
ИСПОЛЬЗОВАНИЕ ВЕРОЯТНОСТНЫХ ПРИЗНАКОВ И АЛГОРИТМОВ, ОСНОВАННЫХ НА ПРИНЦИПЕ ЧАСТИЧНОЙ ПРЕЦЕДЕНТНОСТИ, ДЛЯ ВЫЧИСЛЕНИЯ ЭФФЕКТИВНОСТИ ОБУЧЕНИЯ СИСТЕМЫ РАСПОЗНАВАНИЯ

В. А. Дурденко*, М. А. Ильичев**

* Воронежский институт инновационных систем

** Воронежский институт МВД РФ

Поступила в редакцию 10.03.2010 г.

Аннотация. Проанализированы основные особенности функционирования алгоритмов вычисления оценок и принятия решений по распознаванию. Отмечено существенное затруднение, возникающее при использовании алгоритмов вычисления оценок совместно с вещественнозначными вероятностными признаками. Изложена новая методика вычисления эффективности обучения системы распознавания образов при использовании вещественнозначных вероятностных признаков и алгоритмов принятия решений, основанных на принципе частичной прецедентности. Предложен способ классификации систем распознавания на основе вычисления коэффициента эффективности обучения системы.

Ключевые слова: алгоритм вычисления оценок, системы распознавания образов, вещественнозначные вероятностные признаки, плотность распределения, коэффициент эффективности обучения.

Abstract. Main features of an evaluation algorithm and a decision-making algorithm are analyzed. It has been noticed that there appear some complications when using evaluation algorithms with real-valued probabilistic characteristics. A new method which allows to evaluate the effectiveness of education of an object recognition system when it uses real-valued probabilistic characteristics and decision-making algorithms based on partial precedent principle has been suggested. A new approach to object recognition systems classification which is founded upon evaluation of the education of an object recognition system efficiency coefficient has been offered.

Key words: functional models, pattern recognition systems, process-functional models of pattern recognition systems, methodologies of systems projection.

ВВЕДЕНИЕ

В различных областях человеческой деятельности (экономике, финансах, медицине, бизнесе, геологии, химии, и др.) повседневно возникает необходимость решения задач анализа, прогноза и распознавания, выявления скрытых зависимостей и поддержки принятия оптимальных решений. Анализируя сложившиеся к настоящему времени подходы к решению задачи распознавания, можно отметить, что вне зависимости от предметной области, разработчики систем пользуются методом, который в общем случае предполагает следующие этапы решения указанной задачи.

Обосновывается алфавит признаков и последовательно формируется априорный и рабо-

чий алфавит признаков. Затем проводится классификация распознаваемых объектов в соответствии с задачами, решаемыми системой управления. Выбирается алгоритм распознавания, обеспечивающий отнесение распознаваемого объекта к тому или другому классу или их совокупности. Наконец разрабатываются специальные алгоритмы управления работой системы с тем, чтобы процесс функционирования системы распознавания (СР) был в определенном смысле оптимальным.

Указанный подход к решению проблемы создания систем распознавания в различных предметных областях наряду со своей явной очевидностью обладает и рядом весомых недостатков, которые проявляются не только при разработке, но и на этапе функционирования СР.

Данная работа посвящена изложению основ новой методики вычисления эффективности

© Дурденко В. А., Ильичев М. А., 2010

обучения системы распознавания образов при использовании вещественнозначных вероятностных признаков и алгоритмов принятия решений, основанных на принципе частичной прецедентности.

1. ПОСТРОЕНИЕ АЛГОРИТМОВ РАСПОЗНАВАНИЯ

Теоретические основы алгоритмов распознавания, основанных на принципе частичной прецедентности описаны в многочисленных научных публикациях [1—3]. К данным алгоритмам относятся такие алгоритмы, как тестовый, алгоритм распознавания с представительными наборами и, как более общий случай, алгоритм вычисления оценок (АВО). Принципиальная идея этих алгоритмов основана на сравнении распознаваемого объекта с эталонными по различным наборам признаков (частичных прецедентов), после чего применяется процедура голосования. Суть процедуры голосования состоит в том, что голоса, отданные за каждый класс в который входят объекты по признакам, суммируются и полученное значение нормируется по числу эталонов класса. Отнесение объекта к одному из классов осуществляется пороговым правилом на основе полученных оценок.

