Representative Itemset Mining

Hong Huang and Barry O'Sullivan
Insight Centre for Data Analytics, University College Cork, Ireland

Centre for Data Analytics

Insight

Abstract

Frequent itemset mining is one of the most common of data mining tasks. In its simplest form one is given a table of data in which the columns represent attributes and each row specifies a value for each attribute, each attribute-value pair being referred to as an item. The task is to find sets of these items that occur frequently in the data, where frequency is specified as a minimum occurrence threshold. Such frequent sets of items are referred to as "frequent itemsets". Many efficient techniques have been developed for finding all frequent itemsets. However, a practical problem is that the results sets can be exponentially large in the number of items. We propose representative frequent itemset mining in which the set of itemsets returned provide examples of the space of all possible frequent itemsets. One can see the representative frequent itemset framework as a generalisation of traditional frequent itemset mining that provides an additional parameter for controlling the size of the result set. Specifically, one has access to the traditional frequency threshold, but also the maximum arity of the tuples of itemsets being exemplified.

Background

Frequent itemset mining is one of the most ubiquitous problems in data mining [1, 2]. The problem has its origins in the analysis of transaction data in supermarkets. Each transaction is defined by a set of items, e.g. a transaction might define a set of products that were purchased together. The classic itemset mining problem involves identifying sets of items that appear frequently with each other. Many algorithms have been developed for efficiently enumerating the set of all itemsets that occur more frequently than a specified number of times, e.g. LCM and ECLAT[4, 5, 6].

A standard approach to reducing the number of frequent itemsets to be computed is to consider only those that are maximal. A maximal frequent itemset is one such that no extension of it is also frequent. By focusing on maximal itemsets we can avoid computing all subsets of a frequent itemsets which, by definition, will also be frequent. However, we will see in the experimental results section of this paper that the number of maximal frequent itemsets for a given transaction database can also be impractically large.

Representative Itemsets

We propose an alternative approach to computing a compact but *representative* subset of the maximal frequent itemsets from a transaction database. Specifically, we propose that one considers representative frequent itemset mining in which a very significantly reduced set of itemsets is computed that provides examples of the space of all possible frequent itemsets. Our approach is inspired by the notion of representativeness of explanations [3]. Specifically, every item that appears in a frequent itemset at least once is shown in at least one representative itemset. If there are frequent itemsets without a particular item, one such example will be presented. Our representative approach gives a sense of how items occur in frequent itemsets.

Experiments

The objective of our experiments was to demonstrate the following two features of our representative itemset mining approach. Firstly, we sought to evaluate the size of the representative itemsets one can find in practice. Secondly, we investigated whether one can benefit from the significantly smaller size of a representative set of itemsets from a computational perspective. Our experiments show that the size of a representative itemset can be five orders of magnitude smaller the the number of maximal frequent itemsets, and that they can be computed significantly faster than a state-of-the-art itemset miner.

We based our experiments on a set of benchmarks taken from the CP4IM project database. Our results are presented in Table 1. We list the datasets studied, specifying in each case the number of transactions (#trans) and number of items (#items) for each. Using a 10% frequency threshold (we list the corresponding frequency used in each case as 'thres') we compute the full set of maximal itemsets (#maximal) using the state-of-the-art itemset miner, LCM [4, 5]. Table 1 clearly shows that the size of the set of maximal frequent itemsets can be extremely large, e.g. in excess of 2.5 million in the case of the `australian-credit` dataset.

Results

Table: Summary of our experiments. Representative itemsets are significantly smaller than the set of maximal frequent itemsets - in fact they are usually significantly smaller than the number of items in the dataset, consistent with the theory. They can also be found consistently faster using our proposed algorithm than the alternative generate-and-filter approach. All times reported are in seconds.

Benchmark Details					Minimum	Generate & Filter		RepresentativeItemset	
Dataset	#trans	#items	thres	#maximal	Rep	size	time	size	time
zoo-1	101	36	10	230	5	**5**	0.0092	6	**0.0004**
vote	435	48	44	2,636	7	**8**	0.0419	9	**0.0026**
tic-tac-toe	958	27	96	165	16	18	0.0073	**16**	**0.0047**
splice-1	3.190	287	319	988	214	221	**0.9850**	**215**	1.3887
soybean	630	50	63	331	18	**20**	0.0173	21	**0.0072**
primary-tumor	336	31	34	2,043	7	9	0.0485	**8**	**0.0015**
mushroom	8,124	119	812	453	19	24	**0.0765**	**21**	0.1564
lymph	148	68	15	5,191	14	18	0.2127	**16**	**0.0023**
kr-vs-kp	3,196	73	320	1,984,963	–	17	158.8402	**17**	**0.0345**
hypothyroid	3,247	88	325	2,925,833	–	16	413.4975	**13**	**0.0353**
hepatitis	137	68	14	189,205	–	18	6.7921	**16**	**0.0023**
heart-cleveland	296	95	30	1,647,364	–	22	88.2854	**18**	**0.0058**
german-credit	1,000	112	100	232,107	–	38	20.4248	**35**	**0.0343**
australian-credit	653	125	65	2,580,684	–	23	213.6676	**18**	**0.0146**
anneal	812	93	81	15,977	–	12	2.5915	**9**	**0.1079**

Future Work

We will explore more general notions of representatives, e.g. computing representative k-tuples of itemsets.

References

- C. C. Aggarwal and J. Han, Eds., *Frequent Pattern Mining*. Springer, 2014. [Online]. Available: http://dx.doi.org/10. 1007/978-3-319-07821-2
- Guns, T.; Nijssen, S.; and Raedt, L. D. 2011. Itemset mining: A constraint programming perspective. *Artif. Intell.* 175(12-13):1951-1983.
- O'Sullivan, B.; Papadopoulos, A.; Faltings, B.; and Pu, P. 2007. Representative explanations for over-constrained problems. In *Proceedings of the Twenty-Second AAAI Conference on Artificial Intelligence, July 22-26, 2007, Vancouver, British Columbia, Canada*, 323-328.
- Uno, T.; Kiyomi, M.; and Arimura, H. 2004. LCM ver. 2: Efficient mining algorithms for frequent/closed/maximal itemsets. *In FIMI '04, Proceedings of the IEEE ICDM Workshop on Frequent Itemset Mining Implementations, Brighton, UK, November 1, 2004.*
- Uno, T.; Kiyomi, M.; and Arimura, H. 2005. LCM ver. 3: Collaboration of Array, Bitmap and Prefix Tree for Frequent Itemset Mining. *Proc. 1st Open Source Data Mining Workshop on Frequent Pattern Mining Implementations (OSDM 2005, Chicago, IL), 77-86.*
- Zaki, M.J.; Parthasarathy, S.; Ogihara, M.; and Li, W. 1997. New Algorithms for Fast Discovery of Association Rules *Proc. 3rd ACM SIGKDD Int. Conf. on Knowledge Discovery and Data Mining (KDD'97, Newport Beach, CA)*, 283-296.

The Insight Centre for Data Analytics is supported by Science Foundation Ireland under Grant Number SFI/12/RC/2289

DCU · NUI Galway OÉ Gaillimh · UCC University College Cork, Ireland Coláiste na hOllscoile Corcaigh · UCD Dublin · sfi Science Foundation Ireland