Classification of Datasets with Frequent Itemsets is Wild

Natalia Vanetik
Department of Software Engineering, Sami Shamoon College of Engineering, Beer Sheva, Israel

Keywords: Frequent Itemsets, Dataset Classification.

Abstract: The problem of dataset classification with frequent itemsets is defined as the problem of determining whether or not two different datasets have the same frequent itemsets without computing these itemsets explicitly. The reasoning behind this approach is high computational cost of computing frequent itemsets. Finding well-defined and understandable normal forms for this classification task would be a breakthrough in dataset classification field. The paper proves that classification of datasets with frequent itemsets is a hopeless task since canonical forms do not exist for this problem.

1 INTRODUCTION

Suppose that we are given a dataset that consists of transactions (tuples) each containing one or more items. Frequent itemsets are subsets that appear in a large fraction of dataset tuples, where the exact fraction value is defined by the user and is called *support*. Frequent pattern mining was proposed by Agrawal (Agrawal,Srikant 1994) for shopping basket analysis; both frequent itemsets and association rules were introduced in this paper. Many additional algorithms have been suggested other the years, such as FPGrowth (Han, Pei, Yin 2000), Eclat (Zaki 2000), Genmax (Gouda,Zaki 2005) and many others. This problem has numerous applications in both theoretical and practical knowledge discovery, but its computational complexity is another matter. It has been shown that generating and counting frequent itemsets is #P-complete (see (Yang 2004)).

We focus here on using frequent itemsets in datasets as the means for dataset characterization. Frequent itemsets are important dataset property and they have been used as a classifying feature in virus signature detection (see (Ye et al. 2007)), text categorization (see (Zaïane, Antonie 2002)) and biological data mining (see (Zaki et al. 2010)). A through experimental study of the issue can be found in (Flouvat, De Marchi, Petit 2010). The classifying feature problem is also computationally difficult as it requires computing frequent itemsets for each dataset. In (Palmerini, Orlando, Perego 2004), the authors proposed a statistical property of transactional datasets to characterize dataset density. Paper (Parthasarathy, Ogihara 2000) proposes a similarity measure for homogeneous datasets that is based on frequent patterns appearing in these datasets; this measure is then used to enable dataset clustering. The *diff* operator, proposed in (Subramonian 1998), is another correlation indicator between datasets that captures user beliefs in terms of events and conditional probabilities.

This paper addresses the issue of using frequent itemsets as a classifying feature of datasets. The answer is given in category-theoretic form; we prove that the task of finding well-structured normal forms for frequent itemsets in this case is a hopeless one.

2 PROBLEM STATEMENT

Let D be a dataset composed of transactions $\{t_1,...,t_m\}$, where m is the *dataset size*. Each transaction contains items from a finite set V. The size of V, denoted by n, is called the *cardinality* of the dataset. The number of items may vary from transaction to transaction, but the items in each transaction form a set, i.e. they cannot appear more than once. Additionally, we have a support value $1 \leq S \leq m$. A set I of items (*itemset*) is called *frequent* if it appears in at least S transactions as a subset.

The three main approaches to frequent set generation are Apriori (Agrawal,Srikant 1994), FPGrowth (Han, Pei, Yin 2000) and Eclat (Zaki 2000). This task is computationally expensive as the number of frequent itemsets in D can be exponential in $|V|$. We are asking the following question:

- Is it possible to classify datasets according to their frequent itemsets without explicit computation?

Vanetik N..
Classification of Datasets with Frequent Itemsets is Wild.
DOI: 10.5220/0004167903860389
In *Proceedings of the International Conference on Knowledge Discovery and Information Retrieval* (KDIR-2012), pages 386-389
ISBN: 978-989-8565-29-7

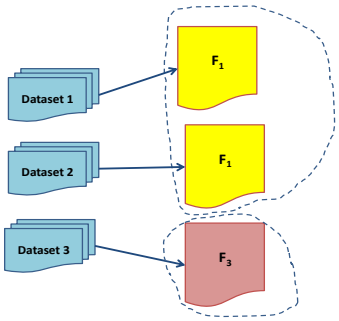

Figure 1: Dataset equivalence as frequent itemset equality.