Недостатком данных алгоритмов является необходимость использования k -значных признаков или введения специальных функций близости (в алгоритмах АВО) для осуществления операций сравнения. Примером данной функции может являться следующая:

$$B_{\{X\}}(S_1, S_2) = \begin{cases} 1, & \sum_{i \in \{X\}} |X_i(S_1) - X_i(S_2)| \leq \varepsilon, \\ 0, & \text{иначе.} \end{cases}$$

где $X_i(S_1)$ и $X_i(S_2)$ — значения i -го признака для объектов S_1 и S_2 , ε — выбранное пороговое значение.

При использовании вещественнозначных вероятностных признаков обоснованный способ выбора величин ε при вычислении функций близости представляет существенные затруднения.

Акцентирование внимания на вещественнозначных вероятностных признаках связано с тем, что системы распознавания образов в подавляющем числе случаев применяются в плохо формализованных предметных областях,

объекты которых характеризуются непрерывным (бесконечным счетным) множеством своих реализаций, являющихся мгновенным отображением состояния объекта. Например, получаемые изображения лица определенного человека при распознавании будут иметь определенные отличия, связанные с углом регистрации, выражением, уровнем освещенности и т.п.

Объекты хорошо формализованных предметных областей имеют конечные счетные множества своих реализаций и, следовательно, признаки распознавания будут детерминированы. Однако, данная задача имеет достаточно тривиальное решение и не представляет особого интереса.

Любую предметную область S в рамках системы распознавания образов можно представить в виде множества классов $S_j, j = \overline{1, Y}$. Классы задаются объектами $s_{ij}, i = \overline{1, N_j}$, а объекты — своими дискретными реализациями $s_{ij}^l, l = \overline{1, L_{ij}}$. Множества дискретных реализаций классов на этапе обучения системы определяют обучающую выборку, т.е. $S_j^* = (S_1^*, \dots, S_Y^*)$ (рис. 1.).

Исходно каждая строка таблицы обучения в алгоритме АВО представляет собой реализацию объекта обучающей выборки, представленную множеством значений признаков.

Модифицируем таблицу обучения следующим образом. Объединим в рамках каждого класса все реализации обучающей выборки по признакам. В результате классы будут представлены множествами диапазонов изменения каждого из признаков по обучающей выборке.

Каждый из диапазонов представляет собой дискретный ряд распределения значений признака, полученный из обучающей выборки по классу. Пусть x_ξ — различные значения признака X_i , полученные после вычислений, а r_ξ — их число. Диапазон наблюдаемых в рамках класса S_j значений признака X_i состоит из $x_\xi - 1$ интервалов. Нормировав значения r_ξ относительно $\sum_{\xi} r_\xi$ — количества изображений класса, получим частоты $p_\xi^* | \sum_{\xi=1}^k p_\xi^* = 1$ (статистический ряд). Имея статистический ряд и воспользовавшись методами интерполяции легко получить эмпирическую функцию распределения $F_j^*(X_i)$, а также плотность распределения $f_j^*(X_i)$. Функция плотности распределения будет наглядно характеризовать характер распределения значений признаков внутри диапазонов по классам.

