

УДК 519.718

С. О. Кузнецов

УСТОЙЧИВОСТЬ КАК ОЦЕНКА ОБОСНОВАННОСТИ ГИПОТЕЗ, ПОЛУЧАЕМЫХ НА ОСНОВЕ ОПЕРАЦИОНАЛЬНОГО СХОДСТВА

Исследуется проблема устойчивости идемпотентной, коммутативной и ассоциативной операции обобщения. Устойчивость понимается как хорошая воспроизводимость на подвыборках обобщения множества всех фактов из выборки. Вводимые понятия родственны идеям, лежащим в основе некоторых математических и прикладных методов, например, jackknife-метода. Устанавливается трудноразрешимость ($\neq P$ -полнота) задачи вычисления индексов устойчивости, а также границы их изменения при пополнении выборки. Приведен алгоритм вычисления индексов устойчивости, линейный от их величины, а также приближенный алгоритм, полиномиальный от размера входа. Приводится описание машинного эксперимента по использованию индексов устойчивости для отбора гипотез в задаче технической диагностики.

1. ОПЕРАЦИЯ СХОДСТВА И ГИПОТЕЗЫ НА ЕЕ ОСНОВЕ

Большинство систем Автоматического Обучения (Machine Learning) — далее АО — использует то или иное понятие сходства как средство, позволяющее выделять закономерности в исследуемых объектах. Сходство также применяется для классификации новых объектов с помощью найденных закономерностей. Сходство обычно понимается как отношение [1, 2], метрически [1, 3, 4], либо как операция, сопоставляющая нескольким исходным объектам подобъект, выражающий их сходство [5–14]. В данной работе сходство будет пониматься как идемпотентная, коммутативная и ассоциативная операция на парах объектов в духе [7, 8]. Эти достаточно естественные свойства операции сходства позволяют однозначно выражать сходство множества объектов через попарные сходства независимо от порядка расположения объектов в базе данных* (например, [15]). Такое определение сходства используется, например, в ДСМ-методе автоматического порождения гипотез (ДСМ-АПГ) [6].

Итак, пусть S — множество объектов, представляющих некоторую предметную область. Операцию на парах объектов из S будем называть операцией сходства, если она задает на множестве S нижнюю полурешетку, т. е. для произвольных объектов x, y, z из S имеют место

* Альтернативой такой операции может служить параметрическое семейство n -местных операций, где n — число объектов, для которых ищется сходство (например, [11]).

соотношения (1)–(4):

$$(1) x \sqcap x = x;$$

$$(2) x \sqcap y = y \sqcap x;$$

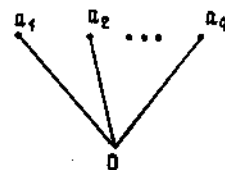
$$(3) x \sqcap (y \sqcap z) = (x \sqcap y) \sqcap z;$$

(4) $x \sqcap s_0 = s_0$ для некоторого s_0 из S , называемого пустым объектом.

Операция \sqcap естественным образом задает на S отношение вложения: $x \sqsubseteq y \Leftrightarrow x \sqcap y = x$ и строгого вложения: $x \sqsubset y \Leftrightarrow (x \sqcap y) \& (x \neq y)$.

Можно указать следующие примеры операций, обладающих свойствами (1)–(4).

1. Булева нижняя полурешетка $\langle 2^U, \cap, \emptyset \rangle$. Такого рода определение сходства соответствует представлению данных в виде множеств дескрипторов, а операции сходства — операции пересечения множеств. Подобное представление используется во многих системах АО и, в частности, в ДСМ-АПГ. В это представление легко переводятся также представления сходства с помощью полурешеток на кортежах фиксированной длины с покомпонентным «векторным» заданием сходства (см. [14], а также п. 6 данной работы):



2. Полурешетка на N -множествах гиперграфов с упорядоченными метками вершин и гиперребер [12—14], в которой результат действия операции схождения на паре множеств гиперграфов \mathcal{S} и \mathcal{H} есть множество всех максимальных по вложению общих подгиперграфов гиперграфов из \mathcal{S} и \mathcal{H} .

3. Интерполяционная полурешетка интервалов [15]. Пусть (i, j) и (k, l) где $U < i, j, k, l < S, i < j, k < l$ — пары чисел, представляющие границы интервалов, U и S — минимальное и максимальное возможные значения. Тогда схождение указанных интервалов есть $(i, j) \wedge (k, l) = (p, q)$, где $p = \min(i, k), q = \max(j, l)$. Нетрудно проверить, что тройка $\langle \{(i, j) \mid U \leq i < j < S\}, \wedge, (U, S) \rangle$ есть нижняя полурешетка, т. е. обладает свойствами (1)—(4). Исходные значения числовых признаков можно задавать при этом парами вида (x, x) , если считается, что данные задаются точно, и парами различных значений вида: нижняя граница возможного значения; верхняя граница возможного значения, — если это не так.

Перейдем теперь к определению гипотез в соответствии с [6]. Пусть нами исследуется некоторое свойство W объектов из S . Множество всех объектов из S , про которые известно, что они обладают свойством W , будем обозначать S^+ , множество объектов, про которые известно, что они не обладают свойством W , будем обозначать S^- . Множество объектов из S , про которые не известно, обладают ли они свойством W или нет, будем обозначать S^c . Таким образом, $S^c = S \setminus (S^+ \cup S^-)$.

Определение 1.1. h есть локальное схождение X_1, \dots, X_n из S , если $X_1 \cap \dots \cap X_n = h$.

Определение 1.2. $\langle h, \{X_1, \dots, X_n\} \rangle$ есть глобальное схождение относительно множества $S' \subseteq S$, если $X_1, \dots, X_n \in S', X_1 \cap \dots \cap X_n = h$, и для произвольного $Y: Y \in S' \setminus \{X_1, \dots, X_n\}$ имеет место $Y \cap h \neq h$ (таким образом, h есть локальное схождение объектов из $\{X_1, \dots, X_n\}$, и это множество есть множество всех объектов из S' , содержащих h).

Определение 1.3. $\langle h, \{X_1, \dots, X_n\} \rangle$ есть положительная (или (+)-гипотеза) гипотеза (относительно причины свойства W), если $\langle h, \{X_1, \dots, X_n\} \rangle$ есть глобальное схождение относительно множества S^+ и h не есть под-объект (в смысле \subseteq) какого-либо объекта из S^- . h будем называть головной гипотезы. Отрицательные (или (-)-гипотезы) гипотезы (о причинах отсутствия свойства W) определяются двойственным образом.

Определение 1.3 — это одно из возможных определений гипотез на основе операции схождения — возможно требование выполнения других условий: более сильных, более слабых или не сравнимых по силе с условием из определения 1.3 (называемого в ДСМ-методе условием «с запретом на контрпример»). Это условие выбрано здесь как наиболее типичное для систем автоматического обучения — требующее, чтобы обобщение примеров не имело в качестве частных случаев контрпримеров. Для проблемы устойчивости гипотез важен не конкретный вид условия, а то, что гипотезы суть глобальные схождения.

Гипотезы, полученные в соответствии с определением 1.3, могут, с одной стороны, представлять самостоятельную прикладную ценность, а с другой стороны, могут использоваться в рамках системы для распознавания или прогноза наличия/отсутствия свойства W у объектов из S^c . Приведем простейший вариант правила прогноза — принцип индуктивного обобщения (ПИО) в соответствии с [6].

Определение 1.4. Объект $P \in S^c$ назовем (+)-прогнозом (или (+)-гипотезой II рода в терминологии ДСМ-метода, [6]), если существует (+)-гипотеза $\langle h, \{X_1, \dots, X_n\} \rangle$, такая, что $h \subseteq P$ и для любой (-)-гипотезы $\langle h', \{Y_1, \dots, Y_k\} \rangle$ имеет место $h' \not\subseteq P$.

Отрицательный прогноз ((-)-прогноз) определяется двойственным образом.

