УДК 004.93

ОЦЕНКА ЧИСЛА ТЕСТОВ ДЛЯ БИНАРНЫХ ТАБЛИЦ

Переяславский Виталий Иванович, к.ф.-м.н., доцент, и.о. зав. кафедры математики и естественно-научных дисциплин

Государственное бюджетное образовательное учреждение высшего профессионального образования Московской области Финансово-технологическая академия, г. Королев Московской области.

*Изучаются вопросы, связанные с возможностью линейного разделения двух подмножеств вершин многомерного единичного куба. Объектом исследования являются пары прямоугольных таблиц, состоящих из чисел нуль и один. Строки таблиц представляют собой координаты двух непересекающихся подмножеств вершин многомерного единичного куба. Приводится понятие информационного веса признака для таких таблиц. Рассматривается линейный алгоритм, разделяющий вершины многомерного единичного куба на два подмножества. Изучаются характеристики этого линейного алгоритма. Даются оценки числа тестов, показывающие вычислительную сложность алгоритма.*

Ключевые слова: дихотомия, многомерный куб, разделение.

A ESTIMATE OF NUMBER OF TESTS FOR BINARY TABLES

Pereyaslavskiy V.I.

State educational institution of higher professional education of Moscow region Financial and technological Academy, Korolev, Moscow region.

*A problem associated with a possibility of a linear separation of two subsets of vertices of a multidimensional singular cube is studied. A subject of research is a pair of rectangular tables with numbers zero and one. Table rows represent coordinates of two disjoint subsets of vertices of the multidimensional singular cube. A concept of an information weight of a feature for such tables is proposed. A linear algorithm dividing vertices of the multidimensional singular cube into two subsets is considered. Then we study characteristics of the linear algorithm.*

*Keywords: dichotomy, multidimensional cube, separation.*

В задачах распознавания образов и кластеризации обучающим материалом часто служат точки в многомерном пространстве. Если признаки бинарные, например, принимающие значения 0 и 1, то эти точки располагаются в вершинах многомерного единичного куба. Каждая гиперплоскость разделяет эти вершины на два подмножества. И наоборот, если заданы два непересекающихся подмножества вершин многомерного единичного куба, то возникают задачи о том, являются ли эти два подмножества вершин линейно разделимыми, а если они линейно разделимы, то разделяет ли их конкретный алгоритм.

Пусть заданы две прямоугольные таблицы и , состоящие из столбцов и и строк соответственно, элементами которых являются числа нуль и единица.

Будем рассматривать такие пары таблиц и , для которых не существует строки, содержащейся в и одновременно.

Пусть - набор столбцов. Обозначим через (ν), = 1,2, таблицу, полученную выбрасыванием всех столбцов таблицы , кроме столбцов из ν.

Назовем ν тестом, если не существует строки, содержащейся в (ν) и (ν) одновременно.

Будем говорить, что ν - тупиковый тест, если ν – тест и любой поднабор ν не является тестом.

Число столбцов в наборе ν назовем длиной набора.

Набор столбцов ν будем записывать в виде строки ( ,…, ) из нулей и единиц, такой, что = 1 тогда и только тогда, когда -ый столбец вошел в набор ν.

Пусть - множество всех всюду определенных, а - множество всех частично определенных функций алгебры логики от переменных. Задание пары таблиц и с условием, что нет строки, содержащейся в и одновременно, эквивалентно заданию функции из такой, что строки первой таблицы будут единицами функции, а строки второй – нулями.

Рассмотрим отображения φ: → и : → . Отображение φ: → (: → ) сопоставляет функции из функцию функции из такую, что = 1 тогда и только тогда, когда набор ( ,…, ) задает тупиковый тест (тест длины ) для пары таблиц, определяемой функцией .

Пусть задана функция из , не равная тождественному нулю.

Спектром функции назовем вектор , где

*= ,*

а отображение назовем спектральным.

Каждый из векторов и назовем весовым вектором для пары таблиц и , определяемой функцией , и обозначим через произвольный из них, если не указано особо, какой из векторов имеется в виду. Весом признака назовем величину в весовом векторе .

Пусть имеем два класса предметов или явлений и , и пусть элементы этих классов изучались по ряду признаков. Рассмотрим задачу отнесения объекта распознавания к одному из классов.

Предположим, что классы и содержат и элементов соответственно и изучались по признакам ( ,…, ) . Закодируем каждый элемент из классов строкой из нулей и единиц следующим образом: если элемент обладает признаком , то на ом месте в этой строке поставим единицу, в противном случае – нуль. Тогда классы и можно представить в виде двух таблиц и , строки которых соответствуют элементам классов и .

Задача отнесения объектов к одному из классов теперь может быть решена следующим образом. По классам и строится пара таблиц

и . Для и подсчитывается весовой вектор Для каждой строки вычисляется величина , называемая весом строки [1, стр.3]. Пусть , = 1,…, и , j = 1,…, - веса строк таблиц и , а - вес строки , описывающей распознаваемый объект.

