УДК 519.68:681.513.7

ОЦЕНКА VCD R-РЕДУЦИРОВАННОГО ЭМПИРИЧЕСКОГО ЛЕСА

Ю.Ю. Дюличева

Таврический национальный университет им. В.И. Вернадского
факультет математики и информатики
пр-т Вернадского, 4, г. Симферополь, 95007, Украина
e-mail: *email@mail.ru*

Abstract

The *VCD* r-pruned empirical forest *BDF (n,μ,r,q)* estimations

$$max(\mu q, \log n) < VCD(BDF\ (n,\mu,r,q)) < r\mu q \log n - \mu q \log \frac{\mu q}{2}$$

and *VCD(BDF(n,μ,r,q))*=$\Theta(\log n)$, where r - a maximum permissible rank of forest's trees branches, n-an attribute space dimension, q - a maximum forest's trees admissible number, μ-a maximum leaves number in each tree are derived from.

Введение

Бинарные решающие деревья (БРД) широко используются при построении информационных интеллектуализированных систем и применяются для распознавания, формирования понятий, построения логических описаний закономерностей [6], [12], [16], [17], [19].

В процессе практического использования БРД возникла проблема перенастройки (излишне точной подгонки под заданную для обучения выборку). По мнению ряда авторов [14], [21], [22] перенастройка "улавливает"скорее "шум"в данных, чем закономерность. Это определило необходимость *постановки следующей проблемы.* Обосновать правила остановки синтеза деревьев для избежания перенастройки (принципы "отсечения"или редукции [9]) и, разработать принципы использования редуцированных БРД, применяя корректирующие методы, незначительно увеличивающие сложность скорректированной модели принятия решений. Последние связано с оценкой сложности класса получаемых решающих правил *VCD* в современной терминологии, принятой в научной литературе [15],[20]).

Анализ последних достижений и публикаций, посвященных этой проблеме [13], [15], [20], [22], позволяет сделать вывод, что в настоящее время неизвестны оценки *VCD* для редуцированных эмпирических лесов. *Нерешенным является вопрос о том, насколько усложняется класс решающих правил при переходе от отдельных БРД к решающему лесу.*

Целью настоящей работы является оценка *VCD* редуцированного эмпирического леса и ее сравнение с соответствующей оценкой для отдельных БРД и решающих правил, представляемых дизъюнктивными нормальными формами.

Все определения и понятия, используемые ниже без пояснений, хорошо известны специалистам. При необходимости их можно найти в литературе [4], [7], [8], [11].

1. Условия обеспечения равномерной сходимости эмпирических частот событий к вероятностям

Изложим ряд положений теории Вапника-Червоненкиса, которые необходимы для понимания полученных ниже результатов.

Пусть X – множество, S – некоторая система его подмножеств; $X^l = x_1, ..., x_l$-последовательность элементов из X длины l. Каждое множество $A \in S$ определяет подпоследовательность X_A этой последовательности, состоящую из тех элементов, которые принадлежат A. Говорят, что A *индуцирует подпоследовательность* X_A на последовательности X^l [1], [2].

Обозначим $\Delta^S(x_1, ..., x_l)$ - число различных подпоследовательностей, индуцированных всеми множествами $A \in S$. Очевидно, что $\Delta^S(x_1, ..., x_l) \leqslant 2^l$.

Функция

$$m^S(l) = \max_{x_1,...,x_l} \Delta^S(x_1, ..., x_l),$$

где максимум берется по всем последовательностям длины l, называется *функцией роста* системы множеств S. Для случая задач обучения распознаванию, когда используется эмпирическая выборка и некоторое решающее правило, функции роста имеет большое значение и поэтому уточняется ниже.

Если S - множество решающих правил, $S = \{F(x, \alpha), \alpha \in A\}$, где α - параметр и $x_1, ..., x_l$ - выборка из множества допустимых объектов X, то эта выборка может быть разделена на два класса не более чем 2^l способами. Такое разделение всякий раз фиксируется выбранным правилом $F(x, \alpha)$, подпоследовательностью $X_{F(x,\alpha)}$ последовательности X^l и ее дополнением в $x_1, ..., x_l$. Иначе говоря, с помощью правила $F(x, \alpha)$ множество X^l делится на два подмножества: состоящее из таких x, что $F(x, \alpha) = 1$ и (второе множество) таких, что $F(x, \alpha) = 0$.

Число способов разделения выборки зависит и от класса решающих правил $\{F(x, \alpha), \alpha \in A\}$, и от состава выборки $x_1, ..., x_l$. Это число в рассматриваемом случае и есть $\Delta^S(x_1, ..., x_l)$.

