Sampling-based Data Mining Algorithms: Modern Techniques and Case Studies

Matteo Riondato

Brown University, Providence, RI 02912, USA
matteo@cs.brown.edu

Abstract. Sampling a dataset for faster analysis and looking at it as a sample from an unknown distribution are two faces of the same coin. We discuss the use of modern techniques involving the Vapnik-Chervonenkis (VC) dimension to study the trade-off between sample size and accuracy of data mining results that can be obtained from a sample. We report two case studies where we and collaborators employed these techniques to develop efficient sampling-based algorithms for the problems of betweenness centrality computation in large graphs and extracting statistically significant Frequent Itemsets from transactional datasets.

1 Sampling the data and data as samples

There exist two possible uses of sampling in data mining. On the one hand, sampling means selecting a small random portion of the data, which will then be given as input to an algorithm. The output will be an approximation of the results that would have been obtained if all available data was analyzed but, thanks the the small size of the selected portion, the approximation could be obtained much more quickly. On the other hand, from a more statistically-inclined point of view, the entire dataset can be seen as a collection of samples from an unknown distribution. In this case the goal of analyzing the data is to gain a better understanding of the unknown distribution. Both scenarios share the same underlying question: how well does the sample resemble the entire dataset or the unknown distribution? There is a trade-off between the size of the sample and the quality of the approximation that can be obtained from it. Given the randomness involved in the sampling process, this trade-off must be studied in a probabilistic setting. In this nectar paper we discuss the use of techniques related to the Vapnik-Chervonenkis (VC) dimension of the problem at hand to analyze the trade-off between sample size and approximation quality and we report two case studies where we and collaborators successfully employed these techniques to develop efficient algorithms for the problems of betweenness centrality computation in large graphs [8] ("sampling the data" scenario) and extracting statistically significant frequent itemsets [10] ("data as samples" scenario).

2 The sample-size/accuracy trade-off: modern techniques

There exist many probabilistic techniques to study the trade-off between accuracy and sample size: large deviation Chernoff/Hoeffding bounds, martingales, tail

bounds on polynomials of random variables, and many others [3, 5]. These classical results bound the probability that the measure of interest (e.g., the frequency) for a single object (e.g., an itemset) in the sample deviates from its expectation (its true value in the dataset or according to the unknown probability distribution) by more than some amount. An application of the union bound is then needed to get simultaneous guarantees on the deviations for all the objects. The so-obtained sample size or quality guarantee then depends on the logarithm of the number of objects. Due to the number of objects involved in many data mining problems (e.g., all possible itemsets or all nodes in a graph), the sample size may be excessively loose and the benefit of sampling could be lost or not enough information about the unknown distribution may be extracted. In a sequence of works [6–10] we investigated the use of techniques based on the Vapnik-Chervonenkis (VC) Dimension [11] to study the trade-off between accuracy and sample size. The VC-dimension of a data mining task is a measure of the complexity of that problem in terms of the richness of the set of measures that the task requires to compute. The advantage of techniques involving VC-dimension is that they allow to compute sample sizes that only depend on this combinatorial quantity (see below), which can be very small and independent from the number of objects and from the size of the dataset. The techniques related to VC-dimension are widely applicable as we show in our case studies.

Definitions and sampling theorem. A *range space* is a pair $(D, \mathcal{R})$ where D is a domain and $\mathcal{R}$ is a family of subsets of D. The members of D are called *points* and those of $\mathcal{R}$ are called *ranges*. The VC-dimension of $(D, \mathcal{R})$ is the size of the largest $A \subseteq D$ such that $\{R \cap A \ : \ R \in \mathcal{R}\} = 2^A$. If ν is any (unknown) probability distribution over D from which we can sample, then a finite upper bound to the VC-dimension of $(D, \mathcal{R})$ implies a bound to the number of random samples from ν required to approximate the probability $\nu(R) = \sum_{r \in R} \nu(r)$ of each range R simultaneously using the *empirical average* of $\nu(R)$ as estimator.

Theorem 1 ([4, 12]). *Let d be an upper bound to the VC-dimension of $(D, \mathcal{R})$. Given $\varepsilon, \delta \in (0, 1)$, let $\mathcal{S}$ be a collection of independent samples from ν of size*

$$|\mathcal{S}| \geq \frac{1}{\varepsilon^2}\left(d + \ln\frac{1}{\delta}\right) \ . \tag{1}$$

Then, with probability at least $1 - \delta$, we have

$$\left|\nu(R) - \frac{1}{|\mathcal{S}|}\sum_{a \in \mathcal{S}} \mathbb{1}_R(a)\right| \leq \varepsilon, \textit{for all } R \in \mathcal{R} \ .$$

The bound on the deviations of the estimation holds *simultaneously* over all ranges. The sample size in (1) depends on the user-specified accuracy and confidence parameters ε and δ and on the bound to the VC-dimension of the range space. If the latter does not depend on the size of the D, then neither will the sample size. This is a crucial and very intriguing property that allows for the development of

sampling-based data mining algorithms that use small samples and are therefore very efficient. The main obstacles in developing such algorithms are: *1.* formulate the data mining task in terms of range spaces and unknown distributions; *2.* compute (efficiently) an upper bound to the VC-dimension of the task at hand; *3.* have an efficient procedure to sample from the unknown distribution. It is possible but not immediate to overcome these obstacles as we did for different important data mining problems.