2. УСТОЙЧИВОСТЬ: МОТИВАЦИЯ И ПРЕЦЕДЕНТЫ

Определения 1.3 и 1.4 подразумевают, что существенной причиной свойства W может быть все то общее, что есть у какого-либо множества объектов X_1, \dots, X_n , обладающих свойством W . Все то, что есть у них несходного, неявно считается несущественным для образования гипотезы $\langle h, \{X_1, \dots, X_n\} \rangle$. Какова степень обоснованности данного утверждения? Несомненно, что (+)-гипотеза, полученная по определению 1.3, может считаться более обоснованной, чем гипотеза, удовлетворяющая более слабому условию, которое требует лишь того, чтобы голова (+)-гипотезы не была бы сходом (-)-примеров («слабое правило» [6]).

Несомненно также, что более обоснованной представляется гипотеза, соответствующая глобальному сходу большего количества примеров. Но и это не все. В одном случае (скажем, для некоторой гипотезы $H_1 = \langle h_1, \{X_1, \dots, X_n\} \rangle$) эти примеры могут быть в некотором смысле «зависимыми», «слишком похожими», как полученные, например, в одной серии экспериментов одним и тем же экспериментатором.

В другом случае (скажем для некоторой гипотезы $H_2 = \langle h_2, \{Y_1, \dots, Y_n\} \rangle$) примеры могут быть более «независимыми», как полученные, например, разными экспериментаторами или разными методиками и т. п. и, вследствие этого, могут отличаться друг от друга во всем, кроме структурного фрагмента h_2 . Это, в частности, будет также означать, что гипотеза H_2 может быть получена и на основе меньшего числа примеров, так как независимость экспериментов позволяет надежнее и быстрее отделить существенно-причинные подструктуры от несущественных и лишь сопутствующих какому-либо типу экспериментов. Это, в свою очередь, означает, что H_2 будет подтверждаться на большем числе подмножеств исходного исчерпывающего множества $\{Y_1, \dots, Y_n\}$ примеров для H_2 , т. е. H_2 обладает большей «устойчивостью», «воспроизводимостью» при случайном пополнении или уменьшении множества исходных примеров. Поскольку каждый пример получен до некоторой степени случайно, устойчивость к такого рода случайностям говорит в пользу большего правдоподобия гипотезы H_2 .

Идея устойчивости использовалась для анализа достоверности гипотез различной природы. К методам, использующим понятия устойчивости, относятся, несомненно, методы непараметрической статистики — метод складного ножа (jackknife) и бутстреп (bootstrap) [16], а также методы скользящего контроля. Простейшим примером неявного использования представлений об устойчивости является построение экстраполяционных многочленов. Пусть имеется n точек x_1, \dots, x_n пространства R^n , по которым нужно построить многочлен, на графике которого лежат точки x_1, \dots, x_n . В общем случае может быть построен многочлен степени не большей n , удовлетворяющий этим условиям. Но если можно построить многочлен P по некоторому подмножеству точек $\{x_1, \dots, x_k\} \subseteq \{x_1, \dots, x_n\}$, так что на его графике лежат все точки x_1, \dots, x_n , то такой многочлен будет иметь степень не выше k , и P , как гипотеза о закономерности, будет проще, а следовательно, более достоверной, чем гипотеза, соответствующий многочлен которой можно построить лишь по всему множеству $\{x_1, \dots, x_n\}$.

Идея устойчивости, как уже говорилось, лежит в основе некоторых непараметрических методов статистики. Так, в методе складного ножа оценка дисперсий произвольных статистик (функций от выборок) получается следующим образом. Из начальной выборки размера n составляются всевозможные подвыборки размера $n-1$. Для i -й подвыборки вычисляется значение S_i интере-

сущей нас статистики S . Составляется среднее S^* этих значений, а затем вычисляется среднее квадратов отклонений S_i от S^* . Полученная величина (с точностью до несущественных арифметических преобразований) будет оценкой дисперсии статистики S по методу складного ножа. Этот метод в ряде случаев дает оценку лучшую, чем другие, традиционные методы. Возможны модификации описанного метода с использованием всех подвыборок размера $n-2$, $n-3$ и т. д., однако это требует значительных вычислительных затрат.

Оценка дисперсии статистики по методу бутстрепа определяется сходным образом, однако новые выборки, также имея размер n , порождаются из исходной n -кратным выбором с возвратом (таким образом, каждый элемент исходной выборки может входить от 0 до n раз в новую выборку).

В некоторых работах идея устойчивости используется вне вероятностных моделей. Так, в работе [17] ставится следующая задача. Пусть дано множество X_N , состоящее из N точек, для которых известно значение функции $f(x)$. Требуется в заданном классе моделей (функций с параметром Θ) $K = \{\mu(x, \Theta) : \Theta = (\theta_1, \dots, \theta_m), \theta_m \in R^m\}$ выделить такую, которая на множестве $X \subset R^n$ ($X \cap X_N = \emptyset$), для точек которого нужно оценить значения функции f , обладала бы наибольшей устойчивостью. Устойчивость определяется как близость для $x \in X$ значений функций вида $\mu(x, \Theta)$, построенных на подвыборках выборки X_N , к значению функции, построенной по всей выборке X_N . Результаты экспериментов, приведенные в [17], свидетельствуют о хорошей избирательной способности предложенного метода.

Идеи, родственные идеям устойчивости, лежат также в основе некоторых вероятностных логик [18]. В работах этого направления, начавшегося, по-видимому, с работ Р. Карнапа [19] по индуктивным логикам, оценка обоснованности утверждения основана на количестве миров (или, в более общей постановке, на численной мере миров), в которых это утверждение выводимо.

В перечисленных выше областях идея устойчивости, как воспроизводимости результатов на подвыборках, осуществлялась для числовых данных. В случае, когда нечисловые данные образуют нижние полурешетки, подобные полурешеткам 1 и 2 из п. 1, обоснование использования устойчивости сходства имеет дополнительный аспект. Хорошая воспроизводимость сходства h на подмножествах множества $\{X_1, \dots, X_n\}$ (т. е. существование большого количества подмножеств $\{X_{i_1}, \dots, X_{i_q}\}$, таких, что $X_{i_1} \sqcap \dots \sqcap X_{i_q} = h$) означает, что множества из $\mathcal{E} = \{(X_1 \setminus h), \dots, (X_n \setminus h)\}$ «мало» сходны между собой*. То, что сходство «остатков» (множеств из \mathcal{E}) нетривиально ($\neq 0$), говорит в пользу того, что причиной W является скорее не h , а некоторые $h_1, \dots, h_q \sqsupset h$: в силу свойств (1)–(4) операции сходства \sqcap , сходства меньшего количества объектов суть меньшие (в смысле отношения \sqsupset) объекты.

Таким образом, плохая воспроизводимость гипотез на подмножествах примеров или низкая устойчивость говорит о том, что в «остатках» из множества X содержится подструктуры или их части, существенные для проявления свойства W , а h может быть причиной свойства W лишь в сочетании с некоторыми «добавками» (дополняющими h до h_1, \dots, h_q).

3. ОСНОВНЫЕ ОПРЕДЕЛЕНИЯ

Все определения будем приводить для положительных гипотез, для отрицательных гипотез определения устойчивости носят двойственный характер. Итак, пусть $H = \langle h, \{X_1, \dots, X_n\} \rangle$ есть (+)-гипотеза I рода. Обо-

* \setminus есть либо булева разность для решетки 1, либо псевдоразность для решетки 2 [12–14].

значим

$$\langle H \rangle_j = \{ \{X_{i_1}, \dots, X_{i_j}\} \mid \{X_{i_1}, \dots, X_{i_j}\} \subset \{X_1, \dots, X_n\},$$

$$X_{i_1} \sqcap \dots \sqcap X_{i_j} = h \};$$

$$\langle H \rangle_\Sigma = \bigcup_{j=2}^{n-1} \langle H \rangle_j;$$

$$g_j(j, H) = \# \langle H \rangle_j, \quad g_\Sigma(\Sigma, H) = \# \langle H \rangle_\Sigma.$$

В тех случаях, когда понятно, о какой гипотезе идет речь, аргументы у $g_j(j, H)$ будем опускать, т. е. писать g_j (или g_Σ).

