An Asymptotically Tighter Bound on Sampling for Frequent Itemsets Mining

Shiyu Ji
shiyu@cs.ucsb.edu

Abstract

In this paper we present a new error bound on sampling algorithms for frequent itemsets mining. We show that the new bound is asymptotically tighter than the state-of-art bounds, i.e., given the chosen samples, for small enough error probability, the new error bound is roughly half of the existing bounds. We also give a new approximation algorithm, which is much simpler compared to the existing approximation algorithms, but can also guarantee the worst approximation error with precomputed sample size.

1 Introduction

Frequent Itemsets (FI) mining has been popular in research recently [1, 4, 7]. The goal of FI mining is to find out the items that most frequently appear in the observed transactions, e.g., the researchers who are the most prolific in writing papers with others, the patterns that appear frequently in a long piece of genetic code, etc. In the era of big data, to compute the exact frequencies can be very time consuming. Thus in many cases approximated values are also acceptable [7]. The state-of-art progressive sampling based FI approximation algorithms [7] need an upper bound of the approximation error for the worst case, i.e., the maximum error the algorithm can generate among all the items. The algorithms keep taking new samples until the upper bound is less than the acceptable threshold. Hence how to bound the maximum error as tightly as possible is an interesting problem. The current bounds use some results of Rademacher average in statistical learning theory [9, 10, 3, 2]. However, we find that based on the idea given by [3], we can develop a new upper bound *without* Rademacher average. We also find that this new bound asymptotically outperforms the existing bounds, i.e., given the chosen samples, as the allowed error probability approaches zero, the new bound is roughly only half of the existing ones. This implies that by using the new bound, a progressive sampling based FI approximation algorithm can reach the guaranteed accuracy with much fewer samples. We also notice that there is no parameter in the new bound that needs to be progressively computed. Hence the sample size that will guarantee the worst error can be precomputed.

2 Preliminaries

2.1 Frequency of Itemset

In this paper we use the notations and definitions from Riondato and Upfal's pioneering work [7]. Let $\mathcal{I}$ be the set of items. A transaction τ is a subset of $\mathcal{I}$ (i.e., $\tau \subseteq \mathcal{I}$). An itemset A is a set of items that appear together in a transaction τ, i.e., $A \subseteq \tau$. Clearly any itemset is also a subset of $\mathcal{I}$. Let transactional dataset $\mathcal{D}$ be the set of all the transactions. In this paper we always assume $\mathcal{D}$ is a finite set. Denote by $T_{\mathcal{D}}(A)$ the set of all the transactions in $\mathcal{D}$ that contain the itemset A. $T_{\mathcal{D}}(A)$ is also known as the support set of A in $\mathcal{D}$. If $\mathcal{D}$ is a finite set, we can define the frequency of itemset A in $\mathcal{D}$ as the fraction of transactions in $\mathcal{D}$ that contain A.

$$f_{\mathcal{D}}(A) = |T_{\mathcal{D}}(A)|/|\mathcal{D}|.$$

Clearly $0 \leq f_{\mathcal{D}}(A) \leq 1$ for any $A \subseteq \mathcal{I}$.

The goal of our sampling algorithm is to approximate $f_{\mathcal{D}}(A)$ given an itemset A as accurately as possible.

2.2 Approximation Algorithms

An (ϵ, δ)-approximation algorithm of the frequencies $f_{\mathcal{D}}(\cdot)$ takes as input all the items $\mathcal{I}$ and outputs a sampled average $f_{\mathcal{S}}(A)$ for each $A \subseteq \mathcal{I}$ such that with probability at least $1 - \delta$,

$$\max_{A \subseteq \mathcal{I}} |f_{\mathcal{D}}(A) - f_{\mathcal{S}}(A)| \leq \epsilon.$$

We often use progressive sampling [7, 8], i.e., to keep taking more samples until a stopping condition is reached. A stopping condition usually takes the form $\Delta(n, \delta) \leq \epsilon$, where n is the number of samples that have been taken, and Δ is an upper bound of the worst approximation error given by statistical learning theory. Note that Δ is usually a function of n and δ.

There is a variant called top-k approximation, which returns the k most frequent itemsets among the observed ones based on the approximated frequencies. This is quite popular in practice since we are often only interested in the most common itemsets.

