In: Encyclopedia of Data Warehousing and Mining, J. Wang (Ed.),
Idea Group Reference, 2005, pp. 207-211.

Condensed representations for data mining

Jean-François Boulicaut
INSA de Lyon
LIRIS CNRS FRE 2672
Bâtiment Blaise Pascal
F-69621 Villeurbanne cedex
France
voice: +33 4 7243 8905
fax: +33 4 7243 8713
email: jean-francois.boulicaut@insa-lyon.fr

Condensed Representations for Data Mining

Jean-François Boulicaut, INSA de Lyon, France

INTRODUCTION

Condensed representations have been proposed in (Mannila & Toivonen, 1996) as a useful concept for the optimization of typical data mining tasks. It appears as a key concept within the inductive database framework (Imielinski & Mannila, 1996; Boulicaut et al., 1999; de Raedt, 2002) and this paper introduces this research domain, its achievements in the context of frequent itemset mining (FIM) from transactional data and its future trends.

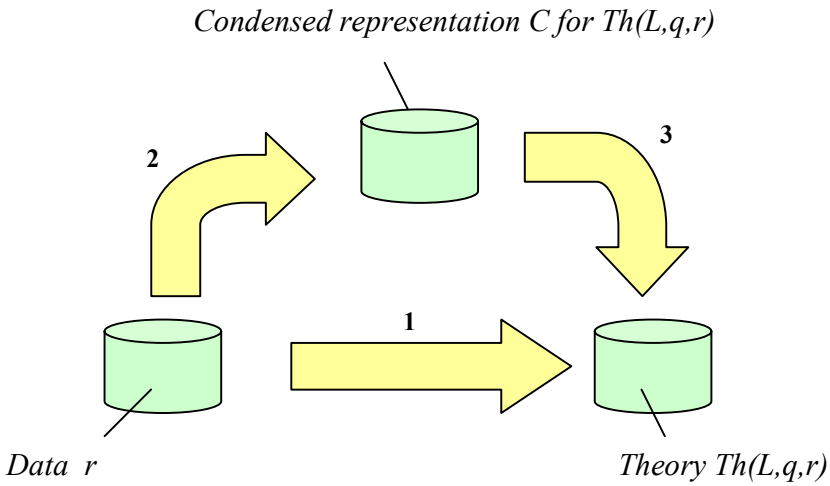
Within the inductive database framework, knowledge discovery processes are considered as querying processes. Inductive databases (IDBs) contain not only data, but also patterns. In an IDB, ordinary queries can be used to access and manipulate data, while inductive queries can be used to generate (mine), manipulate, and apply patterns. To motivate the need for condensed representations, let us start from the simple model proposed in (Mannila & Toivonen, 1997). Many data mining tasks can be abstracted into the computation of a theory. Given a language L of patterns (e.g., itemsets), a database instance r (e.g., a transactional database) and a selection predicate q which specifies whether a given pattern is interesting or not (e.g., the itemset is frequent in r), a data mining task can be formalized as the computation of $Th(L,q,r) = \{\phi \in L \mid q(\phi,r) \text{ is true}\}$. This can be also considered as the evaluation for the inductive query q. Notice that it specifies that every pattern which satisfies q has to be computed. This completeness assumption is quite common for local pattern discovery tasks but is generally not acceptable for more complex tasks (e.g., accuracy optimization for predictive model mining). The selection

predicate q can be defined in terms of a Boolean expression over some primitive constraints (e.g., a minimal frequency constraint used in conjunction with a syntactic constraint which enforces the presence or the absence of some sub-patterns). Some of the primitive constraints generally refer to the "behavior" of a pattern in the data by using the so-called evaluation functions (e.g. frequency).

To support the whole knowledge discovery process, it is important to support the computation of many different but correlated theories.

It is well known that a "generate and test" approach that would enumerate the sentences of L and then test the selection predicate q is generally impossible. A huge effort has been made by data mining researchers to make an active use of the primitive constraints occurring in q to achieve a tractable evaluation of useful mining queries. It is the domain of constraint-based mining, see, e.g., the seminal paper (Ng et al., 1998). In real applications, the computation of Th(L,q,r) can remain extremely expensive or even impossible and the framework of condensed representations has been designed to cope with such a situation. The idea of ε-adequate representations was introduced in (Mannila & Toivonen, 1996; Boulicaut & Bykowski, 2000). Intuitively, they are alternative representations of the data which enable to answer to a class of query (e.g., frequency queries for itemsets in transactional data) with a bounded precision. At a given precision ε, one can be interested in the smaller representations which are then called concise or condensed representations. It means that a condensed representation for Th(L,q,r) is a collection $C \subset Th(L,q,r)$ such that every pattern from Th(L,q,r) can be derived efficiently from C. In the database mining context where r might contain a huge volume of records, we assume that efficiently means without further access to the data. The following figure illustrates that we

can compute Th(L,q,r) either directly (Arrow 1) or by means of a condensed representation (Arrow 2) followed by a regeneration phase (Arrow 3).