Предположим, что ≥ .Отнесение строки к одной из таблиц, а значит, и элемента к одному из классов, будем проводить с помощью правил:

1. если ( + ) / 2, то отнесем к таблице (классу ),
2. если ( + ) / 2, то отнесем к таблице (классу ).

В предлагаемом в данной работе методе распознавания используется идея построения разделяющей гиперплоскости. При этом способ кодирования объектов меняется в ходе построения этой гиперплоскости. Кроме того, в предлагаемом алгоритме использована стохастическая процедура нахождения весов, что позволяет работать с таблицами большой размерности.

В работе [2, стр.8] был приведен стохастический алгоритм вычисления весового вектора, построенного по таблице тупиковых тестов. В данной статье рассматривается стохастический алгоритм нахождения весового вектора , определяемого отображением .

Пусть - устройство, которое выдает случайный набор признаков длины , причем появление любого набора (их всего ) равновероятно. Проведем с каждым набором проверку его на тестовость. Исключим из рассмотрения нетестовые наборы.

Получим устройство , реализующее случайные тесты длины . Можно сказать, что на очередном шаге реализует вектор ξ , где

=

Так как появление любого теста длины в устройстве равновероятно, то вероятность

= 1) = = ,

а математическое ожидание случайной величины равно .

Будем находить путем оценки . Реализуем раз случайную величину . По центральной предельной теореме получаем:

< α ) →

а так как , то

Здесь были использованы обозначения: – сумма реализаций случайной величины , = , .

Из этой формулы можно найти число испытаний , необходимое для получения с данной точностью β и достоверностью γ.

Например, если β = 0,01, γ = 0,99, то а если β = 0,03, γ = 0,097, то .

При выборе значения – длины тестов, используемых в алгоритме, можно опираться на оценку, приведенную ниже.

Теорема 1. Для пары таблиц (( длина тупикового теста не превышает .

Доказательство. Проведем доказательство индукцией. Основание индукции. Рассмотрим случай, когда одна из таблиц представляет собой строку. Пусть, например, , а длина тупикового теста равна . Рассмотрим таблицы (ν) и (ν) . (ν) - это строка (. Так как набор (0,1,…,1) длины не тест для таблиц (ν) и (ν) , то в таблице (ν) есть строка, отличающаяся от ( только в первом признаке из тупикового теста , т.е. в (ν) есть строка ( . Аналогично, в таблице (ν) обязана быть строка ( , откуда в таблице (ν), а значит, и в таблице должно быть не менее, чем строк. Получили, что , что и требовалось показать при .

Предположим, что для пар таблиц, у которых , теорема доказана, и пусть для существуют такие таблицы и , у которых длина тупикового теста не менее . Построим граф с вершинами – строками из - и строками из - . Пусть ν – тупиковый тест длины не менее, чем . Тогда ν \ {t} – не тест, т.е. существуют номера такие, что строки и совпадают на всех признаках набора , кроме –го признака. Соединим вершины и ребром в этом случае и пометим это ребро числом – номером признака . Если есть несколько пар ( ( различающихся только по –му признаку из набора ν , то выберем из них произвольно одну и соединим указанным образом ребром. Таким образом, получим граф, у которого ребер столько же, сколько признаков в тупиковом тесте ν. Доказательство проведем отдельно для трех случаев.

1. Есть вершина графа без ребер. Рассмотрим таблицы и , которые получаются из и выбрасыванием строки, не имеющей в графе ребер. Для этих таблиц , а ν продолжает быть тупиковым тестом. Его длина не меньше, чем , что противоречит предположению индукции.
2. Есть вершина графа с одним ребром. Из таблиц и исключим строку, которая соответствовала вершине с одним ребром, а также исключим столбец, номером которого было помечено это ребро. Получим таблицы и с числом строк и длиной тупикового теста не менее . И в этом случае имеем противоречие с предположением индукции.
3. Все вершины графа имеют не меньше двух ребер. Рассмотрим замкнутую цепь без замкнутых подцепей в этом графе. Все ребра в этой цепи имеют разные номера. Рассмотрим строки пар . Они различаются по -му признаку. С другой стороны по -му признаку совпадают. Это же можно сказать про любую пару ( и ( . Получаем, что совпадают по -му признаку. Значит, случай 3 встретиться не может. Теорема доказана.

Так как тупиковых тестов длины более чем *m*1 + *m*2 − 1 нет, то весовой вектор лучше вычислять на тестах длины *r*  ≤ *m*1 + *m*2 – 1.

Пусть *Nr* ─ число тестов длины *r*, а *Mr* – число нетупиковых тестов длины *r*. Рассмотрим зависимость между *Mr*+1 и *Nr*.

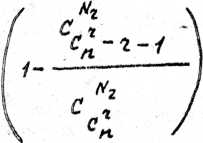
Построим граф с вершинами – тестами длины *r.* Соединим ребром два теста, если расстояние Хемминга между ними равно двум. Ребро между вершинами и пометим у вершины номером столбца, где у стоит нуль, а у стоит единица; у вершины – номером столбца, где у стоит нуль, а у стоит единица.