В задачах обучения распознаванию каждый объект x выборки снабжается меткой ω класса, которому он принадлежит. Положим, $\omega \in \{0, 1\}$ и рассмотрим систему событий

$$S(a) = \{(x, \omega) : (\omega - F(x, \alpha))^2 = 1\},$$

соответствующую ошибкам решающего правила $F(x, \alpha)$ на объектах $x \in X$. Обучающая выборка представляется в виде $(x_1, \omega_1), ..., (x_l, \omega_l)$, и система событий $S(\alpha)$ индуцирует на ней $\Delta(S(\alpha); (x_1, \omega_1), ..., (x_l, \omega_l)) = \Delta^S(x_1, ..., x_l)$ различных подвыборок.

Определение 1. [1], [2] Функция $m^S(l) = \max_{x_1,...,x_l} \Delta^S(x_1, ..., x_l)$, где максимум берется по всевозможным выборкам длины l, называется *функцией роста системы событий, образованной решающими правилами* $\{F(x, \alpha), \alpha \in A\}$.

Теорема 1. [1], [2] *Функция роста либо тождественно равна* 2^l, *либо при* $l > h$ *мажорируется функцией* $1.5l^h/h!$, *то есть*

$$m^S(l) \begin{cases} = 2^l, & \text{если } h{=}l\,, \\ < 1.51\dfrac{l^h}{lh!}, & \text{если } h{<}l\,, \end{cases}$$

где $h{+}1$ *- минимальный объем выборки, при котором нарушается условие* $m^S(l) = 2^l$.

Определение 2. [1], [2] Класс решающих функций имеет *емкость* h, если справедливо неравенство $m^S(l) < 1.5l^h/h!, \quad l > h$. Если же $m^S(l) = 2^l$, то говорят, что емкость класса решающих правил бесконечна.

Например, для класса линейных решающих правил в $\mathbb{R}^n$ максимальное число точек h, которые можно с помощью гиперплоскости разбить на два класса всеми возможными способами, равно n. Поэтому при $l > h$ функция роста оценивается неравенством

$$m^S(l) < 1.5\frac{l^n}{n!},$$

и класс линейных отделителей имеет емкость n.

В зарубежной литературе емкость h класса $\{F(x,\alpha), \alpha \in A\}$ решающих правил называют VCD (Vapnik-Chervonenkis Dimension) класса $\{F(x,\alpha), \alpha \in A\}$.

Если $P(\alpha)$-вероятность ошибки при -использовании правила $F(\alpha)$, а $\nu(\alpha)$ - частота этой ошибки на эмпирической выборке, то имеет место неравенство [1], [2]

$$P\{\sup_{\alpha\in A}|P(\alpha) - \nu(\alpha)| > \varepsilon\} < 6m^S(2^l)e^{-\frac{\varepsilon^2 l}{4}},$$

где $\varepsilon > 0$. Эта оценка может быть нетривиальной, когда VCD класса $\{F(x,\alpha), \alpha \in A\}$ конечна и равна h. Тогда $m^S(l) < 1.5\frac{l^h}{h!}$ и

$$P\{\sup_{\alpha\in A}|P(\alpha) - \nu(\alpha)| > \varepsilon\} < 9\frac{(2l)^h}{h!}e^{-\frac{\varepsilon^2 l}{4}} \to 0 \quad \text{при} \quad l \to \infty$$

Таким образом, конечная величина VCD обеспечивает равномерную сходимость частот ошибок к вероятностям при росте длины обучающей выборки и позволяет оценить скорость этой сходимости, поэтому получение VCD для различных классов решающих правил представляет значительный интерес.

Если класс решающих правил конечен,$|\{F(x,\alpha), \alpha \in A\}| = M < \infty$ то, очевидно, $VCD(\{F(x,\alpha), \alpha \in A\}) \leqslant \log M$. Поэтому оценивание мощности конечных классов решающих правил позволяет, используя последнее неравенство, получить для них оценку VCD, а при детерминистской постановке задачи обучения распознаванию для случая нулевой эмпирической ошибки использовать неравенство [1], [2]

$$P\{\sup_{\alpha\in A}P(\alpha) > \varepsilon\} < M(1-\varepsilon)^l,$$

или, с учетом соотношения $(1-\varepsilon) < e^{-l}, \varepsilon > 0$,

$$P\{\sup_{\alpha \in A} P(\alpha) > \varepsilon\} < Me^{-\varepsilon l}.$$

Заметим, что никакая нижняя оценка $L < M$ числа решающих правил M конечного класса $\{F(x,\alpha), \alpha \in A\}$, вообще говоря, не дает возможности оценки его VCD.

