РАЗДЕЛ VII. ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

Лавров В.В., Немыкин О.И., Прохоров М.Е., Рындин Ю.Г.

Выбор классификатора при наличии обучающих данных ограниченного объёма

ПАО Межгосударственная акционерная корпорация «Вымпел»
(Россия, Москва)

doi: 10.18411/sr-10-12-2020-32

idsp: sciencerussia-10-12-2020-32

Аннотация

В рамках адаптивного байесова подхода предлагается метод выбора алгоритма распознавания (классификатора) при наличии обучающих данных ограниченного объёма, которые представляют собой совокупность отдельных непрерывных признаков распознавания, полученных при обработке исходных оптических изображений (ОИ), формируемых в наземных телескопах при наличии сложной фоновой обстановки, с целью выделения ОИ отдельных информационных объектов. В качестве таких объектов можно рассматривать движущиеся ИСЗ, а также их отдельные конструктивные элементы.

Для решения задачи выбора наилучшего классификатора из числа возможных с учётом низкой вероятности появления таких объектов на исходном изображении вводится характеристика распознавания и скалярный показатель качества (эффективность), которые используются при синтезе соответствующего алгоритма.

Ключевые слова: адаптивный байесов подход, классификатор, характеристика распознавания, эффективность.

Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (РФФИ) (проекты № 19-07-00225А и 19-07-00197А).

Постановка задачи поиска наилучшего классификатора и выбор предпочтительных характеристики и показателя качества (эффективности) распознавания

В работах [1, 2] показано, что в рамках адаптивного байесова подхода задачу совместной проверки M гипотез при аддитивной функции потерь можно свести к независимой проверке M двуальтернативных гипотез с использованием алгоритма формирования по каждой гипотезе нерандомизированного решающего правила

$$d_j(x, c_j) = \begin{cases} 1 & \text{при } p(j|x) > c_j, j = \overline{1, M} \\ 0 & \text{иначе} \end{cases} \quad (1)$$

и характеризовать качество каждого принятого решения с использованием известной в радиолокации характеристики обнаружения [1]. Здесь $d_j(x, c_j)$ – индикаторная функция при выбранном по критерию Неймана-Пирсона порогу c_j ; $p(j|x)$ – апостериорная вероятность j класса в точке признаков $x \in \Omega_x$, равная отношению плотности вероятности выбранного класса $p(x, j)$ к безусловной плотности вероятности получения значения x $p(x) = p(x, j) + (\sum_{i=1}^M p(x, i); i \neq j)$. Введём обозначения для плотностей вероятностей выделенного класса $p(x, 1)$ и альтернативы $p(x, 0)$, а для $d_j(x, c_j) – d(x, c)$.

Заметим, что в силу условий наблюдения, введенные нами плотности неизвестны, а объем обучающей выборки ограничен и не позволяет получить их асимптотически сходящиеся оценки. При этом вероятность наблюдения объектов выбранного класса $p_1(0) = \int_{\Omega_x} p(x, 1) d(x, 0)$ в общей выборке мала по сравнению с $p_0(0) = \int_{\Omega_x} p(x, 0) d(x, 0)$.

В таких условиях поставим задачу поиска наилучшего классификатора a^* из совокупности алгоритмов $a_i \in \Omega_A$ (необязательно байесовых), структура которых удовлетворяет (1).

С целью выбора функционала качества введем в рассмотрение характеристику распознавания $\gamma(p_{\pi}) = p_1(c(p_{\pi})) / (p_1(c(p_{\pi})) + p_0(c(p_{\pi})))$, которая представляет собой отношение условной вероятности принять правильное решение по выбранной гипотезе p_{π} к полной вероятности принятия положительного решения по выбранному классу. При этом p_{π} – условная вероятность правильного. Значение порога $c(p_{\pi})$ вычисляется численно с использованием интегрального уравнения

$$p_{\pi} = \int_{\Omega_x} p(x, 1) d(x, c(p_{\pi})) dx / p_1(0),$$

в структуру которого входят вероятностные характеристики выбранного класса и принимаемое конкретным алгоритмом решение.

На рисунке 1 закрашена область локализации возможных характеристик распознавания, ограниченная двумя отрезками прямых, параллельных оси абсцисс. Кривая 1 соответствует случаю безошибочной классификации объектов выбранного класса, а кривая 2 – случаю отнесения к выбранному классу всех объектов выборки (значение ординат совпадает с априорной вероятностью наличия в выборке объектов выбранного класса).