Определение 3.1. Индекс устойчивости гипотезы $\langle H, \{X_1, \dots, X_n\} \rangle$ j -го уровня для $2 \leq j \leq n-1$ есть

$$I_j = \frac{g_j}{\binom{n}{j}}.$$

Определение 3.2. Интегральный индекс устойчивости гипотезы $\langle H, \{X_1, \dots, X_n\} \rangle$ есть

$$I_\Sigma = \frac{g_\Sigma}{2^n - n - 2}.$$

Определение 3.3. Усредненный индекс устойчивости гипотезы $\langle H, \{X_1, \dots, X_n\} \rangle$ есть

$$I_m = \frac{1}{n-2} \left(\sum_{j=2}^{n-1} I_j \right).$$

Индексы устойчивости соотносятся с операцией сходства также как выборочное среднее соотносится с выборочной дисперсией (подсчитанной по методу складного ножа, см. приведенное выше выражение для оценки дисперсии *jackknife-методом*) в статистике. Видно также, что определение оценок устойчивости в духе бутстрепа не оправдано. В самом деле, взяв i раз в бутстреп-выборку размера n какого-либо объекта из $\{X_1, \dots, X_n\}$ обозначало бы в силу идемпотентности (свойства (1)) операции \sqcap , просто сокращение количества других объектов в бутстреп-выборке. В результате мы имели бы также те же типы выборок, как и при отбрасывании примеров, но эти выборки вносили бы разные вклады в индексы устойчивости. Выборки с большим ($\sim n$) и малым (~ 2) числом примеров обладали бы меньшей вероятностью появления в новых выборках, чем выборки со средним числом примеров. Это предположение не представляется оправданным, так как априори среди выборок нет более предпочтительных.

Следующее свойство индексов устойчивости является следствием простого свойства монотонных булевых функций*: относительное число единиц монотонной булевой функции в $(j+1)$ -м слое булева гиперкуба больше чем в j -м слое.

Лемма 3.1. Для произвольной гипотезы $\langle n, \{X_1, \dots, X_n\} \rangle$ имеет место

$$I_2 \leq \dots \leq I_{n-1}.$$

Доказательство. Рассмотрим семейства $\langle H \rangle_j$, $\langle H \rangle_{j+1}$, а также двудольный граф B , образованный j -м и $(j+1)$ -м слоями булева гиперкуба $2^{\{X_1, \dots, X_n\}}$. В графе B

* А таковой для фиксированной гипотезы $H = \langle h, \{X_1, \dots, X_n\} \rangle$ является функция

$$Y \subset \{X_1, \dots, X_n\} \Rightarrow f(Y) = \begin{cases} 1, & \text{если } Y = \{X_{i_1}, \dots, X_{i_j}\} \text{ и } X_{i_1} \sqcap \dots \sqcap X_{i_j} = h; \\ 0, & \text{если } Y = \{X_{i_1}, \dots, X_{i_j}\} \text{ и } X_{i_1} \sqcap \dots \sqcap X_{i_j} \neq h. \end{cases}$$

каждая из $\binom{n}{j+1}$ вершин $(j+1)$ -го слоя соединена с $(j+1)$ -й вершиной j -го слоя, а каждая из $\binom{n}{j}$ вершин j -го слоя соединена с $n-j$ вершинами $(j+1)$ -го слоя. Выделим в графе B вершины, соответствующие семействам $\langle H \rangle_j$ и $\langle H \rangle_{j+1}$. Так как любое подмножество размера $j+1$ множества из $\langle H \rangle_j$ есть множество из $\langle H \rangle_{j+1}$, то число ребер в графе, соединяющих вершины, соответствующие множествам из $\langle H \rangle_j$, с вершинами, соответствующими множествам из $\langle H \rangle_{j+1}$, есть $e = g_j(n-j)$. В то же время, так как, вообще говоря, не каждое подмножество размера j множества из $\langle H \rangle_{j+1}$ есть множество из $\langle H \rangle_j$, число ребер e не больше $g_{j+1}(j+1)$. Следовательно, $g_j(n-j) < g_{j+1}(j+1)$. Так как $\frac{\binom{n}{j+1}}{\binom{n}{j}} = \frac{n-j}{j+1}$, то $\frac{g_j}{g_{j+1}} < \frac{j+1}{n-j} = \frac{\binom{n}{j}}{\binom{n}{j+1}}$ и $I_j = \frac{g_j}{\binom{n}{j}} < \frac{g_{j+1}}{\binom{n}{j+1}} = I_{j+1}$. \square

$$\frac{g_j}{g_{j+1}} < \frac{j+1}{n-j} = \frac{\binom{n}{j}}{\binom{n}{j+1}} \text{ и } I_j = \frac{g_j}{\binom{n}{j}} < \frac{g_{j+1}}{\binom{n}{j+1}} = I_{j+1}. \quad \square$$

4. ИЗМЕНЕНИЕ УСТОЙЧИВОСТИ ПРИ ПОПОЛНЕНИИ МНОЖЕСТВА ПРИМЕРОВ

Пусть множество исходных примеров S пополняются новым примером E .

Определение 4.1. Новый $(-)$ -пример E опровергает $(+)$ -гипотезу $\langle h, \{X_1, \dots, X_n\} \rangle$, если из множеств примеров S^+ , $S \cup \{E\}$ $(+)$ -гипотеза с головой h не выводится (т. е. $h \not\sqsubseteq E$).

Определение 4.2. Новый $(+)$ -пример E подтверждает $(+)$ -гипотезу $\langle h, \{X_1, \dots, X_n\} \rangle$, если $h \sqsubseteq E$.

Индексы устойчивости гипотезы с головой h после появления k новых примеров будем обозначать с верхним индексом k , например I_j^k . Положим также, для удобства записи,

$$I_n = 1,$$

$$I_j = 0 \text{ для } j \in Z \setminus \{2, \dots, n\}.$$

Теорема 4.1. Пусть множество примеров было пополнено k новыми $(+)$ -примерами, подтверждающими гипотезу $\langle h, \{X_1, \dots, X_n\} \rangle$. Тогда индексы устойчивости гипотезы с головой h будут лежать в следующих пределах

$$\frac{1}{\binom{n+k}{j}} \left[g_j \div \binom{k}{1} g_{j-1} + \dots + \binom{k}{k-1} g_{j-k+1} + g_{j-k} \right] < I_j^k < \frac{1}{\binom{n+k}{j}} \left[g_j + \binom{n}{j-1} + \dots + \binom{n+k-1}{j-1} \right]$$

при $(2 \leq j \leq n+k-1)$;

$$\frac{2^k \cdot g_{\Sigma} + 2^k - 1}{2^{n+k} - (n+k+2)} < I_{\Sigma}^k < \frac{g_{\Sigma} + 2^n(n^k - 1) - k}{2^{n+k} - (n+k+2)}$$

Доказательство. 1. Нижние границы для индексов I_j^k (или \bar{I}_j^k), $j \in \{2, \dots, n+k-1\}$. При заданном начальном значении индекса I_j значение I_j^k будет минимальным, если любое подмножество из множества k новых подтверждающих примеров дает h в пересечении лишь с теми множествами примеров из $\{X_1, \dots, X_n\}$,

которые сами дают h в пересечении. Более строго,

$$I_j^k = \frac{g_j^k}{\binom{n+k}{j}}$$

будет принимать минимальное значение, если для любых $p, q, 1 < p < k, 1 < q < n, \bar{X}_{i_1} \cap \dots \cap \bar{X}_{i_p} \cap (\bar{X}_{i_{p+1}} \cap \dots \cap \bar{X}_{i_q}) = h$ тогда и только тогда, когда $X_j \cap \dots \cap X_{j_q} = h$. Здесь $\bar{X}_{i_1}, \dots, \bar{X}_{i_p} \in \{\bar{X}_1, \dots, \bar{X}_k\}$, $\{X_{i_{p+1}}, \dots, X_{i_q}\}$ — множество новых примеров, $X_{i_1}, \dots, X_{i_q} \in \{X_1, \dots, X_n\}$.