2. Оценка VCD деревьев решений ограниченного ранга

Ранг БРД T определяется индуктивно.

Определение 3 (14)**.** , [20]Если БРД T содержит единственную вершину, то его ранг $r = 0$. Если T содержит корень, левое поддерево T_0 ранга r_0 и правое поддерево T_1 ранга r_1 , то

$$rank(T) = \begin{cases} 1 + r_0, & \text{если } r_0 = r_1 \ , \\ max\{r_0, r_1\}, & \text{если } r_0 \neq r_1 \ , \end{cases}$$

БРД T имеет ранг больше чем r тогда и только тогда, когда полное бинарное дерево глубины $r+1$ может быть "встроено"в T [20].

Обозначим rDT_n, следуя работе [20], класс булевых функций от n переменных, представимых в виде БРД ранга не более r, и $VCD(rDT_n)$-емкость этого класса.

В качестве примера на рис.1 представлены два БРД одинакового ранга 2, имеющие различную сложность, традиционно оцениваемую как число внутренних вершин или листьев дерева.

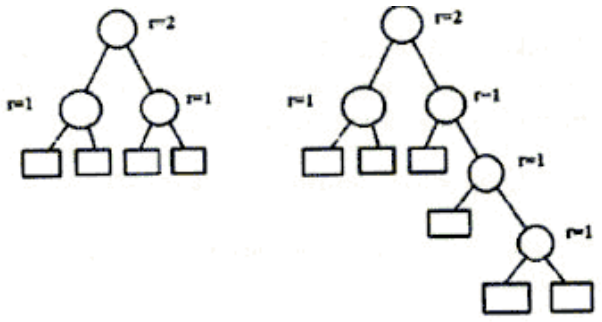

Рис. 1. БРД одинакового ранга 2, имеющие различную сложность

Пример показывает, что такая характеристика, как ранг, не является эффективным показателем качества БРД. При достаточной большой размерности признакового пространства n может оказаться, что дерево небольшого ранга $r \ll n$ будет иметь глубину, близкую к n, что должно приводить к редукции некоторых ветвей. Поэтому результат работы [20] не дает возможности оценивать БРД, построенные при помощи алгоритмов обучения распознаванию, адекватно их сложности и нечувствителен к переподгонке.

Теорема 2. [20]

$$VCD(rDT_n) = \sum_{i=0}^{r} \binom{n}{i}$$

Для дальнейшего изучения деревьев решений понадобится следующее

Следствие 1. *$VCD(rDT_n) = \Theta(n^r)$ при любой заданной константе r и $n \to \infty$.*

Доказательство. Используем теорему 2 и оцениваем сверху и снизу $VCD(rDT_n)$: $\frac{(n-1)^r}{lr!} < \sum_{i=0}^{r} \binom{n}{i} < \frac{n^{r+1}}{n-1}$, следовательно, $\sum_{i=0}^{r} \binom{n}{i} = O(n^r)$ и $n^r = O\left(\sum_{i=0}^{r} \binom{n}{i}\right)$ при $n \to \infty$ и любой зафиксированной константе r, что доказывает следствие. □

3. Оценка VCD бинарного РД с двумя классами

БРД, использующие булевы переменные и метки только двух классов, обеспечивают вычисление соответствующих им булевых функций (б.ф.). Очевидно, что любая б.ф. может быть представлена некоторым БРД [4], поэтому конечный класс БРД как решающих правил очень широк и имеет неограниченную емкость. Конструктивные ограничения, накладываемые на БРД, позволяют выделять конечные подклассы БРД конечной емкости. Таким являлось рассмотренное выше ограничение на величину ранга деревьев.

Рассмотрим конечное множество БРД с не более чем μ листьями и обозначим его соответствующий этому множеству класс булевых функций $BDT(\mu, n)$.

Теорема 3. $\max\{\mu, \log n\} \leqslant VCD(BDT(\mu, n)) < \mu \log(n\mu)$.

Доказательство. Покажем, что $VCD(BDT(\mu, n)) \geqslant \mu$. Действительно, корректная обучающая выборка состоит из попарно различных булевых векторов, поэтому при зафиксированном числе листьев μ для любых μ векторов можно получить любую классификацию, если каждому вектору соответствует единственный лист дерева, и тогда можно любыми 2^μ способами расставить метки классов.

Теперь покажем, что $VCD(BDT(\mu, n))$ может быть больше, чем μ. Пусть $\mu = 2$. Укажем три булевых вектора, которые могут быть расклассифицированы любым способом при $n = 8$. Каждый из вариантов классификации получается, если в единственной внутренней вершине поместить одну из переменных $x_1 - x_3$.

