Неравенства концентрации вероятностной меры в трансдуктивном обучении и PAC-Байесовском анализе

Толстихин Илья Олегович

Диссертация на соискание ученой степени
кандидата физико-математических наук
05.13.17 — теоретические основы информатики

Научный руководитель — д. ф.-м. н. Воронцов Константин Вячеславович

Москва — 2014

Содержание

1. **Неравенства концентрации вероятностной меры**
   Новые неравенства для выборки без возвращения

2. **Теория статистического обучения**
   Локальные меры сложности в трансдуктивном обучении

3. **Комбинаторная теория переобучения**
   Точные оценки для подмножеств шара в Булевом кубе

4. **PAC-Байесовский анализ**
   PAC-Байесовское эмпирическое неравенство Бернштейна

Содержание

Неравенства концентрации вероятностной меры

- Измеримое пространство $(\mathcal{X}, \mathcal{A})$
- Измеримое отображение $g\colon \mathcal{X}^n \to \mathbb{R}$
- Последовательность случайных величин $X_1, \dots, X_n \subset \mathcal{X}$
- Случайная величина $Q = g(X_1, \dots, X_n)$

Задача: получить верхние оценки для

$$Q - \mathbb{E}[Q] \quad \text{и} \quad \mathbb{E}[Q] - Q,$$

справедливые с большой вероятностью.

- **Суммы** $S_n = \frac{1}{n}\sum_{i=1}^{n} X_n$ ограниченных с.в. $X_1, \dots, X_n \subset [0, 1]$.
- **Супремум эмпирических процессов**

$$Q_n = \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} f(X_i),$$

где $\mathcal{F}$ — счетный класс *измеримых* отображений $f\colon \mathcal{X} \to [-1, 1]$, таких что $\mathbb{E}[f(X_1)] = 0$.

Обзор: независимые случайные величины

Случай, когда с. в. $X_1, \ldots, X_n$ независимы, хорошо изучен.

Пусть $\delta \in [0, 1]$. Тогда с вер-ю не меньше $1 - \delta$:

- Для **сумм S_n**:

$$|S_n - \mathbb{E}[S_n]| \leq \sqrt{\frac{\log \frac{2}{\delta}}{2n}}; \qquad \text{(W. Hoeffding, 1963)}$$

$$|S_n - \mathbb{E}[S_n]| \leq \sqrt{\frac{2\,\mathbb{D}[X_1] \log \frac{2}{\delta}}{n}} + \frac{2 \log \frac{1}{\delta}}{3n}. \qquad \text{(С. Н. Бернштейн, 1911)}$$

- Для **супремумов эмпирических процессов Q_n**:

$$|Q_n - \mathbb{E}[Q_n]| \leq \sqrt{\frac{2 \log(2/\delta)}{n}}; \qquad \text{(C. McDiarmid, 1989)}$$

$$Q_n - \mathbb{E}[Q_n] \leq \sqrt{\frac{2v \log(1/\delta)}{n}} + \frac{2 \log(1/\delta)}{3n}, \qquad \text{(M. Talagrand, 1996)}$$

где $v = \sup_{f \in \mathcal{F}} \mathbb{D}[f(X_1)] + 2\mathbb{E}[Q_n]$.

Основная мысль: учет дисперсии может вести к *более точным оценкам*.

Обзор: выборки без возвращения

Пусть элементы $X_1, \ldots, X_n$ выбраны равномерно без возвращения из некоторого конечного множества $\mathcal{C} = \{c_1, \ldots, c_N\}$, где $N \geq n$.

Пример: процедуры скользящего контроля.

- **Суммы** S_n для $\mathcal{C} \subset [0,1]$:

  (Hoeffding, 1963) Неравенства Хефдинга и Бернштейна остаются справедливыми для S_n.

- **Супремумы эмпирических процессов** Q_n:

  (El-Yaniv, Pechyony, 2009) Для $\delta \in [0,1]$ с вер-ю не меньше $1 - \delta$:

$$|Q_n - \mathbb{E}[Q_n]| \leq \sqrt{\left(\frac{N-n}{N-1/2}\right) \frac{1}{\Delta(n,N)} \frac{2\log(2/\delta)}{n}},$$

где $\Delta(n,N) = 1 - \frac{1}{2\max\{n, N-n\}} \approx 1$.

Проблема: отсутствие аналога неравенства Талаграна для выборок без возвращения.

