ЭМПИРИЧЕСКОЕ ПРЕДСКАЗАНИЕ И РАСПОЗНАВАНИЕ ОБРАЗОВ
(Вычислительные системы)
1976 год Выпуск 67

УДК 518.74

МЕТОД ОБНАРУЖЕНИЯ ЗАКОНОМЕРНОСТЕЙ И МЕТОД ПРЕДСКАЗАНИЯ

Е.Е.Витяев

I. Ф о р м у л и р о в к а з а д а ч и. Многие реальные практические задачи содержат признаки, измеренные в разнотипных шкалах. При решении таких задач все признаки обычно сводят к одной шкале. В случае, когда все признаки сводятся к наиболее бедной шкале, например к шкале наименований, часть информации теряется. В случае, когда все признаки сводятся к наиболее богатой шкале, например шкале отношений, вводится бессмысленная информация, которая может существенно исказить имеющиеся данные. Представляет интерес построение метода обнаружения закономерностей и метода предсказания, которые наиболее полно и без искажений учитывают эмпирическую информацию, содержащуюся в признаках (т.е. методов, работающих с разнотипными признаками), и позволяют, используя эту эмпирическую информацию, предсказывать заданную величину (признак), измеряемую в произвольной шкале, также с наиболее полным учетом той информации, которая в этой шкале содержится. В случае, когда предсказываемый признак измерен в шкале наименований, мы приходим к задаче распознавания образов, где образами являются имена признака. Существуют алгоритмы, предсказывающие величину, измеряемую в шкале отношений, на основании признаков, также измеренных в шкале отношений [8,9].

Относительно признаков может быть известна некоторая априорная информация помимо той, которая определяется типом шкал. Например, информация, которая определяется физическими зависимостями, существующими между признаками. Желательно поэтому, чтобы метод предсказания имел возможность учитывать также и такого рода априорную информацию.

Вообще говоря, метод предсказания может быть построен без предварительного нахождения множества закономерностей. Однако предварительное нахождение закономерностей имеет преимущество в том, что хорошо видно, на основании чего получено предсказание. Кроме того, найденные закономерности могут представлять и самостоятельный интерес, например, для специалиста соответствующей области знания.

Относительно множества закономерностей сформулируем следующие пожелания:

– класс рассматриваемых закономерностей должен быть достаточно широк;

– закономерности должны быть устойчивы относительно случайных отклонений значений признаков;

– закономерности должны быть, в некотором смысле, независимы между собой.

Метод обнаружения закономерностей и метод предсказания должны работать с данными, имеющими пробелы.

Цель этой работы заключается в построении метода обнаружения закономерностей и метода предсказания, учитывающих приведенные выше пожелания.

Посмотрим, на основе чего можно было бы удовлетворить всем этим пожеланиям. Возможность общего представления любой шкалы дает теория измерений [7]. Признаки $x_0, x_1, \ldots, x_{n'}$, измеренные в каких-то шкалах, можно, согласно теории измерений, адекватно представить в виде набора многоместных отношений P^i_j, $i = 0,1,\ldots,n'$, $j = 1,2,\ldots,n_i$, и аксиом, записанных в терминах этих отношений. Упомянутую дополнительную априорную информацию также будем записывать в терминах этих отношений. Достаточно общий вид априорной информации будет достигаться за счет того, что будут допускаться любые универсальные формулы, записанные в терминах этих отношений. Шкалы всех признаков, а также дополнительная априорная информация будут представлены в п.2 эмпирической гипотезой [5] H_0.

Предсказание значения признака x_0 на некотором объекте по известным значениям признаков $x_1, \ldots, x_n$, на этом же объекте будет получаться благодаря использованию априорной

информации и предварительно установленного множества закономерностей. Закономерностями являются зависимости между отношениями $P^i_j, i=1,2,\dots,n'; j=1,2,\dots,n_i$, характеризующими признаки $x_1,\dots,x_{n'}$, и отношениями P^0_j, $j=1,2,\dots,n_0$, характеризующими признак x_{n_0}. В качестве закономерностей будут рассматриваться только формулы вида (3). Некоторую гарантию того, что рассматриваемый класс закономерностей достаточно широк, даёт разложение универсальных формул, приведенное в п.2. Согласно этому разложению, любую детерминированную закономерность, представимую в виде универсальной формулы, можно выразить с помощью совокупности формул вида (3). Это же разложение позволяет априорную информацию, записанную в виде совокупности универсальных формул, привести к более простому виду – совокупности формул вида (3).

Устойчивость закономерностей относительно случайных отклонений достигается благодаря определённым вероятностным предположениям (см. пп. 2,3). Независимость закономерностей получается за счет введения понятия существенности отношений в закономерности (см. п.2) и за счет организации самой процедуры обнаружения закономерностей (см. п.3).

Возможность работы с данными, имеющими пробелы, появляется благодаря использованию, помимо значений "истинность" и "ложность", значения "не определено" *) для отношений P^i_j.

