Ученые записки Таврического национального университета им. В.И. Вернадского

Серия "Математика. Механика. Информатика и Кибернетика" N 2 (2003) 143–147

УДК 519.95

В.И. Донской

О МЕТРИЧЕСКИХ СВОЙСТВАХ КРАТЧАЙШИХ ЭМПИРИЧЕСКИХ КОНЪЮНКТИВНЫХ ЗАКОНОМЕРНОСТЕЙ

1. В задачах обучения распознаванию объектов, описанных булевыми переменными-признаками, начальная информация представлена двумя подмножествами: $\mathfrak{M}_1 \subset B^n$, $\mathfrak{M}_0 \subset B^n$; $\mathfrak{M}_1 \cap \mathfrak{M}_0 = \emptyset$; $m = |\mathfrak{M}_1 \cup \mathfrak{M}_0| = o(n)$, $n \to \infty$, где $B^n \triangleq \{0,1\}^n$ - множество вершин единичного n-мерного куба. Считается, что конечное множество $\mathfrak{M}_1 \cup \mathfrak{M}_0$ является результатом выбора m элементов из генеральной совокупности $(\mathcal{K}_1 \cup \mathcal{K}_0) \subseteq B^n$; существующие, но заранее неизвестные множества $\mathcal{K}_1$ и $\mathcal{K}_0$, называют *классами*; $\mathcal{K}_1 \cap \mathcal{K}_0 = \emptyset$. Обучающая выборка в виде множества $\mathfrak{M}_0 \cup \mathfrak{M}_1$ предоставляется вместе с достоверной информацией о принадлежности подмножеств $\mathfrak{M}_1$ и $\mathfrak{M}_0$ классам $\mathcal{K}_1$ и $\mathcal{K}_0$: $\mathfrak{M}_1 \subset \mathcal{K}_1$, $\mathfrak{M}_0 \subset \mathcal{K}_0$. Требуется, используя указанную начальную информацию, построить *решающее правило* (функцию) $\hat{F} : B^n \to \{0,1\}$, как можно менее "отличающуюся" от существующей, но неизвестной функции $F(\tilde{x}) = \begin{cases} 1, & \tilde{x} \in \mathcal{K}_1 \\ 0, & \tilde{x} \in \mathcal{K}_0 \end{cases}$. Степень такого отличия для булевых функций может оцениваться, например, выражением $D(F,\hat{F}) = \sum_{\tilde{x} \subset \mathcal{K}_1 \cup \mathcal{K}_0} |F(\tilde{x}) - \hat{F}(\tilde{x})|$. Корректность функции $\hat{F}$ относительно обучающей выборки означает, что $\sum_{\tilde{x} \in \mathfrak{M}_1 \cup \mathfrak{M}_0} |F(\tilde{x}) - \hat{F}(\tilde{x})| = 0$. Если $D(F,\hat{F}) = 0$, то построенное правило является точным; но значение $D(F,\hat{F})$ невозможно вычислить, заранее не зная функцию F (или описание классов $\mathcal{K}_1$ и $\mathcal{K}_0$). Построенные решающие правила обычно проверяют на контрольных выборках вида $\mathfrak{M}_1^* \cup \mathfrak{M}_0^*$: $\mathfrak{M}_1^* \subset \mathcal{K}_1$; $\mathfrak{M}_0^* \subset \mathcal{K}_0$; $(\mathfrak{M}_1^* \cup \mathfrak{M}_0^*) \cap (\mathfrak{M}_1 \cup \mathfrak{M}_0) = \varnothing$, получая оценку $\hat{D}(F,\hat{F}) = \sum_{\tilde{x} \in \mathfrak{M}_1^* \cup \mathfrak{M}_0^*} |F(\tilde{x}) - \hat{F}(\tilde{x})|$. Оценка $\hat{D}(F,\hat{F})$ приближается к $D(F,\hat{F})$ с ростом $m^* = |\mathfrak{M}_1^* \cup \mathfrak{M}_0^*|$.

