Association Discovery in Two-View Data

(Extended Abstract)

Matthijs van Leeuwen
LIACS, Leiden University
Email: m.van.leeuwen@liacs.leidenuniv.nl

Esther Galbrun
INRIA Nancy – Grand Est
Email: esther.galbrun@inria.fr

***Abstract*—Two-view datasets are datasets whose attributes are naturally split into two sets, each providing a different view on the same set of objects. We introduce[1] the exploratory data mining task of finding small and non-redundant sets of associations that describe how the two views are related. To achieve this, we propose a novel approach in which sets of rules are used to *translate* one view to the other and vice versa. Our models, dubbed *translation tables*, contain both unidirectional and bidirectional rules that span both views and provide lossless translation from either of the views to the opposite view.**

To be able to evaluate different translation tables and perform model selection, we present a score based on the Minimum Description Length (MDL) principle. Next, we introduce three TRANSLATOR algorithms to find good models according to this score. The first algorithm is parameter-free and iteratively adds the rule that improves compression most. The other two algorithms use heuristics to achieve better trade-offs between runtime and compression. The empirical evaluation on real-world data demonstrates that only modest numbers of associations are needed to characterize the two-view structure present in the data, while the obtained translation rules are easily interpretable and provide relevant insight into the data.

***Keywords*—Association discovery, Two-view data, Minimum description length, Association rule mining, Redescription mining**

I. INTRODUCTION

Two-view datasets are datasets whose attributes are split into two sets, providing two alternative views on the same set of objects. Two-view data is a form of multi-view data, which has an arbitrary number of views. In practice, a data analyst is often given different sets of descriptors on the same set of objects, and asked to analyze associations across these views.

In the medical domain, for example, persons could be the objects of interest, and one could have both demographic and medical data. The two views represent clearly different *types* of information. Alternatively, products could be the objects, and one could have both product information and aggregated customer data (e.g., sales, churn, sentiment). Or consider movies, for which we could have properties like genres and actors on one hand and collectively obtained tags on the other hand. In each of these examples, there are two views that convey different information concerning the same objects. An obvious question to a data analyst would be: *what associations are present in these views?* This is a typical *exploratory data mining* [1] question: the task is to discover patterns that together describe the structure of the data. In particular, we are interested in associations that span both views. For instance, certain demographic properties might imply a certain medical condition with high probability. Sometimes, such an association might hold in both directions, implying that the two observations occur mostly together.

It is important to note that we explicitly aim to find a *compact* and *non-redundant* set of such associations, to avoid overwhelming the analyst with a plethora of discoveries. On the other hand, the set should also be *complete* with respect to the structure in the data it describes. Furthermore, we are primarily interested in scenarios where the two views are expressed over *different, typically disjoint, sets of attributes*, rather than two sets of tuples over the same attributes.

II. RELATED WORK

Numerous association discovery and pattern mining techniques exist, but these were not designed to be used with multi-view data. As a consequence, these methods cannot be directly applied on two-view data, while merging the two views would result in the loss of the distinction between the views. Association rule mining [2] algorithms, for example, can be modified to return only rules that span two views of a dataset, but these methods suffer from the infamous *pattern explosion*: the number of rules found is enormous and it is therefore impracticable for a data analyst to manually inspect and interpret them. Acknowledging this problem, methods have been proposed to discover smaller sets of rules, for example via closed itemsets [3] or statistical testing [4]. Other pattern set mining methods, such as KRIMP [5], also address the pattern explosion, but no existing techniques target the (symmetric) two-view setting that we consider.

Both Exceptional Model Mining (EMM) [6] and Redescription Mining (RM) [7], [8] are concerned with finding patterns in two-view data. However, EMM is highly asymmetric, with one side used for descriptions and the other purely as target. Redescription Mining, on the other hand, aims at finding pairs of queries, one for each view, that are satisfied by almost the same set of objects. RM treats both sides equally, but unlike in our approach, associations are required to hold in both directions and are judged individually, so that the complete set of redescriptions is often redundant in practice.

III. APPROACH AND CONTRIBUTIONS

To provide accurate and complete descriptions of the associative structure across a Boolean two-view dataset, we take an approach that combines a new pattern-based model with model selection based on the Minimum Description Length.

[1] M. van Leeuwen & E. Galbrun. "Association discovery in two-view data," *TKDE*, vol. 27, no. 12, pp. 3190–3202, 2015.

Our first main contribution is the *introduction of pattern-based models for Boolean two-view data*. So-called *translation tables* consist of translation rules and can be used to reconstruct one side of the data given the other, and vice versa.

For this, we consider sets of objects characterized by two Boolean datasets over two disjoint item vocabularies. Without loss of generality, we refer to these as left-hand side and right-hand side datasets and denote them by $\mathcal{D}_\mathbf{L}$ (over $\mathcal{I}_\mathbf{L}$) and $\mathcal{D}_\mathbf{R}$ (over $\mathcal{I}_\mathbf{R}$) respectively. In this context, consider a *rule* $r = X \rightarrow Y$, where X is an itemset over $\mathcal{I}_\mathbf{L}$ and Y is an itemset over $\mathcal{I}_\mathbf{R}$. Such a rule can be interpreted as indicating that if X occurs in a transaction of $\mathcal{D}_\mathbf{L}$, then Y is likely to occur in the corresponding transaction of $\mathcal{D}_\mathbf{R}$. In other words, given the left-hand side of the data, rules provide information about occurrences of items in the right-hand side. Thus, they can be used to *translate* $\mathcal{D}_\mathbf{L}$ to $\mathcal{D}_\mathbf{R}$ and are therefore dubbed *translation rules*. Similarly, we define rules in the other direction, and symmetric rules for which both directions hold.