Данный метод обнаружения закономерностей и метод предсказания являются дальнейшей разработкой алгоритма эмпирического предсказания [6].

2. Предварительные определения. Опишем в точности, что является входной информацией метода. Набор шкал признаков $x_0, x_1,\dots,x_{n'}$, а также дополнительная априорная информация будут представлены эмпирической гипотезой $H_0 = \langle v_{H_0}, Int_{H_0}, T_{H_0} \rangle$, где $v_{H_0} = \{ P^i_j,\ i=0,1,\dots,n';\ j=1,\dots,n_i \}$, $Int_{H_0} = \{ P^i_j,\ i=0,\dots,n';\ j=1,\dots,n_i \}$, T_{H_0} – тестовой алгоритм [5]. Отношения, или будем говорить предикатные символы P^i_j, переобозначим для простоты следующим образом: $v_{H_0} = \{ P_1,\dots,P_{n'}, P^0_1,\dots,P^0_{n_0} \}$. Аналогично $Int_{H_0} = \{ P_1,\dots,P_{n'}, P^0_1,\dots,P^0_{n_0} \}$. Будем предполагать, что для T_{H_0} всегда найдется такое множество универсальных формул [1] σ_{H_0} в словаре v_{H_0}, что $T_{H_0}(pr) = 1$ тогда и только тогда, когда все формулы из σ_{H_0} истинны на pr. Определение истинности формулы на pr приведено ниже. Не ограничивая общности, потребуем, чтобы среди формул σ_{H_0} не было тождественно-истинных, которые на самом деле не вносят никакого вклада в содержание H_0.

Фиксируем некоторую гипотезу H_0 и множество объектов **U**. С помощью некоторого случайного процесса выберем множество $\mathbf{B} = \{ a_1,\dots,a_m \}$ объектов из генеральной совокупности **U**. Возьмем в качестве входной информации для метода пару $\langle H_0, pr_0 \rangle$, где H_0 – фиксированная гипотеза, pr_0 – запись результата применения экспериментальных процедур из Int_{H_0} к множеству **B**. Будем предполагать, что $T_{H_0}(pr_0) = 1$, так как в противном случае гипотеза H_0 опровергнута экспериментом pr_0 и пара $\langle H_0, pr_0 \rangle$ не может быть входной информацией метода.

Построим разложение универсальных формул. Пусть Φ – универсальная, не тождественно-истинная формула $\Phi = \forall x_1,\dots,x_n\, A(x_1,\dots,x_n)$, где $A(x_1,\dots,x_n)$ – бескванторная формула. Тогда $A(x_1,\dots,x_n)$ – не тождественно-истинная формула исчисления высказываний, где $P_i(x_1,\dots,x_{n_i})$, входящие в формулу Φ, рассматриваются как булевы переменные. Используем тождества:

$$C_1 \vee C_2 \vee \dots \vee C_n \equiv \bar{C}_2 \& \bar{C}_3 \& \dots \& \bar{C}_n \to C_1 \equiv$$

$$\equiv \bar{C}_1 \& \bar{C}_3 \& \dots \& \bar{C}_n \to C_2 \equiv \dots \equiv \bar{C}_1 \& \bar{C}_2 \& \dots \& \bar{C}_{n-1} \to C_n, \qquad (1)$$

$$\forall x_1,\dots,x_n\ B_1 \& B_2 \equiv \forall x_1,\dots,x_n B_1 \& \forall x_1,\dots,x_n B_2, \qquad (2)$$

где $C_1,\dots,C_n$ – некоторые булевы переменные, B_1, B_2 – произвольные формулы. Так как $A(x_1,\dots,x_n)$ не тождественно-истинна, то её можно привести к сокращенной конъюнктивной нормальной форме [2,3]. В полученной сокращенной конъюнктивной нормальной форме заменим каждый конъюнктивный член на конъюнкцию всех им-

*) В работе [5] третьим значением является значение "не имеет смысла". Для целей данной работы третье значение будем интерпретировать как значение "не определено".

пликаций из (I), тождественных данному конъюнктивному члену. В результате получится конъюнкция вида $C_1^{\varepsilon_1} \& \ldots \& C_n^{\varepsilon_n}$, где $\varepsilon = 0(I)$, если переменная берется с отрицанием (или без него). Применив тождество (2), получим конъюнкцию членов вида

$$\forall x_1,\ldots,x_n (P_{i_1}^{\varepsilon_1}(x_1,\ldots,x_{m_{i_1}}) \& \ldots \& P_{i_n}^{\varepsilon_n}(x_1,\ldots,x_{m_{i_n}}) \to P_{i_0}^{\varepsilon_0}(x_1,\ldots,x_{m_0}). \quad (3)$$

Получим разложение формулы Φ. Обозначим через σ_Φ множество конъюнктивных членов вида (3) разложения Φ.