Рассмотрим слагаемые, из которых складывается значение g_j^k . Первое слагаемое, т. е. g_j соответствует подмножествам размера j -го множества исходных примеров $\{X_1, \dots, X_n\}$. Остальные слагаемые имеют вид $g_{j-s} \binom{k}{s}$. Они получаются в силу того, что произвольное множество $\{X_{i_1}, \dots, X_{i_{j-s}}\}$ из $\langle H \rangle_{j-s}$ может быть пополнено любыми новыми подтверждающими примерами $\bar{X}_{r_1}, \dots, \bar{X}_{r_s}$ до множества $\{X_{i_1}, \dots, X_{i_{j-s}}, X_{r_1}, \dots, X_{r_s}\}$, также дающего h в пересечении всех своих элементов: $X_{i_1} \cap \dots \cap X_{i_{j-s}} \cap \bar{X}_{r_1} \cap \dots \cap \bar{X}_{r_s} = h$.

Значит, $g_j^k = \sum_{s=0}^k g_{j-s} \binom{k}{s}$ и нижняя оценка для I_j^k

доказана.

2. Нижние границы для индекса I_{Σ} (или \bar{I}_{Σ}). При поступлении одного нового подтверждающего примера каждое значение $g_i^1, 2 \leq i \leq n-1$, стало не меньше того, что указано в предыдущем пункте. Вся последовательность значений $g_2, g_2+g_3, \dots, g_{n-2}+g_{n-1}, g_{n-1}+1$, а их сумма $g_{\Sigma} = g_2 + (g_2+g_3) + \dots + (g_{n-2}+g_{n-1}) + (g_{n-1}+1) = 2g_{\Sigma} + 1$. Значит, после поступления k новых подтверждающих примеров значение g_{Σ} будет не меньше $2^k \cdot g_{\Sigma} + 2^{k-1} - 1$ (т. е. $g_{\Sigma}^k \leq 2^k \cdot g_{\Sigma} + 2^k - 1$) и

$$\frac{I_{\Sigma}^k}{2^{n+k} - (n+k+2)} > \frac{2^k \cdot g_{\Sigma} + 2^k - 1}{2^{n+k} - (n+k+2)}$$

3. Верхние границы для индексов I_j^k (или \bar{I}_j^k), $j \in \{2, \dots, n+k-1\}$ получаются при рассмотрении последовательности новых, подтверждающих гипотезу H примеров $\bar{X}^1, \dots, \bar{X}^k$ следующего вида: \bar{X}^i в пересечении с любым ранее имеющимся примером, подтверждающим H , дает само h , точнее $\bar{X}^i \cap X = h$ для $X \in \{X_1, \dots, X_n, \bar{X}^1, \dots, \bar{X}^{i-1}\}$. В этом случае для $k=1$ g_j^1 есть сумма числа старых примеров g_j и новых, образованных из $j-1$ старых примеров и одного нового: $g_j^1 = g_j + \binom{n}{j-1}$. Пусть для $k=t$ $g_j^t = g_j + \binom{n}{j-1} + \dots + \binom{n+t-1}{j-1}$. Тогда для $k=t+1$ g_j^{t+1} есть сумма числа старых примеров g_j^t и новых, образованных из $j-1$ старых примеров и одного нового:

$$g_j^{t+1} = g_j^t + \binom{n+t}{j-1} = g_j + \binom{n}{j-1} + \dots + \binom{n+t-1}{j-1} + \binom{n+t}{j-1}$$

4. Верхние границы для индекса I_{Σ} (или \bar{I}_{Σ}^k). Исходно значения g_i были $g_2, \dots, g_{s-1}, g_s, g_{s+1}, \dots, g_{n-2}$.

g_{n-1} , а после поступления X^1 стали, соответственно,

$$g_2 + \binom{n}{1}, \dots, g_s + \binom{n}{s-1}, \dots, g_{n-1} + \binom{n}{n-2}, \quad g_n + \binom{n}{n-1}, \text{ где } g_n = 1. \text{ При этом}$$

$$g_\Sigma^1 = (g_2 + \dots + g_{n-1}) + \left(\binom{n}{1} + \dots \right)$$

$$\dots + \binom{n}{s-1} + \dots + \binom{n}{n-1} + 1 = g_\Sigma + 2^n - 1.$$

Значит $g_\Sigma^k = g_\Sigma + 2^{n+k-1} + \dots + 2^n - k = g_\Sigma + 2^n(2^k - 1) - k$
и $\bar{I}_\Sigma^k = \frac{g_\Sigma + 2^n(2^k - 1) - k}{2^{n+k} - (n+k+2)}$. \square

Являются ли указанные в теореме оценки точными? Укажем последовательности, для которых эти оценки будут точными на достаточно большом числе новых примеров k .

Рассмотрим булев случай: $S = 2^U$, $U = \{d_1, \dots, d_n\}$. Пусть $\langle h, \{X_1, \dots, X_n\} \rangle$ — произвольная (+)-гипотеза. Пусть $X^1 = X_i \setminus \{h\}$, $U^1 = X_i^1 \cup \dots \cup X_n^1$, $U^2 = U \setminus (U^1 \cup n)$, $U^3 = \{d_1, \dots, d_1 k\}$.

Последовательность новых примеров, реализующая верхнюю границу изменения I_Σ , строится следующим образом. p -й новый пример будет иметь вид $X_p = h \cup \{d_i\}$. Очевидно, что любая пара новых примеров и любой новый пример с любым старым в пересечении образует h . Максимальное количество таких новых примеров есть $k = |U^2| = |U| - |U^1| - |h|$.

Последовательность новых примеров, реализующую нижние границы, построим сходным образом. p -й новый пример будет иметь вид: $X_p = h \cup U^1 \cup \{d_i\}$. Как нетрудно убедиться, новые примеры дают h в пересечении лишь с теми старыми примерами, которые сами дают в пересечении h , и таким образом реализуются нижние границы индексов устойчивости, указанные в теореме 4.1. При этом максимальное число новых примеров k , реализующих в точности нижнюю границу есть также $k = |U^2| = |U| - |U^1| - |h|$.

Рассмотрим теперь пределы нижних и верхних границ индексов устойчивости при $k \rightarrow \infty$.

Нижние границы поуровневых индексов ведут себя по-разному: для индексов верхних уровней они стремятся в пределе к 1, а для нижних уровней — к 0. В самом деле, в силу теоремы 4.1, имеет место $\bar{I}_{n+k-1} =$

$$\frac{1}{\binom{n+k}{n+k-1}} \left(g_{n-1} + \binom{k}{k-1} \right) \text{ и } \lim_{k \rightarrow \infty} \bar{I}_{n+k-1} = 1. \text{ С дру-}$$

$$\text{гой стороны, } \underline{I}_2 = \frac{1}{\binom{n+k}{2}} \cdot g_2 \text{ и } \lim_{k \rightarrow \infty} \underline{I}_2 = 0.$$

Открытым остается вопрос поведения нижних границ индексов средних уровней и усредненного индекса I_m^k .

Предел нижней границы интегрального индекса устойчивости строго больше нуля и меньше единицы.

