УДК 621.396.6

БАЙЕСОВСКАЯ ПРОВЕРКА МНОГОАЛЬТЕРНАТИВНЫХ ГИПОТЕЗ С ИХ ПРЕДВАРИТЕЛЬНОЙ ИЕРАРХИЧЕСКОЙ СЕЛЕКЦИЕЙ

к.т.н. В.Е. Саваневич
(представил д.т.н., проф. Д.В. Голкин)

Вводится понятие и ставится задача оптимизации параметров байесовской процедуры проверки многоальтернативных гипотез с их предварительной иерархической селекцией. Показано, что существуют случаи, в которых способ реализации байесовского разбиения пространства наблюдений на области принятия решений, указанный в классической процедуре, не является наилучшим с точки зрения вычислительных затрат.

Введение. По мере накопления знаний и развития техники растет число различаемых объектов различной природы. При этом возрастает сложность процедур, систем классификации объектов по результатам проведенного эксперимента. Значительная часть статистических классификационных задач может быть сведена к задаче проверки многоальтернативных гипотез в соответствии с одним из критериев байесовской группы. В ряде случаев существенны вычислительные затраты на реализацию данных процедур. При этом существуют условия, при которых возможно сокращение указанных затрат на порядки без потерь в показателях качества принимаемых решений.

Анализ публикаций. Методика синтеза байесовских процедур проверки многоальтернативных гипотез (ПМАГ) известна [1]. В ее рамках находится наилучшее разбиение пространства наблюдений (ПН), а также указывается способ определения области ПН, которой принадлежит рассматриваемая выборка. Таким способом является, например, вычисление всех (по числу гипотез Q) вероятностей формирования рассматриваемой выборки при условии, что она соответствует конкретной гипотезе с последующим выбором гипотезы с максимальной условной вероятностью, минимальным апостериорным риском и т.д. При этом для принятия решения необходимо формирование совокупности статистик из Q или $Q-1$ элементов.

Наилучшему разбиению ПН соответствует наименьший средний риск, называемый байесовским риском. Это безусловный факт. Однако нет никаких доводов в пользу того, что указанный способ реализации процедур

определения области принадлежности выборки является наилучшим.

На практике часто результат эксперимента считается неделимым «квантом» данных. Однако существует возможность рассмотрения результата эксперимента как набора признаков, обоснованная хотя бы дискретной формой его представления. Каждый бит кода можно считать бинарным признаком. Законы распределения выборочных значений и выборки с учетом дискретной формы их представления могут иметь иерархическую форму записи [2].

Целью статьи является введение понятий классов гипотез и их кортежей, а также разработка формального похода к выбору процедур байесовского разбиения ПН с минимальными затратами.

Постановка задачи. Синтез решающего правила (РП) ПМАГ по критерию минимума среднего риска осуществляется при заданных функции правдоподобия, априорных вероятностях гипотез и матрице потерь. Выборка представляется в дискретной форме и содержит R параметров. Предполагается, что каким-либо образом можно упорядочить гипотезы, а в одну и туже точку ПН могут попасть выборки, соответствующие только совокупности соседних гипотез. Необходимо найти способ вычислительной реализации байесовского разбиения ПН с минимальными затратами.

Представление выборки. При использовании простой функции потерь и равных априорных вероятностях гипотез РП ПМАГ имеет вид

$$\hat{j} = \arg\max_{j} P_{y/\theta}(Y/\theta_j), \qquad (1)$$

где $P_{y/\theta}(Y/\theta_j)$ – вероятность получения выборки с параметрами Y при истинности j-й гипотезы с параметрами θ_j.

При использовании других функций потерь и априорных вероятностей гипотез вид РП (1) становится более громоздким, однако его суть и структура не меняются.

Пусть область определения каждого из R параметров выборки разделена на M_r ($r = \overline{1,R}$)интервалов с границами α_{ri} и α_{ri+1} ($i = \overline{0,M_r-1}$). Вводятся простые решающие функции (ПРФ):

$$\varphi_{ri} \equiv \begin{cases} 0 \text{ при } y_r < \alpha_{ri}, \\ 1 \text{ при } y_r \geq \alpha_{ri}, \end{cases},$$

где y_{kr} – значение r-го параметра выборки до дискретизации.

