ЗАСТОСУВАННЯ ГІПЕРПАРАМЕТРІВ У ПРОБЛЕМІ ВИБОРУ МОДЕЛІ НА ОСНОВІ ПРИНЦИПУ ІНДУКЦІЇ СТРУКТУРНОЇ МІНІМІЗАЦІЇ РИЗИКУ

Олександр Галкін

Національний транспортний університет,
01010, вул. Суворова 1, Київ, Україна; e-mail: oleksandr.galkin@mail.ru

APPLICATION OF THE HYPERPARAMETERS IN THE MODEL SELECTION PROBLEM BASED ON THE STRUCTURAL RISK MINIMIZATION INDUCTION PRINCIPLE

Oleksandr Galkin

National Transport University,
01010, Syvorova st., 1, Kyiv, Ukraine

АНОТАЦІЯ. Об'єктом дослідження є методологія вибору моделі, як одна із проблем інтелектуального аналізу даних. Метою дослідження є вивчення проблеми перенавчання, коли доступною є множина гіперпараметрів або моделей. В статті розглядається методологія оптимізації гіперпараметрів з використанням методу градієнтного спуску, а також проблема перенавчання у виборі моделі з використанням критерію помилки перевірки.
Ключові слова: *вибір моделі, проблема перенавчання, оптимізація гіперпараметрів.*

АННОТАЦИЯ. Объектом исследования является методология выбора модели, как одна из проблем интеллектуального анализа данных. Целью исследования является изучение проблемы переобучения, когда доступной является множество гиперпараметров или моделей. В статье рассматривается методология оптимизации гиперпараметров с использованием метода градиентного спуска, а также проблема переобучения в выборе модели с использованием критерия ошибки проверки.
Ключевые слова: *выбор модели, проблема переобучения, оптимизация гиперпараметров.*

SUMMARY. ***Purpose.*** *When a lot of models are available, risk bounds become larger and not taking this fact into account can result in an overfitting phenomenon in the model selection stage. The object of study is the methodology of model selection as one of the problems of data mining. The aim of the investigation is to study the problem of over-fitting when a set of hyperparameters or models is available.* ***Methodology/approach.*** *The paper contains the methodology of optimizing of hyperparameters using the gradient descent and also the problem of over-fitting in model selection using the criterion of validation error.* ***Findings.*** *It is possible to apply the theoretical results for the investigation and estimation the over-fitting problem.* ***Research limitations/implications.*** *The present study provides a starting point to the problem of overfitting in model selection using a validation error criterion.* ***Originality/value.*** *The presented results allow conclude that the additional estimation error grows as the square root of the number of hyperparameters.*
Key words: *model selection, over-fitting problem, optimizing hyperparameters.*

Подано 10.06.2013; прийнято 16.07.2013

ВСТУП

В рамках статистичної теорії навчання, навчальні дані генеруються незалежно з ідентичним розподілом з деякого невідомого розподілу $P(x,y)$, в якому закодована залежність між входом x та виходом y. Тобто, якщо вхідні дані відповідають ймовірнісному розподілу $P(x)$, а вихідними даними є функція $f(x)$, що знаходиться під впливом дисперсії гауссівського шуму σ^2, тоді:

$$P(x,y) = P(x)\mathrm{N}_\sigma(f(x)-y),$$

де N_σ є розподілом Гаусса із щільністю $\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{t^2}{2\sigma^2}\right)$.

Якість функції, яка моделює дане співвідношення, вимірюється шляхом використання очікування функції втрат відносно $P(x,y)$. Розглядаючи задачу класифікації даних, функція втрат буде мати наступний вигляд:

$$\ell(f(x),y) = \mathrm{I}_{f(x)\neq y},$$

де I є функцією-показником, тобто $\mathrm{I}_A = 1 \Leftrightarrow A$ є істинним. Оскільки, одне з припущень в рамках даного дослідження є те, що дані, вихідні значення яких повинні бути заздалегідь передбачені, також гене-