Определим истинность формулы Φ на pr. Формула Φ - истинна на pr тогда и только тогда, когда на pr истинны все формулы из σ_Φ разложения Φ. Формула Ψ из σ_Φ вида (3) истинна на pr тогда и только тогда, когда она истинна на всех наборах $\langle a_1,\ldots, a_n \rangle$, $a_i \in B(pr)$ [5], на которых определены все предикатные символы из Ψ. При проверке истинности формулы Ψ на наборе $\langle a_1,\ldots,a_n \rangle$, предикатный символ $P_i(a_{j_1},\ldots,a_{j_{m_i}})$ считается истинным (ложным), если соответственно $P_i(a_{j_1},\ldots,a_{j_{m_i}}) \in pr$ ($\bar{P}_i(a_{j_1},\ldots,a_{j_{m_i}}) \in pr$). Если формула вида (3) не определена на каждом наборе $\langle a_1,\ldots,a_n \rangle$, $a_i \in B(pr)$, то она также считается истинной на pr.

Применим полученное разложение к σ_{H_0}. Это всегда можно сделать, так как $A \in \sigma_{H_0}$ не тождественно-истинна. Каждую формулу $A \in \sigma_{H_0}$ разложим в σ_A. Обозначим объединения полученных множеств $\bigcup_{A \in \sigma_{H_0}} \sigma_A$ через σ_0. Таким образом, каждая формула $A \in \sigma_0$ имеет вид (3). В дальнейшем ограничимся рассмотрением только формул вида (3) с кванторами или без них.

Множество всех возможных формул вида (3) в словаре ν_{H_0}, взятых без кванторов, обозначим через M. Посылку формулы $A \in M$ обозначим через Π_A, а заключение - через C_A.

ОПРЕДЕЛЕНИЕ I. Для $A \in M$ переменную x_{i_1}, встречающуюся в Π_A, назовем связанной с переменной x_{i_n} из C_A, если найдутся переменные $x_{i_2},\ldots, x_{i_{n-1}}$ в A (индексы не обязательно все различны) такие, что каждая пара переменных $\langle x_{i_1}, x_{i_2} \rangle$, $\langle x_{i_2}, x_{i_3} \rangle, \ldots, \langle x_{i_{n-1}}, x_{i_n} \rangle$ встречается в каком-нибудь одном из предикатных символов, входящих в A.

В дальнейшем ограничимся рассмотрением только таких формул из M, в которых каждая переменная из посылки связана с какой-нибудь переменной из заключения. Множество таких формул обозначим через MC.

Метод обнаружения закономерностей и метод предсказания построены таким образом, что сначала обнаруживаются закономерности, связывающие предикатные символы $P_1,\ldots,P_n$ с одним каким-то фиксированным предикатным символом P_j^0. На основании найденных закономерностей предсказывается этот предикатный символ. Повторяя процедуру обнаружения закономерностей и предсказания для остальных предикатных символов из $\{P_1^0,\ldots,P_{n_0}^0\}$, мы получим соответствующее множество закономерностей и предсказание по нему всех предикатных символов $P_1^0,\ldots,P_{n_0}^0$. Полученное предсказание будет окончательным, если оно согласуется с гипотезой H_0.

Опишем сначала метод обнаружения закономерностей и метод предсказания для какого-то одного фиксированного предикатного символа P_j^0, который будем обозначать через P_0. Обозначим через MCP_0 множество формул из MC, в заключении которых стоит P_0, и через σ^{P_0} - множество формул из σ_0, в заключении которых стоит P_0, а в посылке встречаются предикатные символы из $\{P_1,\ldots,P_n\}$.

Опишем процедуру, которая составляет элементарный акт анализа pr_0. Зафиксируем какую-нибудь формулу $A \in MCP_0$. Пусть число переменных в A равно k. Выделим из множества $B(pr_0) = \{a_1,\ldots,a_m\}$ случайным образом L групп ($L = \left[\frac{m}{k}\right]$, [ ] - целая часть числа) по k символов объектов в каждой: $\langle a_1^i,\ldots,a_k^i \rangle$, $i = 1,2,\ldots,L$; $a_j^i \in B(pr_0)$. Внутри групп символы объектов расположим в случайном порядке. Сформируем массив целых чисел $M^A[0{:}1,\ldots,0{:}1]$ размерности n+1, содержимое всех элементов $[i_1,\ldots,i_n,i_0]$ которого равно 0, $[i_1,\ldots,i_n,i_0]=0$, $i_j = 0,1$; $j = 0,1,\ldots,n$, где n - число предикатных символов в Π_A. Для каждой из L групп символов объектов проделаем следующее. Сопоставим фиксированной группе с заданным порядком $\langle a_{j_1}^i,\ldots,a_{j_k}^i \rangle$ набор переменных $\langle x_1,\ldots,x_k \rangle$. Подставим в A символы объектов вместо соответствующих переменных. На наборе $\langle a_{j_1}^i,\ldots,a_{j_k}^i \rangle$

