Exploring the Dataset Structure by Means of Delta-Classes of Equivalence. *The Case of the Titanic Dataset*⋆

Aleksey Buzmakov[1], Sergei O. Kuznetsov[1], Tatyana Makhalova[2], and Amedeo Napoli[1,2]

[1] National Research University Higher School of Economics, Russia
[2] LORIA (CNRS – Inria NGE – University of Lorraine), Vandœuvre-lès-Nancy, France
{avbuzmakov,skuznetsov}@hse.ru,
tatiana.makhalova@inria.fr, amedeo.napoli@loria.fr

Abstract. Being able to have a quick look at a dataset is essential in many applications. One way is to summarize the dataset by means of a small set of patterns. In this paper we suggest defining such set of patterns as the closed elements of delta-classes of equivalence. This approach allow us to propose an overview of a dataset and then, if necessary, any delta-class of equivalence can be expanded to provide more detailed information about a certain part of the dataset.
To demonstrate our proposal we deeply studied the Titanic dataset about survival of passengers and showed the connections between the passenger attributes and a possible dataset summary in terms of patterns.

Keywords: FCA · Δ-closure · Δ-concepts · use case · Δ-implications.

1 Introduction

In this paper, we are interested in pattern or itemset mining in tabular data. There is a considerable amount of work on many aspects of this subject, especially regarding algorithms and search for interesting patterns [1]. One recurrent problem in itemset mining is the exponential number of resulting itemsets. Focusing on closed itemsets allows a significant reduction of this number by replacing a whole class of itemsets with the largest one, i.e., the closed itemset, which has the same support [7]. Nowadays, there are very efficient algorithms for computing frequent closed itemsets [6], even for low frequency thresholds. However, the efficient generation of closed itemsets only partially solves the problem of the exponential explosion of itemsets, since the main difficulties appear afterwards, when the generated itemsets are processed.

An alternative to the exhaustive enumeration of itemsets is based on "sampling" [4] and on a gradual search for itemsets according to an interestingness measure or a set of constraints [8]. Such algorithms usually result in a rather

⋆ The reported study was funded by RFBR, project number 20-31-70047

small set of itemsets while they may provide only an approximate solution. Although both approaches use quite different techniques, they rely on the same assumption, namely that the "internal or intrinsic structure" of the dataset under study can be understood by means of subset of selected itemsets.

Then, in each approach a particular set of itemsets is returned, which provides a "multifaceted view" of the intrinsic structure underlying the data.

Our approach is based on Δ-classes of equivalence, the generalization of standard classes of equivalence based on closure operator. A user-set parameter Δ measures how much a closed set can differ from its upper neighbors in the partial order of closed sets.

A Δ-class of equivalence allows one to characterize the distribution underlying the data, i.e., when Δ is large, there are only a few Δ-classes of equivalence whose elements are very stable, while when Δ is small, the number of Δ-classes increases and the related information becomes less stable. Moreover, the Δ-classes of equivalence are very stable for large Δ and do not significantly depend on the data sampling used for the analysis.

In this paper we study Δ-classes for Titanic dataset. In particular, we show what kind of conclusions w.r.t. the passengers of Titanic can be made. Further we study what kind of information can be stored in "implications" of the Δ-classes of equivalence. In particular, such implications can show what information is usually associated with a set of attributes.

The paper has the following structure. First we introduce basic definitions related to generalized closure operator. Then in Section 3 we evaluate this closure operator on Titanic dataset.

2 Δ-classes of equivalence

The proposed approach is introduced in terms of Formal Concept Analysis (FCA) [5] and the following notation. A formal context is a triple (G, M, I), where G and M are sets of objects and attributes correspondingly and $I \subseteq G \times M$ is a relation between them. Derivation operator is denoted with arrows in order to clearly show the range and domain of the corresponding mappings:

$$A^{\uparrow} = \{m \in M \mid (\forall g \in A)(g, m) \in I\}, A \subseteq G \quad (1)$$

$$B^{\downarrow} = \{g \in G \mid (\forall m \in B)(g, m) \in I\}, B \subseteq M \quad (2)$$

We should note that operators $(\cdot)^{\uparrow\downarrow}$ and $(\cdot)^{\downarrow\uparrow}$ form closure operators on 2^G and 2^M, respectively. Hereafter, we focus on the operator $(\cdot)^{\downarrow\uparrow}$ and its generalizations. Let us reformulate the definition in terms of "instance counting". Let B be any set of attributes.

Definition 1. *An attribute set B is closed iff* $(\forall m \in M \setminus B)|(B \cup \{m\})^{\downarrow}| \neq |B^{\downarrow}|$.