2.3 Risk Bounds

We briefly review some risk bounds in statistical learning theory [2] with the background of frequent itemsets mining.

For each itemset $A \subseteq \mathcal{I}$, define the indicator function $\phi_A : 2^{\mathcal{I}} \to \{0, 1\}$ as follows.

$$\phi_A(\tau) = \begin{cases} 1 & \text{if } A \subseteq \tau \\ 0 & \text{otherwise} \end{cases}, \quad \tau \subseteq \mathcal{I}.$$

Clearly, the frequency $f_{\mathcal{D}}(A)$ equals to the *true* average of $\phi_A(\tau)$ where τ goes over all the transactions in $\mathcal{D}$.

$$f_{\mathcal{D}}(A) = \frac{1}{|\mathcal{D}|} \sum_{\tau \in \mathcal{D}} \phi_A(\tau).$$

Similarly let $\mathcal{S}$ be the set of the sampled transactions. Then the *sampled* average of $\phi_A(\tau)$ can be defined as

$$f_{\mathcal{S}}(A) = \frac{1}{|\mathcal{S}|} \sum_{\tau \in \mathcal{S}} \phi_A(\tau).$$

Clearly $f_{\mathcal{S}}(A)$ is the frequency of A appearing in the sampled transactions $\mathcal{S}$.

Assume $|\mathcal{S}| = n$. For each transaction $\tau_i \in \mathcal{S}$, let σ_i be a Rademacher random variable taking value from $\{-1, 1\}$ with uniform probability distribution. The σ_i's are independent. Assuming $\mathcal{I}$ is finite, we define the sample conditional Rademacher average as follows.

$$\mathcal{R}_{\mathcal{S}} = \mathbb{E}_{\sigma} \left[\max_{A \subseteq \mathcal{I}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i \phi_A(\tau_i) \right],$$

where $\mathbb{E}_{\sigma}$ denotes the expectation taken over all the random variables σ_i's, conditionally on the sample $\mathcal{S}$.

The following theorem tells us that Rademacher average can be used to upper bound the approximation error, even for the worst case.

Theorem 1. (Theorem 3.2, [2]) For any $\delta > 0$, with probability at least $1 - \delta$,

$$\max_{A \subseteq \mathcal{I}} |f_{\mathcal{D}}(A) - f_{\mathcal{S}}(A)| \leq 2\mathcal{R}_{\mathcal{S}} + \sqrt{\frac{2 \log(2/\delta)}{n}}.$$

If we want to use the upper bound given in Theorem 1 in an approximation algorithm, we still need to upper bound the $\mathcal{R}_{\mathcal{S}}$. A classical result is given by Massart [6].

Theorem 2. (Lemma 5.2, [6]) Let $\ell = \max_{A \subseteq \mathcal{I}} [\sum_{i=1}^{n} \phi_A(\tau_i)^2]^{1/2}$ where each $\tau_i \in \mathcal{S}$. Then

$$\mathcal{R}_{\mathcal{S}} \leq \frac{\ell}{n} \sqrt{2 \log N},$$

where $N = 2^{|\mathcal{I}|}$ and $n = |\mathcal{S}|$.

Hence we have the following stopping condition for an (ϵ, δ)-approximation sampling algorithm.

$$\Delta_1 := \frac{2\ell}{n} \sqrt{2 \log N} + \sqrt{\frac{2 \log(2/\delta)}{n}} \leq \epsilon.$$

However for many applications the above bound is not tight enough [7, 8]. In the next section we will first review the state-of-art bound on the worst approximation error, and then propose a new bound which seems tighter.

3 Refining the Upper Bound

The reason why the bound given in the previous section is often not tight enough in practice is that the ℓ defined in Theorem 2 can be quite large. Suppose there is an itemset A that almost always appears in every transaction in $\mathcal{D}$. Then no matter which sample the algorithm chooses, ℓ is roughly $\sqrt{n}$. For $\delta = 0.01$, $N = 2^{1000}$, even 100,000 samples are taken, the upper bound is still larger than 0.15. For many applications such an upper bound cannot be acceptable and thus we need to take more samples. Clearly if the upper bound is tighter, a lot of samples can be saved.