предикатные символы $P_1,\dots, P_n, P_0$ из A принимают определенные значения, которые зафиксированы в pr_0. Если в pr_0 хотя бы один предикатный символ принимает значение "не определено", то переходим к следующей группе символов объектов. Если все значения определены, то переменным $i_j = 0,1$; $j=0,1,\dots,n$ приписываем значение I или 0, если соответственно $P_j(a^{i}_{j_1},\dots,a^{i}_{j_{m_j}}) \in pr_0$ или $\overline{P}_j(a^{i}_{j_1},\dots, a^{i}_{j_{m_j}}) \in pr_0$, $j = 0,1,\dots,n$. В значениях переменных i_j, $j = 0,1,\dots, n$, фиксируются значения истинности предикатных символов из A. Значение элемента массива $[i_1,\dots, i_n, i_0]$ увеличиваем на единицу. Переходим к следующей группе символов объектов. В результате этой процедуры в массиве $м^A$ накопятся числа, сообщающие, сколько раз встретилось определённое сочетание значений предикатных символов из A в pr_0. Массив $м^A$ и есть результат элементарного акта анализа pr_0. Очевидно, что для формул A_1 и A_2, имеющих одни и те же предикатные символы, массивы $м^{A_1}$ и $м^{A_2}$ будут одинаковыми с точностью до перестановки индексов.

Сформулируем критерий, который будет применяться при отборе тех формул, которые выражают некоторую закономерность в pr_0. Пусть дан массив $м^A$ [0:1, 0:1], $A = (P_1(x_1,\dots,x_0) \Rightarrow \Rightarrow P_0(x_1,\dots, x_{m_0}))$. Определим $n_{ij}=[i,j]$, $i,j=0,1$; $n= \sum_{ij} n_{ij}$. Пусть $\mathcal{P}_{ij}$, $i,j=0,1$; $\sum_{ij} \mathcal{P}_{ij} = 1$, – вероятности соответствующих событий. Сформулируем гипотезу $\mathcal{H}_0$ о двумерной независимости значений P_1 и P_0, $\mathcal{H}_0: \mathcal{P}_{11} = \mathcal{P}_{1.} \times \mathcal{P}_{.1}$, $\mathcal{P}_{1.} = \mathcal{P}_{11} + \mathcal{P}_{10}$, $\mathcal{P}_{.1} = \mathcal{P}_{11} + \mathcal{P}_{01}$ против альтернативы $\mathcal{H}_1: \mathcal{P}_{11} \neq \mathcal{P}_{1.} \times \mathcal{P}_{.1}$. Эта гипотеза является сложной с одним ограничением и двумя степенями свободы [4, стр.734]. Если $\mathcal{H}_0$ не верна, то значения P_1 и P_2 взаимозависимы, что даёт возможность по значениям P_1 предсказывать значения P_0.

Предположим, что значения $n_{1.}$ и $n_{0.}$, где $n_{1.} = n_{10} + n_{11}$, $n_{0.} = n_{01} + n_{00}$, зафиксированы, а n_{11} и n_{01} – независимые случайные величины. В этом случае сформулированная гипотеза $\mathcal{H}_0$ является гипотезой о равенстве вероятностей в двух совокупностях [4, стр.741], $\mathcal{H}_0: \mathcal{P}_{11}/\mathcal{P}_{.1} = \mathcal{P}_{1.}$. Если $\mathcal{H}_0$ не верна, то

$\mathcal{P}_{11}/\mathcal{P}_{.1} \neq \mathcal{P}_{1.}$ и либо $\mathcal{P}_{11}/\mathcal{P}_{.1} > \mathcal{P}_{1.}$, либо $\mathcal{P}_{11}/\mathcal{P}_{.1} < \mathcal{P}_{1.}$. Если $\mathcal{P}_{11}/\mathcal{P}_{.1} > \mathcal{P}_{1.}$, то тогда $\mathcal{P}_{11}/\mathcal{P}_{1.} > \mathcal{P}_{.1}$, $\mathcal{P}_{00}/\mathcal{P}_{0.} > \mathcal{P}_{.0}$, т.е. посылка увеличивает вероятность заключения в формулах:

$$P_1(x_1,\dots, x_{m_1}) \Rightarrow P_0(x_1,\dots, x_{m_0}),$$

$$\overline{P}_1(x_1,\dots, x_{m_1}) \Rightarrow \overline{P}_0(x_1,\dots, x_{m_0}). \qquad (4)$$

Если $\mathcal{P}_{11}/\mathcal{P}_{.1} < \mathcal{P}_{1.}$, то $\mathcal{P}_{10}/\mathcal{P}_{1.} > \mathcal{P}_{.0}$, $\mathcal{P}_{01}/\mathcal{P}_{0.} > \mathcal{P}_{.1}$, т.е. посылка увеличивает вероятность заключения в формулах:

$$P_1(x_1,\dots, x_{m_1}) \Rightarrow \overline{P}_0(x_1,\dots, x_{m_0}),$$

$$\overline{P}_1(x_1,\dots, x_{m_1}) \Rightarrow P_0(x_1,\dots, x_{m_0}). \qquad (5)$$