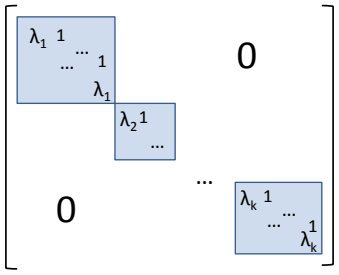

Figure 2: Jordan normal form of a square complex matrix.

3 CLASSIFICATION OF DATASETS

3.1 Canonical Forms

Suppose now that datasets D_1 and D_2 are defined on set V on items, and we are given support constraints S_1 and S_2. We wish to check whether or not datasets D_1 and D_2 are *equivalent* up to their frequent itemsets. In other words, we ask if frequent itemsets F_1 and F_2 of D_1 and D_2 are identical. Figure 1 shows an example of such classification.

There are three main ways of answering this question, some computationally harder than others.

1. **(approach 1)** *Compute both frequent itemset sets and find their symmetric difference.*
   This type of answer is usually the hardest one to compute.
2. **(approach 2)** *Find an algorithm comparing frequent itemset sets without computing them and provide a "yes" or "no" answer.*
   This type of answer can sometimes be easier to achieve.
3. **(approach 3)** *Find canonical representations for both datasets, i.e. find mathematical objects whose equality implies dataset equivalence and inequality implies lack of equivalence.*
   This type of answer tells us everything there is to know about dataset equivalence. Once canonical representations are found, their comparison should be easy and straightforward.

In case of dataset equivalence approach 1 is very hard to implement, since enumeration of frequent itemsets in a dataset is a #P-complete problem by reduction from the problem of determining the number of maximal bipartite cliques in a bipartite graph. This fact is proved in the next claim. Approach 2 to dataset equivalence can provide the answer faster than frequent itemset set computation, since the difference between frequent sets may be determined at early stages of frequent itemset enumeration, thus saving a great deal of time and effort. But do we have a hope of finding a classification for sets of frequent itemsets in various datasets using approach 3? The most prominent advantage of this method is that such a representation needs to be computed only once for each dataset, thus eliminating the need is pairwise dataset comparison. However, the answer to this question is **no** and the explanation follows.

The *problem of classification* of objects in a small category is defined as the problem of finding canonical forms for objects in that category. The issue of deciding on equivalence of two objects up to a certain set of transformation in this case reduces to the problem of comparing their canonical forms up to equality. In our case the category of datasets consists of **finite datasets identical up to row permutations**.

For some problems, such as matrix similarity, well-structured canonical forms exist.

Example 1. *Let A,B be $n \times n$ matrices over an algebraically closed field (finite or infinite). The matrices are similar if there exists a matrix S such that $A = S^{-1}BS$. Three approaches to matrix equivalence discovery in case of matrix similarity can be reformulated as follows.*
(1) Test all possible matrices S until the one that ensures $A = S^{-1}BS$ is found. This approach is infeasible over infinite fields and #P-hard over finite fields.
(2) Find an algorithm that attempts to construct the matrix S from A,B. Such algorithms exist, and they are quire efficient (see e.g. (Giesbrecht 1995)).
(3) Find canonical representations for both matrices. Canonical representations of square matrices over fields for similarity exist and are called Jordan normal forms of matrices. There representations are also square matrices over the same field that have special structure. Figure 2 shows Jordan normal form for matrices over $\mathbb{C}$.

Existence of canonical forms is not limited to matrix problems. For instance, the fundamental theorem of finite abelian groups describes canonical forms that do not have matrix form.

3.2 Matrix Problems

Let us, following (Belitskii, Sergeichuk 2003), define a *matrix problem* as a pair $\mathcal{A} = \{\mathcal{A}_1, \mathcal{A}_2\}$ where $\mathcal{A}_1$ is a set of a-tuples $(A_1, ..., A_a)$ of $m \times n$ matrices over an algebraically closed field and $\mathcal{A}_2$ are operations on tuples from $\mathcal{A}_1$. Given two matrix problems $\mathcal{A} = \{\mathcal{A}_1, \mathcal{A}_2\}$ and $\mathcal{B} = \{\mathcal{B}_1, \mathcal{B}_2\}$, $\mathcal{A}$ is said to be *contained* in $\mathcal{B}$ if there exists a b-tuple $\mathcal{T}(x) = \mathcal{T}(x_1, ..., x_a)$ of matrices, whose entries are non-commutative polynomials in $x_1, ..., x_a$, such that