Riondato and Upfal [7] attempted to give a tighter bound of the Rademacher average $\mathcal{R}_{\mathcal{S}}$.

Theorem 3. (Theorem 3, [7], revised) Let $w : \mathbb{R}^+ \to \mathbb{R}^+$ be the function defined as

$$w(s) = \frac{1}{s} \log \sum_{A \subseteq \mathcal{I}} \exp\left(\frac{s^2 \sum_{i=1}^{n} \phi_A(\tau_i)^2}{2n^2} \right).$$

Then $\mathcal{R}_{\mathcal{S}} \leq \min_{s>0} w(s)$.

Remark. Note that in Theorem 3, the summation in $w(s)$ takes *exactly* $2^{|\mathcal{I}|}$ terms. However in the original version in [7], the authors claimed that the summation could take much less than $2^{|\mathcal{I}|}$ terms. We argue that there is a gap between these two versions. Based on the proof given in [7], one can reach the inequality as follows.

$$\exp(s\mathcal{R}_{\mathcal{S}}) \leq \sum_{A \subseteq \mathcal{I}} \exp\left(\frac{s^2 \sum_{i=1}^{n} \phi_A(\tau_i)^2}{2n^2} \right). \tag{1}$$

Note that on the right hand side, each term in the summation is no less than 1. Hence when taking the logarithm on both sides and dividing by s, each of the $2^{|\mathcal{I}|}$ terms cannot be eliminated. Thus the range of the summation cannot be compressed.

Formally, suppose there is a set $\mathcal{V} \subseteq 2^{\mathcal{I}}$, where $2^{\mathcal{I}}$ denotes the power set of $\mathcal{I}$, such that

$$\alpha(s) := \sum_{A \in 2^{\mathcal{I}}} \exp\left(\frac{s^2 \sum_{i=1}^{n} \phi_A(\tau_i)^2}{2n^2} \right) \leq \sum_{A \in \mathcal{V}} \exp\left(\frac{s^2 \sum_{i=1}^{n} \phi_A(\tau_i)^2}{2n^2} \right) := \beta(s).$$

We take the limits as s approaches 0.

$$2^{|\mathcal{I}|} = \lim_{s \to 0} \alpha(s) \leq \lim_{s \to 0} \beta(s) = |\mathcal{V}|.$$

Hence $\mathcal{V} = 2^{\mathcal{I}}$. This implies any summation over only a part of $2^{\mathcal{I}}$ must be less than the summation over all of $2^{\mathcal{I}}$. Thus one cannot use Inequality (1) to reach Theorem 3 in [7].

3.1 New Bound Without Rademacher Average

In statistical learning theory, the upper bound given by Theorem 1 is for the general case, i.e., the set of transactions $\mathcal{D}$ can be infinite or finite. However, for frequent itemsets mining, the transactional data set $\mathcal{D}$ is always finite. Given the assumption that $\mathcal{D}$ is finite, can we establish any upper bound without using the Rademacher average? Following the similar lines given by Boucheron, Bousquet and Lugosi [3], we can give an affirmative answer.

For any $\epsilon > 0$,

$$\begin{aligned}
&\Pr[\max_{A\subseteq\mathcal{I}} |f_{\mathcal{D}}(A) - f_{\mathcal{S}}(A)| > \epsilon] \\
=&\Pr[\exists A \subseteq \mathcal{I}, f_{\mathcal{D}}(A) - f_{\mathcal{S}}(A) > \epsilon \vee f_{\mathcal{D}}(A) - f_{\mathcal{S}}(A) < -\epsilon] \\
\leq&\Pr[\exists A \subseteq \mathcal{I}, f_{\mathcal{D}}(A) - f_{\mathcal{S}}(A) > \epsilon] + \Pr[\exists A \subseteq \mathcal{I}, f_{\mathcal{D}}(A) - f_{\mathcal{S}}(A) < -\epsilon] \quad \text{(union bound)} \\
\leq&\sum_{A\subseteq\mathcal{I}} \Pr[f_{\mathcal{D}}(A) - f_{\mathcal{S}}(A) > \epsilon] + \sum_{A\subseteq\mathcal{I}} \Pr[f_{\mathcal{S}}(A) - f_{\mathcal{D}}(A) > \epsilon] \quad \text{(union bound)}.
\end{aligned}$$

