ОБОЗРЕНИЕ
ПРИКЛАДНОЙ И ПРОМЫШЛЕННОЙ
Том 3 МАТЕМАТИКИ Выпуск 1
1996

ГРУШО А. А.

СТАТИСТИЧЕСКИЕ КРИТЕРИИ ЗНАЧИМОСТИ ДЛЯ КЛАСТЕРНЫХ СТРУКТУР, ОСНОВАННЫЕ НА ПОПАРНЫХ МЕРАХ БЛИЗОСТИ

1. Введение. Статистический вывод обычно применяется в следующей форме. Возможно некоторое решение и есть статистическая процедура, дающая основание для этого решения. Обоснованием того, насколько можно доверять статистической процедуре, занимается классическая математическая статистика, «идеология» которой в том, чтобы минимизировать риск путем увеличения количества наблюдений и выбором оптимального решающего правила.

Несколько иную задачу перед статистиком ставит «разведочный анализ» в терминологии Тьюки [5], при котором из данных надо получить информацию об их «типичной» структуре (сформировать «образ»), а на фоне этих данных выделить такие, которые не являются «типичными» (будем называть их «выделяющимися») и поэтому несут новую информацию исследователю. Иногда можно предположить модель для описания «типичных» наблюдений. Тогда выявление данных, которые не являются «типичными», возможно путем оценки вероятности их появления в рамках модели. Если эта вероятность мала, рассматриваемая часть данных относится к «выделяющимся» наблюдениям. Получается критерий значимости, о качестве которого можно говорить лишь в той мере, в какой дальнейшие исследования подтвердят правильность выбора «выделяющихся» данных.

Если рассматривать кластерные методы как некоторую форму статистического вывода, то они могут быть отнесены скорее к «разведочному анализу». Построим вероятностно-статистическую модель, связывающую идеи Тьюки и некоторые методы кластерного анализа.

Тьюки предлагает в полученных данных определить «среднее» и «близкое к среднему», а затем исключить все наблюдения, которые можно классифицировать как «близкие к среднему». Новую информацию могут дать только наблюдения, выходящие за границы «близкого к среднему». По мнению автора, точка зрения Тьюки удачно отражает механизм поиска закономерностей в наблюдавшихся данных, которые позволяют в дальнейшем строить адекватные модели, получающие подтверждение классическими методами математической статистики. Од-

нако сглаживающие процедуры построения «среднего» и определение «выделяющихся» наблюдений, в основном, разработаны для числовых величин в векторных пространствах малой размерности. Для объектов нечисловой природы применяется метод сведения к числовым параметрам при помощи частот встречаемости и гистограмм. Однако эти методы основаны на устойчивости частот и требуют большого числа испытаний, особенно в пространствах большой размерности. В то же время определение в «сложных» пространствах меры близости элементов и использование кластерных методов позволяют расширить число случаев, когда возможно различать «средние» и «выделяющиеся» наблюдения.

Пусть $(\Omega, \mathfrak{A})$ — измеримое пространство. На пространстве элементарных исходов определена мера близости $\rho(X, Y)$, $X, Y \in \Omega$ (например, метрика), которая определяет на Ω топологию τ. Пусть также $\mathfrak{A}$ является σ-алгеброй борелевских множеств относительно τ, т. е. это — минимальная σ-алгебра, порожденная множествами из τ. Обозначим $R_x(t)$ открытый шар радиуса t с центром в $x \in \Omega$. Каждый открытый шар лежит в Ω, и поэтому все открытые шары — измеримые множества. Пусть $\mathbf{P}_0$ есть σ-конечная мера на $(\Omega, \mathfrak{A})$ и $\mathbf{P}_1$ — вероятностная мера на $(\Omega, \mathfrak{A})$.

О п р е д е л е н и е. Точка $x \in \Omega$ лежит в информативном множестве меры $\mathbf{P}_1$ относительно меры $\mathbf{P}_0$, если $\exists t_0 \mid \forall t < t_0$

$$\mathbf{P}_1(R_x(t)) > \mathbf{P}_0(R_x(t)).$$

Если меры $\mathbf{P}_1$ и $\mathbf{P}_0$ имеют непрерывные плотности p_1 и p_0, соответственно, относительно некоторой σ-конечной меры μ, то информативное множество можно определить как множество точек из Ω, для которых $p_1 > p_0$.