1. $\mathcal{T}(A) = \mathcal{T}(A_1, ..., A_a) \in \mathcal{B}_1$ if $A = (A_1, ..., A_a) \in \mathcal{A}_1$;
2. for every $A, A' \in \mathcal{A}_1$, A reduces to A' by transformations $\mathcal{A}_2$ if and only if $\mathcal{T}(A)$ reduces to $\mathcal{T}(A')$ by transformations $\mathcal{B}_2$.

Example 2. *For the problem of matrix similarity, $\mathcal{A}_1$ contains square matrices over $\mathbb{C}$ and $\mathcal{A}_2$ contains an operation of multiplication of a matrix by an invertible matrix S and its inverse on left and right.*

The problem of ***simultaneous similarity*** for a pair of $n \times n$ matrices over a field (i.e. pairs (A, B) and (C, D) is defined as follows. Pairs (A, B) and (C, D) are simultaneously similar if and only if there exists S such that $C = S^{-1}AS$ and $D = S^{-1}BS$). In this case, canonical forms do not exist (see (Drozd 1980) and (Gabriel 1972) for detailed explanation).

Problems containing the matrix problem of simultaneous similarity for pairs of matrices are called **wild** as opposed to **tame** problems for which canonical forms exist. In other words, instances of some problems cannot be packed into convenient and self-explainable classes.

Example 3. *The problem of classifying finite groups (even if they are 2-nilpotent) is wild since it contains the problem of classifying pairs of matrices up to simultaneous similarity (see (Sergeichuk 1975)).*

3.3 Dataset Classification

One surprising corollary from the work (Sergeichuk 1975) is the following.

Corollary 4 ((Vanetik, Lipyanski 2010))**.** *The problem of classifying graphs up to isomorphism is wild, i.e. well-defined canonical forms for graphs up to isomorphism do not exist.*

Corollary 4 implies that while the hope of finding an efficient algorithm that distinguishes graphs up to isomorphism still exists, such an algorithm cannot take advantage on some special canonical form of graphs, since these canonical forms do not exist. Unfortunately, this is also the case for frequent itemset sets, as the following shows.

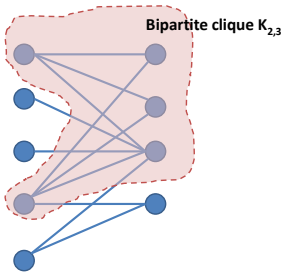

Figure 3: Maximal bipartite clique in a bipartite graph.

Claim 5. *Frequent itemset equivalence is as hard as the graph isomorphism problem.*

Proof. Equivalence of datasets up to frequent itemsets is at least as hard as comparing the number of maximal bipartite cliques in a bipartite graph (see (Zaki 2000) for proof), even if we are talking about closed or maximal itemsets only. A maximal bipartite clique in bipartite graph G is a maximal complete bipartite subgraph $K_{i,j}$ of G (see Figure 3 for an illustration). Finding the size of maximal bipartite clique in a bipartite graph is an NP-complete problem (see (Kuznetsov 1989)) and therefore there exists a polynomial-time reduction from the subgraph isomorphism problem to the maximal bipartite clique problem. Reduction from the graph isomorphism problem to subgraph isomorphism is straightforward. □

Theorem 6. *The problem of classifying datasets up to frequent itemsets is wild.*

Proof. Wildness of the classifying problem for graphs follows from the famous result by V. Sergeichuck (see (Sergeichuk 1975)) and has been proved in Corollary 4. Since the problem of classifying datasets up to their frequent itemsets contains (in the matrix sense) the problem of classifying graphs up to isomorphism, by Claim 5 the classifying problem for datasets up to frequent itemsets is wild. □

Corollary 7. *Dataset classification problem is wild for maximal or closed frequent itemsets as well.*

Note that wildness of a problem does not necessarily imply nonexistence of an efficient algorithm telling whether or not two objects are equivalent up to a certain set of transformations. It does, however, imply that no well-defined representation (representation with normal forms) for equivalence classes of these objects exists.