Recall Hoeffding's inequalities [5]. Let $X_1, \cdots, X_n$ be independent random variables bounded by the intervals $[a_i, b_i]$. Define the sampled average of them as

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Then for any $t > 0$,

$$\Pr[\overline{X} - \mathbb{E}[\overline{X}] > t] \leq \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right),$$

and

$$\Pr[\mathbb{E}[\overline{X}] - \overline{X} > t] \leq \exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right).$$

Note that if we set $X_i = \phi_A(\tau_i)$, then X_i's are independent since τ_i's are independent, and thus $f_{\mathcal{D}}(A) = \mathbb{E}[\overline{X}]$ and $f_{\mathcal{S}}(A) = \overline{X}$. Based on Hoeffding's inequalities,

$$\Pr[f_{\mathcal{D}}(A) - f_{\mathcal{S}}(A) > \epsilon] \leq \exp\left(-2n\epsilon^2\right),$$

and

$$\Pr[f_{\mathcal{S}}(A) - f_{\mathcal{D}}(A) > \epsilon] \leq \exp\left(-2n\epsilon^2\right).$$

Putting the above results together, we have

$$\begin{aligned}
&\Pr[\max_{A\subseteq\mathcal{I}} |f_{\mathcal{D}}(A) - f_{\mathcal{S}}(A)| > \epsilon] \\
\leq&2\sum_{A\subseteq\mathcal{I}} \exp\left(-2n\epsilon^2\right) \\
=&2N\exp\left(-2n\epsilon^2\right),
\end{aligned}$$

where $N = 2^{|\mathcal{I}|}$. Equivalently for any $\delta > 0$, with probability at least $1 - \delta$,

$$|f_{\mathcal{D}}(A) - f_{\mathcal{S}}(A)| \leq \sqrt{\frac{\log(2N) + \log(1/\delta)}{2n}} =: \Delta_2.$$

Note that the above bound Δ_2 is very similar to the result in Section 3.4, [3]. Now the new bound Δ_2 can generate a new stopping condition for an approximation algorithm.

Recall the classical upper bound given in Section 2.3.

$$\Delta_1 := \frac{2\ell}{n}\sqrt{2\log N} + \sqrt{\frac{2\log(2/\delta)}{n}}.$$

Clearly $\lim_{\delta\to 0}\Delta_1/\Delta_2 = 2$, i.e., when δ is very small, the bound Δ_1 is roughly twice of Δ_2 given the sample size n. This assures us that the new bound Δ_2 is highly competitive.

Theorem 3 can give another upper bound on the worst approximation error. However, since the number of terms in the summation grows exponentially on $|\mathcal{I}|$, to find the minimum is computationally infeasible. Furthermore, even if the minimum $w(s^*)$ is found, let Δ_1' be the upper bound of this variant defined as

$$\Delta_1' := w(s^*) + \sqrt{\frac{2\log(2/\delta)}{n}}.$$

By fixing the sample $\mathcal{S}$, we still have $\lim_{\delta\to 0}\Delta_1'/\Delta_2 = 2$. For small δ, the new bound without Rademacher average still outperforms the existing ones.

4 Approximation Algortihm with Precomputed Sample Size

We observe the new bound given in the previous section:

$$\Delta_2 := \sqrt{\frac{\log(2N) + \log(1/\delta)}{2n}},$$

where $N = 2^{|\mathcal{I}|}$. The upper bound Δ_2 can be treated as a function of allowed error probability δ, sample size n and $N = 2^{|\mathcal{I}|}$, all of which are already given. A good news is that there is no parameter that needs to be progressively computed (e.g., ℓ in Δ_1). Thus to guarantee an worst approximation error at most ϵ, we only need to make sure $\Delta_2 \le \epsilon$. By solving it we have

$$n \ge \frac{1}{2}(\log 2N + \log(1/\delta)).$$

Hence an (ϵ, δ)-approximation algorithm takes a very simple form.

Frequent Itemsets Approximation Algorithm

Input: items $\mathcal{I}$, transactional dataset $\mathcal{D} \subseteq 2^{\mathcal{I}}$, $\epsilon > 0$, $\delta > 0$.