Каждую величину $\tilde{x} = (x_1, \ldots, x_i, \ldots, x_n) \in B^n$ будем называть *точкой* или булевым *набором*; $x_1, \ldots, x_i, \ldots, x_n$ - булевыми переменными или координатами. Будем допускать и логическую, и арифметическую интерпретацию булевых $(0,1)$ переменных.

Расстояние Хэмминга между точками $\tilde{\alpha}$ и $\tilde{\beta}$ в B^n определяется соотношением $\rho(\tilde{\alpha},\tilde{\beta}) = \sum_{i=1}^n \alpha_i \oplus \beta_i$ и равно числу несовпадающих значений переменных в булевых наборах $\tilde{\alpha}$ и $\tilde{\beta}$. Функция $\rho : B^n \times B^n \to \mathbb{N}_0$ является метрикой, вместе с которой множество B^n образует метрическое пространство. *Сферой* радиуса r в B^n с центром в точке $\tilde{\alpha}$ называется множество $Sp^r(\tilde{\alpha},\tilde{x}) \triangleq \{\tilde{x} \in B^n : \rho(\tilde{\alpha},\tilde{x}) = r\}$ и, аналогично, *шаром* радиуса r - множество $S^r(\tilde{\alpha},\tilde{x}) \triangleq \{\tilde{x} \in B^n : \rho(\tilde{\alpha},\tilde{x}) \leq r\}$. *Соседними* называют пару $\tilde{\alpha}$, $\tilde{\beta}$ точек в B^n таких, что $\rho(\tilde{\alpha},\tilde{\beta}) = 1$. *Нормой* $\|\tilde{x}\|$ точки $\tilde{x}$ называется суммарное число её единичных координат.

Говорят, что набор $\tilde{\alpha}$ *предшествует* набору $\tilde{\beta}$ ($\tilde{\alpha} \preceq \tilde{\beta}$), если $\forall i \quad (\alpha_i \leq \beta_i)$. Множество $\mathfrak{N}$ называется *интервалом*, если $\mathfrak{N} = \{\tilde{x} \in B^n : \tilde{\alpha} \preceq \tilde{x} \preceq \tilde{\beta}\}$ для некоторых точек $\tilde{\alpha} \in B^n$ и $\tilde{\beta} \in B^n$; число $\rho(\tilde{\alpha}, \tilde{\beta}) = k$ называют *размерностью*, а $(n-k)$ - рангом интервала $\mathfrak{N}$.

Элементарной конъюнкцией (э.к.) называют формулу $x_{i_1}^{\sigma_{i_1}} \wedge \ldots \wedge x_{i_r}^{\sigma_{i_r}}$, в которой все литералы, определяемые как $x^\sigma = \begin{cases} x, & \sigma = 1, \\ \bar{x}, & \sigma = 0, \end{cases}$ различны. Число r называют *рангом* конъюнкции. Каждая э.к. K определяет булеву функцию $K(\tilde{x})$:

$$K(\tilde{x}) = \begin{cases} 1, & \text{если } x_{i_1} = \sigma_{i_1}, \ldots, x_{i_r} = \sigma_{i_r}; \\ 0, & \text{в противном случае.} \end{cases}$$

Тогда множество $\mathfrak{N}_K = \{\tilde{x} : K(\tilde{x}) = 1\}$ является интервалом ранга r, поскольку $\mathfrak{N}_K = \{\tilde{x} : \tilde{\alpha} \preceq \tilde{x} \preceq \tilde{\beta}\}$, где $\tilde{\alpha}$ такой набор переменных, что $\alpha_i = \sigma_i$ при $i \in \{i_1, \ldots i_r\}$ и $\alpha_i = 0$ при $i \notin \{i_1, \ldots i_r\}$, а $\tilde{\beta}$ - такой, что $\beta_i = \sigma_i$ при $i \in \{i_1, \ldots i_r\}$ и $\beta_i = 1$ при $i \notin \{i_1, \ldots i_r\}$.