В самом деле, функция $f(x) = 2^n - \frac{n+x+2}{2^x}$ монотонно возрастает при $x > 0$, так как

$$f'(x) = -\frac{-2^x + (n+x+2) \cdot 2^x \cdot \ln 2}{(2^x)^2} > 0$$

при $x > 0$ и, значит, функция $\bar{I}_\Sigma^k(k)$ монотонно убывает при увеличении k и стремится к $\lim_{k \rightarrow \infty} \bar{I}_\Sigma^k(k) =$

$$= \frac{g_\Sigma + 1}{2^n} > 0.$$

Верхние границы индексов устойчивости ведут себя единообразно: они монотонно возрастают и в пределе стремятся к 1. В самом деле,

$$\bar{I}_\Sigma^k = \frac{1}{\binom{n+k}{2}} \left[g_2 + \binom{n}{1} + \dots + \binom{n+k-1}{1} \right] =$$

$$= \frac{g_2}{\binom{n+k}{2}} + \frac{\binom{n+k}{2} - \binom{n}{2}}{\binom{n+k}{2}} \rightarrow 1$$

и, в силу леммы 3.1, то же самое имеет место для $\bar{I}_\Sigma^k, \dots, \bar{I}_{n+k-1}^k, \bar{I}_m^k$.

$$\bar{I}_\Sigma^k(k) = \frac{2^{n+k} - k - (2^n - g_\Sigma)}{2^{n+k} - (n+k+2)} =$$

$$= \frac{f(k) - d}{f(k) - b} = \left(1 - \frac{d-b}{f(k) - b} \right),$$

где $f(k) = 2^{n+k} - k$ — строго монотонно возрастающая функция для $k > 0$, а $d = (2^n - g_\Sigma) > (n+2) = b$, так как $g_\Sigma \leq 2^n - n - 2$. $f(k) - b = 2^{n+k} - k - n + 2 > 0$ для любых $k, n > 0$, $d - b = \text{const}$. Значит, $\bar{I}_\Sigma^k(k)$ строго монотонно возрастает и $\lim_{k \rightarrow \infty} \bar{I}_\Sigma^k(k) = 1$.

Заметим, что монотонное изменение нижних и верхних границ $I_\Sigma^k(k)$ позволяет усилить формулировку теоремы 4.1 в той ее части, которая относится к этому индексу: указанные границы справедливы не только для подтверждающих, но и для любых непроверяющих примеров. В самом деле, новый (—)-пример не может изменить устойчивость гипотезы — он может лишь опровергнуть саму гипотезу. (+)-примеры, не подтверждающие (+)-гипотезу $\langle h, \{X_1, \dots, X_n\} \rangle$, не могут, в силу указанной монотонности, сделать интегральный индекс устойчивости меньше нижней границы и больше верхней границы.

Анализ асимптотики позволяет делать предположения о поведении лишь индекса I_Σ : скорее всего, при поступлении новых примеров, значение этого индекса будет возрастать, так как уменьшаться ему практически некуда (рис. 1).

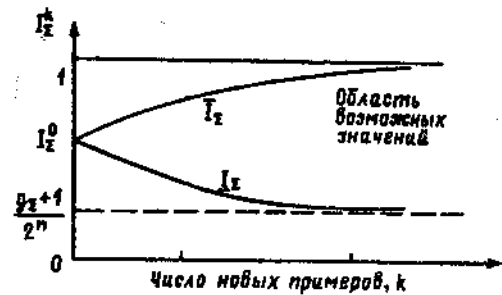


Рис. 1

Например, для гипотезы $\langle h, \{X_1, \dots, X_n\} \rangle$, $n=4$, соотношение исходного значения I_Σ^0 к предельному $\lim_{k \rightarrow \infty} \bar{I}_\Sigma^k$ для среднего ($\sim 1/2$) значения I_Σ^0 будет

$$\frac{g_\Sigma}{2^4 - 4 - 2} \cdot \frac{2^4}{g_\Sigma + 1} \approx 1.3.$$

При $n=8$

$$\frac{g_\Sigma}{2^8 - 8 - 2} \cdot \frac{2^8}{g_\Sigma + 1} \approx 1.04.$$

т. е. отличие очень мало по сравнению с возможным двукратным увеличением $\lim_{k \rightarrow \infty} I_2^k$ по отношению к I_2^0 .

Теперь, если мы представим множество примеров какой-либо гипотезы $\langle h, \{X_1, \dots, X_n\} \rangle$ как возникшее в результате пополнения некоторого исходного множества размера $r \ll n$, мы можем сделать вывод о том, что у гипотез с большим значением n скорее всего I_2 будет больше чем у гипотез с малыми n . Такая «мягкая» зависимость I_2 от числа примеров позволяет отдать предпочтение интегральному индексу устойчивости как наиболее емкому из индексов с известной асимптотикой нижних границ: с одной стороны, явно, I_2 отражает устойчивость гипотезы, а с другой стороны, — неявно — количество подтверждающих примеров.

5. АЛГОРИТМИЧЕСКАЯ СЛОЖНОСТЬ ЗАДАЧИ ВЫЧИСЛЕНИЯ ИНДЕКСОВ УСТОЙЧИВОСТИ

К сожалению, задача точного вычисления индексов устойчивости, по всей видимости (если $P \neq \#P$), даже в случае представления примеров множествами, не может быть решена за время, полиномиальное от размера гипотезы $\langle h, \{X_1, \dots, X_n\} \rangle$, т. е. от $|h| + n$. Это позволяет установить теорема 5.1. Напомним, что $\#P$ [20] есть класс задач, вычисляемых считающей недетерминированной машиной Тьюринга за полиномиальное время, а $\#P$ -полной называется задача из $\#P$, к которой сводима (по Тьюрингу) любая задача из $\#P$. Среди $\#P$ -полных задач есть не только перечислительные задачи, соответствующие NP -полным задачам распознавания (задачи о числе гамильтоновых путей в графе, о числе наборов, выполняющих КНФ и т. д.), но и задачи перечисления, для которых соответствующие задачи распознавания полиномиально разрешимы (например, задача о числе максимальных паросочетаний).

Теорема 5.1. Пусть в качестве представления данных используется булева алгебра (см. п. 1). Тогда задача IU_2 вычисления индекса устойчивости I_2 произвольной гипотезы $\langle h, \{X_1, \dots, X_n\} \rangle$ является $\#P$ -полной.

Для доказательства потребуются введение нескольких вспомогательных задач.

Задача ЧВП (число вершинных покрытий).

Дано: Граф $G = \langle V, E \rangle$.

Найти: Число вершинных покрытий, т. е.

$\# \{V' \subseteq V \mid \text{если } (u, v) \in E, \text{ то } u \in V' \text{ или } v \in V'\}$.

Задачи ЧИ (число импликантов).

Дано: Монотонная 2-КНФ, т. е. формула $F = C_1 \wedge \dots \wedge C_r$, где

$$C_i = (x_i \vee x_{i_2}) \wedge x_{i_1} \wedge x_{i_3} \in X = \{x_1, \dots, x_n\}.$$

Найти: $\#P \{Y \mid Y \subseteq X, \bigwedge_{x_j \in Y} x_j \rightarrow F\}$.

Задача ЧФП (число подсемейств с фиксированным пересечением).

Дано: Конечное множество \mathcal{U} и $\mathcal{Z} \subseteq 2^{\mathcal{U}}$ — семейство различных множеств $\mathcal{Z} = \{X_1, \dots, X_k\}$, причем выполняется $X_1 \cap \dots \cap X_k = h$.

Найти: Число подсемейств \mathcal{Z}' семейства \mathcal{Z} таких, что пересечение всех членов подсемейства \mathcal{Z}' есть h , т. е.

$$\# \{ \mathcal{Z}' \subseteq \mathcal{Z} \mid \mathcal{Z}' = \{X_{i_1}, \dots, X_{i_s}\} \text{ и } X_{i_1} \cap \dots \cap X_{i_s} = h \}.$$

Теперь докажем леммы, позволяющие доказать теорему 5.1.

Лемма 5.2. Задача ЧИ $\#P$ -полна

Доказательство. Сведем к задаче ЧИ задачу «число булевых наборов, удовлетворяющих 2-КНФ $F = C_1 \wedge C_2 \wedge \dots \wedge C_r$ », где $C_i = (y_i \vee y_{i_2}) \wedge y_{i_1} \in X$.

$\#P$ -полнота этой задачи доказана в [20]. Пусть $|X| = n$ и $A = (a_1, \dots, a_n)$ — булев набор, удовлетворяющий F . Пусть $\{j_1, \dots, j_k\}$ — номера единичных компонент булева набора A . Образует конъюнкцию $Y_j = y_{j_1} \wedge \dots \wedge y_{j_k}$, где $y_{j_i} \in X$. Очевидно, что Y_j есть импликант 2-КНФ F . И наоборот, каждому импликанту $Y_m = y_{m_1} \wedge \dots \wedge y_{m_s}$ формулы F соответствует булев набор $A^m = (a_1^m, \dots, a_n^m)$, у которого позиции m_1, \dots, m_s единичные, а остальные — нулевые, причем булев набор A^m выполняет F . Сводимость доказана. Поскольку запись ЧИ не может превышать размера $\log 2^{|X|} = |X|$ и сводимость полиномиальна, лемма 5.2 доказана.