Возьмем для проверки гипотезы $\mathcal{H}_0$ точный критерий независимости, предложенный Фишером [4, стр.736]. Этот критерий является равномерно наиболее мощным, несмещённым критерием как в случае, если проверяется двумерная независимость, так и в случае, если проверяется равенство вероятностей в двух совокупностях [4, стр.740]. Таким образом, применив критерий с некоторым уровнем α, мы получим, что либо нужно принять, что значения P_1 и P_0 независимы и P_1 не увеличивает вероятность предсказания P_0, либо, что P_1 и P_0 зависимы и P_1 можно использовать для предсказания P_0, используя приведенные формулы, так как, согласно тому же критерию, значения P_1 увеличивают вероятность соответствующих значений P_0. Если P_1 и P_0 зависимы, то по соотношениям

$$n_{11} > (n_{1.} \times n_{.1})/n,$$
$$n_{11} < (n_{1.} \times n_{.1})/n \qquad (6)$$

будем знать, применять ли для предсказания формулы (4) или (5) соответственно. В дальнейшем всегда будем предполагать, что зафиксирован некоторый уровень α, с которым применяется критерий.

Рассмотрим $M^A[0:1,\dots,0:1]$, $A \in МСР_0$.

$$A = (P_1^{\varepsilon_1}(x_1,\dots,x_{m_1})\&\dots\&P_n^{\varepsilon_n}(x_1,\dots,x_{m_n}) \to P_0^{\varepsilon_0}(x_1,\dots,x_{m_0})).$$

Нужно выяснить, все ли предикатные символы $P_1^{\varepsilon_1},\dots,P_n^{\varepsilon_n}$ или их совокупности существенны для предсказания $P_0^{\varepsilon_0}$. Рассмотрим произвольное, не пустое подмножество $P_{i_1}^{\varepsilon_{i_1}},\dots,P_{i_k}^{\varepsilon_{i_k}}$ множества $\{P_1^{\varepsilon_1},\dots,P_n^{\varepsilon_n}\}$. Определим массив $M^A_{i_1,\dots,i_k}[i,j]$, $i,j = 0,1$;

$$M^A_{i_1,\dots,i_k}[1,j] = M^A[\varepsilon_1,\dots,\varepsilon_n,j],$$

$$M^A_{i_1,\dots,i_k}[0,j] = \sum_{\varepsilon^1_{i_s}=0,1;\ i_s=i_1,\dots,i_k} M^A[\varepsilon^1_1,\dots,\varepsilon^1_n,j] - M^A[\varepsilon_1,\dots,\varepsilon_n,j],$$

$$\varepsilon^1_{i_s}=\varepsilon_{i_s};\ i_s \neq i_1,\dots,i_k.$$

Этот массив даёт возможность узнать, как влияют предикатные символы $P_{i_1}^{\varepsilon_{i_1}},\dots,P_{i_k}^{\varepsilon_{i_k}}$, рассматриваемые как один предикатный символ вида $P_{i_1}^{\varepsilon_{i_1}}\&\dots\&P_{i_k}^{\varepsilon_{i_k}}$, при фиксированных значениях остальных предикатных символов на увеличение вероятности предсказания $I_0^{\varepsilon_0}$.

ОПРЕДЕЛЕНИЕ 2. Подмножество $P_{i_k}^{i_1},\dots,P_{i_k}^{i_k}$ существенно в A, если конъюнкция $P_{i_1}^{\varepsilon_{i_1}}\&\dots\&P_{i_k}^{\varepsilon_{i_k}}$ взаимозависима с P_0 согласно критерию, примененному к $M^A_{i_1,\dots,i_k}$, и значение ε_0 согласно с первой формулой из (4) или (5) в соответствии с (6).

ОПРЕДЕЛЕНИЕ 3. Формулу $A \in МСР_0$ назовем неприводимой, если любое не пустое подмножество предикатов из $П_A$ существенно в A.

ОПРЕДЕЛЕНИЕ 4. Формулу $A \in МСР_0$ назовем максимальной неприводимой, если нет неприводимой формулы $A' \in МСР_0$ такой, что $C_A \equiv C_{A'}, П_{A'} \equiv П_A \& P_{n+1}^{\varepsilon_{n+1}} \&\dots\& P_{n'}^{\varepsilon_{n'}}$.

Пусть дана произвольная формула A.

ОПРЕДЕЛЕНИЕ 5 [3]. Простым следствием формулы A называется такая не содержащая повторений и не тождественно-истинная элементарная дизъюнкция, которая, будучи логическим следствием A, не поглощается никаким более сильным следствием того же вида, т.е. после отбрасывания какого-нибудь из её членов, перестаёт быть следствием.

СЛЕДСТВИЕ 1. Сокращенная конъюнктивная нормальная форма формулы A есть конъюнкция всех её простых следствий.

СЛЕДСТВИЕ 2. Любая $A \in \sigma^{P_0}$ является такой импликацией, что из $П_A$ нельзя убрать ни один предикат или их совокупность так, чтобы A осталось следствием σ_{H_0}.