Э.к. K называется *допустимой* (или *элементарным классификатором*) относительно обучающей информации $\mathfrak{M}_1 \cup \mathfrak{M}_0$, если $\exists \tilde{x} \in \mathfrak{M}_1 \quad (K(\tilde{x}) = 1)$ и $\forall x \in \mathfrak{M}_0 \quad (K(\tilde{x}) = 0)$.

Э.к. K называется *кратчайшей*, если удаление из K хотя бы одного литерала приводит к конъюнкции K', не являющейся допустимой относительно $\mathfrak{M}_1 \cup \mathfrak{M}_0$.

В большинстве работ, посвященных синтезу логических решающих правил распознавания, предпочтение отдаётся кратчайшим элементарным классификаторам (к.э.к) [2].

Целью настоящей работы является изучение таких к.э.к., которые, будучи найденными в эмпирических наблюдениях (выборках вида $\mathfrak{M}_1 \cup \mathfrak{M}_0$), представляют собой *кратчайшие эмпирические конъюнктивные закономерности* (э.к.з.). *Результатом работы* является обоснование того, что для типичных задач обучения распознаванию (так называемых "задач, *удовлетворяющих условию компактности*"), использование кратчайших э.к.з. неизбежно влечёт ошибки при распознавании объектов, не участвовавших в обучении.

Актуальность рассматриваемого вопроса определяется важностью построения индуктивных моделей классов в виде дизъюнктивных нормальных форм, в которые явно могут входить э.к.з. [3], [4].

Новизна работы состоит в установлении ранее неизвестных свойств кратчайших э.к.з. (теоремы 1, 2).

2. Введём ряд определений, необходимых для выяснения метрических свойств кратчайших эмпирических конъюнктивных закономерностей.

Определение 1. Интервал $\mathfrak{N}$ называется *максимальным для множества* $\mathfrak{M}_1$, если $\mathfrak{N} \cap \mathfrak{M}_1 \neq \emptyset$, $\mathfrak{N} \subseteq B^n \backslash \mathfrak{M}_0$ и не существует другого интервала $\mathfrak{N}_1$ такого, что $\mathfrak{N} \subset \mathfrak{N}_1 \subseteq B^n \backslash \mathfrak{M}_0$.

Определение 2. Точка $\tilde{\alpha} \in B^n$ называется *соседней ко множеству* $\mathfrak{M} \subset B^n$, если $\tilde{\alpha} \notin \mathfrak{M}$ и найдется точка $\tilde{\beta} \in \mathfrak{M}$ такая, что $\rho(\tilde{\alpha}, \tilde{\beta}) = 1$.

Теорема 1. *Любой максимальный для множества* $\mathfrak{M}_1$ *интервал содержит хотя бы одну точку, соседнюю ко множеству* $\mathfrak{M}_0$.