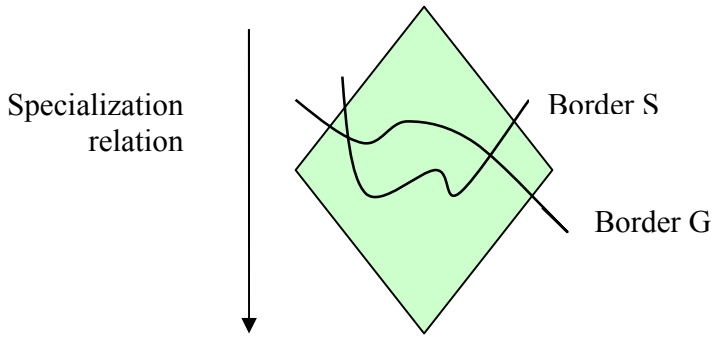
We know several examples of condensed representations for which Phases 2 and 3 are much less expensive than Phase 1. We now introduce the background for understanding condensed representations in the well studied context of FIM.

BACKGROUND

In many cases (e.g., itemsets, inclusion dependencies, sequential patterns) and for a given selection predicate or constraint, the search space L is structured by an anti-monotonic specialization relation which provides a lattice structure. For instance, in transactional data, when L is the power set of items and the selection predicate enforces a minimal frequency, set inclusion is such an anti-monotonic specialization relation. Anti-monotonicity means that when a sentence does not satisfy q (e.g., an itemset is not frequent) then none of its specializations can satisfy q (e.g., none of its supersets are frequent). It becomes possible to prune huge parts of the search space which can not contain interesting sentences. This has been studied a lot within the

« learning as search » framework (Mitchell, 1982) and the generic levelwise algorithm from (Mannila & Toivonen, 1997) has inspired many algorithmic developments. It computes Th(L,q,r) levelwise in the lattice by considering first the most general sentences (e.g., the singleton in the FIM problem). Then, it alternates candidate evaluation (e.g., frequency counting) and candidate generation (e.g., building larger itemsets from discovered frequent itemsets) phases. The algorithm stops when it can not generate new candidates or, in other terms, when the most specific sentences have been found (e.g., all the maximal frequent itemsets). This collection of the most specific sentences is called a positive border in (Mannila & Toivonen, 1997) and it corresponds to the S set of a Version Space in Mitchell's terminology. The Apriori algorithm (Agrawal et al., 1996) is clearly the most famous instance of this levelwise algorithm.

The dual property of monotonicity is interesting as well. A selection predicate is monotonic when its negation is anti-monotonic, i.e., when a sentence satisfies it all its specializations satisfy it as well. In the itemset pattern domain, the maximal frequency constraint or a syntactic constraint which enforces that a given item belongs to the itemsets are two monotonic constraints. Thanks to the duality of these definitions, a monotonic constraint gives rise to a border G which contains the minimally general sentences w.r.t. the monotonic constraint (see the following figure).



When the predicate selection is a conjunction of an anti-monotonic part and a monotonic part, the two borders define the solution set: solutions are between S and G, and (S,G) is a Version Space. For this conjunction case, several algorithms can be used (de Raedt & Kramer, 2001; Jeudy & Boulicaut, 2002; Bucila et al., 2002; Bonchi, 2003). When arbitrary Boolean combinations of anti-monotonic and monotonic constraints are used (e.g., disjunctions), the solution space is defined as a union of several Version Spaces, i.e., unions of couples of borders (de Raedt et al., 2002).

Borders appear as a typical case of condensed representation. Assume that the collection of the maximal frequent itemsets in r is available (i.e., the S border for the minimal frequency constraint), this collection is generally several orders of magnitude smaller than the complete collection of the frequent itemsets in r while all of them can be generated from S without any access to the data. However, in most of the applications of pattern discovery tasks, the user not only wants to get the interesting patterns but also the results of some evaluation functions about these patterns. This is obvious for the FIM problem: these patterns are generally exploited in a post-processing step to derive more useful statements about the data, e.g., the popular frequent association rules which have a high enough confidence (Agrawal et al., 1996). This can be done efficiently if we compute not only the collection of frequent itemsets but also their frequencies.

