ЖУРНАЛ ВЫЧИСЛИТЕЛЬНОЙ МАТЕМАТИКИ И МАТЕМАТИЧЕСКОЙ ФИЗИКИ, 2003, том 43, № 9, с. 1438–1447

УДК 519.710.2

ПЕРЕСТАНОВОЧНЫЙ ТЕСТ В МЕТОДЕ ОПТИМАЛЬНЫХ РАЗБИЕНИЙ

© 2003 г. О. В. Сенько

(119991 Москва, ул. Вавилова, 40, ВЦ РАН)
e-mail: Senko@ccas.ru

Поступила в редакцию 26.03.2002 г.
Переработанный вариант 03.03.2003 г.

Рассматривается метод изучения и описания зависимостей по информации, содержащейся в эмпирических таблицах. В основе метода лежит поиск системы закономерностей, характеризующих существующую зависимость некоторой величины от набора независимых прогностических переменных. Под закономерностью понимается такая область пространства независимых переменных, для которой существует тенденция существенного отклонения значений зависимой величины от ее средних значений по генеральной совокупности. Для поиска закономерностей предлагается метод, основанный на построении оптимальных разбиений пространства независимых переменных. Статься посвящена статистической верификации найденных закономерностей с помощью перестановочного теста. Библ. 6. Фиг. 1. Табл. 1.

ВВЕДЕНИЕ

Метод оптимальных разбиений (см. [1]–[3]) предназначен для поиска по имеющейся выборке прецедентов наиболее полной системы закономерностей, характеризующих зависимость некоторой величины ζ от набора потенциальных прогностических переменных $X_1, \ldots, X_n$. В качестве ζ могут выступать объекты различной природы: бинарные индикаторные функции классов, векторы непрерывных переменных, кривые выживаемости [4] и т.д. Под закономерностью понимается такая подобласть многомерного пространства $\mathbb{R}^n$, для которой существует тенденция существенных отклонений значений величины ζ от ее средних значений по генеральной совокупности. Методы анализа данных и прогнозирования, основанные на оптимальных разбиениях пространства независимых переменных, получили значительное распространение в последние годы. Следует отметить наиболее известный алгоритм CART, основанный на построении регрессионных деревьев (см. [5], [6]). Отличительной особенностью рассматриваемого в настоящей работе подхода является его цель, заключающаяся в наиболее полном и достоверном описании существующих зависимостей. Данная задача на самом деле не является тождественной обычно рассматриваемой задаче построения наиболее точного прогнозирующего алгоритма.

Описания исследуемых прецедентов (объектов) могут быть представлены в виде пар $(\zeta, \mathbf{x})$, где $\mathbf{x}$ принадлежит к некоторому множеству $M_{\mathbf{x}}$ точек многомерного пространства $\mathbb{R}^n$, а зависимая величина ζ принимает значения из некоторого множества M_ζ. Иными словами, множество допустимых объектов исследования M_0 может быть представлено как декартово произведение $M_0 = M_\zeta \times M_{\mathbf{x}}$. Считается, что на множестве M_0 задана σ-алгебра Σ_0, являющаяся декартовым произведением σ-алгебр $\Sigma_\zeta \times \Sigma_{\mathbf{x}}$, заданных на M_ζ и $M_{\mathbf{x}}$. Считается, что на Σ_0 определена вероятностная мера $\mathbf{P}_0$. Эмпирическая оценка условного математического ожидания $\mathbf{M}(\zeta|a)$, где $a \in \Sigma_{\mathbf{x}}$, может быть осуществлена по эмпирической выборке $\tilde{S}_i$, состоящей из объектов, для которых вектор $\mathbf{x}$ принадлежит множеству a. Эту эмпирическую оценку далее будем обозначать через $\hat{\zeta}(\tilde{S}_i)$. Предполагается, что на множестве $M_\zeta \times M_\zeta$ задана функция расстояний $\rho(\zeta', \zeta'')$, соответствующая существующим представлениям о взаимной удаленности объектов и обладающая следующими свойствами:

$$\rho(\zeta', \zeta'') \geq 0, \quad \rho(\zeta', \zeta') = 0, \quad \rho(\zeta', \zeta'') = \rho(\zeta'', \zeta') \; \forall \zeta', \zeta'' \in M_\zeta.$$

Одним из способов эмпирического оценивания является выбор элемента M_ζ, наименее удаленно-

го от объектов $\tilde{S}_i$, а именно

$$\hat{\zeta}(\tilde{S}_i) = \arg\min_{\zeta' \in \mathbf{M}_\zeta} \left[\sum_{S_j \in \tilde{S}_j} \rho(\zeta_j, \zeta') \right].$$

Предлагаемый метод разбиений предназначен для исследования зависимости описания ζ от независимых переменных $X_1, \ldots, X_n$ по эмпирической обучающей выборке $\tilde{S}_0 = \{ s_1^0 = (\boldsymbol{\zeta}_1^0, \mathbf{x}_1^0), \ldots, s_m^0 = (\boldsymbol{\zeta}_m^0, \mathbf{x}_m^0) \}$, где $\mathbf{x}_j^0$ – вектор независимых переменных, а $\boldsymbol{\zeta}_j^0$ – часть описания, связанная с прогнозируемой величиной ζ. В большинстве случаев $\boldsymbol{\zeta}_j^0$ представляет собой просто значение ζ для j-го объекта. Однако для случая прогноза кривых выживаемости в качестве $\boldsymbol{\zeta}_j^0$ выступает пара (t_j, α_j), где t_j – время последнего наблюдения за объектом, а α_j – индикатор, указывающий, существовал ли объекта в момент t_j или в это время была зафиксирована его гибель.