4 CONCLUSIONS

This paper addresses the problem of classifying

datasets with (maximal, closed, all) frequent itemsets. We show that high computational cost is not the only problem of this approach and that there is in fact a deeper reason to why the approach fails. We use category-theoretic results to prove that well-described normal forms for the dataset classification problem do not exist.

REFERENCES

Agrawal, R., Srikant, R. (1994). Fast algorithms for mining association rules in large databases. Proceedings of the 20th International Conference on Very Large Data Bases, VLDB, pages 487–499, Santiago, Chile.

Belitskii, G., Sergeichuk, V. (2003). Complexity of matrix problems, Linear Algebra Appl. 361, pp. 203-222.

Calders, T. (2004). Computational complexity of itemset frequency satisfiability. In: Proc. 23rd ACM PODS 04, pp. 143–154, ACM Press.

Drozd, J. (1980). Tame and wild matrix problems. Lecture Notes in Mathematics, Volume 832, 242–258.

Flouvat, F., De Marchi, F., Petit, JM. (2010), A new classification of datasets for frequent itemsets. J. Intell. Inf. Syst. 34, pp. 1–19.

Friedland, S. (1983). Simultaneous similarity of matrices, Adv. Math. 50 pp. 189–265.

Gabriel, P. (1972). Unzerlegbare Darstellungen I, Manuscripta Math. 6 pp. 71–103.

Giesbrecht, M. (1995). Nearly Optimal Algorithms For Canonical Matrix Forms, SIAM Journal on Computing, v.24 n.5, pp.948–969.

Gouda, K., Zaki, J. M. (2005). GenMax: An Efficient Algorithm for Mining Maximal Frequent Itemsets. Data Mining and Knowledge Discovery: An International Journal, 11(3) pp.223–242.

Han, J., Pei, J., Yin, Y. (2000). Mining frequent patterns without candidate generation. In: Proceeding of the 2000 ACM-SIGMOD international conference on management of data (SIGMOD00), Dallas, TX, pp 1–12.

Kuznetsov, S. O. (1989). Interpretation on Graphs and Complexity Characteristics of a Search for Specific Patterns, Nauchn. Tekh. Inf., Ser. 2 (Automatic Document Math Linguist), vol. 23, no. 1, pp. 23–37.

Palmerini, P., Orlando, S., Perego, R. (2004). Statistical properties of transactional databases. In H. Haddad, A. Omicini, R. L. Wainwright, L. M. Liebrock (Eds.), SAC (pp. 515–519). New York: ACM.

Parthasarathy S., Ogihara, M. (2000). Clustering Distributed Homogeneous Datasets. in Proceedings PKDD 2000, pp. 566–574

Sergeichuk, V. (1977), The classification of metabelian p-groups (Russian), Matrix problems, Akad. Nauk Ukrain. SSR Inst. Mat., Kiev, pp. 150–161.

Subramonian, R. (1998). Defining diff as a Data Mining Primitive. in Proceedings of KDD1998, pp. 334–338

Lipyanski, R., Vanetik N.(2010). The classification problem for graphs and lattices is wild, in Proceeding of the International Conference on Modern Algebra and its Applications, pp. 107-111, Batumi, September 20th–26th.

Yang, G. (2004). The complexity of mining maximal frequent itemsets and maximal frequent patterns. Proc. Int. Conf. Knowl. Discov. Data Mining, pp. 344–353.

Ye, Y., Wang, D., Li, T., Ye, D. (2007). IMDS: intelligent malware detection system, Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1043–1047.

Zaïane, O. R., Antonie, M. L. (2002). Classifying Text Documents by Associating Terms With Text Categories, Australasian Database Conference.

Zaki, M. J. (2000) Scalable algorithms for association mining, IEEE TransKnowl Data Eng 12:372–390.

Zaki, M. J., Carothers, C. D., Szymanski B. K. (2010). VOGUE: A Variable Order Hidden Markov Model with Duration based on Frequent Sequence Mining. ACM Transactions on Knowledge Discovery in Data, 4(1).