Output: approximated frequencies $\hat{f}_{\mathcal{D}}(A)$ for each $A \subseteq \mathcal{I}$ s.t. with probability at least $1-\delta$, $|\hat{f}_{\mathcal{D}}(A)| \le \epsilon$ for any $A \subseteq \mathcal{I}$.

1. $n \leftarrow \lceil \frac{1}{2}(\log(2^{|\mathcal{I}|+1}) + \log(1/\delta)) \rceil$.
2. $\mathcal{S} \leftarrow \emptyset$.
3. If $n \ge |\mathcal{D}|$, $\mathcal{S} \leftarrow \mathcal{D}$; otherwise, choose n itemsets in $\mathcal{D}$ at uniformly random and add them to $\mathcal{S}$.
4. Label the transactions in $\mathcal{S}$: $\mathcal{S} = \{\tau_1, \cdots, \tau_n\}$.
5. For each $A \subseteq \mathcal{I}$, compute $\hat{f}_{\mathcal{D}}(A) \leftarrow \frac{1}{n}\sum_{i=1}^{n}\phi_A(\tau_i)$.
6. Return all the $\hat{f}_{\mathcal{D}}(A)$'s for $A \subseteq \mathcal{I}$.

Note that the algorithm above estimates the frequency for every subset in $\mathcal{I}$. Hence it is inefficient when $|\mathcal{I}|$ is large. In practice, we often only care about the frequencies of a few itemsets, e.g., celebrities in a community, influential papers in a field, etc. For this case, we do not have to estimate the frequencies of the rest itemsets. Denote by $\mathbf{Ob}$ the set of the itemsets to be observed. Then the worst approximation error is defined as the maximum error on every itemset in $\mathbf{Ob}$. By the same reasoning in the derivation of Δ_2, we have the adjusted new bound:

$$\Delta_2' := \sqrt{\frac{\log(2|\mathbf{Ob}|) + \log(1/\delta)}{2n}}.$$

Note that since **Ob** is a subset of $2^{\mathcal{I}}$, this bound Δ_2' is tighter than Δ_2. The approximation algorithm will also be revised as follows.

Frequent Itemsets Approximation Algorithm with Observed Items

Input: all the items $\mathcal{I}$, the observed items $\mathbf{Ob} \subseteq 2^{\mathcal{I}}$, transactional dataset $\mathcal{D} \subseteq 2^{\mathcal{I}}$, $\epsilon > 0$, $\delta > 0$.
Output: approximated frequencies $\hat{f}_{\mathcal{D}}(A)$ for each $A \in \mathbf{Ob}$ s.t. with probability at least $1 - \delta$, $|\hat{f}_{\mathcal{D}}(A)| \leq \epsilon$ for any $A \in \mathbf{Ob}$.

1. $n \leftarrow \lceil \frac{1}{2}(\log(2|\mathbf{Ob}|) + \log(1/\delta)) \rceil$.
2. $\mathcal{S} \leftarrow \emptyset$.
3. If $n \geq |\mathcal{D}|$, $\mathcal{S} \leftarrow \mathcal{D}$; otherwise, choose n itemsets in $\mathcal{D}$ at uniformly random and add them to $\mathcal{S}$.
4. Label the transactions in $\mathcal{S}$: $\mathcal{S} = \{\tau_1, \cdots, \tau_n\}$.
5. For each $A \in \mathbf{Ob}$, compute $\hat{f}_{\mathcal{D}}(A) \leftarrow \frac{1}{n} \sum_{i=1}^{n} \phi_A(\tau_i)$.
6. Return all the $\hat{f}_{\mathcal{D}}(A)$'s for $A \in \mathbf{Ob}$.

Note that we do not have to estimate for any itemset which is out of the observed ones **Ob**.