ОПРЕДЕЛЕНИЕ 6. Формулы из σ^{P_0} будем называть априорно максимальными неприводимыми.

3. Метод обнаружения закономерностей. Он заключается в основном в нахождении максимальных неприводимых формул. В результате обучения с уровнем α будет найдено множество F_α формул, которые будем называть закономерностями уровня α на pr_0. Введем некоторые обозначения:

$$МП_A \overset{df}{=} \{P_1^{\varepsilon_1},\dots,P_n^{\varepsilon_n}\}, \text{ где } П_A = P_1^{\varepsilon_1}\&\dots\&P_n^{\varepsilon_n};$$

$$M_{H_0} \overset{df}{=} \{A' \mid A' \in МСР_0\ A \in \sigma^{P_0}\ МП_{A'} \subset МП_A \text{ или } МП_A \subset МП_{A'}\};$$

$$M_{об} \overset{df}{=} МСР_0 \setminus M_{H_0},\quad M^t_{об} \overset{df}{=} \{\, A \mid A \in M_{об}, |МП_A| = t \,\}.$$

Метод имеет следующие этапы:

0. $F_\alpha \overset{df}{=} \sigma^{P_0}$.

I. Если $A \in M^1_{об}$ неприводима, то включаем её в F_α.

t. В общем случае t-й этап, $t = 1,2,\ldots$. Если есть $A' \in M^t$, $A \in F_\alpha$, $МП_A \subset МП_{A'}$ и $МП_{A'} \setminus МП_A$ существенно в A', то формулу A выбрасываем из F_α. Добавляем к F_α все неприводимые формулы $A' \in M^t_{об}$.

Процедура заканчивается на таком этапе t, что ни для одной формулы из $M^t_{об}$ при данном уровне α и при данном обучающем материале $рг_0$ не может в принципе оказаться, что она неприводима, т.е. F_α в принципе не может быть пополнена.

Поясним процедуру обучения. Если нашлись $A' \in M^t_{об}$, $A \in F_\alpha$, $МП_A \subset МП_{A'}$ и $МП_{A'} \setminus МП_A$ существенно в A', то это значит, что предсказание формулы A можно уточнить с помощью предикатов из $МП_{A'} \setminus МП_A$. Если при этом A' неприводима, то это уточнение фиксируется в F_α включением A'. Может оказаться, что в A' содержится не вся "полезная" информация, которая есть в A. "Остаток" информации можно записать в виде множества формул:

$$\{A'' \mid \Pi_{A''} = \Pi_A \,\&\, (\overline{P_i^{\varepsilon_i}}),\ P_i^{\varepsilon_i} \in МП_{A'} \setminus МП_A \,\}.$$

Но эти формулы, в свою очередь, также будут рассмотрены. Если A' не окажется неприводимой, то, хотя $МП_{A'} \setminus МП_A$ существенно в A', тем не менее есть подмножество Q в $МП_{A'}$, не существенное в A' и, значит, "дублирующее" информацию из $МП_A$. Следовательно, можно ограничиться рассмотрением формулы A''' с $МП_{A'''} = МП_{A'} \setminus Q$, которая также рассматривается отдельно.

Для окончания процедуры обучения нужно найти оценки условной вероятности следствия при истинности посылки. Для каждой формулы $A \in F_\alpha$, $A \notin \sigma^{P_0}$, используя M^A, можно вычислить с некоторым фиксированным коэффициентом доверия $1 - \beta$, используя один из известных методов доверительных интервалов, нижний доверительный предел D^β_A для вероятности $\mathcal{P}(C_A / \Pi_A)$, $\mathcal{P}(\mathcal{P}(C_A / \Pi_A) \geq D^\beta_A) \geq 1 - \beta$. Нам надо, в отличие от приведенного доверительного предела, вычислить доверительный предел D^γ_A, учитывающий одновременное выполнение аналогичных неравенств для всех формул из $M^{(t)} \overset{df}{=} M^1_{об} \cup \ldots \cup M^t_{об}$, где t – последний этап обучения. Пусть $\overline{M}^{(t)}_{об} = N$. Предположим, что найдены доверительные пределы для всех событий F_i из некоторого множества $F = \{ F_i \}$, $i = 1,2,\ldots,N_F$; $\mathcal{P}(\mathcal{P}_{F_i} \geq D^\beta_{F_i}) \geq 1 - \beta$, $i = 1,2,\ldots,N_F$. Тогда, применяя теорему о сложении вероятностей, получим

$$\mathcal{P}((\mathcal{P}_{F_1} < D^\beta_{F_1}) \vee (\mathcal{P}_{F_2} < D^\beta_{F_2}) \vee \ldots \vee (\mathcal{P}_{F_{N_F}} < D^\beta_{F_{N_F}})) < N_F \cdot \beta$$

или, что то же самое,

$$\mathcal{P}((\mathcal{P}_{F_1} \geq D^\beta_{F_1}) \,\&\, \ldots \,\&\, (\mathcal{P}_{F_{N_F}} \geq D^\beta_{F_{N_F}})) \geq 1 - N_F \cdot \beta.$$

Таким образом, для того чтобы учесть одновременное выполнение неравенств для всех формул из $M^{(t)}_{об}$, достаточно подсчитать доверительный предел D^γ_A для $A \in F_\alpha$, $A \notin \sigma^{P_0}$ с уровнем $\gamma = \beta / N$. Для $A \in \sigma^{P_0}$ положим $D^\gamma_A = 1$.