It can be seen that $(\cdot)^{\downarrow\uparrow}$ maps any set of attributes to a closed set of attributes. Let us reformulate Definition 1 in the following equivalent way. B is closed iff

$$(\forall m \in M \setminus B)[|B^{\downarrow}| - |(B \cup \{m\})^{\downarrow}| \geq 1].$$

This form allows changing the threshold 1 at the end of the formula to any other positive value. The larger this value, the less attribute sets would pass a test similar to the one from Definition 1. This leads to the definition of Δ-closedness of an attribute set [2].

Definition 2. *A set of attributes B is called Δ-closed if for any $m \in M$:*

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

In [2] it was shown that Δ-closedness is a closure operator. In particular it means that from a computational point of view, given a non Δ-closed set of attributes B, i.e., $\exists m \in M(|B^{\downarrow}| - |(B \cup \{m\})^{\downarrow} < \Delta)$, it can be closed by iteratively changing B to $B \cup \{m\}$, for any m violating (3) until such attribute is not found. The corresponding closure operator is denoted by $(\cdot)^{\Delta}$. Since it is a closure operator, it divides all sets of attributes 2^M into classes of equivalences having the same closure.

Definition 3. *Given an attribute set B, its equivalence class $Equiv_{\Delta}(B)$ is the set of all attribute sets with the closure equal to the closure of B, i.e.,*

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

Moreover, since according to Definitions 1 and 2 if a set of attributes is Δ-closed than it is necessary closed. Thus, these Δ-classes of equivalence are joins of several closure-based classes of equivalence. It allows introducing a new derivation operator related to Δ-closure.

$$A^{\uparrow_{\Delta}} = (A^{\uparrow})^{\Delta}, A \subseteq G \tag{5}$$

$$B^{\downarrow_{\Delta}} = (B)^{\Delta\downarrow}, B \subseteq M \tag{6}$$

The new Δ-derivation operator allows defining Δ-concepts ordered within a lattice in the similar way to the classical formal concepts. Moreover, since any Δ-closed set of attributes is a closed set of attributes, than the set of Δ-concepts are subset of formal concepts and the order of Δ-concepts is a suborder of the corresponding formal lattice.

Finally, we should discuss Δ-implications since they provide a useful tool for finding associations between attribute sets.

Definition 4. *A rule $A \underset{\Delta}{\rightarrow} B$ is called Δ-implication if $B = A^{\Delta} \setminus A$, i.e., A and $A \cup B$ are from the same Δ-class of equivalence.*

Since Δ-closure can change the support of the attribute set a Δ-implication is not necessary an implication. However, the set of Δ-implications is subset of all association rules, thus they are more easy analysable. In particular, given an implication $A \rightarrow B$, the set B is the set of attributes that are associated with A in most samples from the underlying distribution.

In the next section we experimentally show how such lattices of Δ-concepts and the corresponding implications can be used for data analysis. Moreover,

since Δ-concepts can be found in polynomial time[3] [3], such analysis is suitable for processing really big data.

3 Evaluation

3.1 Dataset

In this paper we use Titanic dataset downloaded (train dataset) from Kaggle[4]. This is one of the most known datasets with easily interpretable patterns that do not require deep diving into the domain knowledge. The dataset describes 891 passengers of the last Titanic ship travel. Every passenger is described with name, age, sex, the number of parents and/or children and the number of spouse and/or siblings travelled together with the passenger. The ticket price and the ticket class is also known as well as the survival state of the passenger after the Titanic shipwreck.

All numerical data is divided into 5 percentiles and then *inter-ordinal scaling* is used on top of these percentiles. For example, for the quantity `Age` it is known from the data that $\frac{1}{5}$ of the passengers were below 19 years old, the next $\frac{1}{5}$ between 19 and 25, then between 25 and 32 and then between 32 and 41 and finally the last $\frac{1}{5}$ of the passengers were above 41 years old. Then new binary attributes are formed based on these limits (19, 25, 32,41), these eight attributes are "`Age` ≥ 19," "`Age` ≤ 19". "`Age` ≥ 25," "`Age` ≤ 25". "`Age` ≥ 32," "`Age` ≤ 32". "`Age` ≥ 41," "`Age` ≤ 41".

Additionally, from the "Name" field the social status is extracted, including "Mr", "Mrs", "Master", etc. It makes in total 49 attributes.

3.2 Concept lattice navigation

Even for such relatively simple data the total number of concepts is 9002. It is not hard to build such a lattice. However, analysis of the lattice is quite hard. It is hardly possible to draw the whole lattice and the only way is to navigate it from the top or from the bottom concepts. However, Δ-classes of equivalence give another means for such analysis and navigation.