Метод основан на построении оптимальных разбиений интервалов допустимых значений одиночных переменных или совместных областей допустимых значений групп переменных в рамках априори заданных моделей. Причем разбиение считается оптимальным, если оно индуцирует разбиение $\tilde{S}_0$ на несколько групп с возможно минимальными расстояниями между объектами внутри одной и той же группы и возможно максимальными расстояниями между объектами из разных групп. Данные различия между группами и внутри групп описываются с помощью специального функционала качества разбиений. Задача при этом сводится к поиску разбиений из рассматриваемых моделей, на которых достигается максимум функционала качества.

1. ФУНКЦИОНАЛ КАЧЕСТВА

Предположим, что R – разбиение обучающей выборки $\tilde{S}_0$ на подвыборки $\tilde{S}_1, \ldots, \tilde{S}_q$. Тогда интегральный функционал качества определяется как сумма

$$F_I(\tilde{S}_0, R) = \left[\sum_{i=1}^{q} \rho(\hat{\zeta}(\tilde{S}_i), \hat{\zeta}(\tilde{S}_0)) m_i \right] D_0^{-1},$$

где m_i – число объектов в подвыборке $\tilde{S}_i$,

$$D_0 = \left[\sum_{i=1}^{m} \rho(\zeta_j, \hat{\zeta}(\tilde{S}_0)) \right] (m-1)^{-1}.$$

Наряду с интегральным функционалом качества может быть использован также локальный функционал качества, в котором оценка проводится по подвыборке, максимально отличающейся от исходной обучающей выборки $\tilde{S}_0$:

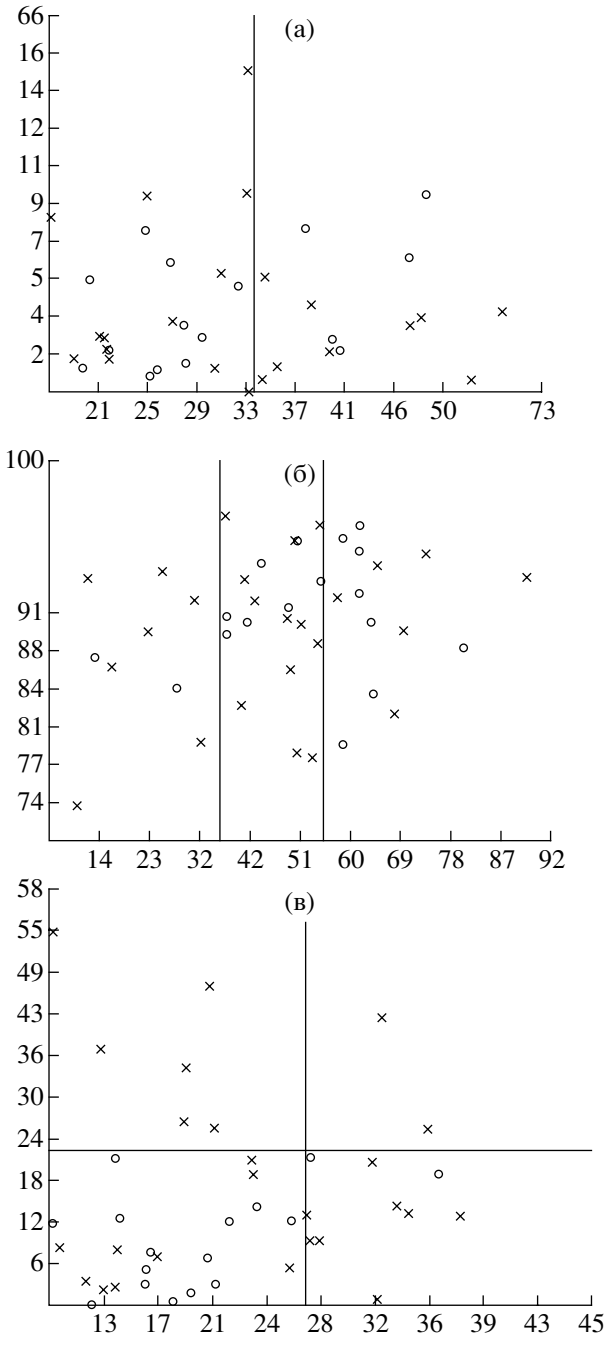
$$F_L(\tilde{S}_0, R) = \max_{i \in \{1, 2, \ldots, q\}} \{ \rho[\hat{\zeta}(\tilde{S}_i), \hat{\zeta}(\tilde{S}_0)] m_i \} D_0^{-1}.$$

2. МОДЕЛИ РАЗБИЕНИЙ

Под моделью разбиения мы понимаем множество разбиений с числом элементов, не превышающим некоторое заранее фиксированное число, которые строятся с помощью априори заданного алгоритма.

Примеры моделей разбиений приведены на фигуре (соответственно, на графиках (а), (б), (в) – модели I–III).

Модель I включает все разбиения интервалов допустимых значений одиночных переменных с числом элементов (подобластей) не более двух, которые разделены с помощью одной граничной точки. Модель II включает все разбиения интервалов допустимых значений одиночных переменных с числом элементов не более трех, которые разделены с помощью не более двух граничных точек. Модель III включает все разбиения области допустимых значений пары перемен-



Фигура.

ных с числом элементов не более четырех. Причем при построении разбиения используется не более одной граничной точки для каждой из двух переменных.

3. ВЕРИФИКАЦИЯ ПОЛУЧЕННЫХ РЕЗУЛЬТАТОВ

Важнейшим свойством результатов анализа является их статистическая достоверность. Задача значительно упрощается, если объем исходной эмпирической информации достаточен для формирования двух независимых выборок $\tilde{S}_0$ и $\tilde{S}_c$. Причем выборка $\tilde{S}_0$ используется для поиска оптимального разбиения R_0 внутри одной из заранее заданных моделей, а выборка $\tilde{S}_c$ – для оценки статистической значимости выявленных закономерностей, которая трактуется как оценка статистической значимости различий между группами объектов $\tilde{S}_c$, формируемыми разбиением R_0. Для получения такой оценки могут быть использованы стандартные статистические критерии: χ^2, Стьюдента, Уилкоксона, logrank-тест и др.