Доказательство. Пусть $\mathfrak{N}$ - произвольный максимальный для множества $\mathfrak{M}_1$ интервал. Согласно определению, $\mathfrak{N} \cap \mathfrak{M}_1 \neq \emptyset$ и $\mathfrak{N} \cap \mathfrak{M}_0 = \emptyset$, и тогда $\forall \tilde{\alpha} \in \mathfrak{N} \quad \forall \tilde{\beta} \in \mathfrak{M}_0 \quad (\rho(\tilde{\alpha}, \tilde{\beta}) > 0)$. Предположим противное: пусть максимальный для множества $\mathfrak{M}_1$ интервал $\mathfrak{N}$ не содержит ни одной точки, соседней ко множеству $\mathfrak{M}_0$. Тогда должно выполняться неравенство $\forall \tilde{\alpha} \in \mathfrak{N}\ \forall \tilde{\beta} \in \mathfrak{M}_0\ (\rho(\tilde{\alpha}, \tilde{\beta}) \geq 2)$. Поэтому для любой пары точек $\tilde{\alpha} \in \mathfrak{N}$ и $\tilde{\beta} \in \mathfrak{M}_0$ найдётся номер переменной $i \in \{1, 2, \ldots, n\}$ такой, что $\alpha_i \neq \beta_i$. Зафиксируем один такой номер i. Для каждой точки $\tilde{\alpha} \in \mathfrak{N}$ построим соседнюю к ней точку $\tilde{\alpha}^* = (\alpha_1, \ldots, \alpha_{i-1}, \bar{\alpha}_i, \alpha_{i+1}, \ldots, \alpha_n)$. Очевидно, построенная совокупность точек $\tilde{\alpha}^*$ образует интервал. Обозначим его $\mathfrak{N}^*$. Любая точка $\tilde{\alpha}^* \in \mathfrak{N}^*$ находится на сфере радиуса 1 с центром в некоторой точке $\tilde{\alpha} \in \mathfrak{N}$. Тогда из предположения $\forall \tilde{\alpha} \in \mathfrak{N}\ \forall \tilde{\beta} \in \mathfrak{M}_0\ (\rho(\tilde{\alpha}, \tilde{\beta}) \geq 2)$ следует, что $\mathfrak{N}^* \cap \mathfrak{M}_0 = \emptyset$. Из того, что $\mathfrak{N} \cap \mathfrak{M}_0 = \emptyset$, следует, что $(\mathfrak{N} \cup \mathfrak{N}^*) \cap \mathfrak{M}_0 = \emptyset$. Объединение $\mathfrak{N} \cup \mathfrak{N}^* = \mathfrak{N}_1$ образует интервал такой, что $\mathfrak{N}_1 \cap \mathfrak{M}_0 = \emptyset$, а $\mathfrak{N}_1 \cap \mathfrak{M}_1 \neq \emptyset$, т.к. $\mathfrak{N} \cap \mathfrak{M}_1 \neq \emptyset$. Поскольку $\mathfrak{N} \subset \mathfrak{N}_1$, интервал $\mathfrak{N}$ не может быть максимальным для множества $\mathfrak{N}_1$: его можно расширить до интервала $\mathfrak{N}_1$. Полученное противоречие доказывает теорему. □

Определение 3. Будем говорить, что задача удовлетворяет *условию слабой компактности*, если найдутся такие точки $\tilde{\alpha}, \tilde{\beta} \in B^n$, что $S^{r_1}(\tilde{\alpha}, \tilde{x}) \subseteq \mathcal{K}_1$ и $S^{r_0}(\tilde{\beta}, \tilde{x}) \subseteq \mathcal{K}_0$ для некоторых значений радиусов $r_1 < [\frac{n}{2}]$ и $r_0 < [\frac{n}{2}]$.

Любой найденной кратчайшей э.к.з. K взаимно однозначно соответствует максимальный для множества $\mathfrak{M}_1$ интервал $\mathfrak{N}_K$. Рассмотрим *элементарное эмпирическое правило (продукцию)* вида

$$(K(\tilde{x}) = 1) \rightarrow (\tilde{x} \in \mathcal{K}_1) \tag{1}$$

и эквивалентное ему выражение $(\tilde{x} \in \mathfrak{N}_K) \Longrightarrow (\tilde{x} \in \mathcal{K}_1)$.

Учитывая теорему 1, заключаем, что в $\mathfrak{N}_K$ содержатся точки, соседние ко множеству $\mathfrak{M}_0$. Пусть задача удовлетворяет условию слабой компактности и $\mathfrak{M}_0 \subset S^{r_0-1}(\tilde{\beta}, \tilde{x})$. Тогда, очевидно, решающее правило (1) будет давать ошибки на точках сферы $Sp^{r_0}(\tilde{\beta}, \tilde{x})$.

