Data Summarization with Informative Itemsets

Michael Mampaey* Nikolaj Tatti ** Jilles Vreeken **

Universiteit Antwerpen, Belgium

Abstract

Data analysis is an inherently iterative process. That is, what we know about the data greatly determines our expectations, and hence, what result we would find the most interesting. With this in mind, we introduce a well-founded approach for succinctly summarizing data with a collection of informative itemsets; using a probabilistic maximum entropy model, we iteratively find the most interesting itemset, and in turn update our model of the data accordingly. As we only include itemsets that are surprising with regard to the current model, the summary is guaranteed to be both descriptive and non-redundant. The algorithm that we present can either mine the top-k most interesting itemsets, or use the Bayesian Information Criterion to automatically identify the model containing only the itemsets most important for describing the data. Or, in other words, it will 'tell you what you need to know'. Experiments on synthetic and benchmark data show that the discovered summaries are succinct, and correctly identify the key patterns in the data. The models they form attain high likelihoods, and inspection shows that they summarize the data well with increasingly specific, yet non-redundant itemsets.

1 Informative and Succinct Summarization with Itemsets

Knowledge discovery from data is an inherently iterative process. What we already know about the data greatly determines our expectations, and therefore, which results we would find interesting and/or surprising. Early on in the process of analyzing a database, for instance, we are happy to learn about the generalities underlying the data, while later on we will be more interested in the specifics that build upon these concepts. Essentially, this process comes down to summarization: we want to know what is interesting in the data, and we want this to be reported succinctly and without redundancy.

As natural as it may seem to update a knowledge model during the discovery process, few pattern mining techniques actually follow such a dynamic approach of discovering patterns that are surprising with regard to what we have learned so far. That is, while many techniques provide a series of patterns in order of interestingness, most score these patterns using a static model; during this process the model, and hence the itemset scores, are not updated with the knowledge gained from previously discovered patterns. The static approach gives rise to the typical problem of traditional pattern mining: overwhelmingly large and highly redundant collections of patterns.

Our goal therefore is to discover the set of itemsets that provides the most important information about the data, while containing as little redundancy as possible. To model the data, we construct a maximum entropy distribution that allows us to directly calculate the expected frequencies of itemsets. Then, at each iteration, we return the itemset that provides the most information, i.e., for which our frequency estimate according to the model was most off. We update our model with this new knowledge, and continue the process. The non-redundant model that contains the most important information is thus automatically identified. Therefore, we paraphrase our method as 'tell me what I need to know'.

*Michael Mampaey is supported by the Agency for Innovation by Science and Technology in Flanders (IWT).

**Nikolaj Tatti and Jilles Vreeken are supported by Post-Doctoral Fellowships of the Research Foundation—Flanders (FWO).

2 Identifying the Best Summary

Our objective is to find a succinct summary of a binary dataset, that is, to obtain a small, yet high-quality set of itemsets $\mathcal{C}$, that describes key characteristics of the data at hand, D, in order to gain useful insights.

To model the data, we use the powerful and versatile class of maximum entropy models. This is a class of probabilistic models that are identified by the Maximum Entropy principle [1] as those models that make optimal use of the provided information. That is, they rely only on this information and are fully unbiased otherwise. For a collection of itemsets $\mathcal{C}$, we construct the distribution $p^*_{\mathcal{C}}$ which satisfies all the frequency constraints imposed by $\mathcal{C}$, and maximizes the entropy $H(p^*_{\mathcal{C}})$. It can be shown that $p^*_{\mathcal{C}}$ has a log-linear form; the parameters of this distribution can be found using the Iterative Scaling procedure [2]. While solving and querying the maximum entropy model is infeasible in general, we show that in our setting this can be accomplished efficiently, depending on the amount of overlap between the selected patterns. This method groups transactions into blocks, according to an equivalence relation induced by $\mathcal{C}$, and employs the Inclusion-Exclusion principle to efficiently calculate probabilities.

To evaluate the quality of a collection of itemsets as a summary for a dataset, we use the Bayesian Information Criterion (BIC) [4], which favors models that fit the data well with few parameters, and is defined as $s(\mathcal{C}) = -\log p^*_{\mathcal{C}}(D) + 1/2\,|\mathcal{C}|\log|D|$. The smaller this score, the better the model. The first term is simply the negative log-likelihood of the model, while the second term is a penalty on the number of parameters—the number of itemsets in our case. Consequently, the best model is identified as the model that provides a good balance between high likelihood and low complexity. Moreover, we automatically avoid redundancy, since models with redundant itemsets are penalized for being too complex, without sufficiently improving the likelihood.

3 Efficiently Mining the Summary

To mine the summary from the data, we present the MTV algorithm, which mines succinct summaries with Maximally informaTiVe itemsets. The algorithm constructs a summary $\mathcal{C}$ by iteratively adding the itemset which provides the most information, i.e., which decreases the quality score function $s(\mathcal{C})$ the most. The model is then updated to incorporate this new knowledge. We show that this is equivalent to adding the itemset which maximizes the Kullback-Leibler divergence between the two consecutive maximum entropy distributions, and present a computationally efficient heuristic which approximates the Kullback-Leibler divergence, and which expresses the divergence between the itemset's estimated and observed frequency. Since this heuristic is convex, mining the itemset which optimizes it can easily be achieved using the branch and bound technique as proposed by Nijssen et al. [3]. This approach allows us to mine our collection of itemsets on the fly, rather than taking a two-phase approach where we would pick them from a larger candidate set which would have to be mined and stored beforehand. The user can also easily infuse background knowledge into the model (in the form of itemset frequencies, e.g., those of the individual items), to avoid discovering patterns that are redundant with regard to what he or she already knows.

References

[1] I. Csiszár. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability*, 3(1):146–158, 1975.

[2] J. Darroch and D. Ratcliff. Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics*, 43(5):1470–1480, 1972.

[3] S. Nijssen, T. Guns, and L. De Raedt. Correlated itemset mining in ROC space: a constraint programming approach. In *Proceedings of ACM SIGKDD'09*, pages 647–656, 2009.

[4] G. Schwarz. Estimating the dimension of a model. *Annals of Statistics*, 6(2):461–464, 1978.