	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8
$\bar{\alpha_1}$	0	0	1	1	0	1	1	0
$\bar{\alpha_2}$	0	1	0	0	1	1	1	0
$\bar{\alpha_3}$	1	0	0	1	1	0	1	0

С учетом примера БРД при $n = 2$ и $\mu = 4$ (рис.1) очевидно, что достижимая нижняя граница VCD определяется неравенством $\max\{\mu, \log n\} \leqslant VCD(BDT(\mu, n))$.

Учитывая, что

$$(BDT(\mu, n)) < (\mu - 1)! 2^{\mu-1} n^{\mu-1} [3],$$

получаем $VCD(BDT(\mu, n)) < \log(\mu - 1)! + (\mu - 1) + (\mu - 1) \log n$.

Известна оценка $lnn! < (n + \frac{1}{2}) \ln(n + 1) - n$[11] , используя которую получаем:

$$\log(\mu - 1)! < (\mu - \frac{1}{2}) \log \mu - (\mu - 1) \log e < \mu \log \mu - (\mu - 1),$$

$$\log(\mu - 1)! + (\mu - 1) + (\mu - 1) \log n < \mu \log \mu + \mu \log n = \mu \log(n\mu).$$

□

Следствие 2. *При любом $\mu = const$ имеет место оценка*

$$VCD(BDT(\mu, n)) = \Theta(logn).$$

Доказательство. Очевидно, что $\log \leqslant \max\{\mu, \log n\}$. Следовательно, $\log n \leqslant VCD(BDT(\mu, n))$ и $\log n = O(VCD(BDT(\mu, n)))$. Из неравенства (1) легко получить оценку . Поэтому $VCD(BDT(\mu, n)) = \Theta(\log n)$.

Сравнивая порядок с полученным выше результатом , убеждаемся, что оптимизация БРД по параметру - листьев более обоснована с точки зрения теории равномерной сходимости Вапника-Червоненкиса и обеспечивает существенно более высокую скорость сходимости. □

4. VCD эмпирического леса

Далее рассматривается алгоритм синтеза r-редуцированного эмпирического леса "по ссылкам".

Уточним алгоритм построения r-редуцированного эмпирического леса, называемый далее ***DFBSA-Decision Forest Building Sequencing Algoritm.***

1. Для построения леса используется непротиворечивая эмпирическая (обучающая) таблица T_{mn}, содержащая булевы наборы значений переменных с указанной принадлежностью одному из классов. Таблица непротиворечива: в ней нет двух одинаковых наборов, принадлежащих разным классам.
2. По заданному $\varepsilon > 0$ и значениям m, n находится такой допустимый ранг r конъюктивной закономерностью, что вероятность случайного обнаружения закономерности ранга r в случайно выбранной таблице не превысит ε. Получение соответствующей оценки подробно описано в [5],[7].
3. Строится бинарное решающее дерево одним из известных методов [4],[6],[10],[12],[16],[17],[18],[19] с учетом следующего правила отсечения: если при построении БРД ранг ветви оказывается больше r, то в этой ветви остается r внутренних вершин, а листья, исходящие из последней по порядку вершины ветви, помечаются следующим образом. Если какой-нибудь лист

из указанных двух листьев соответствует интервалу куба B^n, в который попадают наборы только одного класса из T_{mn}, то этот лист помечается соответствующей меткой класса. Иначе лист помечается указателем (ссылкой) на корневую вершину следующего дерева, которое предстоит построить. Такое правило отсечения приводит к получению БРД, листья которого помечены либо метками классов, либо ссылками на следующее дерево.