Определение 4. Будем говорить, что задача удовлетворяет *условию сильной компактности*, если найдутся точки $\tilde{\alpha}, \tilde{\beta} \in B^n$ и числа r_1, R_1, r_0, R_0, все - меньшие $[\frac{n}{2}]$, такие что $r_1 \leq R_1$, $r_0 \leq R_0$, $S^{r_1}(\tilde{\alpha}, \tilde{x}) \subseteq \mathcal{K}_1 \subseteq S^{R_1}(\tilde{\alpha}, \tilde{x})$, $S^{r_0}(\tilde{\beta}, \tilde{x}) \subseteq \mathcal{K}_2 \subseteq S^{R_0}(\tilde{\beta}, \tilde{x})$ и $S^{R_1}(\tilde{\alpha}, \tilde{x}) \cap S^{R_0}(\tilde{\beta}, \tilde{x}) = \emptyset$.

Лемма 1. *Пусть* $S^R(\tilde{\alpha}, \tilde{x})$ - *шар и* $\mathfrak{N} = \{\tilde{x} \in B^n : \tilde{\alpha} \preceq \tilde{x} \preceq \tilde{\beta}\}$ - *интервал размерности большей или равной* $2R$. *Тогда* $\{\tilde{x} \in \mathfrak{N} : \rho(\tilde{\alpha}, \tilde{x}) \leq R\} \subseteq S^R(\tilde{\alpha}, \tilde{x})$.

Доказательство. $\{\tilde{x} \in \mathfrak{N} : \rho(\tilde{\alpha}, \tilde{x}) \leq R\} = \bigcup_{r=0}^{R} \{\tilde{x} : (\tilde{\alpha} \preceq x) \wedge (\rho(\tilde{\alpha}, \tilde{x}) = r)\} \subseteq$

$$\subseteq \bigcup_{r=0}^{R} \{\tilde{x} : \rho(\tilde{\alpha}, \tilde{x}) = r\} = \bigcup_{r=0}^{R} Sp^r(\tilde{\alpha}, \tilde{x}) = S^R(\tilde{\alpha}, \tilde{x}).$$

Включение не является строгим, поскольку при $\|\tilde{\alpha}\| = 0$ имеет место равенство множеств.

Теорема 2. *Если задача удовлетворяет условию сильной компактности,* $\max \rho(\tilde{x}, \tilde{y}) \geq 2R_2$ *и* $R_2 > R_1$, $(\tilde{x}, \tilde{y}) \in \mathfrak{M}_1 \times \mathfrak{M}_0$, *то можно указать такую выборку* $\mathfrak{M}_1 \cup \mathfrak{M}_0$, *что найдется правило вида (1), основанное на кратчайшей эмпирической конъюнктивной закономерности ранга* $r \leq n - 2R_2$, *которое будет давать ошибку не менее чем для* $\frac{1}{2}4^{R_2}$ *объектов - булевых наборов длины* n.

Доказательство. Рассмотрим к.э.к. K ранга $r = n - 2R_2$ и соответствующий ей максимальный интервал $\mathfrak{N}_K$, сконструированный так, что $\mathfrak{N}_K = \{\tilde{x} : \tilde{\alpha} \preceq \tilde{x} \preceq \tilde{\beta}\}$, $\mathfrak{N}_K \cap \mathfrak{M}_1 \neq \emptyset$, $\mathfrak{N}_K \cap \mathfrak{M}_0 = \emptyset$ и $\mathfrak{N}_K \cap S^{R_2}(\tilde{\beta}, \tilde{x}) \neq \emptyset$. Из определения сильной компактности и леммы 1 следует, что $\{\tilde{x} \in \mathfrak{N}_K : \rho(\tilde{\beta}, \tilde{x}) \leq R_2\} \cap \mathcal{K}_2 \neq \emptyset$ и $\{\tilde{x} \in \mathfrak{N}_K : \rho(\tilde{\beta}, \tilde{x}) \leq R_2\} \subseteq S^{R_2}(\tilde{\beta}, \tilde{x})$. Поэтому возможен случай, когда все точки множества $\{\tilde{x} \in \mathfrak{N}_K : \rho(\tilde{\beta}, \tilde{x}) \leq R_2\}$ будут неверно классифицироваться правилом (1), согласно которому $\mathfrak{N}_K \subset \mathcal{K}_1$ (рис.1).