Если пространство Ω дискретно и $\mathfrak{A}$ — множество всех подмножеств Ω, то каждая точка является открытым шаром и информативное множество A определяется следующим образом:

$$A = \{x \in \Omega\colon \mathbf{P}_1(x) > \mathbf{P}_0(x)\}.$$

Если Ω — пространство вещественных чисел и $p_0 = \text{const}$, то информативные множества соответствуют событиям, на которых наблюдается превышение плотностью данного уровня (см., например, [9]). Если $\mathbf{P}_0$ — вероятностная мера, то информативное множество является объектом, встречавшимся ранее в задачах определения оптимального статистического решения при двух альтернативах [3]. Если вероятностная мера $\mathbf{P}_0$ отвечает «средним» наблюдениям, а «выделяющиеся» наблюдения получаются в соответствии с другой вероятностной мерой $\mathbf{P}_1$, то это можно интерпретировать в терминах «разведочного анализа» следующим образом. Задача сглаживания (нахождения «среднего» и «близких

к среднему» значений) отвечает оценке распределения $\mathbf{P}_0$ и определению наблюдений, согласующихся с этим распределением. Задача выявления «выделяющихся» значений в данных соответствует нахождению таких элементов, получение которых в мере $\mathbf{P}_0$ маловероятно даже с учетом возможности больших уклонений в большом массиве данных. Расположение таких «выделяющихся» наблюдений возможно только в информативном множестве, а возможность их появления определяется преобладанием вероятностной массы $\mathbf{P}_1$ относительно $\mathbf{P}_0$ на информативном множестве. Как отмечалось выше, оценки меры $\mathbf{P}_0$ можно заменить некоторой (возможно, грубой) теоретической моделью. Тогда появление маловероятных данных интерпретируется как вкрапление, полученное в соответствии с вероятностной мерой $\mathbf{P}_1$, а значения данных указывают местоположение информативного множества (так как только там располагаются выделяющиеся наблюдения). Здесь не требуются проверки соответствия выбранной модели $\mathbf{P}_0$ или оптимальной статистической процедуры, потому что будущий анализ данных подтвердит или опровергнет правильность выбора «выделяющихся» данных.

В случае, когда $\mathbf{P}_1$ и $\mathbf{P}_0$ — вероятностные меры на вещественной прямой, для определения «средних» и «близких к среднему» значений вводятся такие специальные числовые характеристики как среднее, медиана, дисперсия и др. В произвольном пространстве с метрикой (а тем более с произвольной мерой близости) вместо указанных характеристик естественнее определять характеристики нетипичных значений, например, через конфигурации, связанные с так называемым пороговым графом G_t. В пороговом графе G_t вершины соответствуют наблюдаемым данным $(X_1, \ldots, X_n)$, а ребра — это такие пары точек (X_i, X_j), что $\rho(X_i, X_j) < t$ (t — число, называемое порогом). Приведем примеры типичных конфигураций, маловероятное появление которых в пороговом графе (т. е. слабое соответствие с моделью типичных данных $\mathbf{P}_0$) указывает на принадлежность точек из такой конфигурации к информативному множеству.

1. Наличие компоненты связности в пороговом графе G_t размера $\geqslant k$. В этом случае «выделяющиеся» значения — точки этой компоненты.

2. В пороговом графе G_t есть дерево объема k. Вершины этого дерева — «выделяющиеся» значения.

3. Пороговый граф G_t содержит клику (полный подграф) объема k. Тогда вершины этой клики — «выделяющиеся» значения.

4. Функция $f_{k_1}(a)$ ($a \subseteq \{1, \ldots, n\}$, $k_1 \neq k$) называется плотностью ребер на k-элементных подмножествах в пороговом графе G_t, если $f_k(a)$ равна числу ребер в подграфе порогового графа G_t, который получается ограничением G_t на вершины $X_{i_1}, \ldots, X_{i_k}, \{i_1, \ldots, i_k\} = a$. Если максимум плотности ребер в пороговом графе превосходит некоторую границу T, то вершины соответствующего подмножества — «выделяю-

щиеся» значения [1].

В любом случае множество «выделяющихся» данных определяется тем, что некоторое событие, связанное с ними, маловероятно в мере $\mathbf{P}_0$ и весьма вероятно в мере $\mathbf{P}_1$. Проблема состоит в адекватном определении таких событий. Разумеется, мы не знаем свойства меры $\mathbf{P}_1$, но можем предполагать известной меру $\mathbf{P}_0$. Тогда искомые события должны быть маловероятными в $\mathbf{P}_0$, и в этом случае мы смотрим, выполняется ли в данных одно из выбранных таким образом событий, оцениваем вероятность осуществления этого события в мере $\mathbf{P}_0$ и утверждаем, что данные принадлежат информативному множеству, если соответствующая вероятность достаточно мала. Отсюда возникают задачи оценки распределений различных характеристик случайных пороговых графов в различных пространствах с определенной на них мерой близости. Примеры решения таких задач приведены в следующем разделе.