4. Пусть $PREV\_USED$-список, использованных переменных (признаков) при построении $k-1$-го дерева. Для синтеза k-го дерева выделяются все наборы таблицыT_{mn}, которые "попали"в интервалы, соответствующие ветвям, заканчивающимся листьями со ссылками от $k-1$-го дерева (обозначим совокупность всех интервалов, образующих область отказа $k-1$-го дереваN_{otk}^{k-1}). Эти наборы составляют некоторую подтаблицу $T_{|N_{otk}^{k-1}|,n-|PREV\_USED|} \subset T_{mn}$, $|N_{otk}^{k-1}| < m$. Построение k-го дерева осуществляется на основании таблицы $T_{|N_{otk}^{k-1}|,n-|PREV\_USED|}$ с учетом нового порядка отбора признаков для внутренних вершин. Сначала используются признаки, не вошедшие в список$PREV_USED$, упорядоченные по используемому критерию выбора. Если их не хватает для синтеза ветвей допустимого ранга, используются признаки списка $PREV\_USED$. Далее создается список $NEXT\_USED$ признаков, использованных при построении k-го дерева. Затем k-ое дерево достраивается на объектах, попавших в таблицу $T_{mn}\backslash T_{|N_{otk}^{k-1}|,n-|PREV\_USED|}$ и любых признаках, не вошедших в список $NEXT\_USED$. После достройки k-го дерева формируется его список $PREV\_USED$, в который войдут все признаки, использованные как в процессе построения, так и в процессе достройки k-го дерева. После того как k-ое дерево построено, снова проверяется одно из выбранных условий прекращения синтеза эмпирического леса (см.х[5]).
5. Пусть уже построено $k \geqslant 1$ деревьев. Условие прекращения синтеза состоит в следующем: если k-ое БРД не содержит ссылок в листьях (а содержит только метки классов), или, в противном случае, $k = q$, где константа q задает ограничение на возможное число деревьев эмпирического леса, либо уже были использованы все переменные, либо область отказа первого дерева не сужается (т.е. k-ое дерева не "приписывает"метку класса ни одному из объектов, попавших в пересечение областей отказа первого и $k-1$-го деревьев) - процесс синтеза леса завершается. Получается либо корректный (не делающий ошибок на объектах таблицы T_{mn}), либо некорректный относительно обучающей таблицы лес. Если условие прекращения синтеза не выполняется, то начинается синтез следующего дерева.

Суть алгоритма ***DFBSA*** состоит в том, что последовательно строятся q эмпирических деревьев не более чем с $\mu \leqslant 2^r$ листьями с учетом редукции ветвей. В итоге

получается не более чем $q - 2^r$ решающих ветвей (конъюнкций). Если каждое отдельное БРД, входящее в r-редуцированный лес, определяет ортогональную ДНФ, содержащую не более 2^r конъюнкций, то в целом по всему лесу конъюнкции, соответствующие разным деревьям, могут быть и неортогональными. Это становится очевидным, если предположить, например, что при большом числе n переменных два разных дерева, входящие в лес, используют во внутренних вершинах непересекающиеся подмножества этих переменных-признаков.

На рис.2 представлена схема эмпирического леса. Пометки K соответствуют корректным ветвям (с метками классов) и для последнего по порядку синтеза дерева–терминальную метку отказа (если последнее дерево оказалось некорректным).

Поскольку процесс синтеза БРД эквивалентен построению разбиения куба B^n, в процессе достройки деревьев алгоритмов с область отказа не расширяется, даже если последующие деревья имеют большее число некорректных ветвей, чем исходное.

Алгоритм принятия решений на основе построенного эмпирического леса, называемый далее ***DMFSA*** *(**D**ecision **M**aking **F**orest **S**equencing **A**lgorithm) состоит в следующем.*

Предоставляется описание объекта — набор значений n булевых признаков. Указатель устанавливается на первое дерево. Согласно значениям признаков осуществляется "прохождение" ветви и определение соответствующего листа. Если этот лист помечен меткой класса — объект опознан. Иначе, если это ссылка, указатель устанавливается на следующее (определенное порядком синтеза при выполнении алгоритма ***DFBSA*** дерево. Если в итоге будет получена терминальная ссылка без метки класса — принимается решение об отказе опознавания предъявленного объекта.

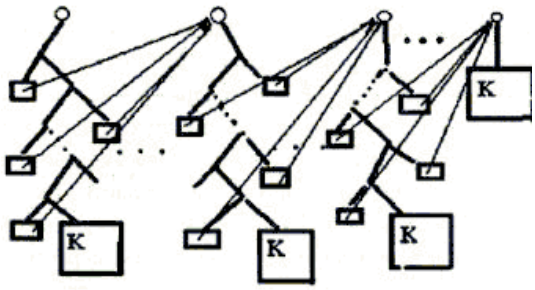

Рис. 2. Схема эмпирического леса

Очевидно, что результат работы алгоритма ***DFBSA*** определяется построенным лесом и предопределенным порядком просмотра деревьев леса. Переход на новое дерево соответствует классификации заново, по измененной системе признаков, другой ветвью. Предыдущий отказ от решения и ссылка на следующее дерево никакого значения для последующей классификации не имеет. Используемые конъюнкции остаются r-редуцированными.

Согласно описанному правилу принятия решений, совокупность ветвей с пометками одного и того же класса эквивалентна заданию некоторой ДНФ, и число различных решающих правил, определяемых r-редуцированным эмпирическим лесом, не больше числа различных ДНФ, состоящих из конъюнкций, каждая из которых имеет ранг не более r.