In fact, the semantics of an inductive query is better captured by extended theories, i.e., collections like $\{(\phi,e) \in L \otimes E \mid q(\phi,r)$ est vrai et $e=\zeta(\phi,r)\}$ where e is the result of an evaluation function ζ in r with values in E. In our FIM problem, ζ denotes the frequency (e is a number in [0,1]) of an itemset in a transactional database r. The challenge of designing condensed representations for an extended theory ThE is then to identify subsets of ThE from which it is possible to generate ThE either exactly or with an approximation on the evaluation functions.

MAIN RESULTS

We emphasize the main results concerning condensed representations for frequent itemsets since this is the context for which it has been studied a lot.

Condensed representations by borders

It makes sense to use borders as condensed representations. For FIM, specific algorithms have been designed for computing directly the S border, e.g., (Bayardo, 1998). Also, the algorithm in (Kramer & de Raedt, 2001) computes borders S and G and has been applied successfully to feature extraction in the domain of molecular fragment finding. In this case, a conjunction of a minimal frequency in one set of molecule (say, the active ones) and a maximal frequency in another set of molecules (say the inactive ones) is used. This kind of research is related to the so-called emergent pattern discovery (Dong & Li, 1999).

Considering the extended theory for frequent itemsets, it is clear that given the maximal frequent sets and their frequencies, we have an approximate condensed representation of the frequent itemsets. Without looking at the data, we can regenerate the whole collection of the frequent itemsets (subsets of the maximal ones) and we have a bounded error on their frequencies: when considering a subset of a maximal σ-frequent itemset, we know that its frequency is in $[\sigma,1]$. Even though more precise bounds can be computed, this approximation is useless in practice. Indeed, when using borders, users have other applications in mind, e.g., feature construction.

The maximal frequent itemsets can be computed in cases where very large frequent itemsets hold such that the regeneration process becomes impossible. Typically, when a maximal frequent itemset has a size 30, it should lead to the regeneration of around 10^{10} frequent sets.

Exact condensed representations of frequent sets and their frequencies

Since (Boulicaut & Bykowski, 2000), a lot of attention has been put on exact condensed representations based on frequent closed sets. According to the classical framework of Galois connection, the closure of an itemset X in r, closure(X,r), is the maximal superset of X which has the same frequency than X in r. Furthermore, a set X is closed if X=closure(X,r). Interestingly, the sets of itemsets which have the same closures constitute equivalence classes of itemsets which have the same frequencies, the maximal one in each equivalence class being a closed set (Bastide et al., 2000).

Frequent closed sets are itemsets which are both closed and frequent. In dense and/or highly correlated data, we can have orders of magnitude less frequent closed sets than frequent itemsets. In other terms, it makes sense to materialize, and, when possible, to look for the fastest computations of the frequent closed sets only. It is then easy to derive the frequencies of every frequent itemset without any access to the data. Many algorithms have been designed for computing frequent closed itemsets, e.g., (Pasquier et al., 1999; Zaki, 2002). Empirical evaluations of many algorithms for the FIM problem are resported in (Goethals & Zaki, 2004). Frequent closed set mining is a real breakthrough the computational complexity of FIM in difficult contexts.

A specificity of frequent closed set mining algorithms is to need for a characterization of (frequent) closed set generators. Interestingly, these generators constitute condensed representations as well. An important characterization is the one of free sets (Boulicaut et al., 2000) which has been proposed independently under the name of key patterns (Bastide et al., 2000). By definition, the closures of (frequent) free sets are (frequent) closed sets. Given the quivalence classes we quoted earlier, free sets are their minimal elements, and the freeness property can lead to efficient pruning thanks to its anti-monotonicity. Computing the frequent free sets plus an extra collection of some non free itemsets (part of the so-called negative border

of the frequent free sets), it is possible to regenerate the whole collection of the frequent itemsets and their frequencies (Boulicaut et al., 2000; Boulicaut et al., 2003). On one hand, we have often much more frequent free sets than frequent closed sets but, on another hand, they are smaller. The concept of freeness has been generalized for other exact condensed representations like the disjunct-free itemsets (Bykowski & Rigotti, 2001), the non derivable itemsets (Calders & Goethals, 2002), and the minimal k-free representations of frequent sets (Calders & Goethals, 2003). Regeneration algorithms and translations between condensed representations have been studied as well, e.g., in (Kryszkiewicz et al., 2004).