рується з того ж самого розподілу, метою статистичного навчання є знаходження наступної функції мінімізації втрат:

$$R(f) = \int \ell(x, f(x)) dP(x, y). \qquad (1)$$

Проблема навчання зводиться до пошуку функції $f \in F$, що мінімізує очікування (1) функції-показника втрат $\ell(f(x), y) = \mathrm{I}_{f(x) \neq y}$. Дане очікування не може бути обчислено, оскільки розподіл $P(x, y)$ є невідомим. Однак, враховуючи навчальну множину $\{(x_i, y_i)\}_{1 \le i \le n}$, ми можемо мінімізувати *емпіричний ризик*:

$$R_{емп}(f) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i), y_i). \qquad (2)$$

МЕТА ДОСЛІДЖЕННЯ

Використовуючи принцип індукції емпіричної мінімізації ризику, вибір множини функцій F має вирішальне значення: якщо дана множина буде надто великою, ми можемо зіткнутися з проблемою *перенавчання*.

ВИКЛАД МАТЕРІАЛУ

У даній статті ми розглядаємо методологію вибору моделі, як одну із проблем інтелектуального аналізу даних. Розглядаючи задачу класифікації даних, F повинно бути обмежено для того, щоб мати відповідну складність. Вибір класу функцій F називається *вибором моделі*.

Вибір моделі є досить складною проблемою, однак в рамках теорії Вапника-Червоненкіса (ВЧ) можна спробувати знайти клас функцій, який мінімізує границю наступного твердження.

Твердження 1. Нехай F є класом функцій розмірності ВЧ h, тоді для будь-якого розподілу P та деякої вибірки $\{(x_i, y_i)\}_{1 \le i \le n}$, що отримана з цього розподілу, справедливою є наступна нерівність з ймовірністю $1-\eta$:

$$\forall f \in F,\ R(f) \le R_{емп}(f) + \sqrt{\frac{h\left(\log\frac{2n}{h}+1\right) - \log\left(\frac{\eta}{4}\right)}{n}} + \frac{1}{n}.$$

Це є ідеєю принципу індукції структурної мінімізації ризику, що представлена на рисунку 1.

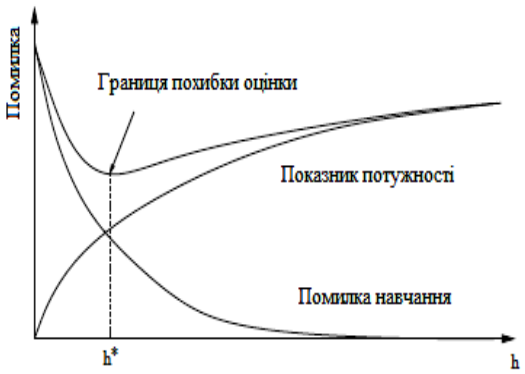

Рис. 1. Гіперплощина відповідає вирівнюванню точок у верхній частині рисунка

Fig. 1. Hyperplane corresponds to the alignment points at the top of the figure

Розглянемо сімейство класу функцій F_i, кожна з яких має розмірність ВЧ h_i. У традиційному формулюванні структурної мінімізації ризику [1], класи функцій є вкладеними ($F_i \subset F_{i+1}$), що призводить до наступного.що призводить до підвищення складності ($h_i \ge h_{i+1}$). Однак, дане припущення не є необхідним.

Нехай для даного i, f_n^i є функцією мінімізації емпіричного ризику по F_i. З твердження 1 ми маємо, що з ймовірністю $1-\eta_i$

$$R(f_n^i) \le R_{емп}(f_n^i) + \sqrt{\frac{\varphi(h_i) - \log(\eta_i / 4)}{n}} + \frac{1}{n}, \qquad (3)$$

де $\varphi(h)$ є показником складності,

$$\varphi(h) = h\left(\log\frac{2n}{h}+1\right).$$

Для одного класу функцій ми зафіксували одне значення η_i. Зафіксуємо рівняння (3) рівномірним по i: з ймовірністю $1-\sum\eta_i$,

$$\forall i,\ R(f_n^i) \le R_{емп}(f_n^i) + \sqrt{\frac{\varphi(h_i) - \log(\eta_i / 4)}{n}} + \frac{1}{n}. \qquad (4)$$

Припустимо, що існує p класів функцій $F_1, ..., F_p$, а також виберемо $\eta_i = \eta / p$.