Число различных конъюнкций ранга не более r равно $\sum_{i=0}^{r} 2^i \binom{n}{i}$. Легко проверить, что

$$\frac{2^r(n-r)^r}{r!} < \sum_{i=1}^{r} 2^i \binom{n}{i} < \frac{(2n)^{r+1}-1}{2n-1}.$$

Оценим мощность и *VCD конечного класса* $DNF(n,\mu,r)$ *решающих правил, образованных дизъюнктивными нормальными формами, содержащими не более* μ *конъюнкций ранга не более* r, *состоящими из литералов* n *переменных.*

Теорема 4.

$$\frac{\left(\left(\frac{2n}{r}-1\right)^r-\mu\right)^\mu}{\mu!} < |DNF(n,\mu,r)| < \frac{1.5^\mu n^{r\mu}}{\mu!}$$

Доказательство. Оценка мощности класса решающих правил $DNF(n,\mu,r)$ снизу получается, если рассмотреть только одинаковые по рангу $r > 1$ конъюнкции с одинаковым числом инверсий равным $[\frac{r}{2}]$. Обозначим множество таких конъюнкций $K(n,r,\frac{r}{2})$. Каждые две различные составленные из этих конъюнкций ДНФ определяют две различные функции из $P_2(n)$. Это следует из того, что для любой конъюнкции из $K(n,r,[\frac{r}{2}])$ можно указать набор значений переменных, на котором она обращается в единицу, а любая отличная от нее конъюнкция обращается в нуль. Действительно, пусть $L = \bar{x}_{i_1} \cdot \ldots \cdot \bar{x}_{i_{[r/2]}} \cdot x_{i_{[r/2]-1}} \cdot \ldots \cdot x_{i_r}$ - произвольная конъюнкция ранга r с ровно $[\frac{r}{2}]$ отрицательными литералами. Пусть H-любая конъюнкция из множества $K(n,r,[\frac{r}{2}])$, отличная от L. Если H состоит из тех же переменных, что и L, то хотя бы одна из переменных $x_{i_1} \cdot \ldots \cdot x_{i_{[r/2]}}$ войдет в конъюнкцию H без инверсии. Не теряя общности, пусть это будет переменная x_{i_1}. Тогда для любой точки $\tilde{\alpha} = \alpha_1, \ldots, \alpha_n$ такой, что $\alpha_{i_1} = 0, \ldots, \alpha_{i_{[r/2]}} = 0, \alpha_{i_{[r/2]+1}} = 1, \ldots, \alpha_{i_r} = 1$, будет иметь место $L(\tilde{\alpha}) = 1$, а $H(\tilde{\alpha}) = 0$. Пусть теперь в H содержится хотя бы одна переменная x_p, не содержащаяся в L. Если переменная x_p входит в H, без инверсии, то для точки $\tilde{\alpha}$ такой, что $\alpha_{i_1} = 0, \ldots, \alpha_{i_{[r/2]}} = 0, \alpha_{i_{[r/2]+1}} = 1, \ldots, \alpha_{i_r} = 1, \alpha_p = 0$, будет иметь место $L(\tilde{\alpha}) = 1$, а $\mathrm{H}(\tilde{\alpha}) = 0$. Пусть теперь в H содержится хотя бы одна переменная x_p, не содержащаяся L. Если переменная x_p входит в H без инверсий, то для точки $\tilde{\alpha}$ такой, что $\alpha_{i_1} = 0, \ldots, \alpha_{i_{[r/2]}} = 0, \alpha_{i_{[r/2]+1}} = 1, \ldots, \alpha_{i_r} = 1, \alpha_p = 0$, будет иметь место $L(\tilde{\alpha}) = 1$, а $\mathrm{H}(\tilde{\alpha}) = 0$. Если же переменная x_p входит в H с инверсией, то для точки $\tilde{\alpha}$ такой, что $\alpha_{i_1} = 0, \ldots, \alpha_{i_{[r/2]}} = 0, \alpha_{i_{[r/2]+1}} = 1, \ldots, \alpha_{i_r} = 1, \alpha_p = 0$, будет иметь место $L(\tilde{\alpha}) = 1$, а $\mathrm{H}(\tilde{\alpha}) = 0$.