К сожалению, стандартные статистические тесты не могут быть использованы в случае ограниченного объема имеющейся информации. Дело в том, что одни и та же выборка не может быть использована и для поиска закономерностей, и для их верификации. Игнорирование этого обстоятельства реально может привести к существенно завышенным оценкам статистической значимости. Поэтому для верификации результатов предлагается перестановочный тест, позволяющий использовать одну и ту же выборку $\tilde{S}_0$ для получения оптимальных решений и для их верификации.

Перестановочный тест основан на проверке нулевой гипотезы $\mathcal{H}_0$ о независимости прогнозируемого описания ζ от переменных, используемых при построении оптимальных разбиений. Напомним, что в качестве меры статистической значимости $\mathcal{H}_0$ используется вероятность $\Pr\{T(\tilde{S}) > T(\tilde{S}_0) \mid \mathcal{H}_0,\ \tilde{S} \in W(\tilde{S}_0)\}$, где $W(\tilde{S}_0)$ – некоторое множество допустимых выборок, содержащее $\tilde{S}_0$, а $T(\tilde{S})$ – функция, определенная на выборке реальных данных и характеризующая степень отклонения этих данных от нулевой гипотезы. В случае использования перестановочного теста для верификации закономерностей, выявленных с помощью метода оптимальных разбиений, в качестве статистики критерия берется оптимальное значение используемого функционала качества разбиений, которое будем обозначать через $F_{\text{opt}}(\tilde{S})$. Выборки данных рассматриваются для простоты как элементы декартового произведения $M_0^m = M_0 \times \ldots \times M_0$. Предполагается, что объекты в выборках независимы друг от друга и одинаково распределены. Иными словами, предполагается, что выборки принадлежат вероятностному пространству

$(M_0^m, \Sigma_0^m, \mathbf{P}_0^m)$, где Σ_0^m – минимальная σ-алгебра, содержащая все элементы вида $\tilde{a} = (a_1, \ldots, a_m)$, где $a_i \in \Sigma_0$. Вероятностная мера $\mathbf{P}_0^m$ для каждого такого элемента удовлетворяет условию

$$\mathbf{P}_0^m(\tilde{a}) = \mathbf{P}_0(a_1)\ldots\mathbf{P}_0(a_m).$$

Сопоставим набору элементов $\bar{\zeta} = (\zeta_1, \ldots, \zeta_m)$ неупорядоченное множество положительных целых чисел $\tilde{l}(\bar{\zeta}) = \{l(\zeta_1^*), \ldots, l(\zeta_k^*)\}$, где $\zeta_1^*, \ldots, \zeta_k^*$ – отличные друг от друга элементы M_ζ, $k \le m$, $l(\zeta_j^*)$ – число элементов из $\bar{\zeta}$, равных ζ_j^*. Обозначим через $\bar{\zeta}(\tilde{S}_0)$ набор элементов $(\zeta_1^0, \ldots, \zeta_m^0)$ из $\tilde{S}_0$. В качестве множества $W(\tilde{S}_0)$ в перестановочном тесте используется множество выборок данных $W^p(\tilde{S}_0)$ вида $\{(\zeta_1', \mathbf{x}_1^0, \ldots, \zeta_m', \mathbf{x}_m^0)\}$, которые удовлетворяют условию $\tilde{l}(\bar{\zeta}') = \tilde{l}[\bar{\zeta}(\tilde{S}_0)]$, где $\bar{\zeta}' = (\zeta_1', \ldots, \zeta_m')$. очевидно, что $W^p(\tilde{S}_0)$ представляет собой конечное множество выборок, каждая из которых может быть получена из $\tilde{S}_0$ путем перестановки элементов ζ относительно фиксированных векторов **x**. Ограничимся случаем, когда множества M_ζ и $M_{\mathbf{x}}$ конечны. Из-за конечной точности любых физических измерений данное ограничение не является существенным для подавляющего числа практических задач. Справедлива следующая

Теорема 1. *Пусть N_0^p – число отличных друг от друга элементов множества $W^p(\tilde{S}_0)$, а $N_1^p[F_{\text{opt}}(\tilde{S}_0)]$ – число отличных друг от друга элементов $W^p(\tilde{S}_0)$ таких, что $F_{\text{opt}}(\tilde{S}') > F_{\text{opt}}(\tilde{S}_0)$. Пусть также справедлива нулевая гипотеза $\mathcal{H}_0$ о независимости ζ от* **x** *и гипотеза о независимости друг от друга отдельных объектов; тогда*

$$\Pr\{F_{\text{opt}}(\tilde{S}') > F_{\text{opt}}(\tilde{S}_0) \,|\, \mathcal{H}_0, \tilde{S} \in W^p(\tilde{S}_0)\} = \frac{N_1^p[F_{\text{opt}}(\tilde{S}_0)]}{N_0^p}.$$