Approximate condensed representations for frequent sets and their frequencies

Looking for approximate condensed representations can be useful for very hard data sets. We mentioned that borders can be considered as approximate condensed representations but with a poor approximation of the needed frequencies. The idea is to be able to compute the frequent itemsets with less counting at the price of an acceptable approximation on their frequencies. One idea is to compute the frequent itemsets and their frequencies on a sample of the original data (Mannila & Toivonen, 1996) but bounding the error on the frequencies is hard. In (Boulicaut et al., 2000), the concept of δ-free set was introduced. In fact, the free itemsets are a special case when $\delta = 0$. When $\delta > 0$, we have less δ-free sets than free sets and thus the representation is more condensed. Interestingly, experimentations on real data sets have shown that the error in practice was much smaller than the theoretical bound (Boulicaut et al., 2000; Boulicaut et al., 2003). Another family of approximate condensed representations has been designed in (Pei et al., 2002) where the idea is to consider the maximal frequent itemsets for different frequency thresholds and then approximate the frequency of any frequent set by means of the computed frequencies.

Multiples uses of condensed representations

Another interesting achievement is that condensed representations of frequent itemsets are not only useful for FIM in difficult cases but also to derive more meaningful patterns. In other terms, instead of a regeneration phase which can be however impossible due to the size of the collection of frequent itemsets, it is possible to use directly the condensed representations. Indeed, closed sets can be used to derive informative or non redundant association rules. Also, frequent and valid association rules with a minimal left-hand side can be derived from δ-free sets and these rules can be used, among others, for association-based classification. It is also possible to derive formal concepts in Wille's terminology and thus applying the so-called Formal Concept Analysis.

FUTURE TRENDS

So far, we consider that two promising directions of research are considered and should provide new breakthrough. First, it is useful to consider new mining tasks and design condensed representations for them. For instance, (Casali et al., 2003) considers the computation of aggregates from data cubes, and (Yan et al., 2003) addresses sequential pattern mining. An interesting open problem is to use condensed representations during model mining (e.g., classifiers). Then, merging condensed representations with constraint-based mining (with various constraints, including constraints that are neither anti-monotonic nor monotonic) seems to be a major issue. One challenging problem is to decide which kind of condensed representation has to be materialized for optimizing sequences of inductive queries (e.g., for interactive association rule mining), i.e., the real context of knowledge discovery processes.

CONCLUSION

We introduced the key concept of condensed representations for the optimization of data mining queries and thus the development of the inductive database framework. We have summarized an up-to-date view on borders. Then, referencing the work on the various condensed representations for frequent itemsets, we have pointed out the breakthrough the computational complexity of the FIM tasks. This is important because the applications of frequent itemsets go much further than the classical association rule mining task. Finally, the concepts of condensed representations and ε-adequate representations are quite general and might be considered with success for many other pattern domains.

REFERENCES

Agrawal, R., Mannila, H., Srikant, R., Toivonen, H. & Verkamo, A.I. (1996). Fast discovery of association rules. In *Advances in Knowledge Discovery and Data Mining*: AAAI/MIT Press, 307-328.

Bastide, Y., Taouil, R., Pasquier, N., Stumme, G., & Lakhal, L. (2000). Mining frequent patterns with counting inference. *SIGKDD Explorations*, 2 (2), 66-75.

Bayardo, R. (1998). Efficiently Mining Long Patterns from Databases. ACM SIGMOD International Conference on Management of Data SIGMOD'98. Seattle, USA, 85-93.

Bonchi, F. (2003). Frequent pattern queries: language and optimizations. Ph.D. Thesis University of Pisa, Italy, TD-10/03.

Boulicaut, J-F., Klemettinen, M., & Mannila, H. (1999). Modeling KDD processes within the inductive database framework. *Data Warehousing and Knowledge Discovery DaWaK'99*. Florence, Italy, 293-302.

Boulicaut, J-F. & Bykowski, A. (2000). Frequent closures as a concise representation for binary data mining. *Knowledge Discovery and Data Mining, Current Issues and New Applications PaKDD'00*. Kyoto, Japan, 62-73.

Boulicaut, J-F., Bykowski, A. & Rigotti, C. (2000). Approximation of frequency queries by means of free-sets. *Principles of Data Mining and Knowledge Discovery PKDD'00*. Lyon, France, 75-85.

Boulicaut, J-F., Bykowski, A. & Rigotti, C. (2003). Free-sets: a condensed representation of boolean data for the approximation of frequency queries. *Data Mining and Knowledge Discovery*, 7 (1), 5-22.

Bucila, C., Gehrke, J., Kifer, D., & White, W.M. (2003). DualMiner: A dual-pruning algorithm for itemsets with constraints. *Data Mining and Knowledge Discovery*, 7 (3), 241-272.