На этом процедура обнаружения закономерностей заканчивается. Результатом является множество формул F_α, для каждой из которых подсчитан предел D^γ_A. Множество F_α – множество закономерностей уровня α – может иметь самостоятельный интерес, помимо использования его в предсказании.

Множество закономерностей F_α ещё не есть общая закономерность для предсказания P_0. Такую общую закономерность, исходя из множества

$$F_\alpha = \{\Pi_{A^1_i} \overset{D^\gamma_i}{\Rightarrow} P_0,\ i = 1,2,\ldots,l_1\} \cup \{\Pi_{A^2_j} \overset{D^\gamma_j}{\Rightarrow} \overline{P}_0,\ j = 1,2,\ldots,l_2\},$$

можно было бы, например, получить, если взять формулу

$$(\Pi_{A^1_1} \,\&\, \Pi_{A^1_2} \,\&\, \ldots \,\&\, \Pi_{A^1_{l_1}} \overset{D^\gamma_1}{\Rightarrow} P_0) \,\&\, (\Pi_{A^2_1} \,\&\, \Pi_{A^2_2} \,\&\, \ldots \,\&\, \Pi_{A^2_{l_2}} \overset{D^\gamma_2}{\Rightarrow} \overline{P}_0),$$

где $D_1^Y = \min_i D_i^Y$, $D_2^Y = \min_j D_j^Y$. Но с помощью такой закономерности можно предсказывать P_0 только с наименьшей оценкой вероятности. Любое подобное обобщение, связанное с уменьшением числа вероятностных характеристик (в данном случае D_j^Y), дает более грубую, хотя, возможно, более простую формулировку обнаруженной закономерности (в данном случае за счет приведения членов в посылке вид закономерности упростится). Поэтому в работе не рассматривается вопрос об общем представлении закономерностей из F_α. Для получения наиболее точного предсказания используется само F_α.

4. Метод предсказания на основании найденных закономерностей. Предположим, что из генеральной совокупности U случайно выбрали объект b. Определим $Int^i_{H_0} = Int_{H_0} \setminus \{P^0_0, \ldots, P^0_{n_0}\}$; pr_0 – протокол наблюдений объектов $a_1, \ldots, a_m, b$ с помощью интенсиональных процедур из $Int^i_{H_0}$, т.е. будем считать, что относительно символа объекта b нам известны значения всех предикатных символов из $\{P_1, \ldots, P_n\}$.

Фиксируем набор $\langle a_{i_1}, \ldots, a_{i_{k-1}}, b, a_{i_{k+1}}, \ldots, a_{i_{m_0}} \rangle$, $a_{i_j} \in B(pr_0)$. Опишем метод предсказания значения $P_0(a_{i_1}, \ldots, b, \ldots, a_{i_{m_0}})$ для этого набора. Для остальных наборов он аналогичен. Переберём последовательно формулы из F_α. Для $A \in F_\alpha$ смотрим, есть ли в Π_A переменные, не входящие в C_A. Если есть, то из $B(pr_0)$ случайно выберем столько символов объектов $\langle a_1, \ldots, a_k \rangle$, сколько таких переменных. Подставим в Π_A объекты из наборов $\langle a_1, \ldots, a_k \rangle, \langle a_{i_1}, \ldots, b, \ldots, a_{i_{m_0}} \rangle$, объекты из второго набора подставляются вместо $x_1, \ldots, x_{m_0}$. Проверим, все ли значения предикатных символов из Π_A имеют смысл и истинна ли Π_A. Если нет, то переходим к следующей формуле. Если да, то делаем предсказание, что $P_0(a_{i_1}, \ldots, b, \ldots, a_{i_{m_0}})$ истинна, если $\varepsilon_0 = 1$, и ложна, если $\varepsilon_0 = 0$, с вероятностью большей, либо равной D_A^Y. Переходим к следующей формуле. После пересмотра всех формул нужно принять решение, предсказывать ли значение "истина", "ложь"

или вообще не делать никаких предсказаний – "отказ". Это решение должно определяться теми дополнительными предположениями, которые мы делаем: например, априорными знаниями о характере задачи, признаков, о требуемой надежности предсказания. Мы сделаем предположение о независимости формул из F_α. Две формулы $\Pi_1 \to P_0$ и $\Pi_2 \to P_0$ будем считать независимыми, если выполняются следующие два требования:

1. $\mathcal{P}(\Pi_1 \& \Pi_2) = \mathcal{P}(\Pi_1) \cdot \mathcal{P}(\Pi_2)$.