Доказательство. При справедливости условий теоремы все элементы $W^p(\tilde{S}_0)$ имеют одну и ту же вероятность $\mathbf{P}_0(\tilde{\zeta}_1^0)\mathbf{P}_0(\tilde{\mathbf{x}}_1^0)\ldots\mathbf{P}_0(\tilde{\zeta}_m^0)\mathbf{P}_0(\tilde{\mathbf{x}}_m^0)$, где $\tilde{\zeta}_j^0$ – множество всех элементов M_0 вида $(\zeta_j^0, \mathbf{x})$, а $\tilde{\mathbf{x}}_j^0$ – множество всех элементов M_0 вида $(\zeta, \tilde{\mathbf{x}}_j^0)$. Пусть $W_1^p(\tilde{S}_0) = \{\tilde{S}' \mid \tilde{S}' \in W_1^p(\tilde{S}_0), T(\tilde{S}') > T(\tilde{S}_0)\}$. Вероятность

$$\Pr\{T(\tilde{S}') > T(\tilde{S}_0) \mid \mathcal{H}_0, \tilde{S} \in W^p(\tilde{S}_0)\} = \frac{\mathbf{P}_0^m[W_1^p(\tilde{S}_0)]}{\mathbf{P}_0^m[W^p(\tilde{S}_0)]} = \frac{N_1^p[T(\tilde{S}_0)]}{N_0^p}.$$

Число элементов N_0^p в множестве $W^p(\tilde{S}_0)$ обычно чрезвычайно велико из-за очень быстрого роста числа возможных перестановок даже при небольшом увеличении объемов выборки. Полный перебор их для большинства практических зада совершенно невозможен. Однако для оценок отношения $N_1^p[T(\tilde{S}_0)]/N_0^p$ может быть использован метод, основанный на случайной генерации множества выборок $W_{\text{мс}}^p(\tilde{S}_0) \subseteq W^p(\tilde{S}_0)$ и вычислении той доли из них, для которой $T(S') > T(\tilde{S}') > T(\tilde{S}_0)$. Недостатком таких оценок является их не очень большая точность, особенно в случаях очень высокого уровня значимости выявленной закономерности. Для повышения точности требуется соответствующее увеличение объема набора $W_{\text{мс}}^p(\tilde{S}_0)$ с соответствующим ростом трудоемкости вычислений. Возникает также вопрос о возможности систематических ошибок, связанных с тем, что числовые последовательности, генерируемые случайными датчиками, на самом деле являются псевдослучайными.

Для ответа на эти вопросы представляется интересным получение точных оценок распределения оптимальных значений функционала качества разбиений. Ниже изложен метод получения таких распределений для случая, когда в качестве зависимой величины ζ выступает бинарная индикаторная функция двух непересекающихся классов.

4. ФУНКЦИОНАЛ КАЧЕСТВА РАЗБИЕНИЙ ПРИ ИЗУЧЕНИИ ЗАВИСИМОСТИ БИНАРНЫХ ИНДИКАТОРНЫХ ФУНКЦИЙ КЛАССОВ

Рассмотрим задачу изучения зависимости принадлежности объектов к двум непересекающимся классам K_1 и K_2 от потенциальной прогностической переменной X. Данная задача может трактоваться как задача изучения зависимости бинарной индикаторной функции ζ, принимающей значение 1 на объектах класса K_1 и значение 0 на объектах класса K_2. Метод разбиений заключается в поиске такого порогового значения δ, чтобы распределения объектов классов K_1 и K_2 в подвыборке $\tilde{S}_1$ с $X < \delta$ и $\tilde{S}_2$ с $X \geq \delta$ различалось максимальным образом. Пусть $\nu_j^i = n_j^i/n_j$ – доля объектов класса K_i в подвыборке $\tilde{S}_j$, n_j – число объектов в подвыборке $\tilde{S}_j$, n_j – число объектов класса K_i в подвыборке $\tilde{S}_j$, $-\nu_j = (n_1^i + n_2^i)/(n_1 - n_2)$ – общая доля объектов класса K_i во всей выборке $\tilde{S}_0 = \tilde{S}_1 \cup \tilde{S}_2$, $i, j \in \{1, 2\}$. В качестве оценки математического ожидания ζ– по подвыборке $\tilde{S}_j$ естественно использовать величину ν_j^i. Для оценки разбиений будем использовать интегральный функционал качества $F_I(\delta, \tilde{S}_0)$, который также можно рассматривать как зависящий от двух подвыборок $\tilde{S}_1$ и $\tilde{S}_2$ и записывать его, соответственно, как $F(\tilde{S}_1, \tilde{S}_2)$. В качестве функции расстояния $\rho[\hat{\zeta}(\tilde{S}_j), \hat{\zeta}(\tilde{S}_0)]$ будет использовано обычная евклидова метрика. Тогда для функционала качества справедлива формула $F_I(\tilde{S}_1, \tilde{S}_2) = [(\nu_1^1 - \nu_1)^2 n_1 + (\nu_2^1 - \nu_1)^2 n_2]/[\nu_1(1 - \nu_1)]$. Знаменатель здесь имеет значение только для сравнения результатов оптимизации для различных задач и никак не влияет на результаты оптимизации в конкретной задаче. Далее будем рассматривать только числитель $F(\tilde{S}_1, \tilde{S}_2) = [(\nu_1^1 - \nu_1)^2 n_1 + (\nu_2^1 - \nu_1)^2 n_2]$.

Далее приводится ряд лемм, касающихся свойств функционала $F(\tilde{S}_1, \tilde{S}_2)$, используемых при дальнейших построениях.

Определение 1. Будем говорить, что подвыборка $\tilde{S}_j$ *предпочтительна по классу* K_i, если $\nu_j^i > \nu_i$, $j, j \in \{1, 2\}$.

Лемма 1. *Пусть подвыборка* $\tilde{S}_1$ *предпочтительна по классу* K_1; *тогда подвыборка* $\tilde{S}_2$ *предпочтительна по классу* K_2.

Доказательство. По определению, из предпочтительности подвыборки $\tilde{S}_1$ по классу K_1 следует, что $n_1^1/n_1 > (n_1^1 + n_2^1)/(n_1 + n_2)$, откуда получаем, что

$$n_1^1 > \frac{n_1}{n_2} n_2^1 \quad \text{или} \quad \nu_1 > \frac{n_2^1 + n_2^1 n_1/n_2}{n_1 + n_2} = \nu_2^1.$$

Лемма 2. *Пусть подвыборка* $\tilde{S}_1$ *предпочтительна по классу* K_1. *Тогда перенос объекта, принадлежащего классу* K_1, *из подвыборки* $\tilde{S}_2$ *в подвыборку* $\tilde{S}_1$ *приводит к увеличению функционала* $F(\tilde{S}_1, \tilde{S}_2)$.