Let us first increase the Δ threshold for the lattice. If $\Delta = 90$, then the lattice size is only 11 and it can be drawn. It is shown in Figure 1. For every concept the corresponding extent size and Δ-measure are shown. Every attribute is shown outside of the concept with an arrow attached to the concept of the first attribute entry. It can be seen that the lattice involves only 9 attributes out of 49. All other attributes are attached to the `BOTTOM` concept and are not shown. It

[3] Being more precise the enumeration procedure can be set in such a way, that it finishes in input-polynomial time. It is achieved by iteratively increasing the threshold θ if the number of the already found patterns is too large. The result is the set of all patterns with $\Delta \geq \theta$. However, if θ is automatically set to be too high, the procedure still finishes in input-polynomial time but the result set is empty.

[4] https://www.kaggle.com/c/titanic

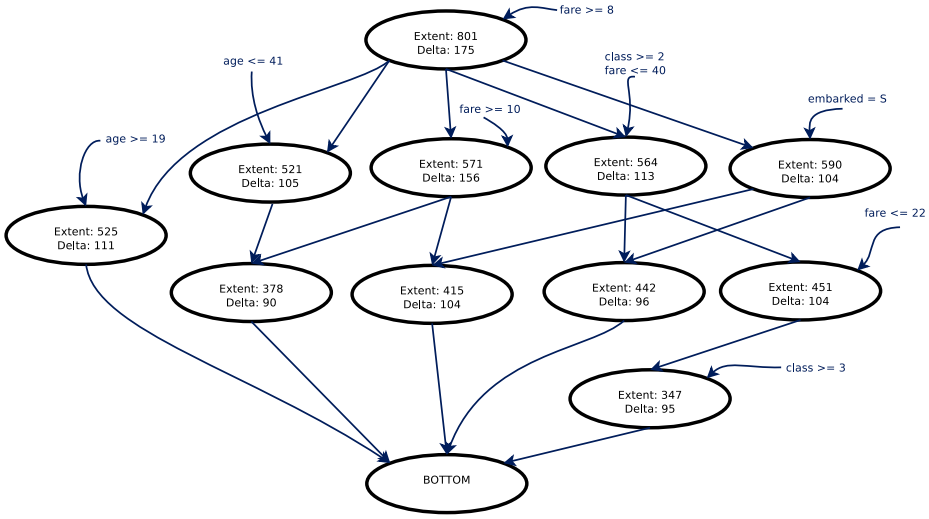

Fig. 1. The lattice of concepts with $\Delta \geq 90$ for Titanic dataset.

means that starting for any attribute a from this set, one has $\Delta(\{a\}^{\downarrow}) < 90$, i.e., smaller than the threshold. For example, the attribute $\texttt{fare} \geq 8$ is introduced to the lattice at the very top concept. It means that $|\emptyset^{\downarrow}| - |\{\texttt{fare} \geq 8\}^{\downarrow}| < 90$, i.e., for most of objects $\texttt{fare} \geq 8$. Moreover, since the top concept in this lattice has the maximal value of Δ, its extent is the most typical extent for all passengers. Since we cannot make it more precise without excluding less objects than 175.

Then we can see that there are 5 concepts below the top concept. They are the only concepts that are significantly different from their children. Accordingly, this concept lattice for $\Delta = 90$ shows the structure of different groups of passengers (formal concepts with Δ-closure for $\Delta = 90$. A concept is only shown if there is no more precise description that contains similar number of objects.

This setting allows finding and prioritizing the association rules related to a certain set of attributes.

3.3 Δ-Association rules

Let us study female subpopulation of the passengers. In particular, we can try to Δ-close the following description $\texttt{sex} = \texttt{female}$. What Δ should be used? One of the reasonable choice is the maximal Δ such that Δ-closure of $\texttt{sex} = \texttt{female}$ is different from M. For example in Figure 1 no concept with $\texttt{sex} = \texttt{female}$ is found. It means that for $\Delta \geq 90$, $\{\texttt{sex} = \texttt{female}\}^{\downarrow\uparrow_\Delta} = M$. If the whole lattice is available it corresponds to the concept with the maximal Δ-measure such that the intent of the concept is a superset of $\{\texttt{sex} = \texttt{female}\}$. In the form

of Δ-implication it can be written as[5]:

$$\{\texttt{sex} = \texttt{female}\} \underset{\Delta}{\rightarrow} \{\texttt{fare} \geq 10\}.$$

It suggests that women rarely buy cheapest tickets. Moreover, since we do not find $\texttt{class} \leq 2$ attribute, it means that they can afford the 3rd class, but nevertheless not the cheapest tickets. What about men?

$$\{\texttt{sex} = \texttt{male}\} \underset{\Delta}{\rightarrow} \{\texttt{title} = \texttt{Mr}, \texttt{fare} \geq 8\}.$$

Now we see that the preference for more expensive tickets is missing. However, we see that most of the men are titled "Mr". It is not the classical closure assuming that all men are `Mr`. Indeed, the correspondence between the title and sex is shown in Table 1.

