

## ИСПОЛЬЗОВАНИЕ $\varepsilon$ -СЕТЕЙ ДЛЯ ЛИНЕЙНОГО РАЗДЕЛЕНИЯ ДВУХ МНОЖЕСТВ В ПРОСТРАНСТВЕ $R^d$

**Аннотация.** Введено понятие  $\varepsilon$ -разделимости двух множеств. Доказаны необходимые и достаточные условия  $\varepsilon$ -разделимости, а также сведение задачи  $\varepsilon$ -разделения двух множеств к задаче разделения их  $\varepsilon$ -сетей, которые не пересекаются.

**Ключевые слова:** эpsilon-сети, разделение множеств, размерность Вапника–Червоненкиса.

### ПОСТАНОВКА ЗАДАЧИ

Пусть заданы два конечных множества:  $A \subset R^d$  и  $B \subset R^d$ , мощности которых  $|A|=n_A$ ,  $|B|=n_B$ . Предположим, что  $A \not\subset \text{conv} B$ ,  $B \not\subset \text{conv} A$ . В простейшем случае, если выпуклые оболочки множеств  $A \subset R^d$  и  $B \subset R^d$  не пересекаются, их можно разделить, т.е. найти гиперплоскость, относительно которой данные множества будут находиться по разные стороны. Предположим, что множества неразделимы, т.е.  $\text{conv}(A) \cap \text{conv}(B) \neq \emptyset$ . Возникает вопрос, в каком случае данные множества можно разделить, исключив из них небольшое количество точек, например  $\varepsilon \in (0, 1)$  частей от общего количества.

На данный момент существует большое количество методов классификации, каждый из которых имеет преимущества и недостатки. Наиболее популярный из них — дискриминантный анализ Фишера [1]. Он широко используется в таких отраслях информатики, как машинное обучение, поиск информации и распознавание образов. Сложность алгоритма линейного дискриминантного анализа оценивается как  $O(ndt + t^3)$ , где  $n$  — количество наблюдений в обучающей выборке,  $d$  — количество признаков,  $t = \min(n, d)$  [2]. Поэтому при больших значениях  $n$  и  $d$  алгоритм использовать невозможно.

Байесовский классификатор [3] оптимальный, легко реализуется программно, на его основе построено много методов классификации. Однако поскольку на практике функции правдоподобия классов восстанавливают по конечным выборкам данных, байесовский классификатор перестает быть оптимальным [4]. Его сложность алгоритма оценивается как  $O(nd)$  [5].

Сравнительно новый метод опорных векторов, известный в литературе как SVM [6] благодаря принципу оптимальной разделяющей гиперплоскости, приводит к максимизации ширины разделяющей полосы между классами. Таким образом, этот метод способствует более уверенной классификации. Однако наряду с этим он нестойкий к шуму в исходных данных. Существенным недостатком метода является отсутствие разработанных общих методов построения выпрямляющих пространств и ядер, которые наилучшим образом подходят к конкретной задаче [7]. Сложность алгоритма метода опорных векторов оценивается как  $O(n^3)$  [8]. Классификация с помощью кластерного анализа, а также вальдовский последовательный анализ с применением информационной меры Кульбака описаны в работе [9].

### ОСНОВНЫЕ ОПРЕДЕЛЕНИЯ

**Определение 1.** Множества  $A$  и  $B$  называются  $\varepsilon$ -разделимыми, если существуют  $A_1 \subset A$ ,  $B_1 \subset B$ , для которых

$$\text{conv}(A \setminus A_1) \cap \text{conv}(B \setminus B_1) = \emptyset, \quad (1)$$

$$|A_1| + |B_1| < \varepsilon(n_A + n_B). \quad (2)$$

**Определение 2.** Гиперплоскость  $L$  называется разделяющей для множеств  $A$  и  $B$ , если

$$\text{conv}_A \subset L^+, \text{conv}_B \subset L^-.$$

**Определение 3.** Гиперплоскость  $L_\varepsilon$  называется  $\varepsilon$ -разделяющей для множеств  $A$  и  $B$ , если

$$\frac{|A \cap L_\varepsilon^+| + |B \cap L_\varepsilon^-|}{n_A + n_B} \geq 1 - \varepsilon.$$

Для решения задачи нахождения  $\varepsilon$ -разделяющей гиперплоскости двух множеств будем использовать  $\varepsilon$ -сети. Рассмотрим ранжированное пространство  $(X, \mathcal{R})$ , где  $X$  — некоторое множество,  $\mathcal{R}$  — совокупность подмножеств множества  $X$  [10].