Доказательство. Отметим, что $n_2 > 1$. Нетрудно показать, что

$$F(\tilde{S}_1, \tilde{S}_2) = (\nu_1^1 - \nu_1)^2 n_1 + (\nu_2^1 - \nu_1)^2 n_2 = (1 + n_1/n_2)(\nu_1^1 - \nu_1)^2 n_1.$$

Предположим, что перенос объекта из класса K_1 из подвыборки $\tilde{S}_2$ в подвыборку G_1 привел к образованию групп $\tilde{S}_1'$ и $\tilde{S}_2'$. Тогда

$$F(\tilde{S}_1', \tilde{S}_2') = \left(1 + \frac{n_1 + 1}{n_2 - 1}\right)\left(\frac{n_1^1 + 1}{n_1 + 1} - \nu_1\right)^2 (n_1 + 1).$$

Поскольку $n_1^1/n_1 > \nu_1$, получаем, что $[(n_1^1 + 1)/(n_1 + 1) - \nu_1]^2 \geq (\nu_1^1 - \nu_1)^2$. Очевидно, что $[1 + (n_1 + 1)/(n_2 - 1)] > (1 + n_1/n_2)$ и $n_1 + 1 > n_1$. Таким образом, функционал F может быть представлен в виде произведения трех сомножителей. Причем для $F(\tilde{S}_1', \tilde{S}_2')$ два сомножителя превышают зна-

чения соответствующих сомножителей для $F(\tilde{S}_1, \tilde{S}_2)$ и один сомножитель для $F(\tilde{S}_1', \tilde{S}_2')$ по крайней мере не меньше соответствующего сомножителя для $F(\tilde{S}_1, \tilde{S}_2)$. Следовательно, справедливо строгое неравенство $F(\tilde{S}_1', \tilde{S}_2') > F(\tilde{S}_1, \tilde{S}_2)$. Лемма доказана.

Лемма 3. *Пусть группа $\tilde{S}_1$ предпочтительна по классу K_1. Предположим, что после переноса одного объекта из K_1 в $\tilde{S}_2$ из $\tilde{S}_1$ подвыборка $\tilde{S}_1$ остается предпочтительной по классу K_1. Тогда данный перенос приводит к уменьшению функционала $F(\tilde{S}_1, \tilde{S}_2)$.*

Доказательство непосредственно следует из леммы 2.

Лемма 4. *Пусть подвыборка $\tilde{S}_1$ предпочтительна по классу K_1. Пусть после переноса одного объекта из K_2 в $\tilde{S}_1$ из $\tilde{S}_2$ подвыборка $\tilde{S}_1$ остается предпочтительной по классу K_1. Тогда данный перенос приводит к уменьшению функционала $F(\tilde{S}_1, \tilde{S}_2)$.*

Доказательство. Из леммы 1 следует, что подвыборка $\tilde{S}_2$ предпочтительна по классу K_2 после переноса объекта класса K_2 в подвыборку $\tilde{S}_1$. Условия леммы 4 тождественны условиям леммы 3 с точностью до перестановки номеров классов (групп). В заключение рассмотрим случай, когда ни одна из групп не является предпочтительной ни по одному из классов.

Лемма 5. *Если обе подвыборки $\tilde{S}_1$ и $\tilde{S}_2$ не являются предпочтительными по классам, то функционал $F(\tilde{S}_1, \tilde{S}_2) = 0$ и перенос произвольного объекта из одной подвыборки в другую ведет к увеличению функционала.*

Доказательство данной леммы достаточно очевидно следует из вида функционала.

5. АЛГОРИТМ РАСЧЕТА РАСПРЕДЕЛЕНИЯ ОПТИМАЛЬНЫХ ЗНАЧЕНИЙ ФУНКЦИОНАЛА F НА МНОЖЕСТВЕ ВСЕВОЗМОЖНЫХ ПЕРЕСТАНОВОК

Целью настоящего раздела является построение эффективного алгоритма, позволяющего вычислять величины $N[\tilde{S}_0, a]$, определенные в разд. 3, при произвольных значениях a. Очевидно, что такая задача может быть решена путем прямого перебора всевозможных перестановок. Однако такая процедура является чрезмерно трудоемкой. Полученные ниже результаты позволяют значительно сократить перебор.

Далее будет предполагаться, что значения переменной X на всех объектах $\tilde{S}_0$ различны. Пусть задана некоторая последовательность T_0, включающая N_1 объектов из класса K_1 и N_2 объектов из класса K_2, пронумерованных в порядке возрастания соответствующих значений переменной X. Номер объекта S из K_2 в данной нумерации обозначим $i_2(S)$. Будем использовать запись $i_2(j)$ для обозначения в нумерации i_2 номера объекта $S \in K_2$, имеющего номер j в нумерации i_1. Пусть l – номер объекта $S \in K_2$ в нумерации i_2, тогда через $i_2^-(S)$ будет обозначать его номер в нумерации i_1.

Из леммы 5 следует, что максимум функционала F всегда строго превышает 0 и достигается на разбиении, при котором каждая из групп является предпочтительной по одному из классов. Предположим, что максимум функционала F достигается для некоторого разбиения с граничной точкой $(j_0, j_0 + 1)$ такого, что группа объектов с $i_1 \le j_0$ является предпочтительной по классу K_1. В этом случае будем говорить, что точка максимума $(j_0, j_0 + 1)$ является точкой максимума левого типа по классу K_1. Если же группа объектов с $i_1 \le j_0$ является предпочтительной по классу K_2, будем говорить, что точка максимума $(j_0, j_0 + 1)$ является точкой максимума правого типа по классу K_1.