Результат каждой ПРФ можно считать признаком выборки. Совокупность значений всех ПРФ полностью характеризует дискретную выборку

$$P_{y/\theta}(Y/\theta_j) = \prod_{n=1}^{N_{ПРФ}} P_{ПРФ}(\varphi_n = i_n/\theta_j), \quad (2)$$

где $$Y = \{\varphi_1,\ldots,\varphi_n,\ldots,\varphi_{N_{ПРФ}}\}; \quad (3)$$

$N_{ПРФ}$, n – количество и номер ПРФ.

Описание (3) для решения задач ПМАГ часто является избыточным. С целью ликвидации данной избыточности предлагается следующее. Один из признаков $\varphi_{\tau 1}$ с номером $\tau_1 = \overline{1, N_{ПРФ}}$ выбирается в качестве первого используемого. Для каждого из возможных его значений выбирается второй признак $\varphi_{\tau 2}$ ($\varphi_{\tau 1}$) и т.д. При этом описание выборки Y (3) можно заменить описанием, построенным указанным выше способом:

$$H(Y_k) = \{h_1, \ldots, h_{n(Y_k)}\}, \quad (4)$$

где $h_1 = \varphi_{\tau 1}$;

$$h_2 = \begin{cases} \varphi_{\tau 21} & \text{при } \varphi_{\tau 1} = 0, \\ \varphi_{\tau 22} & \text{при } \varphi_{\tau 1} = 1, \end{cases} \quad \text{и т.д.}$$

На основе описания (4) может быть синтезировано РП байесовской процедуры ПМАГ:

$$\hat{j} = \arg\max_j P_{ПРФ}(H(Y)/\theta_j). \quad (5)$$

Введение понятия класса гипотез и его кортежа. С учетом положительной определенности вероятности, максимум выражения (5) будет больше нуля. При определенным образом выбранном описании выборки (4) согласно постановки задачи для большей части гипотез вероятность $P_{ПРФ}(h_1 = i_1/\theta_j)$ примет свои граничные значения 0 или 1. Если $P_{ПРФ}(h_1 = i_1/\theta_j) = 0$, то функция правдоподобия $P_{ПРФ}(H(Y)/\theta_j)$ при $h_1 = i_1$ тождественна нулю. Тем самым отпадает необходимость расчета значений еще не использованных признаков (4).

С использованием признака h_1 множество гипотез будет разделено на нулевой и первый классы с $P_{ПРФ}(h_1 = 0) = 1$ и $P_{ПРФ}(h_1 = 1) = 1$ соответственно. Совокупность объектов, для которых $P_{ПРФ}(h_1 = 0) \in]0, 1[$ разбивается следующим образом. Если истинное значение признака соответствует $h_1 = 0$, то гипотеза относится, во-первых, к нулевому классу, а во-вторых к кортежу первого класса. В противном случае гипотезу относят к первому классу и кортежу нулевого класса. Каждая гипотеза на любом уровне иерархии является элементом только одного класса и может принадлежать кортежам нескольких классов. Совокупность гипотез заданного класса и его кортежа можно назвать подмножеством гипотез. Введение признака h_2 приведет к появлению четырех новых подмно-

жеств гипотез, включающих гипотезы с одинаковыми значениями двух используемых признаков, а также кортежи к ним. Использование последующих признаков выборки осуществляется аналогичным образом. При этом на каждом уровне иерархии гипотезы только одного подмножества будут иметь отличную от нуля апостериорную вероятность принадлежности им выборки. В связи с этим введенные подмножества можно назвать подмножествами ранее неотвергнутых гипотез.

Структура байесовского РП ПМАГ с предварительной иерархической селекцией исследуемых гипотез. Рассматриваемое РП является иерархическим. В связи с этим его удобно представить в виде дерева. Висячие вершины графа раскрашены номерами принимаемых гипотез, а внутренние вершины – номерами используемых в них ПРФ. Дополнительно каждой внутренней вершине графа ставится в соответствие подмножество ранее неотвергнутых гипотез. При этом во внутренних вершинах, предшествующих висячим, всегда используются ПРФ байесовского типа. ПРФ байесовского типа является байесовским РП ПМАГ, синтезированным на основе подмножества гипотез соответствующей вершины графа.