Учитывая установленное свойство конъюнкций множества $K(n,r,[\frac{r}{2}])$, получаем

$$\left|K(n,r,[\frac{r}{2}])\right| = \binom{r}{[\frac{r}{2}]}\binom{n}{r},$$

$$|DNF(n,\mu,r)| > \binom{\binom{r}{[\frac{r}{2}]}\binom{n}{r}}{\mu},$$

$$\binom{r}{[\frac{r}{2}]}\binom{n}{r} > \left(\frac{2n}{r}-1\right),$$

$$|DNF(n,\mu,r)| > \binom{\left(\frac{2n}{r}-1\right)^r}{\mu} > \frac{\left(\left(\frac{2n}{r}-1\right)^r-\mu\right)^\mu}{\mu!}.$$

Для оценки мощности класса $DNF(n,\mu,r)$ решающих правил сверху рассмотрим сумму

$$\sum_{i=1}^{r} 2^i\binom{n}{i} < 2n+\frac{(2n)^2}{2!}+\frac{(2n)^3}{3!}+\ldots+\frac{(2n)^r}{r!} < 2n+\frac{(2n)^2}{2!}+\frac{(2n)^3}{3!}+\frac{(2n)^4}{4!}+n^5+\ldots+n^r.$$

Легко проверить, что

$$2n+\frac{(2n)^2}{2!}+\frac{(2n)^3}{3!}+\frac{(2n)^4}{4!} < n+n^2+n^3+n^4 \quad \text{при} \quad n>2.$$

Учитывая что,

$$n+n^2+\ldots+n^r = \frac{n^{r+1}-n}{n-1} < \frac{n^{r+1}}{n-1} = \frac{n}{n-1}\cdot n^r < 1.5n^r \quad \text{при} \quad n>2.$$

$$\binom{1.5n^r}{\mu} = \frac{(1.5n^r)!}{\mu!(1.5n^r-\mu)!} < \frac{(1.5n^r)^\mu}{\mu!} = \frac{1.5^\mu n^{r\cdot\mu}}{\mu!}$$

□

Следствие 3.

$$|DNF(n,\mu,r)| = \Theta(n^{r\cdot\mu}).$$

Следствие 4.

$$VCD(DNF(n,\mu,r)) < r\mu\log n - \mu\log\frac{\mu}{2} = O(\log n).$$

Следствие 5.

$$VCD(DNF(n,\mu,r)) = \Theta(\log n).$$

Доказательство. В следствии 2 получена оценка $\log n = O(VCD(BDT(\mu,n)))$ для любого возможного ранга $r \leqslant \mu-1$, поэтому $\log n = O(VCD(BDT(\mu,n,r)))$. Очевидно, что $BDT(\mu,n,r) \subset DNF(n,\mu,r)$, поэтому $VCD(DNF(n,\mu,r)) = O(\log n)$ по следствию 4. □

Теорема 5.

$$\max(\mu q, \log n) < VCD(BDF(n, \mu, r, q) < r\mu q \log n - \mu q \log \frac{\mu q}{2}.$$

Доказательство. Оценка сверху. Пусть $F - r$-редуцированный эмпирический лес. Для любой точки $\bar{x} \in B^n$ классифицирующая ее конъюнкция, соответствующая некоторой ветви одного из деревьев леса F, определяется структурой леса однозначно. Поэтому решающее правило, определяемое лесом $F \in BDF(n, \mu, r, q)$, соответствует некоторой зафиксированной ДНФ. Следовательно, если решающее правило $\gamma \in \mathrm{B}DF(n, \mu, r, q)$, то $\gamma \in DNF(n, \mu q, r)$, и тогда

$$VCD(BDF(n, \mu, r, q)) < VCD(DNF(n, \mu q, r)) < r\mu q \log n - \mu q \log \frac{\mu q}{2}$$

Оценка снизу. Используем оценку для одного дерева с μq листьями, учитывая, что она была получена при рассмотрении случая лишь одной внутренней вершины. Очевидно, что

$$BDT(\mu, n, 1) \subset BDT(\mu, n, r) \subset BDF(n, \mu, r, q),$$

откуда получаем $\max(\mu q, \log n) < VCD(\mathrm{B}DF(n, \mu, r, q))$ (см. доказательство теоремы 3). □

Следствие 6.

$$VCD(BDF(n, \mu, r, q)) = \Theta(\log n).$$

Следствие 7.

$$VCD(BDF(n, \mu, r, \mu)) < r\mu \log n - \mu \log \frac{\mu}{2}.$$

Заключение

Основным результатом данной статьи является доказанное неравенство

$$\max(\mu q, \log n) < VCD(BDF(n, \mu, r, q)) < r\mu q \log n - \mu q \log \frac{\mu q}{2}$$

и следствие $VCD(BDF(n, \mu, r, q)) = \Theta(\log n)$, согласно которым можно сделать *главный вывод*: переход от БРД к r-редуцированному эмпирическому лесу не изменяет порядка VCD, т.е. не существенно увеличивает сложность расширенного класса решающих правил. В то же время обеспечивается коррекция, позволяющая настроиться по обучающей выборке на правильную классификацию как можно большего числа объектов. Такая возможность объясняется прежде всего тем, что эмпирический лес, вообще говоря, использует различные подсистемы признаков для синтеза отдельных деревьев, не увеличивая при этом рангов решающих конъюнкций.

