They are both characterized by Δ, where the higher value for Δ, the more stable are passkeys/closed itemsets w.r.t. noise. A passkey can be considered as a most concise "definition" of a Δ-closed itemset.

In this section we study how stable Δ-equivalence classes are distributed in the dataset. To analyze the datasets we use the following observations:

(i) the equivalence classes with small passkeys are easier to mine by the algorithm that computes the closure

Algorithm 2: An algorithm for computing Δ of passkeys for every concept.

Data: $\{c_i\}$ is a set of concepts with associated levels, Δ-value for their intent and the most closest child

Result: The value of Δ for passkeys of every concept

```
def ComputePKDeltas({c_i}):
    foreach c ∈ {c_i} do
        Δ_pk(c) = 1;
        Cls(c) = c
    end
    d = 1;
    while d < |G| do
        d = d + 1;
        hasUpdates = True;
        while hasUpdates is True do
            hasUpdates = False;
            foreach c ∈ {c_i} do
                cls = Cls(c);
                if |Ext(Ref(cls))| − |Ext(cls)| < d
                 then
                    Cls(c) = Cls(cls);
                    hasUpdates = True;
                end
            end
        end
        foreach c ∈ {c_i} do
            cls = Cls(c);
            level_Δ(cls) = min(level_Δ(cls), level(c));
        end
        foreach c ∈ {c_i} do
            cls = Cls(c);
            if level(c) == level_Δ(cls) then
                Δ_pk(c) = d;
            end
        end
    end
```

structure [10], and easier to interpret, since Δ-passkeys are the smallest elements in the Δ-equivalence classes,

(ii) the Δ-equivalence classes where passkeys have high Δ are more stable w.r.t. noise.

We analyzed 25 datasets from LUCS-KDD repository (their parameters are given in the supplementary material) paying attention to the properties mentioned above. Among all the datasets under analysis we discovered three main types of behaviors –typical, deterministic, and two kinds of intermediates– which are displayed in the columns of Fig. 3.

B. Visualizing the closure structure

Fig. 3 visualizes Δ-closure-structures for different types of data. The first level at the bottom of each subfigure shows the distribution of passkeys constituted of only one attribute w.r.t. their values of Δ. Actually this first level corresponds to the attribute frequency distribution and does not contain any information about interaction between attributes. By contrast, starting from the second level, the Δ-closure levels include information about interactions between attributes. For example, the second level contains Δ-closed itemsets generated from passkeys of size 2.

TABLE I: Dataset characteristics and computation time

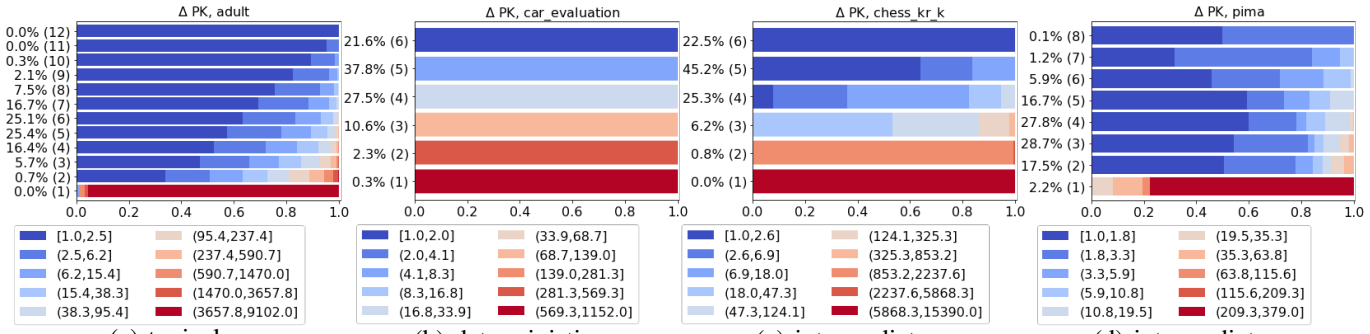
Datasets	$\|G\|$	$\|M\|$	# closed	runtime, sec	
				GDPM	Δ-passkeys
adult	48842	95	359141	984	247
chess kr k	28056	40	84636	146	52
cyl. bands	540	120	39829537	2404	8803
horse col.	368	81	173866	11	21
ionosphere	351	155	23202541	2467	3090
mushroom	8124	88	181945	164	27
nursery	12960	27	115200	46	24
pen digits	10992	76	3605507	12863	5020
soybean	683	99	2874252	379	393

Datasets may have mainly three main types of behaviors. A majority of the datasets which were analyzed, namely "auto", "cylinder bands", "dermatology", "ecoli", "glass", "heart diseases", "hepatitis", "horse colic", "pen digit", "soybean", "wine", "zoo", have a closure structure which is similar to the closure structure of "adult", as displayed in Fig. 3a. We call this type of behavior "typical". For all these datasets, the ratio of the equivalence classes having passkeys with a low Δ increases exponentially with the level number. It means that most of the closed itemsets are not stable and can easily become not closed if new objects are added. The equivalence classes which are the most stable to noise are located at the first levels.

The second type of behavior is given in Fig. 3b and termed as "deterministic". It is observed for "car evaluation" and "nursery" datasets, where each level mostly contains Δ-equivalence classes with passkeys having the same value. Actually these datasets were built from hierarchical decision models. Then each entry in such a dataset usually has a rather "determinististic" behavior, i.e., a much less random nature while higher order interactions between attributes are unusual.

The "intermediate" behavior can be different. In some cases the number of "unstable" Δ-equivalence classes grows less rapidly than for the majority of "typical" datasets. This means that a larger number of "stable" Δ-equivalence classes are present in the upper levels. The other datasets with similar behavior are "iris", "led7", "mushroom", and "tic tac toe".