**Определение 4.** Проекцией  $\mathcal{R}$  на  $A$  называется множество

$$\text{Pr}_A(\mathcal{R}) = \{r \cap A : r \in \mathcal{R}\}.$$

**Определение 5.** Считают, что  $A$  дробится с помощью  $\mathcal{R}$ , если  $\text{Pr}_A(\mathcal{R}) = 2^A$ .

**Определение 6.** Размерностью Вапника–Червоненкиса для ранжированного пространства  $(X, \mathcal{R})$  называется мощность (возможно, бесконечная) наибольшего подмножества из  $X$ , которое дробится с помощью  $\mathcal{R}$  :

$$VC(X, \mathcal{R}) := \max\{m : \exists A \subset X, A \text{ дробится с помощью } \mathcal{R}\}.$$

**Теорема Радона.** Каждое множество из  $d+2$  или более точек в  $R^d$  может быть представлено как объединение двух непересекающихся множеств, выпуклые оболочки которых имеют общую точку [11].

**Следствие** [10]. Размерность Вапника–Червоненкиса для ранжированного пространства  $(R^d, H^d)$ , где  $H^d$  — совокупность всех полупространств в  $R^d$ , определяется как  $VC(R^d, H^d) = d+1$ .

**Определение 7.** Пусть заданы ранжированное пространство  $(X, \mathcal{R})$ , множество  $A \subset X$  и  $\varepsilon \in R$ ,  $0 < \varepsilon < 1$ , тогда подмножество  $N \subset A$  называется  $\varepsilon$ -сетью для множества  $A$ , если  $\forall r \in \mathcal{R} : |r \cap A| \geq \varepsilon |A| \Rightarrow N \cap (r \cap A) \neq \emptyset$ .

**Теорема Вельцля–Хаусслера** [12]. Пусть  $VC(X, \mathcal{R}) = \delta < \infty$ , тогда  $\forall A \subset X$ ,  $|A| = n$ ,  $\forall \varepsilon \in (0, 1)$  существует  $\varepsilon$ -сеть  $N$  множества  $A$ , мощность которой не зависит от мощности множества  $A$ , кроме того  $|N| \leq \frac{8\delta}{\varepsilon} \log_2 \frac{8\delta}{\varepsilon}$ .

#### $\varepsilon$ -РАЗДЕЛЕНИЕ МНОЖЕСТВ В ПРОСТРАНСТВЕ $R^d$

Рассмотрим ранжированное пространство  $(R^d, H^d)$ , где  $H^d$  — совокупность всех полупространств в  $R^d$ . В  $(R^d, H^d)$  будем строить  $\varepsilon$ -сети множеств  $A$  и  $B$ . Покажем, что задачу  $\varepsilon$ -разделения множеств  $A$  и  $B$  можно свести к задаче разделения выпуклых оболочек их непересекающихся  $\varepsilon$ -сетей. Сформулируем основную теорему данной статьи.

**Теорема 1.** Чтобы множества  $A$  и  $B$  были  $\varepsilon$ -разделимыми, необходимо и достаточно существования  $\varepsilon_A, \varepsilon_B$  и соответствующих им  $\varepsilon$ -сетей  $N_{\varepsilon_A}^A, N_{\varepsilon_B}^B$  в  $(R^d, H^d)$ , для которых выполняются соотношения

$$\varepsilon_A n_A + \varepsilon_B n_B < \varepsilon(n_A + n_B), \quad (3)$$

$$\text{conv}_{\varepsilon_A} N_{\varepsilon_A}^A \cap \text{conv}_{\varepsilon_B} N_{\varepsilon_B}^B = \emptyset. \quad (4)$$

**Доказательство. Необходимость.** Предположим, что множества  $A$  и  $B$   $\varepsilon$ -разделимы. Обозначим  $|A_1| = n_{A_1}$ ,  $|B_1| = n_{B_1}$ . Пусть  $\varepsilon_A = \frac{n_{A_1}}{n_A} + \delta$ ,  $\varepsilon_B = \frac{n_{B_1}}{n_B} + \delta$ , где  $\delta$  выбрано из условия  $(n_{A_1} + n_{B_1}) + 2\delta < \varepsilon(n_A + n_B)$ . Тогда в качестве сетей можно использовать  $N_{\varepsilon_A}^A = A \setminus A_1$ ,  $N_{\varepsilon_B}^B = B \setminus B_1$ . Согласно (1) выполняется соотношение (4), согласно (2) — неравенство (3).

