Some theoretical open problems in data mining

December 20, 2009

Abstract

This document contains a selection of open problems provided by participants of the data mining PhD reading seminar at the IT University of Copenhagen in fall 2009. It is not intended to showcase the "most significant" open problems in the field, but merely problems found interesting by the participants. In particular, the theoretical angle of the seminar is reflected.

1 Frequent itemsets — a relaxed version

Suggested by Rasmus Pagh.

This is an attempt to formulate an "easiest possible" version of the frequent itemsets problem that would have interesting (theoretical) implications for association mining. Given a collection of sets $S_1, \ldots, S_n \subseteq U$, define $\text{sup}(S) = |\{i \mid S \subseteq S_i\}|$. For parameters $\varepsilon > 0$, $k, \Delta, t \in \mathbf{N}$, where it is promised that $|\{S \in \binom{U}{k} \mid \text{sup}(S) \geq (1+\varepsilon)\Delta\}| \geq t$ the task is to output t sets $T_1, \ldots, T_t \in \binom{U}{k}$ where $\text{sup}(T_i) \geq \Delta$. The following may be assumed:

- Parameter restrictions: $|U| = O(n^2)$; $\Delta = O(\log n)$ for constant ε.
- k and $1/\varepsilon$ are a small numbers, so exponential dependence is acceptable.
- It is acceptable that the output satisfies the guarantee only with probability $1/2$.

The problem is to come up with algorithms and/or hardness results for the above. We note that hardness results for related problems were shown in [11].

2 Deterministic frequent items [2, Q4]

Suggested by Anders Schack-Nielsen.

Given a stream of insertions and deletions of items (the set of items is $\{1, 2, \ldots, n\}$) we wish to estimate the frequencies of each item within an ϵ-factor of the L_1-norm of the frequencies (m). A space lower bound for this problem is $\Omega(\epsilon^{-1} \log(m) \log(\epsilon n))$ [10]. The Count-Min algorithm is a randomized $O(\epsilon^{-1} \log(m) \log(\delta^{-1}))$-space algorithm that succeeds with probability $1 - \delta$ [6]. In [10] a deterministic algorithm is given that uses space

$$O(\epsilon^{-2} \log^2_{\epsilon^{-1}}(n) \log(\log_{\epsilon^{-1}}(n)) \log m).$$

Does there exist deterministic algorithms with space bounds as good as Count-Min or can we find stronger lower bounds? See Cormode et al. [4] for recent progress. The case where it is possible to delete remains open.

3 Boolean matrix decomposition (BMD)

Suggested by Andrea Campagna.

The problem of Exteded Boolean Matrix Decomposition (EBMD) can be, in general, written as: given a boolean Matrix $A_{m\times n}$, the operator $||Q||_1 := \sum_{i=1}^{m}\sum_{j=1}^{n}|q_{ij}|$, find two matricies: $C_{m\times k}$ and $X_{k\times n}$ such that: C is boolean and contains a subset of columns of A, $\forall i, j$, x_{ij} in X belongs to $\{-1, 0, 1\}$, $A = C \times X$ and $||A - C \times X||_1$ is minimized. If the input is only A and k, the problem is called EBMD, if A and C are given, the problem is called EBU. In the reconstruction of the matrix A, two kinds of errors can occur: a 1 becoming a 0 and the other way around. It turns out that (all problems are to be intended as decision versions): EBMD, EBU, 1-0 free EBU are NP complete; no poly time algorithms for 1-0 free EBMD unless P=NP; 0-1 free EBU $\in$ P; no claims about 0-1 free EBMD.
The questions to pose are many: what about the complexity of the optimization versions? What about 0-1 free EBMD? Are the hard problems FPT?

References: [12, 15, 16]

4 Grammar approximation in sequences

Suggested by Christian Theil Have.

Members of an RNA family are normally found using comparative analysis where a candidate sequence is aligned to known RNA families. The RNA primary sequence folds into a secondary structure which is essential for its function and more powerful alignment methods utilizes structure information. Comparative analysis is biased towards finding instances which resemble already known families, which leaves the problem of discovering novel families.

We define the problem as follows. We are given a sequence $S = s_1 \dots s_n$, $s_i \in \{A, G, C, T/U\}$ and consider any subsequence $S_{i\dots j}$ to be a candidate RNA sequence, where the length of the subsection $|S_{i\dots j}|$ is above a given significance threshold, $|S_{i\dots j}| > l_{significant}$. We then wish to find the k such subsequences $S_1 \dots S_k$ which have the best alignment on both sequence and structure level.

Subsequences and alignments are scored according to a model which assigns a probability to each subsequence or alignment of such. Furthermore, it provides selection of the best alignment of sequences. The problem is then to select k subsequences which maximizes this best alignment score. The alignment probability scores depend on the parameterization of the alignment model, which could initially be estimated from known RNA families.

A short sketch for an sampling algorithm based on [3] and similar to [18] is given here. Initially the alignment model parameters are estimated from known RNA families. Then the following process is iterated until convergence: 1. (sampling): Randomly select a set of k subsequences, such that the chance that a subsequence is included is proportional to its probability

according to the model. 2. (reestimation): At random leave one of the k sequences out and estimate model probabilities using the remaining $k-1$ subsequences.