Bykowski, A. & Rigotti, C. (2001). A condensed representation to find frequent patterns. *ACM Principles of Database Systems PODS'01*. Santa Barbara, USA, 267-273.

Calders, T. & Goethals, B. (2002). Mining all non-derivable frequent itemsets. *Principles of Data Mining and Knowledge Discovery PKDD'02*. Helsinki, Finland, 74-85.

Calders, T. & Goethals, B. (2003). Minimal k-free representations of frequent sets. *Principles of Data Mining and Knowledge Discovery PKDD'03,* Dubrovnik, Croatia, 71-82.

Casali, A., Cicchetti, R., & Lakhal, L. (2003). Cube Lattices: A Framework for Multidimensional Data Mining. *SIAM International Conference on Data Mining SDM'03*. San Francisco, USA.

de Raedt, L. (2002). A perspective on inductive databases. *SIGKDD Explorations*, 4(2):66-77.

de Raedt, L. & Kramer, S. (2001). The levelwise version space algorithm and its application to molecular fragment finding. *International Joint Conference on Artificial Intelligence IJCAI'01*. Seattle, USA, 853-862.

de Raedt L., Jäger, M., Lee, S.D. & Mannila, H. (2002). A theory of inductive query answering. *IEEE International Conference on Data Mining ICDM'02*. Maebashi City, Japan, 123-130.

Dong, G. & Li., J. (1999). Efficient mining of emerging patterns: discovering trends and differences. *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining SIGKDD'99*, Dan Diego, USA, 43-52.

Goethals, B., & Zaki, M.J. (2004). Advances in frequent itemset mining implementations: report on FIMI'03. *SIGKDD Explorations,* 6 (1), 109-117.

Imielinski, T. & Mannila, H. (1996). A database perspective on knowledge discovery. *Communications of the ACM*, 39 (11), 58-64.

Jeudy, B. & Boulicaut J-F. (2002). Optimization of association rule mining queries. *Intelligent Data Analysis*, 6 (4), 341-357.

Kryszkiewicz, M., Rybinski, H., & Gajek, M. (2004). Dataless transitions between concise representations of frequent patterns. *Intelligent Information Systems,* 22 (1), 41-70.

Mannila, H. & Toivonen, H. (1996). Multiple uses of frequent sets and condensed representations. *International Conference on Knowledge Discovery and Data Mining KDD'96.* Portland, USA, 189-194.

Mannila, H. & Toivonen, T. (1997). Levelwise search and borders of theories in knowledge discovery. *Data Mining and Knowledge Discovery*, 1 (3), 241-258.

Mitchell, T.M. (1982). Generalization as search. *Artificial Intelligence*, 18, 203-226.

Ng, R., Lakshmanan, L.V.S., Han, J. & Pang, A. (1998). Exploratory mining and pruning optimizations of constrained associations rules. *ACM SIGMOD International Conference on Management of Data SIGMOD'98*. Seattle, USA, 13-24.

Pasquier, N., Bastide, Y., Taouil, R., & Lakhal, L. (1999). Efficient mining of association rules using closed itemset lattices. *Information Systems* 24 (1), 25-46.

Pei, J., Dong, G., Zou, W. & Han, J. (2002). On computing condensed frequent pattern bases. *IEEE International Conference on Data Mining ICDM'02*. Maebashi City, Japan, 378-385.

Yan, X., Han, J., & Afshar, R. (2003). CloSpan: Mining Closed Sequential Patterns in Large Databases. *SIAM International Conference on Data Mining SDM'03*. San Francisco, USA.

Zaki, M.J., & Hsiao, C.J. (2002). CHARM: an efficient algorithm for closed itemset mining. *SIAM International Conference on Data Mining SDM'02*. Arlington, USA, 33-43.

TERMS AND DEFINITIONS

Condensed representations: Alternative representations of the data which preserve crucial information for being able to answer some kind of queries. The most studied example concerns frequent sets their frequencies. Their condensed representations can be several orders of magnitude smaller than the collection of the frequent itemsets.

Constraint-based data mining: It concerns the active use of constraints which specify the interestingness of patterns. Technically, it needs for strategies to "push" the constraints, or at least part of them, deeply into the data mining algorithms.

Inductive databases: An emerging research domain where knowledge discovery processes are considered as querying processes. Inductive databases contain both data and patterns or models which hold in the data. They are queried by means of more or less ad-hoc query languages.

Pattern domains: A pattern domain is the definition of a language of patterns, a collection of

evaluation functions which provide properties of patterns in database instances, and the kinds of constraints which can be used to specify pattern interestingness.