Table 1. Correspondence between passenger sex and title

	Miss	Mrs	Mr	Master	Other
female	182	128	0	0	4
male	0	0	525	40	12

Let us dive deeper into the title. What do we know about `Master`-title?

$$\{\texttt{title} = \texttt{Master}\} \underset{\Delta}{\rightarrow} \{\texttt{sex} = \texttt{male}, \texttt{age} \leq 19, \texttt{fare} \geq 10, \texttt{class} \geq 2\}.$$

We can see, that `Master` corresponds to young men from 2nd and 3rd class but not with the cheapest tickets. It is a quite strange combination and a deeper investigation is needed. However, it is not an artefact of the procedure. If we check the original dataset all findings are supported, i.e., they are mostly from the 2nd and 3rd class, but not so cheap. The proposed procedure was useful here only for highlighting such finding. In contrast, for `Miss`-title no new information is found.

Let us now formulate a question about women that had cheap tickets:

$$\{\texttt{sex} = \texttt{female}, \texttt{fare} \leq 10\} \underset{\Delta}{\rightarrow} \{\texttt{title} = \texttt{Miss}, \texttt{class} = 3, \texttt{fare} \leq 8\}.$$

In fact if decision is to travel cheap, then it is the cheapest option. Similar, answer will be given for men traveled cheap. However, the group of men that traveled cheep is about 7 times larger than the group of women. Thus, if we would have just requested "who traveled cheap", than the result would be the group of men.

Finally, let us ask how age affects the travel behavior.

$$\{\texttt{age} \leq 25\} \underset{\Delta}{\rightarrow} \{\texttt{class} \geq 2, 8 \leq \texttt{fare} \leq 40\}.$$

[5] For simplicity, the attribute sets are shown with reduction, i.e., if by knowing that $\texttt{fare} \geq 10$ we can conclude that $\texttt{fare} \geq 8$, the last attribute is not shown.

Thus, young people usually travel in the 2nd or 3d class. Similarly, for people aged more than 41 years.

$$\{\texttt{age} \geq 41\} \underset{\Delta}{\rightarrow} \{\texttt{class} \leq 2, \texttt{fare} \geq 22\},$$

i.e., such people usually prefer the 1st or 2nd class and the fare is more than 22.

Let us finally show that we can be interested also in combinations of attributes. For example, what about women of age more than 41 years?

$$\{\texttt{age} \geq 41, \texttt{sex} = \texttt{female}\} \underset{\Delta}{\rightarrow} \{\texttt{class} = 1, \texttt{fare} \geq 40, \texttt{title} = \texttt{Mrs}\}$$

So in contrast to generally aged people, women usually travel in the 1st class and they are titled `Mrs`.

4 Conclusion

Based on Titanic dataset Δ-classes of equivalence are shown to be useful for exploratory data analysis. In particular, it can be used to systematize the dataset and to prioritize association rules related to certain requests, e.g., every attribute can be associated with the most frequent attributes taken into account co-occurrences of the attributes. Since, Δ-classes of equivalence can be found in polynomial time [3], such approach is suitable for very large datasets. Moreover, such approach focuses on the distribution, where the dataset is taken from, rather than the dataset itself.

References

1. Aggarwal, C.C., Han, J.: Frequent pattern mining. Springer (2014)
2. Boley, M., Horváth, T., Wrobel, S.: Efficient discovery of interesting patterns based on strong closedness. Statistical Analysis and Data Mining: The ASA Data Science Journal **2**(5-6), 346–360 (2009)
3. Buzmakov, A., Kuznetsov, S.O., Napoli, A.: Fast generation of best interval patterns for nonmonotonic constraints. In: Proceedings of the Joint European Conference on Machine Learning and Knowledge Discovery in Databases. pp. 157–172. Springer (2015)
4. Dzyuba, V., van Leeuwen, M., De Raedt, L.: Flexible constrained sampling with guarantees for pattern mining. Data Mining and Knowledge Discovery **31**(5), 1266–1293 (2017)
5. Ganter, B., Wille, R.: Formal concept analysis–mathematical foundations (1999)
6. Hu, Q., Imielinski, T.: Alpine: Progressive itemset mining with definite guarantees. In: Proceedings of the SIAM International Conference on Data Mining. pp. 63–71. SIAM (2017)
7. Pasquier, N., Bastide, Y., Taouil, R., Lakhal, L.: Efficient mining of association rules using closed itemset lattices. Information systems **24**(1), 25–46 (1999)
8. Smets, K., Vreeken, J.: Slim: Directly mining descriptive patterns. In: Proceedings of the 12 SIAM International Conference on Data Mining, Anaheim, California. pp. 236–247 (2012)