Согласно введенному правилу в каждой внутренней вершине, начиная с корня, производится проверка признаков выборки, соответствующая его ПРФ. По результатам данной проверки выборка направляется в один из потомков внутренней вершины. После серии таких проверок выборка попадает в одну из внутренних вершин с ПРФ байесовского типа, а затем – в одну из висячих вершин, соответствующую принятой правилом гипотезе.

Оптимизация параметров РП ПМАГ с предварительной иерархической селекцией исследуемых гипотез. К ранее укомплектованным исходным данным добавляется множество ПРФ, допустимых к использованию, с указанием затрат на их применение. В качестве методов оптимизации правила (графа) иерархической селекции гипотез целесообразно использовать методы синтеза статистических алгоритмов минимальной сложности [3].

Пример. По результатам наблюдения ИСЗ наблюдательные средства контроля космического пространства за сутки формируют более 80000 измерений. Каждое измерение необходимо отнести к одному из более чем 4000 каталогизированных объектов [4]. Параметры измерений представляются в станционной системе координат (ССК). Параметры движения ИСЗ – в оскулирующих элементах орбиты (ОЭО). Для классификации каждого измерения необходимо последовательно спрогнозировать ОЭО каталогизированных ИСЗ на момент привязки измерения, пе-

ресчитать их в ССК, вычислить невязки между параметрами орбиты и измерения и определить их взвешенную сумму. Самоочевидна трудоемкость такой вычислительной схемы. Автором совместно с Е.В. Ветлугиным была разработана модельная программа классификации измерений с предварительной иерархической селекцией гипотез об их принадлежности. Оптимизация процедуры предварительной иерархической селекции производилась методом целенаправленного отбора вариантов с отбраковкой [2]. На первых двух уровнях иерархии проверялись параметры плоскости орбиты, получаемые путем пересчета параметров измерения в ОЭО. Область определения этих параметров была разбита на 10000 подобластей, а из каталогизированных ИСЗ было составлено 10^4 (10^5) подмножеств гипотез о принадлежности измерений объектам. Среднее число гипотез в непустых подмножествах составило 8,36 (7,18). При этом средняя трудоемкость разработанной процедуры (с учетом необходимости формирования подмножеств гипотез) по сравнению с использованием процедуры ПМАГ общего вида снизилась на два порядка.

Выводы. Введено РП ПМАГ с предварительной иерархической селекцией гипотез, которому всегда соответствует байесовское разбиение ПН на области. Доказывается это тем, что в байесовском РП не может быть принята за истинную гипотеза с нулевой апостериорной вероятностью. Кроме того, на принадлежность любой точки ПН той либо иной области принятия решения гипотезы с нулевой апостериорной вероятностью не влияют. Введенное РП ПМАГ в ряде случаев позволяет на порядки сократить вычислительные затраты. Следовательно, существуют случаи, в которых способ реализации байесовского разбиения ПН на области принятия решений, указанный в классической процедуре ПМАГ (разработанной без учета вычислительных затрат), не является наилучшим с точки зрения вычислительных затрат.

ЛИТЕРАТУРА

1. *Леман Э. Проверка статистических гипотез: Пер. с англ. / Под ред. Ю.В. Прохорова. – М.: Наука, 1979. – 408 с.*
2. *Саваневич В.Е. Постановка задачи синтеза алгоритмов минимальной сложности // Системи обробки інформації. – Х.: НАНУ, ПАНМ, ХВУ. – 2002. – Вип. 4 (20). – С. 67– 69.*
3. *Саваневич В.Е., Ветлугин Е.В. Введение и оптимизация параметров иерархической формы представления функции правдоподобия при классификации локационной информации // Системи обробки інформації. – Х.: НАНУ, ПАНМ, ХВУ. – 2002. – Вип. 1 (17). – С. 27 – 33.*
4. *Хуторовский З.Н. Ведение каталога космических объектов // Космические исследования. – 1993. – Вып. 4, т. 31.– С. 101 – 114.*

Поступила 12.08.2004

САВАНЕВИЧ Вадим Евгеньевич *– канд. техн. наук, доцент, докторант ХВУ. В 1986 году окончил Харьковское ВУРЭ. Область научных интересов – обработка локационной информации, информметрия.*