Новые результаты

Пусть $(X_i)_{i=1}^n$ и $(Z_i)_{i=1}^n$ выбраны равномерно без и с возвращением соответственно из конечного множества $\mathcal{C} = \{c_1, \ldots, c_N\}$.

$$Q_n = \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n f(X_i), \quad Q_n^{iid} = \sup_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^n f(Z_i), \quad \sigma_{\mathcal{F}}^2 = \sup_{f \in \mathcal{F}} \mathbb{D}[f(Z_1)].$$

Теорема (Толстихин, ИОИ 2012; Tolstikhin et al., COLT 2014)

Для всех $\delta \in [0,1]$ с вер-ю не меньше $1-\delta$:

$$|Q_n - \mathbb{E}[Q_n]| \le 2\sqrt{\frac{2\sigma_{\mathcal{F}}^2 \log(2/\delta)}{n} \left(\frac{N}{n}\right)}.$$

Теорема (Tolstikhin et al., COLT 2014)

Для всех $\delta \in [0,1]$ с вер-ю не меньше $1-\delta$:

$$Q_n - \mathbb{E}[Q_n^{iid}] \le \sqrt{\frac{2(\sigma_{\mathcal{F}}^2 + 2\mathbb{E}[Q_n^{iid}]) \log(1/\delta)}{n}} + \frac{\log(1/\delta)}{3n}.$$

Содержание

1. **Неравенства концентрации вероятностной меры**
   Новые неравенства для выборки без возвращения

2. **Теория статистического обучения**
   Локальные меры сложности в трансдуктивном обучении

3. **Комбинаторная теория переобучения**
   Точные оценки для подмножеств шара в Булевом кубе

4. **PAC-Байесовский анализ**
   PAC-Байесовское эмпирическое неравенство Бернштейна

Теория статистического обучения: индуктивная постановка

- Неизвестное вероятностное распределение P на $\mathcal{X} \times \mathcal{Y}$.
- Объекты обучающей выборки $S = \{(X_i, Y_i)\}_{i=1}^n$ н. о. р. из P.
- Эмпирический и средний риск

$$L_n(h) = \frac{1}{n} \sum_{i=1}^{n} \ell\big(h(X_i), Y_i\big), \quad L(h) = \mathbb{E}_{(X,Y)\sim P}\big[\ell\big(h(X), Y\big)\big]$$

для функции потерь $\ell \colon \mathcal{Y} \times \mathcal{Y} \to [0,1]$ и отображений $h \colon \mathcal{X} \to \mathcal{Y}$.

Задача обучения: минимизация $L(h)$ по классу $\mathcal{H}$ на основе S.

Минимизация эмпирического риска (МЭР):

$$\min_{h\in\mathcal{H}} L_n(h) = L_n(\hat{h}_n).$$

Нас интересуют верхние границы для $L(\hat{h}_n) - \inf_{h\in\mathcal{H}} L(h)$.

Теория Вапника-Червоненкиса (В. Н. Вапник, А. Я. Червоненкис, 1968):
Переход к рассмотрению равномерных по классу $\mathcal{H}$ отклонений:

$$L(\hat{h}_n) - \inf_{h\in\mathcal{H}} L(h) \le 2 \sup_{h\in\mathcal{H}} |L(h) - L_n(h)|.$$

Оценки в теории статистического обучения

Отказ от наложения дополнительных условий на распределение P и класс $\mathcal{H}$ ведет к **медленным скоростям сходимости** $O(n^{-1/2})$:

- (В. Н. Вапник, А. Я. Червоненкис, 1968):
  Бинарная функция потерь, VC-dimension
- (В. Н. Вапник, А. Я. Червоненкис, 1974, 1981):
  Ограниченные функции потерь, метрическая энтропия
- (Koltchinskii, Panchenko, 1999; Bartlett, Mendelson, 2001)
  Оценки на основе *глобальных Радемахеровских сложностей*

Эти результаты основаны на неравенствах Хефдинга и МакДиармида.

Однако, учет особенностей конкретных задач часто может вести к **быстрым скоростям сходимости** $o(n^{-1/2})$:

- (Mammen, Tsybakov, 1999):
  Условия малого шума: $|\,\mathsf{P}\{Y = 1|X\} - 0.5| > s$ для $s > 0$.
- (Massart, 2000; Bartlett et al., 2005; Koltchinskii, 2006):
  Локальный подход, локальные Радемахеровские сложности

Эти результаты основаны на неравенстве Талаграна.