2. $\mathcal{P}(\Pi_1 \& \Pi_2 \& \overline{P_0}) = \mathcal{P}(\Pi_1 \& P_0) \cdot \mathcal{P}(\Pi_2 \& \overline{P_0})$.

Из этих условий следует, что

$$\mathcal{P}(P_0/\Pi_1 \vee \Pi_2) = 1 - (1 - (P_0/\Pi_1))(1 - (P_0/\Pi_2)).$$

Пусть на основании формул $A_1, \ldots, A_s$ из F_α было сделано предсказание истинности P_0 с вероятностями большими, либо равными $D^Y_{A_1}, \ldots, D^Y_{A_{s_1}}$. Тогда, используя предположение о независимости, можно оценить вероятность $\mathcal{P}_И$ истинности P_0: $\mathcal{P}_И \geq 1 - \prod_{i=1}^{s_1} (1 - D^Y_{A_i})$. Аналогично можно оценить вероятность $\mathcal{P}_Л$, используя предсказания ложности P_0, сделанные на основании каких-то других формул $A'_1, \ldots, A'_{s_2}$ из F_α: $\mathcal{P}_И \geq 1 - \prod_{i=1}^{s_1} (1 - D^Y_{A'_i})$. В случае, если $D^Y_{A_i} = 1$ для некоторого i, $\mathcal{P}_Л$ полагаем равным 0. В случае, если $D^{Y_i}_{A'_j} = 1$ для некоторого j, $\mathcal{P}_И$ полагаем равным 0. В обоих случаях предположение о независимости не может быть выполнено. Поэтому значения $\mathcal{P}_И$ и $\mathcal{P}_Л$ определяются согласно априорной информации. Используя $\mathcal{P}_И$ и $\mathcal{P}_Л$, можно теперь получить окончательное предсказание.

Для $P_0(a_{i_1}, \ldots, b, \ldots, a_{i_{m_0}})$ предсказывается значение "истина", если $\mathcal{P}_И > \mathcal{P}_Л$, и предсказывается значение "ложь", если $\mathcal{P}_Л > \mathcal{P}_И$. Если $\mathcal{P}_И = \mathcal{P}_Л$ или если не нашлось ни одной формулы, на основании которой можно было бы сделать предсказание, то тогда не делается никакого предсказания – значение "не определено".

Повторяя описанную процедуру предсказания для каждого набора $\langle a_{i_1}, \ldots, b, \ldots, a_{i_{m_0}} \rangle$, $a_1, \ldots, a_{m_0} \in B(pr_0)$, можно предска-

зать все значения предикатного символа P_0 на наборах, включающих новый объект b, т.е. доопределить P_0 на множестве $\{a_1, \ldots, a_m, b\}$.

Итак, мы описали метод обнаружения закономерностей и метод предсказания для одного фиксированного предикатного символа P_0. Применяя описанный метод ко всем остальным предикатным символам из $\{P^0_1, \ldots, P^0_{n_0}\}$, мы доопределим все их на множестве $\{a_1, \ldots, a_m, b\}$. Получив такое предсказание, мы ещё не получим в точном смысле предсказания признака x_0 в соответствующей шкале. Доопределенные значения $P^0_1, \ldots, P^0_{n_0}$ на $\{a_1, \ldots, a_m, b\}$ должны удовлетворять аксиомам из σ_0, содержащим только предикатные символы из $\{P^0_1, \ldots, P^0_{n_0}\}$. Если аксиомы выполняются, то полученное предсказание является окончательным, если нет, то тогда полученное предсказание не удовлетворяет априорным предположениям о свойствах x_0 и не делается никакого предсказания – "отказ".

Автор считает своим приятным долгом выразить благодарность Ю.Г.Косареву, А.А.Москвитину, и К.Ф.Самохвалову за ряд полезных замечаний по работе.

Литература

1. МАЛЬЦЕВ А.И. Алгебраические системы, М., "Наука", 1970.

2. Дискретная математика и математические вопросы кибернетики, под ред. С.В.Яблонского, О.Б.Лупанова, т.I, "Наука", 1974.

3. БРОДСКИЙ И.Н. Элементарное введение в символическую логику, 2-е изд., ЛГУ., 1972.

4. КЕНДАЛ М., СТЬЮАРТ А. Статистические выводы и связи. М., "Наука", 1973.

5. САМОХВАЛОВ К.Ф. О теории эмпирических предсказаний. -В кн.: Вычислительные системы. Вып. 55, Новосибирск, 1973, с.3-35.

6. ВИТЯЕВ Е.Е. Алгоритм эмпирического предсказания. -В кн.: Вычислительные системы. Вып. 61, Новосибирск, 1975, с. 28-36.

7. Психологические измерения. Сб. под ред. Л.Д.Мешалкина. М., "Мир", 1967.

8. Нелинейные и линейные методы в распознавании образов. М., "Наука", 1975.

9. ИВАХНЕНКО А.Г., ЛАПА В.Г. Предсказание случайных процессов. Киев, "Наукова думка", 1971.

Поступила в ред.-изд.отд.
13 января 1976 года