Якщо $\hat{i}$ є моделлю, що вибрана за принципом системної мінімізації ризику, тобто $\hat{i}$ мінімізує праву частину рівняння (3), тоді з рівняння (4) ми маємо, що з ймовірністю $1-\eta$

$$R(f_n^{\hat{i}}) \le \min_{1\le i\le p} R_{емп}(f_n^i) + \\ +\sqrt{\frac{\varphi(h_i)+\log(p)-\log(\eta_i/4)}{n}}+\frac{1}{n}. \quad (5)$$

Додатковим фактором є той факт, що протестованим моделям p відповідає лише $\log(p)$. Якщо значення p є малим, використання цього фактору не має сенсу. Однак, коли кількість моделей є експоненціально великою або навіть нескінченною, необхідно встановити для моделі заздалегідь визначену вагу шляхом вибору постійних величин η_i. Для нескінченного числа моделей, двома можливими варіантами є $\eta_i = \eta 2^{-i}$ або $\eta_i = \eta 2^{-i}/(\pi^2/6)$. Перший варіант призводить до більшого відхилення в сторону першої моделі, оскільки показник $\log(\eta_i)$ в рівнянні (3) є лінійним у першому випадку та логарифмічним в другому випадку.

Як уже було зазначено, у випадку коли доступними є багато моделей, границі ризику стають більшими. Ігнорування даного факту може призвести до феномену перенавчання на стадії вибору моделі. Для вивчення цієї проблеми спочатку узагальнимо процедуру вибору моделі.

У загальному випадку, параметри α та гіперпараметри θ алгоритму навчання повинні бути певним чином оцінені. У випадку емпіричної мінімізації ризику, гіперпараметри θ відповідають за вибір класу функцій F, тоді як параметри α відповідають за опис самої функції в класі.

Процес навчання є процедурою, що складається з двох етапів:

1. При фіксованому значенні θ, необхідно знайти найкращі параметри α^0,
$$\alpha^0(\theta) = \arg\min_{\alpha} T(\alpha,\theta).$$
2. Знайти найкраще значення θ,
$$\theta^0 = \arg\min_{\theta} V(\alpha^0(\theta),\theta).$$

Перший етап полягає в класичній мінімізації ризику: модель, що описана θ є фіксованою, а емпірична мінімізація ризику є випадком, коли T є емпіричною помилкою. Вибір моделі виконується на другому етапі (вибір θ). Для структурної мінімізації ризику, V є верхньою границею, що представлена в твердженні 1. Зауважимо, що виконання класичної моделі вибору може зайняти багато часу, оскільки для кожного протестованого значення θ вимагається мінімізація по α.

Як правило, критерієм V є оцінка або верхня границя помилки узагальнення, що призводить до прямої залежності від виконання алгоритму. Тим не менш, досить ефективним критерієм є помилка перевірки. Цей критерій не залежить від алгоритму навчання, а також досить легко обчислюється. Враховуючи цей факт, даний критерій буде використовуватися нами в подальших дослідженнях.

Припустимо, що моделі $\theta_1,...,\theta_p$ знаходяться в режимі тестування. Для кожної моделі θ_i застосовується алгоритм навчання та виводиться функція f_i. Нехай F^* є множиною функцій $\{f_1,...,f_p\}$. Крок вибору моделі полягає у виборі найкращої функції в F^* за допомогою критерію вибору моделі V.

Припустимо, що V є помилкою перевірки:

$$R_{перев}(f) = \frac{1}{n'}\sum_{i=1}^{n'} \ell(f(x_i'), y_i'),$$

де $\{(x_i', y_i')\}_{1\le i\le n'}$ є незалежною вибіркою, що взята з того ж розподілу, що і навчальна множина.