Теория статистического обучения: трансдуктивная постановка

(В. Н. Вапник, 1998)

- n объектов $\boldsymbol{X}_n$ выбраны равномерно без возвращения из конечной генеральной совокупности $\boldsymbol{X}_N = \{X_i\}_{i=1}^N$.
- Получаем ответы $\boldsymbol{Y}_n$ для объектов $\boldsymbol{X}_n$ с помощью $\varphi\colon \mathcal{X} \to \mathcal{Y}$.
- Обозначим $\ell_h(X) = \ell\big(h(X), \varphi(X)\big)$ для $\ell\colon \mathcal{Y} \times \mathcal{Y} \to [0,1]$.
- Получаем обучающую выборку $S = (\boldsymbol{X}_n, \boldsymbol{Y}_n)$ и объекты контрольной выборки $\boldsymbol{X}_u = \boldsymbol{X}_N \setminus \boldsymbol{X}_u$, где $u = N - n$.
- $\hat{L}_n(h), L_u(h)$ и $L_N(h)$ — среднее выборочное значение риска h на выборках $\boldsymbol{X}_n, \boldsymbol{X}_u$ и $\boldsymbol{X}_N$ соответственно.

Задача обучения: поиск на основе S и $\boldsymbol{X}_u$ отображения $h \in \mathcal{H}$ с малым средним риском на контрольной выборке $\boldsymbol{X}_u$.

$$\hat{h}_n = \arg\min_{h\in\mathcal{H}} \hat{L}_n(h), \quad h_u^* = \arg\min_{h\in\mathcal{H}} L_u(h), \quad h_N^* = \arg\min_{h\in\mathcal{H}} L_N(h).$$

Проблема: отсутствие оценок для $L_u(\hat{h}_n) - L_u(h_u^*)$, ведущих к быстрым скоростям сходимости **в общих предположениях**.

Новые результаты

Рассмотрим $h \in \mathcal{H}$ с малыми дисперсиями избыточных потерь:

$$\mathcal{H}(r) = \left\{ h \in \mathcal{H} \colon \mathbb{E}\left[\left(\ell_h(X) - \ell_{h_N^*}(X) \right)^2 \right] \leq r \right\}.$$

Пусть $L_n^{iid}(h) = \frac{1}{n} \sum_{i=1}^n \ell_h(Z_i)$ для случайных величин $\{Z_i\}_{i=1}^n$, выбранных равномерно с возвращениями из $\boldsymbol{X}_N$.

Теорема (Tolstikhin et al., COLT 2014)

Пусть для некоторой константы $B > 0$ и для всех $h \in \mathcal{H}$ выполнено:

$$\mathbb{D}\left[\ell_h(X) - \ell_{h_N^*}(X) \right] \leq B \cdot \left(L_N(h) - L_N(h_N^*) \right).$$

Пусть, кроме того, существует подкоренная функция $\psi_n(r)$, такая что:

$$B \cdot \mathbb{E}\left[\sup_{h \in \mathcal{H}(r)} L_N(h) - L_n^{iid}(h) - \left(L_N(h_N^*) - L_n^{iid}(h_N^*) \right) \right] \leq \psi_n(r).$$

Пусть r_n^* — неподвижная точка $\psi_n(r)$. Тогда с вер-ю не меньше $1 - \delta$:

$$L_N(\hat{h}_n) - L_N(h_N^*) \leq \frac{901}{B} r_n^* + (16 + 25B) \frac{\log(1/\delta)}{3n}.$$

Новые результаты

Теорема (Tolstikhin et al., COLT 2014)

В условиях прошлой теоремы, с вер-ю не меньше $1-\delta$:

$$L_u(\hat{h}_n) - L_u(h_u^*) \le \frac{901N}{B}\left(\frac{r_n^*}{u} + \frac{r_u^*}{n}\right) + \frac{(32+50B)N\log(2/\delta)}{3nu}.$$

Известны используемые на практике случаи, когда $r_n^* = o(n^{-1/2})$:

- (Massart, 2000)
  Для задач с бинарной функцией потерь и $\mathrm{VCD}(\mathcal{H}) = d < \infty$ справедливо $r_n^* \sim \frac{d\log n}{n}$.

Условия теорем выполнены, например, в следующих случаях:

- (Lee, Bartlett, Williamson, 1998)
  Для задач с равномерно ограниченным выпуклым классом $\mathcal{H}$ и квадратичной функцией потерь $\ell(y', y'') = (y' - y'')^2$;
- (Mammen, Tsybakov, 1999)
  Для задач с бинарной функцией потерь при $\varphi \in \mathcal{H}$.