Рис. 1. Область локализации характеристик распознавания

Параметрами, определяющими поведение любой характеристики, являются структура классификатора и используемые признаки распознавания.

Важным свойством характеристики распознавания является то, что она монотонная невозрастающая функция. Докажем это для байесова решающего правила. Преобразуем выражение для $\gamma(p_{\pi})$ следующим образом

$$\gamma(p_{\pi}) = 1 / (1 + p_0(c(p_{\pi})) / p_1(c(p_{\pi})))$$

Для оптимальных байесовых классификаторов и классификаторов, синтезированных с использованием критерия Неймана-Пирсона, последнее слагаемое в знаменателе неубывающая функция параметра p_{π} . Отсюда следует вышеупомянутое свойство характеристики распознавания.

Из вышесказанного следует, что показателем качества распознавания разумно выбрать значение площади под характеристикой распознавания [3] $S = \int_0^1 \gamma(p_{\pi}) dp_{\pi}$. Условимся называть эту величину эффективностью.

Очевидно, что чем ближе S к единице, тем лучше качество распознавания. Данный показатель можно рассматривать в качестве аналога среднего риска для двальтернативной задачи, его значение не зависит от значения аддитивной функции потерь. В условиях априорной неопределенности при наличии единственной обучающей выборки не представляет труда получить эмпирическую оценку эффективности [3].

Из вышесказанного следует критерий выбора оптимального алгоритма распознавания a^* из совокупности проверяемых алгоритмов $a \in \Omega_A$ с одновременной оценкой их параметров:

$$a^* = \arg \max_{a \in \Omega_A} S(a)$$

при дополнительном требовании к поведению характеристики распознавания (монотонная невозрастающая функция) для алгоритма a^* . При ограниченном объеме обучающих данных мы имеем дело с оценками параметров анализируемых алгоритмов принятия решения. Поэтому требует внимания выбор информативного состава признаков. Введем следующее определение информативности совокупности признаков распознавания (ПР) размерности L по отношению к совокупности ПР размерности $L_1 < L$, при условии, что последняя содержится в первой и выбран байесов алгоритм принятия решения: $I(L | L_1) = S(L) - S(L_1)$. Введенное определение корректно, поскольку значение эффективности распознавания при отсутствии признаков на входе алгоритма классификации равно априорной вероятности появления объектов заданного класса.

Вопросы реализации алгоритма по эмпирическим данным ограниченного объема с использованием методов многомерного статистического анализа [4] рассмотрены в [3] и в данной работе не затрагиваются.

Заключение.

В работе рассмотрена задача поиска наилучшего классификатора и используемого состава признаков распознавания из числа доступных алгоритмов принятия решения в условиях ограниченного состава обучающей выборки и неполного вероятностного описания содержащейся в ней информации.

В рамках статистической проверки двувальтернативных гипотез предложены показатель качества и характеристика распознавания, использование которых позволяет найти оптимальный классификатор при наличии соответствующей совокупности алгоритмов формирования решения и процедуру его обучения по доступной обучающей выборке.

1. Тартаковский Г.П. Теория информационных систем. М.: Физматкнига, 2005.
2. Рындин Ю.Г., Тартаковский Г.П., Тюрин В.С. //Радиотехника и электроника, 2004, Т.49, вып. №9, стр. 1066.
3. Прохоров М.Е. Оценка качества распознавания при наличии единственной обучающей выборки ограниченного объема. //Радиотехника и электроника, 2007, Т.52, вып. №1, стр. 28-32.
4. Эфрон Б. Нетрадиционные методы многомерного статистического анализа. М.: Финансы и статистика, 1988.

Шульгин П.В., Муллин В.В., Ханмагомедов Р.Х.

Оценка параметров кодирования при использовании алгоритма Шеннона-Фано

*ФГКВОВУ ВО «Пермский военный институт войск национальной гвардии
Российской Федерации»
(Россия, Пермь)*

doi: 10.18411/sr-10-12-2020-33

idsp: sciencerussia-10-12-2020-33

Аннотация

Статья посвящена оценке параметров кодирования при использовании алгоритма Шеннона-Фано. Проведенные численные исследования показали, что применение данного алгоритма кодирования способствует уменьшению объема передаваемого сообщения, что позволяет сократить время передачи отправляемого сообщения.

Ключевые слова: Шеннон, кодирование, алгоритм, алфавит, вероятность, численное исследование.