Лемма 5.3. Задача ЧВП $\#P$ -полна.

Доказательство. Принадлежность задачи ЧВП к классу $\#P$ не вызывает сомнений, так как размер записи решения задачи не превосходит $\log 2^{|V|} = |V|$ бит. $\#P$ -полноту задачи докажем сведением к ней $\#P$ -полной задачи ЧИ. Так же, как и в [21], по произвольной 2-КНФ F строим граф $G = \langle V, E \rangle$ с $V = \{u_1, \dots, u_n\}$ (где каждое u_i соответствует переменной x_i из F), и $E = \{(u_i, u_j) \mid (x_i \vee x_j) \text{ входит в конъюнкцию } F\}$. При этом любое вершинное покрытие G будет соответствовать импликанту F и наоборот: любой импликант F соответствует некоторому вершинному покрытию G . Сводимость осуществляется за линейное от размера F время. \square

Лемма 5.4. Задача ЧФП $\#P$ -полна.

Доказательство. Сведем по Тьюрингу к частному случаю этой задачи (при $h = \emptyset$) задачу ЧВП. Пусть нам дан произвольный граф $G = \langle V, E \rangle$ без изолированных вершин (что не нарушает общности), где $V = \{a_1, \dots, a_n\}$, $E \subseteq V \times V$. Обозначим через $N(V)$ множество ребер из E , инцидентных вершине v ; $\bar{N}_E(V) = E \setminus N(V)$. Для некоторых $v_i, v_j \in V$, имеет место $N(v_i) = N(v_j) = \emptyset$ тогда и только тогда, когда $N(v_i) = N(v_j) = \{(v_i, v_j)\}$, т. е. v_i и v_j инцидентны одному и тому же ребру, не связанному с остальными вершинами графа G . Будем вычислять ЧВП графа следующим образом: выделим в графе G изолированные ребра (т. е. ребра вида $(v_i, v_j) \in E$, для которых $N(v_i) = N(v_j)$). Предположим, что в графе G k таких ребер. Вершинное покрытие этих ребер может быть осуществлено 3^k способами (3 способа на каждое ребро: оно может быть покрыто какой-либо из вершин, либо обеими). Если остальные ребра графа G могут быть покрыты d различными способами, то все ребра графа имеют $d \cdot 3^k$ покрытий. Итак, осталось определить число вершинных покрытий неизоллированных ребер графа (обозначим это множество через E'). Пусть ребрам из E инцидентны вершины из множества $V' \subseteq V$; тогда для любых двух различных вершин $v_i, v_j \in V'$ имеет место $N(v_i) \neq N(v_j)$. По определению, вершины $v_1, \dots, v_r \in V'$ образуют покрытие E' тогда и только тогда, когда $N(v_1) \cup \dots \cup N(v_r) = E'$ или, по закону де-Моргана, $\bar{N}_{E'}(v_1) \cap \dots \cap \bar{N}_{E'}(v_r) = \emptyset$. Таким образом, множество $\{v_1, \dots, v_r\}$ образует вершинное покрытие графа $G = \langle V', E' \rangle$ тогда и только тогда, когда пересечение всех множеств семейства $\bar{N}_{E'}(v_1), \dots, \bar{N}_{E'}(v_r)$ пусто. Значит мы свели задачу ЧВП для $G = \langle V, E \rangle$ к задаче ЧФП с $\mathcal{Z} = \{\bar{N}_{E'}(a_1), \dots, \bar{N}_{E'}(a_n)\}$. Сводимость полиномиальна, так как размер множества \mathcal{Z} не превышает

$n \binom{n}{2} = n(n-1)$, т. е. $O(n^2)$. \square

Теорема 5.1 является простым следствием леммы 5.4. В условиях, когда гипотеза H не противоречива и множество всех примеров, ее породивших, есть \mathcal{Z} ,

ЧФП есть g_x .

Следствие. Задача определения I_j $\#P$ -полна.

В самом деле, если известны значения I_j для $2 \leq j \leq n-1$, то значение I_n вычисляется как

$$I_n = \frac{\sum_{j=1}^{n-1} I_j \binom{n}{j}}{2^n - n - 2}.$$

Дадим описание и сложностной анализ алгоритма, вычисляющего индексы устойчивости гипотезы $\langle h, \{X_1, \dots, X_n\}$ для булева случая.

Будем считать, что все примеры и гипотезы представлены булевыми векторами размера $|U|=m$ и две битовые строки размера b перемножаются или сравниваются на совпадение за b машинных операций.

1-й уровень.

1. Порождаются подсемейства семейства $\{X_1, \dots, X_n\}$ размера $n-1$, т. е. семейство $\mathcal{F}_1: \{X_1, \dots, X_{n-1}\}, \{X_1, \dots, X_{n-2}, X_n\}, \dots, \{X_2, \dots, X_n\}$. При этом каждое из подсемейств может быть представлено множеством чисел от 1 до n . На этот шаг тратится $O(\log n \cdot (n-1) \cdot n)$ машинных операций и $O(\log n \cdot (n-1) \cdot n)$ памяти.

2. Для каждого множества из семейства \mathcal{F}_1 вычисляется пересечение всех его членов и проверяется на совпадение с h . Подсчитывается число пересечений, совпадающих с h . На этот шаг тратится $O(n(n-1)m)$ машинных операций и $O(\log n + m)$ памяти.

1-й уровень.

Входом i -го уровня является семейство $\mathcal{F}_{n-i+1}: \{Y_1, \dots, Y_{t_{i-1}}\}$ подсемейств размера $n-i+1$ семейства $\{X_1, \dots, X_n\}$, причем члены указанных подсемейств в пересечении дают h (и, следовательно, $t_{i-1} = I_{n-i+1} \binom{n}{i-1}$, т. е. $t_j = g_{n-j}$).

1. Для каждого Y_s из \mathcal{F}_{n-i+1} порождаем $n-i+1$ подсемейство $Y_s^1, \dots, Y_s^{n-i+1}$ размера $n-i$. Порождаемые семейства представляем лексикографически упорядоченными наборами чисел: каждому X_i из $\{X_1, \dots, X_n\}$ ставится в соответствие число i , а подсемейству Y_s^r — набор чисел, соответствующих всем множествам из Y_s^r .

На порождение множеств Y_s используется $O(t_{i-1} \cdot \log n \cdot (n-i+1) \cdot n)$ машинных операций. Этим множеств будет не более $t_{i-1}(n-i+1)$, поэтому на их размещение в лексикографическом порядке потребуется не более $O(t_{i-1} \log n \cdot (n-i+1) \cdot \log(t_{i-1}(n-i+1)))$ операций. Объем памяти, необходимый для этого, не превосходит $O(t_{i-1}(n-i+1) \cdot \log n)$.

2. Пересекаем все множества в каждом Y_s^r и сравниваем полученные результаты с h . На это тратится не более $O(t_{i-1}(n-i+1) \cdot m \cdot (n-i))$ машинных операций, где $t_{i-1}(n-i+1)$ — максимальное число семейств Y_s , а $(n-i)$ — максимальный размер Y_s .

3. Одновременно с пересечением и сравнением подсчитывается число совпадающих с h пересечений. На это требуется не более $O(t_{i-1}(n-i+1) \cdot m)$ операций и $O(\log(t_{i-1}(n-i+1)m))$ памяти.