Содержание

Комбинаторная теория переобучения

Основная задача: получение точных оценок вер-ти переобучения

$$\mathsf{P}\Bigl\{ L_u\bigl(\mu(S)\bigr) - \hat{L}_n\bigl(\mu(S)\bigr) \geq t \Bigr\}$$

для различных методов обучения $\mu\colon S \to \mathcal{H}$ и семейств $\mathcal{H}$.

Главные отличия от трансдуктивного обучения и VC-теории:

- Отказываемся от рассмотрения чрезмерно общих постановок.
- Отказываемся от часто завышенных неравенств концентрации.
- Учет «геометрии» множества **бинарных** векторов ошибок отображений из $\mathcal{H}$ на **конечной** генеральной совокупности $\boldsymbol{X}_N$:

$$A = \Bigl\{ \bigl([h \text{ ошибается на } X_i]\bigr)_{i=1}^{N},\ h \in \mathcal{H} \Bigr\} \subseteq \{0,1\}^N.$$

Главные результаты и подходы:

- Оценки на основе порождающих и запрещающих множеств
- Оценки расслоения-связности
- **Теоретико-групповой подход** (А. И. Фрей, 2009)

Новые результаты

Для $a \in A$ и $\boldsymbol{X} \subseteq \boldsymbol{X}_N$ обозначим:

$$L(a, \boldsymbol{X}) = \frac{1}{|\boldsymbol{X}|} \sum_{X \in \boldsymbol{X}} I(a, X), \quad A(\boldsymbol{X}) = \arg\min_{a \in A} L(a, \boldsymbol{X}).$$

Рандомизированный метод МЭР μ на основе обучающей выборки $\boldsymbol{X}_n$ выбирает вектор ошибок случайно и равномерно из $A(\boldsymbol{X}_n)$.

Симметрическая группа перестановок S_N и подгруппа симметрий A:

$$S(A) = \{\pi \in S_N \colon \pi A = A\}.$$

Γ и Ω: множества орбит действия $S(A)$ на A и $\{\boldsymbol{X} \subset \boldsymbol{X}_N \colon |\boldsymbol{X}| = n\}$.

Теорема (Толстихин, ИОИ 2010; Фрей, Толстихин, ДАН 2014, JMLDA 2014)

Для рандомизированного МЭР μ справедливо:

$$\mathsf{P}\left\{L\bigl(\mu(S_n), X_u\bigr) - L\bigl(\mu(S_n), X_n\bigr) \geq t\right\} =$$

$$\sum_{\gamma \in \Gamma} \sum_{\omega \in \Omega} \frac{|\gamma|}{C_N^n} \bigl|\{\boldsymbol{X} \in \omega \colon a_\gamma \in A(\boldsymbol{X})\}\bigr| \frac{\bigl[L(a_\gamma, \boldsymbol{X}_N \backslash \boldsymbol{X}_\omega) - L(a_\gamma, \boldsymbol{X}_\omega) \geq t\bigr]}{|A(\boldsymbol{X}_\omega)|},$$

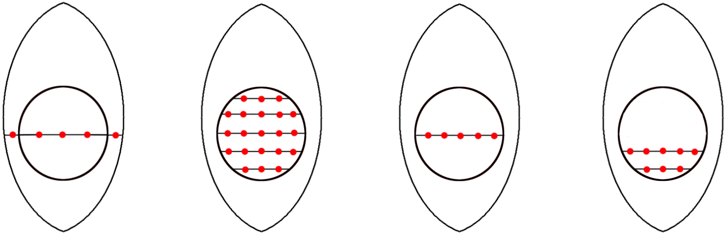
где a_γ и $\boldsymbol{X}_\omega$ — произвольные элементы орбит γ и ω.