Оскільки для всіх $f \in F^*$, $R_{перев}(f)$ є незміщеною оцінкою істинного ризику $R(f)$, стандартним способом виконання вибору моделі є вибір функції f_i, що мінімізує $R_{перев}(f)$.

Емпіричний ризик $R_{емп}$ не є незміщеною оцінкою істинного ризику, оскільки функції в F^* вибираються з використанням навчальних прикладів. У цьому і полягає причина, чому "невидимі" приклади

необхідні для того, щоб мати незміщену оцінку ризику.

Нехай f^* є мінімізатором помилки перевірки, тобто

$$f^* = \arg\min_{f \in F^*} R_{перев}(f).$$

Як і для емпіричної мінімізації ризику, $R_{навч}(f^*)$ не є незміщеною оцінкою $R(f^*)$. Для визначення верхньої границі $R(f^*)$, ми повинні мати єдиний аргумент збіжності. Для цього виведемо верхню границю:

$$P\left\{\sup_{f \in F^*} |R(f) - R_{навч}(f)| > \varepsilon\right\}.$$

Використовуючи нерівність Хефдінга [2], ми стверджуємо, що

$$\forall f \in F^*,\ \mathrm{P}\left\{|R(f) - R_{навч}(f)| > \varepsilon\right\} < 2\exp(-2n'\varepsilon^2).$$

Оскільки потужність множини F^* дорівнює p, об'єднання границь призводить до наступного:

$$P\left\{\sup_{f \in F^*} |R(f) - R_{навч}(f)| > \varepsilon\right\} \le 2p\exp(-2n'\varepsilon^2).$$

З цього ми можемо зробити висновок, що з ймовірністю $1-\eta$, ми маємо:

$$R(f^*) \le R_{перев}(f) + \sqrt{\frac{\log(p) - \log(\eta/2)}{2n'}}. \quad (6)$$

Поки значення p є не надто великим, $R_{навч}(f)$ буде ефективною оцінкою $R(f)$, а зведення до мінімуму помилки перевірки буде мати сенс. Однак, якщо значення p є великим (тобто $\log(p)$ має порядок n'), може виникнути проблема перенавчання. Дану проблему можна порівняти з класом функцій, що є досить великим при проведенні емпіричної мінімізації ризику. Два етапи є фактично еквівалентними: етап вибору моделі полягає в проведенні емпіричної мінімізації з використанням множини перевірки на множині F^*. Таким чином, якщо значення F^* є надто великим (тобто існує надто багато моделей), перенавчання буде відбуватися під час цього етапу [3].

Використовуючи результати попередніх досліджень, оптимізація параметрів моделі займає багато часу, а використання різних моделей має непомірно високу обчислювальну складність. Використовуваний нами підхід полягає в оптимізації гіперпараметрів з використанням методу градієнтного спуску. У цьому випадку, як і для емпіричної мінімізації ризику, значення p в рівнянні (6) не повинно бути числом моделей, що проходять тестування (наприклад, число кроків градієнта). Однак, значення p може бути числом різних можливих значень гіперпараметрів, яке є дуже великим [4].

При намаганні оптимізувати m гіперпараметрів $(\beta_1, ..., \beta_m)$ на множині перевірки, кожен з яких може приймати q різних значень, число функцій в F^* буде дорівнювати $p = q^m$, а рівняння (6) буде вказувати на те, що додаткова похибка оцінки з етапу вибору моделі буде мати порядок $\sqrt{m/n'}$, тобто можна записати, що

$$R(f^*) \le R_{перев}(f) + \sqrt{\frac{m\ln(q) - \ln(\eta/2)}{2n'}}. \quad (7)$$

У випадку, якщо гіперпараметри приймають неперервні значення, побудова верхньої границі є більш складним завданням, однак це можна зробити за допомогою чисел покриття та традиційних границь ВЧ.

Оскільки вибір моделі представляє собою виконання емпіричної мінімізації ризику на F^* з використанням множини перевірки, можна зробити висновок, що $R_{емр}(f)$ може бути замінено на $R_{перев}(f)$, а h буде вказувати на розмірність ВЧ F^*.