Лемма 6. *Пусть точка максимума $(j_0, j_0 + 1)$ является точкой максимума левого типа. Тогда объект с номером j_0 принадлежит классу K_1, а объект с номером $i_0 + 1$ – классу K_2.*

Доказательство. Предположим, что объект с номером j_0 принадлежит классу K_2. Поскольку точка $(j_0, j_0 + 1)$ является точкой максимума левого типа, то группа с $i_1 > j_0$ является предпочтительной по классу K_2. Переход из точки $(j_0, j_0 + 1)$ в точку $(j_0 - 1, j_0)$ будет соответствовать переносу объекта класса в группу, которая уже была по этому классу предпочтительной. Следовательно, согласно лемме 2, значение функционала F в $(j_0 - 1, j_0)$ превосходит значение F в точке

$(j_0, j_0 + 1)$, что противоречит условию леммы. Совершенно аналогично рассматривается предположение о том, что объект с номером $j_0 + 1$ принадлежит классу K_1.

Лемма 7. *Пусть точка максимума* $(j_0, j_0 + 1)$ *является точкой максимума правого типа. Тогда объект с номером* j_0 *принадлежит классу* K_2, *а объект с номером* $j_0 + 1$ *– классу* K_1.

Доказательство совершенно аналогично доказательству леммы 6.

Из лемм 6 и 7 следует

Теорема 2. *Разбиение, при котором достигается максимум функционала, задается границей, расположенной между объектами, принадлежащими разным классам.*

Сопоставим последовательности T последовательность T' такую, что на j-м месте в T' стоит объект, стоящий на месте $n - j$ в T. Последовательность T' далее будем называть обратной последовательности T. Предположим что точка $(j_0, j_0 + 1)$ является точкой максимума левого типа по классу K_1. Тогда точка $(n - j_0, n - j_0 - 1$ является точкой максимума правого типа по классу K_1. Предположим, что некоторая последовательность имеет точки максимума только левого типа. Тогда обратная последовательность будет иметь ровно столько же точек максимума правого типа. Очевидно, что справедливо и обратное.

Следовательно, справедлива следующая

Лемма 8. *Процедура обращения задает взаимно однозначное соответствие между последовательностями, имеющими точки максимума функционала* F *по классу* K_1 *только левого или только правого типа. Причем оптимальные значения* F *на соответствующих друг другу последовательностях равны.*

Будем говорить, что объект S из класса K_1 находится в нулевой позиции, если $i_1(S) < i_1(S')$, где $S' \in K_2$, и $i_2(S') = 1$. Будем говорить, что объект S из класса K_1 находится в N_2-й позиции, если $i_1(S) < i_1(S')$, где $S' \in K_2$, и $i_2(S') = N_2$. Наконец, будем говорить, что объект S из класса K_1 находится в j-й позиции, если $i_1(S') > i_1(S) > i_1(S'')$, где $S', S'' \in K_2$, и $i_2(S') = j$, $i_2(S') = j + 1$. Вектором позиций по классу K_2 или просто вектором позиций будем называть целочисленный вектор $\mathbf{h}_T$ размерности $N_2 + 1$, причем j-я компонента равна числу объектов класса K_1 в последовательности T в j-й позиции.

Определение 2. Будем говорить, что последовательность T'' *превышает последовательность* T' $(T'' > T')$ при условиях:

а) существует целое l_0 такое, что $\mathbf{h}_{T''}[l_0] > \mathbf{h}_{T'}[l_0]$;

б) для любого $l_0 < l \le N_2 + 1$ соответствующие компоненты вектора позиций равны между собой, т.е. $\mathbf{h}_{T''}[l_0] > \mathbf{h}_{T'}[l_0]$.

Определение 3. Пусть T – некоторая последовательность. Будем говорить, что последовательность T' является *максимальной для последовательности* T по позиции l, если $\mathbf{h}_T[l'] = 0\ \forall l' < l$ и $\mathbf{h}_T[l'] = = \mathbf{h}_{T'}[l]\ \forall l' > l$.

Пусть максимум функционала F на некоторой последовательности T_0 достигается в точке $(j_0, j_0 + 1)$, являющейся точкой максимума левого типа по классу K_1. Тогда, согласно лемме 1, объект s_0 такой, что $i_1(s_0) = j_0 + 1$, принадлежит K_2. Пусть $i_2(s_0) = l + 1$. Перенос объектов класса K_1 из позиции с номером l' в позицию с номером l'', где $l', l'' \le l$ и $l' < l''$, будем называть переносом в интервале $(0, l)$ слева направо.

Теорема 3. *Пусть последовательность* T_0' *может быть получена из* T_0 *путем ряда переносов слева направо в интервале* $(0, l)$. *Тогда реализуется только один из двух вариантов*:

а) *точка* $(j_0, j_0 + 1)$ *является точкой максимума левого типа по классу* K_1 *на последовательности* T_0';

б) *по крайней мере одна из точек из множества* $\{(j_0'', j_0'' + 1)|$ *объект с номером* j_0'' *принадлежит классу* K_2 *и* $j_0'' < j_0 - 1\}$ *является точкой максимума правого типа со значением функционала* F, *превышающим значение, достигнутое в точке* $(j_0, j_0 + 1)$ *на последовательностях* T_0 *и* T_0'.