Суммарная сложность всех шагов вычисления I_j с учетом того, что $\log t_{i-1} \leq \log 2^n = n$, будет не более

$O\left(\left(\sum_{i=1}^{n-1} t_i\right) \cdot m \cdot n^2\right)$ машинных операций при использовании не более $O\left(\left(\sum_{i=1}^{n-1} t_i\right) \cdot m \cdot n \cdot \log^2 n\right)$ памяти. На нахождение I_n . I_m тратится не более $O\left(\left(\sum_{i=1}^{n-1} t_i\right) \cdot m \cdot n^3\right)$ машинных операций при использовании не более

$O\left(\left(\sum_{i=1}^{n-1} t_i\right) \cdot m \cdot n^3\right)$ машинных операций при использовании не более

$O\left(\left(\sum_{i=1}^{n-1} t_i\right) \cdot m \cdot n \cdot \log^2 n\right)$ памяти. Таким образом, верна Теорема 5.5.

1. Подсчет индекса устойчивости $I_j = \frac{g_j}{\binom{n}{j}}$ $2 < j < n-1$ возможен за время, линейное от $\max_{2 < j < n-1} g_j$.

2. Подсчет интегрального индекса устойчивости $I_n = \frac{\sum_{j=2}^{n-1} g_j}{2^n - n - 2} = \frac{g_n}{2^n - n - 2}$ возможен за время, линейное от g_n .

Этот результат в совокупности с результатом о $\#P$ -полноте задач вычисления индексов устойчивости говорит о том, что предложенный алгоритм оптимален с точностью до полиномиального множителя $O(n^3)$ (в смысле определения оптимальности алгоритмов для $\#P$ -полных задач из [20]).

Можно предложить эффективные алгоритмы вычисления нижних и верхних оценок интегрального I_n и усредненного I_m индексов устойчивости для булевого случая $(S = \langle 2^U, \Pi, \emptyset \rangle)$ на основе следующих утверждений.

Теорема 5.5. Для произвольной (+)-гипотезы $H = \langle h, \{X_1, \dots, X_n\} \rangle$ имеет место

$$\frac{g_2 + \dots + g_k}{\binom{n}{2} + \dots + \binom{n}{k}} < I_n < \frac{g_{n-r} + \dots + g_{n-1}}{\binom{n}{n-r} + \dots + \binom{n}{n-1}}.$$

Теорема 5.6. Для произвольной (+)-гипотезы $H = \langle h, \{X_1, \dots, X_n\} \rangle$ имеет место

$$\frac{1}{k} \left(\frac{g_2}{\binom{n}{2}} + \dots + \frac{g_k}{\binom{n}{k}} \right) < I_m < \frac{1}{r} \left(\frac{g_{n-r}}{\binom{n}{n-r}} + \dots + \frac{g_{n-1}}{\binom{n}{n-1}} \right).$$

Докажем сначала следующую лемму.

Лемма 5.7. Для произвольных последовательностей (a_i) , (b_i) таких, что $a_i > 0$, $b_i > 0$ и $\frac{a_i}{b_i} < \frac{a_{i+1}}{b_{i+1}}$ и для произвольного $s > 1$ имеет место

$$\frac{a_1 + \dots + a_s}{b_1 + \dots + b_s} < \frac{a_1 + \dots + a_{s+1}}{b_1 + \dots + b_{s+1}} < \frac{a_2 + \dots + a_{s+1}}{b_2 + \dots + b_{s+1}} < \frac{a_{s+1}}{b_{s+1}}.$$

Проведем индукцией по s доказательство первого неравенства $s=1$.

Пусть $\frac{a_1}{b_1} < \frac{a_2}{b_2}$, тогда $\frac{a_1}{b_1} < \frac{a_1 + a_2}{b_1 + b_2} < \frac{a_2}{b_2}$. В самом деле,

$$\frac{a_1 + a_2}{b_1 + b_2} - \frac{a_1}{b_1} = \frac{a_1 b_1 + a_2 b_1 - a_1 b_1 - a_1 b_2}{b_1 (b_1 + b_2)} > 0,$$

$$\frac{a_1 + a_2}{b_1 + b_2} - \frac{a_2}{b_2} = \frac{a_1 b_2 + a_2 b_2 - a_2 b_1 - a_2 b_2}{b_2 (b_1 + b_2)} < 0.$$

Пусть для $s < m$ утверждение доказано, и $a_1 + \dots + a_m = A$, $b_1 + \dots + b_m = B$, тогда

$$\frac{a_1 + \dots + a_{m+1}}{b_1 + \dots + b_{m+1}} - \frac{a_1 + \dots + a_m}{b_1 + \dots + b_m} = \frac{A + a_{m+1}}{B + b_{m+1}} - \frac{A}{B} =$$

$$= \frac{AB + a_{m+1}B - BA - b_{m+1}A}{B(B + b_{m+1})} = \frac{a_{m+1}B - b_{m+1}A}{B(B + b_{m+1})} = \Delta.$$

Но $\frac{a_{m+1}}{b_{m+1}} > \frac{a_m}{b_m}$, а в силу индуктивного предположения, $\frac{a_m}{b_m} > \frac{A}{B}$ и числитель Δ не отрицателен. \square

Другие неравенства доказываются аналогично. Доказательство теоремы 5.5 следует непосредственно из леммы 3.1 и 5.7. Аналогично доказываются теорема 5.6. Теоремы 5.5 и 5.6 позволяют вычислять верхние и нижние границы I_x и I_m за время $O(|U| \cdot n^{k+r})$.

Практика показывает, что даже при небольших количествах подтверждающих гипотезу примеров, точное вычисление индексов становится затруднительным. В то же время вычисления в соответствии с теоремами 5.5 и 5.6 не всегда дают хорошие приближения: при тех значениях параметров k и r из этих теорем, которые позволяют использовать машинные ресурсы (см. раздел 6), верхние и нижние оценки часто далеки друг от друга. По-видимому, эффективное вычисление хороших приближений индексов устойчивости будет возможным с использованием методов Монте-Карло.

6. МАШИННЫЙ ЭКСПЕРИМЕНТ

Экспериментальное изучение индексов устойчивости проводилось при поиске средствами ДСМ-системы потенциальных причин дефектов полиамидов, производимых НПО «Пластмассы».

Описание сырьевых и технологических параметров производства полимеров состояло из восьми компонент: характеристик массовой доли летучих веществ, кислотности, цветности сырья; номера автоклава, в которых производился полимер, времени суток и номера смены, во время которых производился полимер; времени от окончания дегазации до выгрузки, времени простоя.

Дефекты конечного продукта были связаны с отклонениями от нормы четырех из следующих характеристик полимера: соответствие ТУ, соответствие ГОСТу, отсутствие непрорубов (некоторая характеристика вязкости продукта), отсутствие инородных включений.

Сходство описаний технологических условий синтеза определялось покомпонентно следующим образом. Компоненты описания, соответствующие номеру автоклава, времени суток, номеру смены, принимали неупорядоченные значения. Сходство значений этих компонент описания определялось так:

$$x \wedge y = \begin{cases} x, & \text{если } x=y; \\ 0, & \text{если } x \neq y. \end{cases}$$

Для компонент описания, которые принимали числовые значения, задание сходства предполагало предварительное разбиение экспертами всего множества возможных значений на «интервалы сходства», внутри которых значения считаются сходными, а вне — различными. Более точно, пусть L — множество всех интервалов сходства для некоторой компоненты j , принимающей числовые значения.

Тогда

$$x \wedge y = \begin{cases} 1, & x \in i, y \in i, i \in L; \\ 0, & x \in i, y \in k, i \in L, k \in L, i \neq k. \end{cases}$$

Подобным образом задавалось сходство значений компонент, соответствующих массовой доле летучих веществ, кислотности, цветности, времени простоя, времени от окончания дегазации до выгрузки.

В ходе эксперимента каждое из четырех свойств полимера исследовалось отдельно. Каждая из 253 ситуа-

ций была отнесена для каждого свойства к классу положительных или отрицательных примеров.

Гипотез, полученных с помощью правила «запретом на контрпример», было довольно много (от 500 до 1500 по разным свойствам полимера).