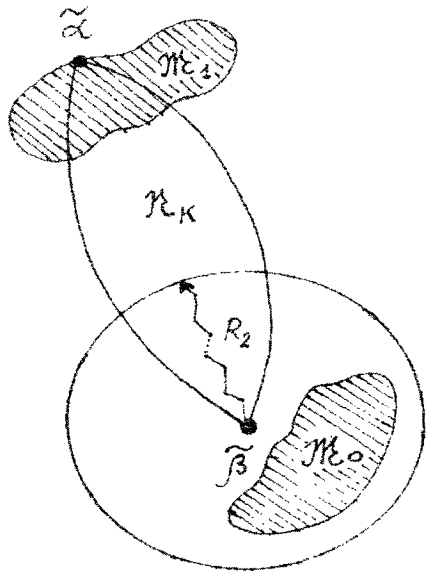

Рис. 1

$$|\{\tilde{x} \in \mathfrak{N}_K : \rho(\tilde{\beta}, \tilde{x}) \leq R_2\}| = \sum_{i=0}^{R_2} C_{2R_2}^{R_2} > \frac{1}{2}4^{R_2},$$

поскольку размерность интервала $\mathfrak{N}_K$ равна $n - r = 2R_2$.

Следствие 1. При выполнении условий теоремы 2 и $R_2 > \frac{n}{2} - \delta$, $0 < \delta < \frac{n}{2}$, ошибка решающего правила на основе кратчайшей э.к.з. может иметь место на подмножестве, являющемся $\frac{1}{2^{2\delta+1}}$ частью множества B^n.

Доказательство. $|\{\tilde{x} \in \mathfrak{N}_K : \rho(\tilde{\beta}, \tilde{x}) \leq R_2\}| > \frac{1}{2}4^{R_2} > \frac{1}{2}4^{\frac{n}{2}-\delta} = \frac{1}{2}4^{-\delta}2^n = \frac{1}{2^{2\delta+1}}2^n$.

Основной результат данной работы состоит в том, что использование кратчайших эмпирических конъюнктивных закономерностей для задач распознавания, удовлетворяющих условиям сильной и слабой компактности, может привести к большому числу ошибок (более чем на $\frac{1}{2^{2\delta+1}}$ части конечного метрического пространства B^n, где $0 < \delta < \frac{n}{2}$). Из этого результата можно сделать *вывод* о нецелесообразности использования в задачах индуктивного моделирования, удовлетворяющих гипотезе компактности, подхода, основанного на минимизации частично заданных булевых функций, без ограничения ранга отыскиваемых э.к.з. снизу.

Представляется перспективным получение не только верхних [1], но и нижних оценок рангов э.к.з., используемых в задачах обучения распознаванию и индуктивного моделирования.

Список литературы

[1] Донской В.И., Дюличева Ю.Ю. *Индуктивная модель r-корректного эмпирического леса* // Труды Международной конференции по индуктивному моделированию. - 2002. - Львов. - С.54-58.

[2] Журавлёв Ю.И. *Избранные научные труды.* - М.:Магистр, 1998. - 403 с.

[3] Donskoy V.I. *Case-, Knowledge-, and Optimization-Based Hybrid Approach in AI* // Proc.11th. Int.Conf. IEA-AIE-98.-Lecture Notes in Computer Science. - Springer. - 1998. - Vol.I. - P.520-527.

[4] Donskoy V. *Pseudo-Boolean Scalar Optimization Models with Incomplete Information* // GMOOR Newsletter. - 1996. - №1/2. - P.20-26.

E-mail: donskoy@ccssu.crimea.ua