5 Approximating Top-k Frequent Itemsets

In practice we often need to find out the top-k frequent itemsets among the given candidates **Ob**. We can slightly revise the algorithm given in the previous section to approximate the k most frequent itemsets. A new problem here is how to determine the approximation error ϵ, which should be less than the distance between the k-th and $(k+1)$-th most frequent itemsets. It is also possible that these two itemsets have the same frequency. Combining these cases, a progressive sampling approximation algorithm can be given as follows:

Top-k Frequent Itemsets Approximation Algorithm with Observed Items

Input: all the items $\mathcal{I}$, the observed items $\mathbf{Ob} \subseteq 2^{\mathcal{I}}$, transactional dataset $\mathcal{D} \subseteq 2^{\mathcal{I}}$, sampling increase Δn, $k > 0$, $\delta > 0$.
Output: k itemsets among **Ob** that have the highest frequencies with probability at least $1 - \delta$.

1. $n \leftarrow 0$. $\hat{f}_{\mathcal{D}}(A) \leftarrow 0$ for any $A \in \mathbf{Ob}$.
2. $\mathcal{S} \leftarrow \emptyset$.
3. If $n \geq |\mathcal{D}|$, $\mathcal{S} \leftarrow \mathcal{D}$ and go to Step 8; otherwise, choose Δn itemsets from $\mathcal{D}$ at uniformly random, and denote by $\Delta\mathcal{S}$ the chosen Δn itemsets.
4. Label the transactions in $\Delta\mathcal{S}$: $\Delta\mathcal{S} = \{\tau_1, \cdots, \tau_{\Delta n}\}$.
5. For each $A \in \mathbf{Ob}$, compute $\hat{f}_{\mathcal{D}}(A) \leftarrow \frac{1}{n+\Delta n} \left(n \cdot \hat{f}_{\mathcal{D}}(A) + \sum_{i=1}^{\Delta n} \phi_A(\tau_i) \right)$.
6. $n \leftarrow n + \Delta n$. $\mathcal{S} \leftarrow \mathcal{S} \cup \Delta\mathcal{S}$
7. $\epsilon \leftarrow \sqrt{\frac{\log(2|\mathbf{Ob}|) + \log(1/\delta)}{2n}}$.
8. If $\epsilon < d$, where d is the distance between the approximated frequencies $\hat{f}_{\mathcal{D}}(\cdot)$ of the k-th and $(k+1)$-th most frequent itemsets in **Ob**, or $\mathcal{S} = \mathcal{D}$, then return the top-$k$ results (tie-breaking when necessary). Otherwise, go to Step 3.

Note that in out top-k approximation algorithm, the stopping condition depends on the frequency distance

between the k-th and $(k+1)$-th most frequent itemsets. If these two frequencies tie, it is likely that many samples will be needed since we cannot distinguish them based on approximated frequencies. Hence we require the number of samples should not exceed $|\mathcal{D}|$. If $|\mathcal{D}|$ samples are needed, we can compute the exact frequencies and apply tie-breaking rules on the ranked results.

References

[1] Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. Mining association rules between sets of items in large databases. In *Acm sigmod record*, volume 22, pages 207–216. ACM, 1993.

[2] Stéphane Boucheron, Olivier Bousquet, and Gábor Lugosi. Theory of classification: A survey of some recent advances. *ESAIM: probability and statistics*, 9:323–375, 2005.

[3] Olivier Bousquet, Stéphane Boucheron, and Gábor Lugosi. Introduction to statistical learning theory. In *Advanced lectures on machine learning*, pages 169–207. Springer, 2004.

[4] Jiawei Han, Hong Cheng, Dong Xin, and Xifeng Yan. Frequent pattern mining: current status and future directions. *Data Mining and Knowledge Discovery*, 15(1):55–86, 2007.

[5] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

[6] Pascal Massart. Some applications of concentration inequalities to statistics. In *Annales de la Faculté des sciences de Toulouse: Mathématiques*, volume 9, pages 245–303, 2000.

[7] Matteo Riondato and Eli Upfal. Mining frequent itemsets through progressive sampling with rademacher averages. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1005–1014. ACM, 2015.

[8] Matteo Riondato and Eli Upfal. Abra: Approximating betweenness centrality in static and dynamic graphs with rademacher averages. *arXiv preprint arXiv:1602.05866*, 2016.

[9] Vladimir Vapnik. *Statistical learning theory*, volume 1. Wiley New York, 1998.

[10] Vladimir Vapnik. *The nature of statistical learning theory*. Springer Science & Business Media, 2013.