Доказательство. Пусть $j_0' \ge j_0$ и $(j_0', j_0' + 1)$ – потенциальная точка максимума для последовательности T_0'. Поскольку переносы слева направо в интервале $(0, l)$ не могут привести к изменениям долей объектов классов K_1 и K_2 в группах, задаваемых граничной точкой $(j_0', j_0' + 1)$, то

значение функционала F в данной точке также не меняется. Следовательно, значение функционала F в точке $(j_0, j_0 + 1)$ на последовательности T_0' остается равным максимальному значению f_0 функционала F на последовательности T_0, достигнутому в точке $(j_0, j_0 + 1)$, а значение функционала F на последовательности T_0' в произвольной точке $(j_0', j_0' + 1)$ при $j_0' > j_0$ не превышает f_0.

Пусть $l_0 = i_2(j_0 + 1)$. Тогда, согласно лемме 1, потенциальная граничная точка левого типа на последовательности T_0' может быть представлена в виде $(i_2^-(l) - 1, i_2^-(l))$, где $l < l_0$. Пусть G_l^- – группа объектов T_0', имеющих номер в нумерации i_1, меньше $i_2^-(l)$, а $G_l^+ = T_0'/G_l^-$. Пусть также G_{0l}^- – группа объектов T_0, имеющих номер в нумерации i_1 меньше $i_2^-(l)$, а $G_{0l}^+ = T_0'/G_l^-$. Из условия теоремы следует, что группа G_l^+, G_l^- получены путем переноса объектов из класса K_1 из группы G_{0l}^- в группу G_{0l}^+. Предположим, что группа G_l^- является предпочтительной по классу K_1. Тогда, согласно лемме 3, значение функционала F в точке $(i_2^-(l) - 1, i_2^-(l))$ на последовательности T_0' меньше значения функционала F в этой же точке на последовательности T_0 и не превышает f_0.

Таким образом, оказывается, что для последовательности T_0' не существует точки максимума левого типа со значением функционала F, превышающим f_0, а точка максимума правого типа со значением функционала F, превышающим f_0, должна принадлежать (если она существует) к множеству граничных точек $\{(j_0'', j_0'' + 1) \mid$ объект с номером j_0'' принадлежит классу K_2 и $j_0'' < j_0 - 1\}$. Теорема доказана.

Полученные теоретические результаты позволяют предложить алгоритм, вычисляющий распределение значений функционала F при гипотезе о равновероятности всех последовательностей, которые могут быть получены с помощью перестановок объектов. На каждом шаге алгоритм генерирует последовательности в порядке их возрастания.

Очевидно, что искомое распределение оптимальных значений функционала может быть найдено, если это оптимальное значение вычислять на каждом шаге. Однако использование теоремы 2 позволяет существенно сократить объемы вычислений.

Действительно, пусть точка $(j_0, j_0 + 1)$ является точкой максимума левого типа по классу K_1 на последовательности T_0 с оптимальным значением функционала $F(T_0) = f_0$. Пусть $l_0 = i_2(j_0 + 1)$. Через $T_{l_0}^m$ обозначим последовательность, максимальную для последовательности T_0 по позиции l_0. Для произвольной последовательности T, удовлетворяющей следующему набору условий (точка $(j_0, j_0 + 1)$ является точкой максимума левого типа со значением функционала $F(T_0) = f_0$):

а) $T_0 < T < T_{l_0}^m$ или $T = T_{l_0}^m$;

б) последовательность T может быть получена из последовательности T_0 путем ряда переносов слева направо в интервале $(0, l_0)$;

в) пусть $0 < l' < l_0$ и $j' = i_2^-(l')$, имеем $F(T, j') \le f_0$.

Поскольку последовательности генерируются в порядке возрастания их векторов позиций, то для всех последовательностей от T_0 до $T_{l_0}^m$ не требуется поиска оптимальных значений функционала F в случае, если выполнены условия б) и в).

Проверка условия в) достаточно очевидна. Для каждого числа $0 < l' < l_0$ можно заранее найти суммарное число $\chi(l')$ объектов класса K_1, которое должно быть перенесено из позиций последовательности T_0 с номером из интервала $0 < l < l'$ в позиции с номером l, удовлетворяющим неравенствам $l' < l \le l_0$, для того чтобы было нарушено неравенство $F(T, j') \le f_0$. Для произвольной последовательности T, удовлетворяющей условию а), для того чтобы убедится в выполнении условия в), необходимо и достаточно, чтобы для произвольного числа из интервала $0 < l' < l_0$ было выполнено неравенство

$$\sum_{i = l' + 1}^{l_0} (\mathbf{h}_T[i] - \mathbf{h}_{T_0}[i]) \ge \chi(l').$$

Для проверки условия б) может быть использована следующая

Теорема 4. *Последовательность T может быть получена из последовательности T_0 путем ряда переносов слева направо в интервале $(0, l_0)$ в том и только том случае, если $\forall l' \in \{1, 2, \ldots, l_0 - 1\}$ выполнено неравенство*

$$\sum_{i = l' + 1}^{l_0} (\mathbf{h}_T[i] - \mathbf{h}_{T_0}[i]) \geq \mathbf{h}_{T_0}[l'] - \mathbf{h}_T[l'].$$

Доказательство. *Необходимость*. Предположим, что существует такой набор переносов объектов класса K_1 из позиций с меньшим номером в позиции с бóльшим номером, что в результате из последовательности T_0 получается последовательность T. Пусть $l' \in \{1, 2, \ldots, l - 1\}$. Обозначим через $N_{l'}$, число объектов K_1, перенесенных из позиций с номером, меньшим l', в позиции с номерами, не меньшими l'. Поскольку переносы возможны только слева направо (из позиции с меньшим номером в позицию с бóльшим номером), то $\forall l' \in \{1, 2, \ldots, l_0 - 1\}$ справедливо неравенство $N_{l'} \geq 0$. Очевидно, что

$$N_{l'} = \sum_{i = l'}^{l_0} (\mathbf{h}_T[i] - \mathbf{h}_{T_0}[i]) = \sum_{i = l' + 1}^{l_0} (\mathbf{h}_T[i] - \mathbf{h}_{T_0}[i]) + \mathbf{h}_T[l'] - \mathbf{h}_{T_0}[l']$$