Представляется перспективным дальнейшее изучение и разработка алгоритмов синтеза редуцированных решающих лесов, которые существенно зависят от способов выбора переменных для построения БРД, составляющих лес.

Список литературы

1. Вапник В.Н. Восстановление зависимостей по эмпирическим данным. - М.: Наука, 1979. - 447с.
2. Вапник В.Н., Червоненкис А.Я. Теория распознавания образов - М.: Наука, 1974. - 416 с.
3. Донской В.И. Асимптотика числа бинарных решающих деревьев // Ученые записки Таврического национального университета им. В.И. Вернадского, Серия "Математика". -2001. - Т. 14(53),№ 1. - С.36-38.
4. Донской В.И., Башта А.И. Дискретные модели принятия решений при неполной информации - Симферополь: Таврия, 1992. - 166 с.
5. Донською В.Й., Дюлічева Ю.Ю. Бінарні розв'язуючі дерева в задачах інтелектуального аналізу інформації // Наукові вісті Національного технічного університету України "Київський політехнічный Інститут". - 2001. - вип.5. - с.12-18.
6. Донской В.И., Дюличева Ю.Ю. Деревья решений с k-значными переменными // Труды Междунар. конфю "Знание - Диалог - Решение". - Том 1. - Санкт-Петербург: Изд-во "Лань". - 2001.- С.201-207.
7. Донской В.И., Дюличева Ю.Ю. Индуктивная модель r-корректного эмпирического леса // Труды Междунар. конф. ио индуктивному моделированию. - Львов. - 2002. - С.54-58.
8. Дюличева Ю.Ю. Принятие решений на основе индуктивной модели эмпирического леса // Искусственный интеллект. - 2002. - №2. - С.110-115.
9. Дюличеиа Ю.Ю. Стратегии редукции решающих деревьев(обзор) // Таврический вестник информатики и математики. - 2002. - №1. - С.10-17.
10. Норушис А. Построение логических (древообразных) классификаторов методами нисходящего поиска (обзор) // Статистические проблемы управлении. - 1990. - вып.93. - С.131-157.
11. Яблонский С.В. Введение в дискретную математику. - М.: Наука, 1986. - 384 с.
12. Breslow L.A.,D.W.Aha Simplifying Decision Trees: A Survey // Knowledge Engineering Review. - 1997. - Vol.12. - P.1-40.
13. Devroye L., Gyorfi L., Lugosi G. A Probabilistic Theory of Pattern Recognition. - Springer-Verlag, NY, 1996. - 636 p.
14. Elomaa T. The Biases of Decision Tree Pruning Strategies // In D. Hand, J. Kok. M. Berthold (eds.), Advances in Intelligent Data Analysis, Proc. 3rd IDA, Lecture Notes in Computer Science 1642 - Springer. - 1999. - P.63-74.
15. Golea M., Barlett P.L., Lee W.S., Mason L. Generalization in decision trees and DNF: Does size matter? // Advances in Neural Information Processing Systems. - MIT Press. - 1998. - Yol.10. - P.259-265.
16. Kothari R., Dong M. Decision Trees for Classification: A Review and Some New Results Pattern Recognition: From Classical to Modern Approaches, S. R. Pal, and A. Pal (Eds.). - World Scientific. - 2001. - Chapter 6. - P.169.184.
17. Malerba D., Esposito F., Semeraro G. A Further Comparison Methods of Decision Tree Induction // Learning From Data: Artificial Intelligence and Statistics V, D. Fisher and H. Lenz, eds., Lecture Notes in Statistics. - Berlin: Springer. - 1996. - № 112. - P.365-374.
18. Mingers J. An empirical comparison of selection measures for decision-tree induction // Machine Learning.- 1989.- Vol.3.- P.319-342.
19. Safavian S.R., Landgrebe D. A Survey of Decision Tree Classifier Methodology // IEEE Transactions on Systems, Man, and Cybernetics. - 1991. - Vol.21, №3 - P.660-674.
20. Simon H. The Vapnik-Chervonenkis Dimension of Decision Trees with Bounded Rank // Information Processing Letters. - 1991. -39. - P.137-141.

21. Schaffer C. When Does Overfitting Decrease Prediction Accuracy in Induced Decision Trees and Rule Sets? // In Proceedings of the European Working Session of Learning (EWSL-91). - Berlin. - 1991. - P.192-205.
22. Schaffer C. Overfitting Avoidance as Bias // Machine Learning. - 1993. - Vol 10 - P.153-178.