The approach has inherent complexity problems. Commonly used alignment models such as covariance models [9] (essentially PCFGs) have $O(n^3)$ parsing complexity. In addition when aligning multiple sequences, the complexity raise the exponent proportional to the number of sequences aligned. Progressive alignment techniques deals with this at the cost of suboptimal alignments. Lower complexity approximate models [17] which filters away unpromising candidates before applying more costly models also seems as a promising way of dealing with the large complexity of the problem.

5 KEMENY-RANK-AGGREGATION for 3 permutations

Suggested by Nina Sofia Taslaman.

Given k permutations $\pi_1, ..., \pi_k$ of the same set of elements, a *Kemeny optimal permutation* is defined as an ordering π which minimizes the total number of adjacent transpositions required to transform π into each of the π_i's. For $k = 2$ one may simply take either $\pi = \pi_1$ or $\pi = \pi_2$. For $k \geq 4$ the problem is known to be NP-hard (Dwork et al. [8], Biedl et al. [5]), with a randomized 11/7-approximation proposed by Ailon et al. [1]. For $k = 3$ the problem is still open. (In [7], Dwork et al. give a reduction to the problem of finding minimum feedback edge sets on unweighted tournament graphs, but the proof seems to be wrong.)

6 Concurrent and interleaved episodes in event sequences

Suggested by Afsaneh Dohryab.

Consider a sequence of events (*e,t*), where *e* is the type of the event and *t* is the time when the event occured. Given a set *E* of event types, one can try to find which event types occure frequently together. An episode α is a partially ordered set of elements from *E*. An episode might, e.g., state that events of type *A* and *B* occure (in either order) before an event of type *C*. Given a sequence S=(e_1,t_1),...,(e_n,t_n), a slice s_t of S of width *W* consists of those events (e_i,t_i) of S such that $t \leq t_i \leq t+W$. An episode α occurs in s_t, if there are events in s_t corresponding to the event types of α and they occure in an order following the partial order of α [14, 13]. An open question is how to find/recognize concurrent and interleaved episodes from a sequence of events.

References

[1] N. Ailon, M. Charikar, and A. Newman. Aggregating inconsistent information: Ranking and clustering. *J. ACM*, 55(5), 2008.

[2] Open Problems in Data Streams and Related Topics IITK Workshop on Algorithms for Data Streams '06. http://www.cse.iitk.ac.in/users/sganguly/data-stream-probs.pdf, December 2006. Compiled and edited by Andrew McGregor.

[3] T. L. Bailey. Discovering novel sequence motifs with MEME. *Current Protocols in Bioinformatics*, 2002.

[4] R. Berinde, G. Cormode, P. Indyk, and M. J. Strauss. Space-optimal heavy hitters with strong error bounds. In J. Paredaens and J. Su, editors, *PODS*, pages 157–166. ACM, 2009.

[5] T. C. Biedl, F.-J. Brandenburg, and X. Deng. Crossings and permutations. In *Graph Drawing*, pages 1–12, 2005.

[6] G. Cormode and S. Muthukrishnan. An improved data stream summary: The count-min sketch and its applications. *Journal of Algorithms*, 55(1):58–75, 2005.

[7] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation revisited.

[8] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation methods for the web. In *WWW*, pages 613–622, 2001.

[9] S. R. Eddy and R. Durbin. RNA sequence analysis using covariance models. *Submitted to Nucleic Acids Research*, unknown(unknown):unknown, 1994.

[10] S. Ganguly and A. Majumder. CR-precis: A deterministic summary structure for update data streams. *Lecture Notes in Computer Science*, 4614:48–59, 2007.

[11] M. Hamilton, R. Chaytor, and T. Wareham. The parameterized complexity of enumerating frequent itemsets. In H. L. Bodlaender and M. A. Langston, editors, *IWPEC*, volume 4169 of *Lecture Notes in Computer Science*, pages 227–238. Springer, 2006.

[12] H. Lu, J. Vaidya, and V. Atluri. Optimal boolean matrix decomposition: Application to role engineering. In *Proceedings of the 24th International Conference on Data Engineering, ICDE 2008, April 7-12, 2008, Cancún, México*, pages 297–306. IEEE, 2008.

[13] H. Mannila and H. Toivonen. Discovering generalized episodes using minimal occurences.

[14] H. Mannila, H. Toivonen, and A. Inkeri Verkamo. Discovery of frequent episodes in event sequences. *Data Mining and Knowledge Discovery*, 1(3):259–289, September 1997.

[15] P. Miettinen. The boolean column and column-row matrix decompositions. *Data Min. Knowl. Discov.*, 17(1):39–56, 2008.

[16] P. Miettinen, T. Mielikäinen, A. Gionis, G. Das, and H. Mannila. The discrete basis problem. *IEEE Trans. Knowl. Data Eng.*, 20(10):1348–1362, 2008.

[17] Weinberg and Ruzzo. Faster genome annotation of non-coding RNA families without loss of accuracy. In *Annual International Conference on (Research in) Computational (Molecular) Biology*, volume 8, 2004.

[18] Z. Yao, Z. Weinberg, and W. L. Ruzzo. CMfinder - a covariance model based RNA motif finding algorithm. *Bioinformatics*, 22(4):445–452, 2006.