**Достаточность.** Предположим, что выполняются условия (3), (4) теоремы 1. Докажем, что множества  $A$  и  $B$  являются  $\varepsilon$ -разделимыми. Пусть множества  $A$  и  $B$  не являются  $\varepsilon$ -разделимыми. Это значит, что условие (1) может выполняться только в том случае, когда

$$n_{A_1} + n_{B_1} \geq \varepsilon(n_A + n_B). \quad (5)$$

Поскольку выполняется соотношение (4), то для  $\text{conv } N_{\varepsilon_A}^A$  и  $\text{conv } N_{\varepsilon_B}^B$  существует разделяющая гиперплоскость. Предположим, что  $\text{conv } N_{\varepsilon_A}^A \in L_+$  и  $\text{conv } N_{\varepsilon_B}^B \in L_-$ . Обозначим  $A^N = \{x \in A: x \in L_+\}$ ,  $B^N = \{x \in B: x \in L_-\}$ . Множества  $A \setminus A^N$  и  $B \setminus B^N$  согласно определению  $\varepsilon$ -сетей удовлетворяют соотношениям  $|A \setminus A^N| < \varepsilon_A n_A$ ,  $|B \setminus B^N| < \varepsilon_B n_B$ .

Рассмотрим множества, которые исключаем:  $A_1 = A \setminus A^N$ ,  $B_1 = B \setminus B^N$ . Тогда для этих множеств согласно условию (3)  $n_{A_1} + n_{B_1} < \varepsilon(n_A + n_B)$ . Таким образом, неравенство (5) для данных множеств не выполняется. Получаем противоречие.

Достаточность доказана. Теорема 1 доказана.

Введем обозначения  $\eta_A = \frac{|A \cap \text{conv } B|}{n_A}$ ,  $\eta_B = \frac{|B \cap \text{conv } A|}{n_B}$  и  $\varepsilon_A = \eta_A + \frac{1}{n_A}$ ,  $\varepsilon_B = \eta_B + \frac{1}{n_B}$ .

**Теорема 2.** Пусть

$$\varepsilon > \frac{\eta_A n_A + \eta_B n_B}{n_A + n_B},$$

тогда для любой  $\varepsilon$ -разделяющей гиперплоскости  $L_\varepsilon$  множеств  $A$  и  $B$  существуют  $\varepsilon$ -сети  $N_{\varepsilon_A}^A$  и  $N_{\varepsilon_B}^B$ , для которых  $L_\varepsilon$  — разделяющая гиперплоскость.

**Доказательство.** Будем строить  $\varepsilon$ -сеть  $N_{\varepsilon_A}^A$  таким образом, чтобы в нее не попали те точки множества  $A$ , которые принадлежат выпуклой оболочке  $\text{conv } B$ . Поскольку  $\varepsilon_A > \eta_A$ , то  $\forall r: |r \cap A| \geq \varepsilon_A n_A$  существует точка  $x$  такая, что  $x \in (r \cap A)$ , но  $x \notin \text{conv } B$ . При построении  $\varepsilon$ -сети  $N_{\varepsilon_A}^A$  в качестве представителя множества  $r$  будем использовать точку  $x$  такую, что  $x \notin \text{conv } B$ . Благодаря полупространствам  $r$ , для которых  $|r \cap A| = \eta_A n_A + 1$  и  $|(r \cap A) \cap \text{conv } B| = \eta_A n_A$ ,  $\varepsilon$ -сеть  $N_{\varepsilon_A}^A$  содержит все те точки множества  $A$ , которые являются ближайшими соседями к пересечению  $A \cap \text{conv } B$ .

Поскольку  $\varepsilon$ -сети  $N_{\varepsilon_A}^A$  и  $N_{\varepsilon_B}^B$  содержат всех ближайших соседей к пересечениям  $A \cap \text{conv } B$  и  $B \cap \text{conv } A$  и не содержат точек, принадлежащих этим пересечениям, можно утверждать, что среди разделяющих гиперплоскостей для множеств  $N_{\varepsilon_A}^A$  и  $N_{\varepsilon_B}^B$  найдется гиперплоскость, которая  $\varepsilon$ -разделяет множества  $A$  и  $B$ .