Индексов устойчивости I_m , I_x вычислялись для гипотез с помощью программы, созданной В. Г. Ивашко для IBM PC AT на языке С. Имевшиеся ресурсы памяти (около 270 КБ свободной оперативной памяти) позволяли программе вычислять точные значения для гипотез с $n < 16$. Максимальное время счета (для $n=15$) составляло 10 с (процессор Intel 80386, 20 мГц). Для гипотез с большими n за то же время вычислялись верхние оценки в соответствии с теоремами 5.5, 5.6. Эти оценки, однако, мало отличались от 1. Таким образом, адекватное представление об индексах устойчивости можно было получить лишь для $n < 16$, что для указанной задачи было вполне достаточным — очень мало гипотез имело число подтверждающих примеров больше 15.

На данном экспериментальном материале подтвердилось предположение (раздел 4) о росте в среднем значений I_x с ростом n : по всем четырем указанным свойствам среднее по всем гипотезам значение I_x либо монотонно возрастало, либо имело незначительные участки убывания. Значение I_m было несколько больше значений I_x при малых n и становилось меньше при больших (для указанного интервала (3,15)) значениях n . На рис. 2 изображен типичный график зависимости

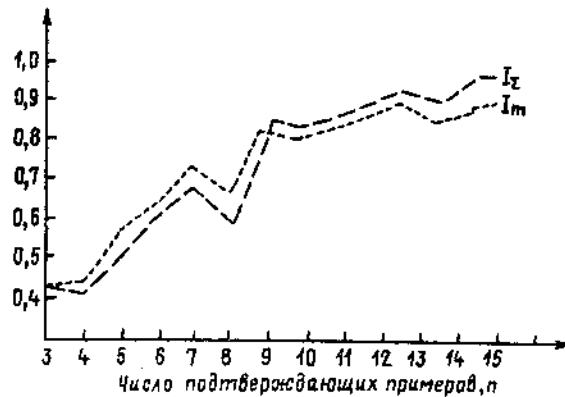


Рис. 2

средних значений I_m , I_x от n (для гипотез относительно причин отсутствия непрорубов). У всех восьми гипотез (относительно различных свойств), выделенных экспертами как наиболее осмысленные, значения интегрального и усредненного индексов устойчивости были выше средних, что позволяет говорить о хорошей избирательной способности этих индексов для данной задачи.

7. ДРУГИЕ ВОЗМОЖНЫЕ ОПРЕДЕЛЕНИЯ УСТОЙЧИВОСТИ

Идея устойчивости может быть воплощена в различных индексах устойчивости. При различных обстоятельствах, выяснение которых представляет самостоятельный интерес, некоторые из них могут оказаться предпочтительней других. Приведем некоторые возможные другие определения индексов устойчивости гипотезы $\langle H, \{X_1, \dots, X_n\} \rangle$:

1) «Средний слой»

$$I_M = \begin{cases} \frac{I_n}{2}, & \text{если } n \text{ четно;} \\ \frac{I\left[\frac{n}{2}\right] + I\left[\frac{n}{2}+1\right]}{2}, & \text{если } n \text{ нечетно;} \end{cases}$$

2) «Минимальное покрытие»

$$I_c = I_j, \text{ где } j = \min_{2 < i < n-1} (I_i \neq 0);$$

3) «Максимальное антипокрытие»

$$I_s = I_j, \text{ где } j = \max_{2 < i < n-1} (I_i \neq 1);$$

4) Устойчивость прогноза в духе работы [17].

Пусть S^+ , S^- — множества исходных примеров, $S^+ \cap S^-$ — множество вопросов (т. е. объектов из S^+ , для которых нужно провести прогноз), P^+ , P^- — множества всех (+) и (-) прогнозов, полученных на основе всех гипотез, порожденных по S^+ , S^- . Тогда I_s есть доля всех подмножеств множества $S^+ \cup S^-$, для которых множества всех порожденных прогнозов совпадают с прогнозами, полученными по всему множеству примеров $S^+ \cup S^-$.

* * *

Автор считает своим приятным долгом выразить признательность В. К. Финну, М. В. Арапову, О. Г. Горбачеву за плодотворные обсуждения, А. Ю. Когану, указавшему на связь вводимых конструкций с идеями непараметрической статистики и приведшему короткое доказательство Леммы 3.1, Д. П. Скворцову за указание ряда неточностей в первоначальных вариантах текста и В. Г. Ивашку, создавшему программу точного вычисления I_s и I_m .

СПИСОК ЛИТЕРАТУРЫ

1. Классификация и кластер / Под ред. Д. Райзин.— М.: Мир, 1980.
2. Шрейдер Ю. А. Равенство, сходство, порядок.— М.: Наука, 1971.
3. Kodratoff Y., Tecuci G. Learning Based on Conceptual Distance // IEEE Transactions on Pattern Analysis and Machine Intelligence.— 1988.— Vol. 10, N 6.— P. 897—909.
4. Fu King-Sun. A Graph Distance Measure // IEEE Transactions on Pattern Analysis and Machine Intelligence.— 1984.— Vol. SMC, № 14.— P. 398—408.
5. Vere S. A. Inductive learning of relational productions / Waterman D. A., Hayes-Roth F. (Eds.) Pattern-directed inference systems.— N. Y.: Acad. Press, 1978.
6. Финн В. К. Правдоподобные выводы и правдоподобные рассуждения // Итоги науки и техники.

Сер. Теория вероятностей. Математическая статистика. Теоретическая кибернетика. Т. 28.— М.: ВИНТИ, 1988.— С. 3—84.

7. Гусакова С. М., Финн В. К. О формализации локального и глобального сходств // НТИ. Сер. 2.— 1986.— N 6.— С. 16—19.
8. Гусакова С. М., Финн В. К. Сходство и правдоподобный вывод // Изв. АН СССР. Сер. Техническая кибернетика.— 1987.— N 5.— С. 42—63.
9. Akinniyi F. A., Wong A. K. C. A New Product Graph Based Algorithm for Subgraph Isomorphism // Proc. of Pattern Recognition and Computer Vision.— 1983.— June.— P. 457—467.
10. Загоруйко Н. Г., Скоробогатов В. А., Хворостов П. В. Вопросы анализа и распознавания молекулярных структур на основе общих фрагментов // Вычислительные системы.— 1984.— Вып. 103.— С. 26—50.
11. Денищик Е. Ю. Нахождение максимальных общих подграфов в семействе графов // Вычислительные системы.— 1987.— Вып. 103.— С. 85—89.
12. Кузнецов С. О. Определение сходства на гиперграфах как основа правдоподобного вывода на структурированных данных // Тез. докл. Всесоюз. конф. по искусственному интеллекту, Переяславль-Залесский, 1988 г.— М.: ВИНТИ, 1988.— Т. 1.— С. 442—448.
13. Кузнецов С. О., Финн В. К. Распространение процедур экспертных систем типа ДСМ на графы // Изв. АН СССР. Сер. Техническая Кибернетика.— 1988.— N 5.— С. 4—11.
14. Кузнецов С. О. ДСМ-метод как система автоматического обучения // Итоги науки и техники. Сер. Информатика. Т. 15. Интеллектуальные информационные системы.— М.: ВИНТИ, 1991.— В печати.
15. Кононов Б. П. Сходство и отличие: операционное моделирование, вопросы взаимосвязи и единственности // НТИ Сер. 2.— 1983.— N 3.— С. 27—31.
16. Эфрон Б. Нетрадиционные методы многомерного статистического анализа.— М.: Финансы и статистика, 1988.— 263 с.
17. Гусев В. А. Использование подвыборок и понятия устойчивости в задаче определения общего вида искомой зависимости // Заводская лаборатория.— 1987.— Т. 53, N 1.
18. Dantzin E. Algorithms for probabilistic inference // Lecture Notes in Computer Science.— 1990.— Vol. 417.
19. Carnap R. The Logical Foundations of Probability— Chicago: University of Chicago Press, 1962.
20. Valiant L. G. The Complexity of Computing the Permanent // Theoretical Computer Science.— 1979.— N 8.— P. 189—201.
21. Valiant L. G. The Complexity of Enumeration and Reliability Problems // SIAM J. Comput.— 1979.— Vol. 8, N. 1.— P. 410—421.

Материал поступил в редакцию 27.11.90.