Новые результаты

Теорема (Толстихин, ИОИ 2010; Фрей, Толстихин, ДАН 2014, JMLDA 2014)

Получены точные оценки для следующих модельных семейств A:

- *Хэммингов шар;*
- *Центральный слой Хэммингова шара;*
- *d нижних слоев Хэммингова шара;*



слой | шар | слой шара | нижние слои шара

Содержание

1. **Неравенства концентрации вероятностной меры**
   Новые неравенства для выборки без возвращения
2. **Теория статистического обучения**
   Локальные меры сложности в трансдуктивном обучении
3. **Комбинаторная теория переобучения**
   Точные оценки для подмножеств шара в Булевом кубе
4. **PAC-Байесовский анализ**
   PAC-Байесовское эмпирическое неравенство Бернштейна

PAC-Байесовский анализ (**P**robably **A**pproximately **C**orrect)

PAC-Байесовский анализ изучает стохастические предикторы в индуктивной постановке теории статистического обучения:

- Рассмотрим вероятностное распределение ρ на $\mathcal{H}$
- Для каждого нового $X \in \mathcal{X}$ стохастический предиктор G_ρ:
  1. Выбирает $h \sim \rho$ независимо от X;
  2. Возвращает ответ $h(X)$.
- Для краткости будем обозначать $\langle \cdot, \rho \rangle = \mathbb{E}_{h\sim\rho}[\cdot]$.
- Определим средний и эмпирический риск предиктора G_ρ:

$$L(G_\rho) = \langle L(h), \rho \rangle \,, \quad L_n(G_\rho) = \langle L_n(h), \rho \rangle \,.$$

- **Задача обучения:** поиск на основе обучающей выборки S распределения ρ с малым средним риском $L(G_\rho)$.

Пусть $\mathrm{KL}(\rho \| \pi) = \left\langle \frac{\rho(h)}{\pi(h)}, \rho \right\rangle$ — дивергенция Кульбака–Лейблера между двумя распределениями ρ и π на $\mathcal{H}$.

Обзор: PAC-Байесовские неравенства

Для любого *не зависящего от* S_n распределения π на $\mathcal{H}$, любых $c > 1$ и $\delta > 0$ с вероятностью не меньше $1 - \delta$ (относительно реализации обучающей выборки S_n) следующие неравенства:

Н-во МакАллистера (McAllester, 1998)

$$L(G_\rho) \le L_n(G_\rho) + \frac{1+c}{2}\sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \ln(1/\delta)}{2n}};$$

PAC-Байесовское н-во Бернштейна (Seldin et al., 2012)

$$L(G_\rho) \le L_n(G_\rho) + (1+c)\sqrt{\left\langle \mathbb{D}_{(X,Y)\sim P}[\ell_h(X,Y)], \rho \right\rangle \frac{\mathrm{KL}(\rho\|\pi) + \ln(1/\delta)}{n}};$$

PAC-Байесовское kl-неравенство (Seeger, 2002)

$$L(G_\rho) \le L_n(G_\rho) + \sqrt{2L_n(G_\rho)\frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{n+1}{\delta}}{n}} + \frac{\mathrm{KL}(\rho\|\pi) + \ln\frac{n+1}{\delta}}{n};$$

справедливы одновременно для всех распределений ρ на $\mathcal{H}$.

Обзор: применение PAC-Байесовских неравенств

$$\mathsf{P}\left\{\forall\rho \text{ на } \mathcal{H}\colon L(G_\rho) \le L_n(G_\rho) + C\sqrt{\frac{\mathrm{KL}(\rho\|\pi) + \ln(1/\delta)}{n}}\right\} \ge 1-\delta$$

- Минимизация оценки по ρ ведет к новым алгоритмам обучения с гарантированной обобщающей способностью.
- (Germain et al., 2014)
  Получение оценок для композиций классификаторов.
- (Langford, Shawe-Taylor, 2002)
  Часто средний риск $L(h)$ отображений из $\mathcal{H}$ удается оценить сверху средним риском $L(G_\rho)$ со специально подобранным ρ.
- PAC-Байесовские неравенства ведут к **наиболее точным** на сегодняшний день оценкам для линейных классификаторов.

Часто PAC-Байесовское неравенство Бернштейна существенно точнее kl-неравенства.

Проблема: дисперсия $\left\langle \mathbb{D}_{(X,Y)\sim P}[\ell_h(X,Y)], \rho \right\rangle$ нам неизвестна.