Further, we introduce a translation scheme, illustrated in Figure 1, that takes a Boolean view and translation table as input, and returns a reconstructed opposite view as output. Each individual rule spans both views of the data and hence provides insight in how the two sides are related. We use both bidirectional and unidirectional rules to allow the construction of succinct models that allow for easy interpretation.

Given a dataset, different translations tables will clearly result in different translations and an important question is *how good* a specific translation table is. In general, some items might be missing from the reconstructed view while some might be introduced erroneously. To make translation lossless, we add a *correction table* that corrects both of these types of errors; the larger the reconstruction error, the larger the number of corrections. Given this, we could try to find the model that minimizes the size of the correction table, but this would result in overly complex translation tables.

For this reason, our second main contribution is to introduce *model selection for translation tables based on the Minimum Description Length (MDL) principle* [9]. The MDL principle takes both the complexity of the model and the complexity of the data given the model into account, and is therefore very useful for model selection when a balance between these complexities is desirable. In the current context, we use it to select small sets of rules that provide accurate descriptions of the associative two-view structure.

Having defined our models and a way to score them, we need to search for the optimal translation table with respect to this score. Unfortunately, exhaustive search for the globally optimal translation table is practically infeasible. Nevertheless, it is possible to search for and find the single rule that gives the largest gain in compression given a dataset and current translation table, allowing us to construct a good translation table in a greedy manner.

Our third main contribution are *three* TRANSLATOR *algorithms*, each of which takes a two-view dataset as input and induces a good translation table by starting from an empty table and iteratively adding rules. By introducing an exact method for finding the best rule in each iteration, we have the best possible baseline to which we can compare the heuristic approaches (on modestly sized problem instances).

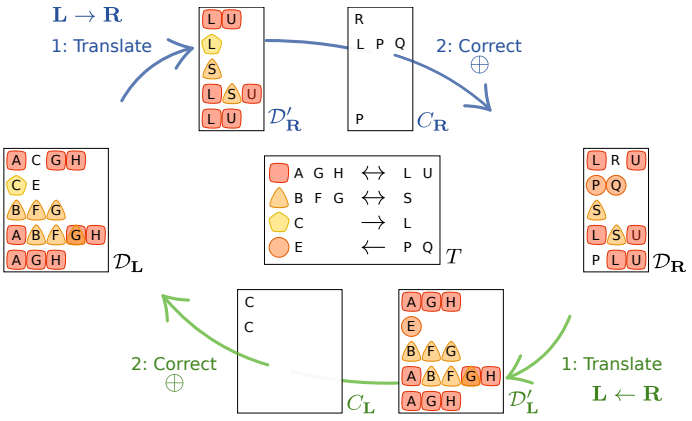

Fig. 1. Translation of a toy dataset, consisting of the two views $\mathcal{D}_\mathbf{L}$ and $\mathcal{D}_\mathbf{R}$, in both directions with translation table T (in the center).

Experimental results indicate that our model and algorithm are able to discover two-view structure in datasets. Both quantitative and qualitative analysis demonstrate that TRANSLATOR discovers more compact and complete models than existing methods mentioned in the previous section. Indeed, in our experiment with a collection of fourteen benchmark datasets, our algorithm obtained compression ratios averaging at 79.73%, while competitor methods REREMI [8], MAGNUM OPUS [4], and modified KRIMP [5] achieved poorer compression ratios, averaging respectively at 97.74%, 123.81% and 300.73%.

IV. CONCLUSIONS

In summary, we introduce the exploratory data mining task of finding small and non-redundant sets of associations that provide insight in how the two sides of two-view datasets are related. To this end, we propose a translation-based approach that uses rules to translate one view into the other and vice versa, introduce a model selection method based on the MDL principle, and present three TRANSLATOR algorithms for inducing high quality sets of patterns, dubbed translation tables. Moreover, the obtained translation rules are easily interpretable and provide relevant insight into the data.

REFERENCES

[1] T. D. Bie and E. Spyropoulou, "A theoretical framework for exploratory data mining: Recent insights and challenges ahead," in *Proc. of ECML/PKDD'13*, 2013, pp. 612–616.

[2] R. Agrawal, T. Imieliński, and A. Swami, "Mining association rules between sets of items in large databases," in *Proc. of SIGMOD'93*, 1993.

[3] M. J. Zaki, "Mining non-redundant association rules," *Data Mining and Knowledge Discovery*, vol. 9, no. 3, pp. 223–248, 2004.

[4] G. I. Webb, "Discovering significant patterns," *Machine Learning*, vol. 68, no. 1, pp. 1–33, 2007.

[5] J. Vreeken, M. van Leeuwen, and A. Siebes, "Krimp: mining itemsets that compress," *Data Mining and Knowledge Discovery*, vol. 23, no. 1, pp. 169–214, 2011.

[6] D. Leman, A. Feelders, and A. Knobbe, "Exceptional model mining," in *Proc. of ECML/PKDD'08*, 2008, pp. 1–16.

[7] L. Parida and N. Ramakrishnan, "Redescription mining: Structure theory and algorithms," in *Proc. of AAAI'05*, 2005, pp. 837–844.

[8] E. Galbrun and P. Miettinen, "From black and white to full color: extending redescription mining outside the boolean world," *Statistical Analysis and Data Mining*, vol. 5, no. 4, pp. 284–303, 2012.

[9] P. D. Grünwald, *The Minimum Description Length Principle*. MIT Press, 2007.