Another type of intermediate behavior was observed for the datasets "ionosphere", "page blocks", and "pima", and is presented in Fig. 3d, where the Δ-closure structure of "pima" is shown. For these datasets, the ratio of the Δ-equivalence classes including passkeys with larger Δ-values may increase with levels, i.e., as the levels increase, the ratio of more robust Δ-equivalence classes increases as well. Intuitively, this kind of behavior is related to datasets where the 1st level of the Δ-closure structure contains a substantial number of attributes of very low frequency. This induces unreliable or unstable itemsets in the upper levels of the Δ-closure structure.



(a) typical (b) deterministic (c) intermediate (d) intermediate

Fig. 3: The distribution of Δ-passkeys by levels based on 10 frequency bins distributed within the Δ-closure levels. The horizontal bars represent closure levels, where the level number is given in parentheses. The leftmost value is the percentage of closed itemsets $|\mathcal{C}_k|/|\mathcal{C}|$ at level k. The width of a color within a bar is proportional to the ratio of Δ-passkeys for Δ in $(v_1, v_2]$ at level $\mathcal{C}_k$.

Finally, the patterns with high Δ-values are even more interesting if they are located at high levels of the closure structure, since they reflect high order interactions between attributes and are very stable to noise.

V. Discussion and Conclusion

In this paper we introduced the "Δ-closure structure" of a dataset, along with their theoretical and practical properties. The closure structure provides a view of the content of a dataset and may be used for guiding data mining. We have found three main types of "dataset behaviors" which are showing the stability and robustness of the corresponding closed itemsets. Here after we synthesize the main aspects of the closure structures:

- The Δ-closure structure is a partition of the whole set of itemsets. It is determined by the Δ-closed itemsets and their equivalence classes. The latter are represented by delta-passkeys, i.e., i.e., the smallest keys in an equivalence class. Accordingly, passkeys can represent any closed itemset of attributes without loss.
- The closure structure is organized around a set of levels depending on the size of the passkeys, which is related to the distribution of the closed itemsets and their passkeys. Moreover, the Δ-closure structure allows us to study the stability and robustness of the closed itemsets.

In future work, further studies are needed for a deeper understanding of the benefits of the Δ-closure structure for dealing with noisy data. More specifically, we should investigate the advantages brought by the Δ-closure structure for taking noise into account (i) in practical applications, and (ii) in interpreting the resulting patterns.

References

[1] C. C. Aggarwal and J. Han, eds., *Frequent Pattern Mining*, Springer, 2014.

[2] M. Boley, T. Horváth, and S. Wrobel, *Efficient Discovery of Interesting Patterns Based on Strong Closedness*, in Proceedings of SDM, SIAM, 2009, pp. 1002–1013.

[3] 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 (2003), pp. 5–22.

[4] J. Cheng, Y. Ke, and W. Ng, *δ-Tolerance Closed Frequent Itemsets*, in Proceedings of ICDM, IEEE, 2006, pp. 139–148.

[5] V. Dzyuba, M. van Leeuwen, and L. D. Raedt, *Flexible constrained sampling with guarantees for pattern mining*, Data Mining and Knowledge Discovery, 31 (2017), pp. 1266–1293.

[6] F. Flouvat, F. D. Marchi, and J. Petit, *A new classification of datasets for frequent itemsets*, Journal of Intelligent Information Systems, 34 (2010), pp. 1–19.

[7] B. Ganter and R. Wille, *Formal Concept Analysis – Mathematical Foundations*, Springer, 1999.

[8] M. Klimushkin, S. A. Obiedkov, and C. Roth, *Approaches to the Selection of Relevant Concepts in the Case of Noisy Data*, in Proceedings of ICFCA, Springer LNCS 5986, 2010, pp. 255–266.

[9] S. O. Kuznetsov and T. P. Makhalova, *On interestingness measures of formal concepts*, Information Sciences, 442-443 (2018), pp. 202–219.

[10] T. Makhalova, A. V. Buzmakov, S. O. Kuznetsov, and A. Napoli, *Introducing the closure structure and the GDPM algorithm for mining and understanding a tabular dataset*, International Journal of Approximate Reasoning, 145 (2022), pp. 75–90.

[11] N. Pasquier, Y. Bastide, R. Taouil, and L. Lakhal, *Efficient mining of association rules using closed itemset lattices*, Information Systems, 24 (1999), pp. 25–46.

[12] R. G. Pensa and J. Boulicaut, *Towards Fault-Tolerant Formal Concept Analysis*, in Proceedings of AI*IA, Springer LNCS 3673, 2005, pp. 212–223.

[13] G. Ramesh, W. Maniatty, and M. J. Zaki, *Feasible itemset distributions in data mining: theory and application*, in Proceedings of PODS, ACM, 2003, pp. 284–295.

[14] M. Riondato and F. Vandin, *Finding the True Frequent Itemsets*, in Proceedings of SDM, SIAM, 2014, pp. 497–505.

[15] K. Smets and J. Vreeken, *Slim: Directly Mining Descriptive Patterns*, in Proceedings of SDM, SIAM, 2012, pp. 236–247.

[16] M. van Leeuwen and A. Ukkonen, *Fast Estimation of the Pattern Frequency Spectrum*, in Proceedings of ECML-PKDD, Springer LNCS 8725, 2014, pp. 114–129.

[17] J. Vreeken and N. Tatti, *Interesting Patterns*, in Frequent Pattern Mining, C. C. Aggarwal and J. Han, eds., Springer, 2014, pp. 105–134.

[18] C. Yang, U. M. Fayyad, and P. S. Bradley, *Efficient discovery of error-tolerant frequent itemsets in high dimensions*, in Proceedings of KDD, ACM, 2001, pp. 194–203.