Однак розмірність F^* взагалі неможливо обчислити, оскільки функції в цій множині є розв'язками задачі оптимізації. З інтуїтивної точки зору, необхідно провести заміну розмірності ВЧ F^* на число гіперпараметрів m, як і в дискретному випадку. Це можна зробити з припущенням того, що функції в F^* "гладко" залежать від параметрів моделі θ. Приведемо лише скорочене доведення.

Нехай Ω є множиною можливих значень параметрів моделі. Для кожного з них, $f_\theta = f_{\alpha^0(\theta)}$ є оптимальною функцією в

моделі θ, тобто мінімізацією навчального критерію T. Ми маємо, що

$$F^* = \{ f_\theta,\ \theta \in \Omega \}.$$

Будемо вважати, що f_θ суттєво не змінюється з θ, тобто виконується наступна умова Ліпшиця:

$$\forall(\theta, \theta') \in \Omega^2,\ \left\| f_\theta - f_{\theta'} \right\|_\infty \le C \left\| \theta - \theta' \right\|_\infty .$$

Для надання сенсу такому припущенню, ми повинні розглянути клас дійсних функцій. Число покриття F^* може бути обмежено постійною величиною (залежно від C), помноженою на число покриття Ω. Нарешті, стандартні результати класифікації даних з використанням дійсних функцій забезпечують оцінки помилки узагальнення в контексті чисел покриття та похибок поля [5-10].

ВИСНОВКИ

Підводячи підсумки викладеного матеріалу, зазначимо, що нашим завданням було визначити в чому полягає небезпека перенавчання, коли доступними є багато гіперпараметрів (моделей). Встановлено, що додаткова похибка оцінки зростає відповідно до квадратного кореня з числа гіперпараметрів. Зауважимо, що в даній статті ми вивчали проблему перенавчання у виборі моделі з використанням критерію помилки перевірки, однак те ж саме відбувається з будь-яким визначеним критерієм. Використовуючи отримані нами результати експериментальних досліджень для проблеми вибору характеристик з використанням методу структурної мінімізації ризику, можна зробити висновок, що коли число вибраних характеристик є великим, границя очікуваного ризику стає більшою.

REFERENCES

1. **Vapnik V., Chervonenkis A., 1974.** Theory of Pattern Recognition. Nauka Publ.
2. **Hoeffding W., 1963.** Probability inequa-lities for sums of bounded random variables. Journal of the American Statistical Association, no. 58(301), 13-30.
3. **Ng A.Y. 1997.** Preventing over-fitting of cross-validation data. In Proceedings of the 14th International Conference on Machine Learning. Morgan Kaufmann.
4. **Bengio Y. 2000.** Gradient-based optimization of hyper-parameters. Neural Computation, no. 12(8).
5. **Bradley P.S., Mangasarian O.L., 1998.** Feature selection via concave minimization and support vector machines. In Proc. 13th International Conference on Machine Learning. San Francisco, CA, 82-90.
6. **Guyon I., Weston J., Barnhill S., Vapnik V., 2002.** Gene selection for cancer classification using support vector machines. Machine Learning, vol. 46(1/3), 289.
7. **Jebara T., Jaakkola T., 2000.** Feature selection and dualities in maximum entropy discrimination. In Uncertainty In Artificial Intelligence.
8. **LeCun Y., Denker J., Solla S., Howard R.E., Jackel L.D., 1990.** Optimal brain damage. Advances in Neural Information Processing Sustems II, San Mateo, CA. Morgan Kauffman.
9. **Bartlett P., Shawe-Taylor J., 1999.** Generalization performance of support vector machines and other pattern classfiers. In Scholkopf B., Burges C., Smola A., editors. Advances in Kernel Methods – Support Vector Learning. MIT Press, Cambridge, MA.
10. **Hsu C.-W. and Lin C.-J. 2001.** A simple decomposition method for support vector machines. To appear in Machine Learning.