Теорема 2 доказана.

Итак, задачу  $\varepsilon$ -разделения двух пересекающихся множеств ( $A$  и  $B$ ) можно свести к тривиальной задаче разделения двух выпуклых непересекающихся множеств ( $\text{conv}N_{\varepsilon_A}^A$  и  $\text{conv}N_{\varepsilon_B}^B$ ).

Обозначим  $n_{\varepsilon_A} = |N_{\varepsilon_A}^A|$ ,  $n_{\varepsilon_B} = |N_{\varepsilon_B}^B|$ . Поскольку сложность алгоритма поиска  $\varepsilon$ -сети линейная [13], а построение выпуклых оболочек для  $\varepsilon$ -сетей  $N_{\varepsilon_A}^A$  и  $N_{\varepsilon_B}^B$ , например, методом Джарвиса оценивается как  $O(n_{\varepsilon_A}^2)$  и  $O(n_{\varepsilon_B}^2)$  [14], то общую сложность алгоритма  $\varepsilon$ -разделения множеств  $A$  и  $B$  с использованием  $\varepsilon$ -сетей можно оценить как  $O(m + m_\varepsilon^2)$ , где  $m = \max(n_A, n_B)$ ,  $m_\varepsilon = \max(n_{\varepsilon_A}, n_{\varepsilon_B})$ .

Таким образом, доказано, что для  $\varepsilon$ -разделимости двух множеств необходимо и достаточно существования некоторых делимых  $\varepsilon$ -сетей этих множеств. Предложен метод построения делимых  $\varepsilon$ -сетей и показано, что задачу разделения двух пересекающихся множеств можно свести к задаче разделения двух непересекающихся выпуклых множеств.

#### СПИСОК ЛИТЕРАТУРЫ

1. Fisher R.A. The use of multiple measurements in taxonomic problems // Annals of Eugenics. — 1936. — N 7. — P. 179–188.
2. Deng Cai, Xiaofei He, Jiawei Han. Training linear discriminant analysis in linear time. — [http://researchweb.iit.ac.in/~nataraj.j/poseSearchReports/icde08\\_dengcai.pdf](http://researchweb.iit.ac.in/~nataraj.j/poseSearchReports/icde08_dengcai.pdf).
3. Айвазян С. А., Бухштабер В. М., Енюков И. С., Мешалкин Л. Д. Прикладная статистика: классификация и снижение размерности. — М.: Финансы и статистика, 1989. — 607 с.
4. Воронцов К.В. Лекции по статистическим (байесовским) алгоритмам классификации: <http://www.ccas.ru/voron/download/Bayes.pdf>.
5. Chris Fleizach, Satoru Fukushima. A naive Bayes classifier on 1998 KDD Cup. — <http://sysnet.ucsd.edu/~cfleizac/cse250b/project1.pdf>.
6. Vapnik V.N. The nature of statistical learning theory. — 2nd ed. — New York: Springer, 2000. — 314 p.
7. Воронцов В.К. Лекции по методу опорных векторов. — <http://www.ccas.ru/voron/download/SVM.pdf>.
8. Ivor W. Tsang, James T. Kwok, Pak-Ming Cheung. Core vector machines: fast SVM training on very large data sets // Journal of Machine Learning Research. — 2005. — N 6. — P. 363–392.
9. Иванчук М.А., Малык И.В. Сравнение методов распределения наблюдений на классы при прогнозировании наличия осложнений у тяжелобольных // Кибернетика и системный анализ. — 2015. — 51, № 2. — С. 164–174.
10. Райгородский А.М. Системы общих представителей в комбинаторике и их приложения в геометрии. — М.: МЦНМО, 2009. — 136 с.
11. Данцер Л., Грюнбаум Б., Кли В. Теорема Хелли. — М.: Мир, 1968. — 162 с.
12. Haussler D., Welzl E.  $\varepsilon$ -nets and simplex range queries // Discrete & Computational Geometry. — 1987. — N 2. — P. 127–151.
13. ICS Theory Group. Computational Statistics. — <https://www.ics.uci.edu/~eppstein/280/cluster.html>.
14. Препарата Ф., Шеймос М. Вычислительная геометрия: Введение. — М.: Мир, 1989. — С. 478.

Поступила 27.04.2015