Рассмотрим тест *η* длины *r*. Добавлением одного признака к тесту *η* можно получить  *n* – *r* тестов длины *r* + 1. Тогда (*n* – *r*)⋅*Nr* – это число нетупиковых тестов длины *r* + 1, если не учитывать того, что при этом тесты длины *r* + 1 могут повторяться. Подсчитаем число повторений. Рассмотрим вершину , и пусть из нее выходит *pij* ребер, помеченных числом *j* у этой вершины. Объединим набор признаков с признаком *j*. Образуется нетупиковый тест, так как – тест, и легко проверить, что число тестов длины *r*, из которых его можно получить добавлением одного признака, равно *pij* + 1, т.е. в сумме (*n* – *r*)⋅*Nr* этот тест учитывался *pij* + 1 раз. Значит, число лишних повторений – *pij*. Так как из *pij* + 1 теста длины *r* можно получить добавлением одного признака этот нетупиковый тест, то вклад каждого теста в лишних повторениях равен . Просуммировав это выражение для вершины по всем *j* и далее по всем вершинам, получим следующую зависимость между *Nr*  и *Mr*+1.

Теорема 2.

где

Величины *Ai* можно назвать вкладом *i* -го теста длины *r* в повторениях тестов длины *r* + 1.

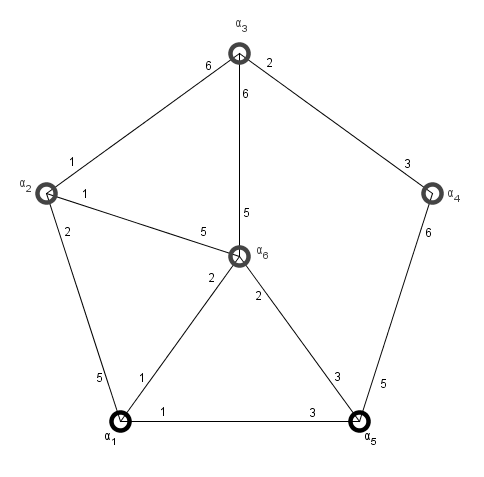
Приведем пример, иллюстрирующий эту теорему. Для пары таблиц 

г+1

 и 

тестами длины 4 будут следующие наборы признаков {2; 3; 4; 6}, {3; 4; 5; 6}, {1; 2; 4; 5}, {1; 2; 4; 6}, {1; 3; 4; 5}, {1; 3; 4; 6}.

Обозначим эти тесты через , , , , , и построим граф (рис.1).



**Рисунок 1 – Граф тестов**

Подсчитаем величины *Aj*:,

Тогда, учитывая, что *N*4 = 6, *n* – *r* = 6 – 4 = 2, из теоремы 2 получаем

.

Исследуем зависимость *Mr*+1 от *Nr* в среднем.

Теорема 3. Если и если появление любого набора из наборов длины *r*  в *Nr* тестовых наборах равновероятно, то



Доказательство. Воспользуемся предыдущей теоремой





Найдем . Для этого подсчитаем вероятность того, что *pij* = *k.*

Эта вероятность равна 

Поэтому 

Отсюда 

Пусть означает, что

Следствие 1. При *r = const, n→ ∞*, если , где *α* – постоянная, такая, что 0 ≤ *α ≤*1, имеет место следующее соотношение:



Это означает, что если рассмотреть *α* – ю часть всех наборов длины *r*, то в среднем нетупиковых тестов будет -я часть от общего числа наборов длины *r* + 1.

Доказательство. По теореме 3







Следствие 2. При *r → ∞, n→ ∞*, , где *α* – постоянная, такая, что 0 < *α ≤*1, если *r*< n – 2, то 

Доказательство. Аналогично доказательству следствия 1 имеем



Используя соотношения  и , так как



получим

,



Это означает, что при больших *r* и *n* из *α*~й части всех наборов длины *r* можно получить в среднем почти все наборы длины *r*+ 1 добавлением одного признака.

Полученные результаты можно использовать следующим образом.

Пусть нам известны числа *Nr* и *Nr+*1. Подсчитаем по *Nr* величину *М*(*Мr*+1). Если окажется, что *М*(*Мr*+1) < *Nr+*1, то это означает, что есть тупиковые тесты длины *r* + 1 и весовой вектор лучше вычислять на тестах длины большей чем *r.*

Библиография.

1. Дмитриев А.Н., Журавлев Ю.И., Кренделев Ф.П. О математических принципах классификации предметов и явлений // Дискретный анализ. Сборник трудов. – Вып.7. – Новосибирск: Наука, Сиб.отд., 1966. – С.3-15.
2. Кузнецов В.Е. Об одном стохастическом алгоритме вычисления информационных характеристик таблиц по методу тестов // Дискретный анализ. Сборник трудов. – Вып.23. – Новосибирск: Наука, Сиб.отд., 1973. – С.8-23.