3 Case studies

In line with the nature of this Nectar paper, we present two case studies where we and collaborators successfully used VC-dimension to develop efficient sampling-based algorithms for important data and graph mining problems.

Betweenness Centrality In [8] we developed a sampling-based algorithm to compute guaranteed high-quality approximations of the betweenness centrality indices of all vertices in a large graph. We defined a range space $(D, \mathcal{R})$ where D is the set of all shortest paths in the graph and $\mathcal{R}$ contains one range R_v for each node v in the graph, where R_v is the set of shortest paths that pass through v. A shortest path p between two nodes u and w is sampled with probability $\nu(p)$ proportional to the number of nodes in the graph and the number of shortest paths between u and w. With this definition of the sampling distribution, we have that $\nu(R_v)$ is exactly the betweenness centrality of the node v. We showed that the VC-dimension of this range space is at most the logarithm of the diameter of the graph. Thus, through Thm. 1, the number of $s-t$-shortest path computations to approximate all betweenness values depends on this quantity rather than on the logarithm of the number of nodes in the graph as previously thought.

True Frequent Itemsets. In [10] we introduced the problem of finding the *True* Frequent Itemsets from a transactional dataset. The dataset is seen as a sample from an unknown distribution ν defined on all possible transactions and the task is to identify the itemsets generated frequently by ν, without reporting false positives (i.e., non-frequently-generated itemsets). We formulated the problem in terms of range spaces and computed its VC-dimension in order to use (a variant of) Thm. 1. The domain D is the set of all possible transaction built on the set of items. For each itemset A we define the range R_A as the set of transactions in D that contain A. The frequency of A in the dataset is now the empirical average of the "true frequency of A", i.e., the probability that ν generates a transaction that contains A. We showed that the (empirical) VC-dimension of $(D, \mathcal{R})$ is tightly bounded from above by a characteristic quantity of the dataset, namely the maximum integer d such that the dataset contains at least d transactions of length at least d forming an antichain. A bound to this quantity can be computed with a single linear scan of the dataset. A more refined bound to the empirical VC-dimension can be computed by solving a variant of a knapsack problem. These bounds allow us to compute a value ε such that the itemsets with frequency in the dataset greater than $\theta + \varepsilon$ have, with high probability, true frequency at least θ. The use of VC-dimension allows us to achieve much higher statistical power

(i.e., to identify more true frequent itemsets) than methods based on the classical bounds and the Bonferroni correction.

4 Future directions and challenges

Sampling will always be a viable option to speed up the analysis of very large datasets. The database research community, often an early adopter of modern storage technologies, is showing a renovated interest in sampling [1, 13]. There is thriving research to develop stronger simultaneous/uniform bounds to the deviations of sets of functions by leveraging modern probability results involving the Rademacher averages, the shatter coefficients, the covering numbers, and the many extensions of VC-dimension to real functions [2]. The major challenges in using these techniques for more and more complex data mining problems are *1.* understanding the best formulation of the problem in order to leverage the best available bounds to the sample size, and *2.* developing bounds to the VC-dimension (or other combinatorial quantities) to be able to use sampling theorems similar to Thm. 1. There is huge room for additional contributions from the data mining community, to show how powerful theoretical results can be used to develop efficient practical algorithms for important data mining problems.

References

[1] Agarwal, S., Mozafari, B., Panda, A., Milner, H., Madden, S., Stoica, I.: BlinkDB: Queries with bounded errors and bounded response times on very large data. In: EuroSys'12

[2] Boucheron, S., Bosquet, O., Lugosi, G.: Theory of classification: A survey of some recent advances. In: ESAIM: Probability and Statistics 9, 323–375 (2005)

[3] Dubhashi, D.P., Panconesi, A.: Concentration of Measure for the Analysis of Randomized Algorithms. Cambridge University Press (2009)

[4] Har-Peled, S., Sharir, M.: Relative (p, ε)-approximations in geometry. Discr. & Computat. Geom. 45(3), 462–496 (2011)

[5] Mitzenmacher, M., Upfal, E.: Probability and Computing: Randomized Algorithms and Probabilistic Analysis. Cambridge University Press (2005)

[6] Riondato, M., Akdere, M., Çetintemel, U., Zdonik, S.B., Upfal, E.: The VC-dimension of SQL queries and selectivity estimation through sampling. In: ECML PKDD'11

[7] Riondato, M., DeBrabant, J.A., Fonseca, R., Upfal, E.: PARMA: A parallel randomized algorithm for association rules mining in MapReduce. In: CIKM'12

[8] Riondato, M., Kornaropoulos, E.M.: Fast approximation of betweenness centrality through sampling. In: WSDM'14

[9] Riondato, M., Upfal, E.: Efficient discovery of association rules and frequent itemsets through sampling with tight performance guarantees. In: ACM Trans. Knowl. Disc. from Data, (in press)

[10] Riondato, M., Vandin, F.: Finding the true frequent itemsets. In: SDM'14

[11] Vapnik, V.N.: The Nature of Statistical Learning Theory. Springer-Verlag, New York, NY, USA (1999)

[12] Vapnik, V.N., Chervonenkis, A.J.: On the uniform convergence of relative frequencies of events to their probabilities. Theory of Prob. and its Appl. 16(2), 264–280 (1971)

[13] Wang, J., Krishnan, S., Franklin, M.J., Goldberg, K., Kraska, T., Milo, T.: A sample-and-clean framework for fast and accurate query processing on dirty data. In: SIGMOD'14