Новые результаты

Рассмотрим выборочную дисперсию потерь отображения h:

$$\mathbb{D}_n(h) = \frac{1}{n-1}\sum_{i=1}^{n}\bigl(\ell_h(X_i, Y_i) - L_n(h)\bigr)^2.$$

Теорема (Tolstikhin, Seldin, NIPS 2014)

Для любого не зависящего от S_n распределения π на $\mathcal{H}$ и любых $c > 1$, $\delta > 0$ с вер-ю не меньше $1-\delta$:

$$\Bigl\langle \mathbb{D}_{(X,Y)\sim P}[\ell_h(X,Y)], \rho \Bigr\rangle$$

$$\leq \langle \mathbb{D}_n(h), \rho\rangle + (1+c)\sqrt{\frac{\langle \mathbb{D}_n(h), \rho\rangle \left(\mathrm{KL}(\rho\|\pi) + \ln\frac{T}{\delta}\right)}{2(n-1)}} + \frac{2c\left(\mathrm{KL}(\rho\|\pi) + \ln\frac{T}{\delta}\right)}{n-1}$$

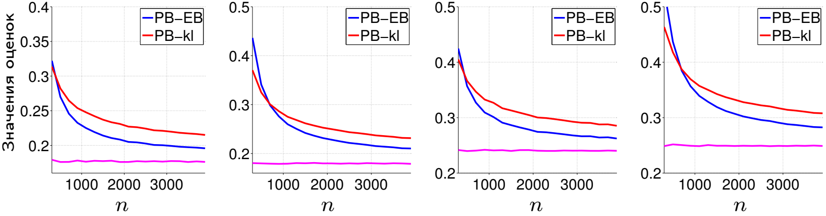
одновременно для всех ρ, где

$$T = \left\lceil \frac{1}{\ln c}\ln\left(\frac{1}{2}\sqrt{\frac{n-1}{\ln(1/\delta)}+1}+\frac{1}{2}\right)\right\rceil.$$

Новые результаты: эксперименты

Эксперименты на модельных выборках

- Задача регрессии $\mathcal{Y} = \mathbb{R}$ с классом линейных функций $\mathcal{H}$.
- Абсолютные потери $\ell(y', y'') = |y' - y''|$.



Эксперименты на выборках из репозитория UCI

Выборка	n	d	Тест	оценка PB-kl	оценка PB-EB
winequality	6497	11	0.106 ± 0.0022	0.175 ± 0.0006	$\mathbf{0.162 \pm 0.0006}$
parkinsons	5875	16	0.188 ± 0.0055	0.266 ± 0.0013	$\mathbf{0.250 \pm 0.0012}$
concrete	1030	8	0.111 ± 0.0038	$\mathbf{0.242 \pm 0.0010}$	0.264 ± 0.0011

Список публикаций по теме диссертации

1. **I. Tolstikhin**, G. Blanchard, M. Kloft. Localized Complexities for Transductive Learning. // *Conference on Learning Theory (COLT)*, 2014.
2. **I. Tolstikhin**, Y. Seldin. PAC-Bayes-Empirical-Bernstein inequality. // *Advances in Neural Information Processing Systems 26 (NIPS)*, 2013. pp. 109–117.
3. Фрей А. И., **Толстихин И. О.**. Комбинаторные оценки вероятности переобучения на основе покрытий множества алгоритмов. //*Доклады РАН*, Т. 455, №. 3, 2014. С. 265-268.
4. Фрей А. И., **Толстихин И. О.**. Комбинаторные оценки вероятности переобучения на основе кластеризации и покрытий множества алгоритмов. // *Machine Learning and Data Analysis*, V. 1, №. 7, 2014. pp. 761-778.
5. Толстихин И. О. Локализация оценок избыточного риска в комбинаторной теории переобучения. // *Межд. конф. ИОИ-9, Москва, Торус Пресс*, 2012, С. 54-57.
6. Толстихин И. О. Вероятность переобучения плотных и разреженных семейств алгоритмов. // *Межд. конф. ИОИ-8, Москва, МАКС Пресс*, 2010, С. 83-86.
7. Толстихин И. О. Точная оценка вероятности переобучения для одного специального семейства алгоритмов. // *Междунар. конф. «Ломоносов 2010», Москва*, 2010, pp. 98-99.

Результаты, выносимые на защиту

1. Получено два новых неравенства типа Талаграна для выборок без возвращения.
2. В задаче трансдуктивного обучения получены первые оценки, основанные на локальных мерах сложности и ведущие к быстрым скоростям сходимости в общих предположениях.
3. В теоретико-групповом подходе комбинаторной теории переобучения предложено рассматривать орбиты действия группы симметрий на множестве разбиений генеральной выборки. На основе этого подхода получены точные оценки вероятности переобучения для трех модельных задач.
4. В PAC-Байесовском анализе теории статистического обучения получено новое PAC-Байесовское эмпирическое неравенство Бернштейна, полностью вычислимое на основе обучающей выборки и во многих случаях ведущее к существенно более точным оценкам по сравнению с другими известными неравенствами.