и

$$N_{l' + 1} = \sum_{i = l' + 1}^{l_0} (\mathbf{h}_T[i] - \mathbf{h}_{T_0}[i]).$$

Поскольку $N_{l'}, N_{l' + 1} \geq 0$, то

$$\sum_{i = l' + 1}^{l_0} (\mathbf{h}_T[i] - \mathbf{h}_{T_0}[i]) \geq \mathbf{h}_T[l] - \mathbf{h}_{T_0}[l].$$

Таблица

Оценки	K_1	K_2	F									
			≤1.0	≤2.0	≤3.0	≤4.0	≤5.0	≤7.0	≤9.0	≤11.0	≤13.0	≤15.0
Точная	13	13	0.0	0.114	0.4441	0.6881	0.8431	0.9419	0.9817	0.9951	0.9989	0.9995
Монте-Карло	13	13	0.0	0.108	0.431	0.682	0.836	0.937	0.980	0.996	0.999	0.999
Точная	15	15	0.003	0.164	0.3978	0.63	0.7747	0.9335	0.9795	0.9912	0.9982	0.9997
Монте-Карло	15	15	0.0	0.082	0.378	0.661	0.811	0.921	0.976	0.996	0.998	0.999
Точная	16	16	0.0	0.1	0.3445	0.6168	0.7711	0.9156	0.9771	0.9927	0.9979	0.9992
Монте-Карло	16	16	0.0	0.082	0.363	0.650	0.809	0.935	0.983	0.996	0.999	0.999
Точная	27	9	0.0075	0.162	0.301	0.6419	0.7222	0.9102	0.9422	0.9853	0.9908	0.9984
Монте-Карло	27	9	0.007	0.140	0.262	0.660	0.734	0.914	0.945	0.985	0.991	0.999
Точная	24	8	0.009	0.173	0.3417	0.6407	0.7312	0.9177	0.9491	0.9878	0.9916	0.9985
Монте-Карло	24	8	0.007	0.153	0.305	0.662	0.757	0.926	0.953	0.991	0.994	0.999
Точная	20	10	0.0031	0.119	0.3848	0.5852	0.7892	0.9339	0.9681	0.9884	0.9973	0.9995
Монте-Карло	20	10	0.002	0.101	0.451	0.597	0.790	0.933	0.967	0.988	0.996	0.999
Точная	22	11	0.0027	0.103	0.401	0.5689	0.7742	0.9253	0.9681	0.9911	0.997	0.9993
Монте-Карло	22	11	0.002	0.095	0.425	0.583	0.783	0.926	0.970	0.992	0.997	1.000
Точная	30	6	0.016	0.208	0.4121	0.5499	0.619	0.9779	0.9198	0.9681	0.9889	0.991
Монте-Карло	30	6	0.013	0.175	0.359	0.483	0.549	0.880	0.920	0.969	0.991	0.992
X^2 (1 с.св)	–	–	0.683	0.843	0.9167	0.9545	0.9746	0.9918	0.9973	0.9991	0.9997	0.9999

Достаточность. Поскольку $\forall l' \in \{1, 2, \ldots, l_0 - 1\}$ справедливо неравенство $N_{l'} \geq 0$, то может быть предложен следующий набор переносов слева направо, переводящих последовательность T_0 в последовательность T: на первом шаге N_1 объектов переносится из позиции с номером 0 в позицию с номером 1, далее N_2 объектов переносится из позиции с номером 1 в позицию с номером 2, процесс продолжается до тех пор, пока N_{l_0} объектов не будет перенесено из позиции $l_0 - 1$ в позицию L_0.

6. ЧИСЛЕННЫЕ ЭКСПЕРИМЕНТЫ И ИХ ОБСУЖДЕНИЕ

Алгоритм расчета кривых распределения максимумов функционала F был реализован программно. Результаты расчетов приведены в таблице. Для сравнения там приведены также оценки Монте-Карло кривых распределения, рассчитанные с использованием 3000 случайных перестановок.

Можно отметить близость кривых, рассчитанных этими двумя способами. Особенно близки кривые на наиболее важном с практической точки зрения участке с оптимальными значениями функционала, превышающими 5, что подтверждает абсолютную правомерность использования оценок Монте-Карло. Сравнение с последней строкой позволяет количественно оценить степень завышения статистической значимости выявленных различий при вычислении статистики критерия χ^2 по той же самой выборке, по которой была рассчитана оптимальная граница.

СПИСОК ЛИТЕРАТУРЫ

1. *Sen'ko O.V., Kuznetsova A.V.* The use of partitions constuctions for stochastic dependencies approximation // Proc. Internat. Conf. Systems and Signals in Intelligent Technol. 1998, Minsk (Belarus). P. 291–297.
2. *Kuznetsova A.V., Sen'ko O.V., Zabotina et al.* The prognosis of survivance in solid tumor patients based on optimal partitions of immunological parameters ranges // J. Theor. Med. 2000. V. 2. P. 317–327.
3. *Sen'ko O.V., Kuznetsova A.V., Echin A.* The method of data analysis dased on partitioning // Proc. Comput. Statistics. Short Communs and Posters. COMPSTAT, 2000. P. 259–260.
4. *Kaplan E.L., Meier P.* Nonparametric estimation from incomplete observations // J. Amer. Stat. Assoc. 1958. V. 53. P. 457–481.
5. *Хардле В.* Прикладная непараметрическая регрессия. М.: Мир, 1993.
6. *Chou P.* Optimal partitioning for classification and regression trees // IEEE Trans. Pattern Analys. and Mach. Intelligente. 1991. V. 13. P. 340–354.