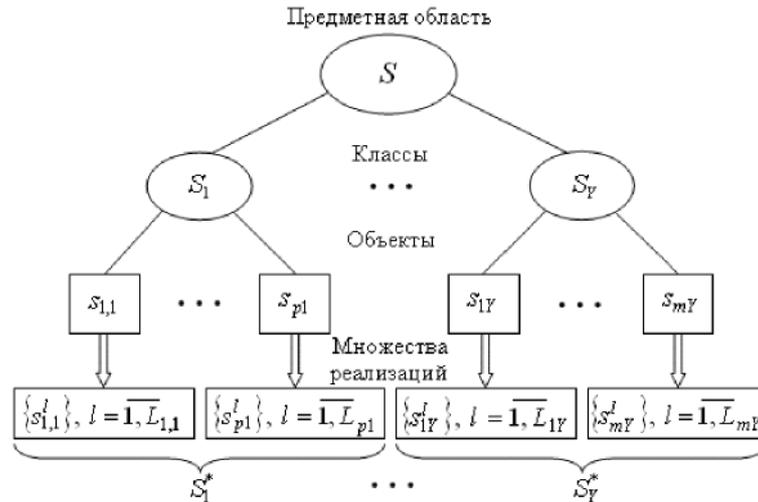


Рис. 1. Иерархическая модель представления предметной области применительно к задаче распознавания.

Тогда оценкой неизвестного объекта s по признаку X_i в модифицированном алгоритме АВО будет выражение

$$G_{X_i}^s = \frac{f_j^*(s)}{\max f_j^*(X_i)},$$

где $\max f_j^*(X_i)$ — максимальное значение эмпирической плотности распределения признака X_i .

2. ОЦЕНКА КОЭФФИЦИЕНТА ЭФФЕКТИВНОСТИ ОБУЧЕНИЯ СИСТЕМЫ РАСПОЗНАВАНИЯ

Рассмотрим особенности вычисления качества обучения системы распознавания, которая использует алгоритмы, основанные на принципе частичной прецедентности.

Пусть имеется один признак и распознается два класса. Будем считать, что плотности распределения значений признаков по классам не известны и по каждому классу даны только диапазоны возможных значений признака.

Пусть после обработки обучающей выборки диапазоны значений признака, соответствующие классам K_1 и K_2 пересекаются (рис. 2.).

При этом, если значение признака произвольного объекта попадает на участки $ab \in K_1$ или $cd \in K_2$, то принимаются однозначные решения (отдается голос за класс). На участке $bc \in K_1 \cap K_2$ решение не определено. Введем следующие обозначения $l_n = bc$, $X^{K_1} = ac$, $X^{K_2} = bd$. Тогда коэффициент эффективности обучения системы для случая, представленного на рис. 1, вычисляется следующим образом:

$$K_{\text{ЭО}}^{1,2} = 1 - \frac{l_n(X^{K_1} + X^{K_2})}{2X^{K_1}X^{K_2}}.$$

Если в системе имеется один признак и Y классов, то коэффициент эффективности обучения будет рассчитываться как:

$$K_{\text{ЭО}}^{i,Y} = 1 - \frac{1}{Y} \sum_{j=1}^Y \frac{l_{nj}}{X^{K_j}}, \quad (1)$$

где l_{nj} — область значений признака, где решение не определено.

Обобщая (1) для случая N признаков и Y классов получим следующее выражение:

$$K_{\text{ЭО}}^{N,Y} = 1 - \frac{1}{N} \sum_{i=1}^N \frac{1}{Y} \sum_{j=1}^Y \frac{l_{nj}}{X_i^{K_j}} = 1 - \frac{1}{NY} \sum_{i=1}^N \sum_{j=1}^Y \frac{l_{nj}}{X_i^{K_j}}.$$

При этом справедливо

$$K_{\text{ЭО}} = \begin{cases} 1, & X^{K_1} \cap X^{K_2} = \emptyset, \\ (0, \dots, 1), & X^{K_1} \cap X^{K_2} \neq \emptyset, \\ 0, & X^{K_1} \cap X^{K_2} = X^{K_1} = X^{K_2}. \end{cases}$$

Пусть имеются плотности распределения значений признаков для классов. Если системе объекты распознавания предъявляются последо-

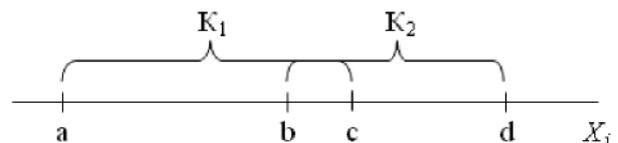


Рис. 2. Диапазоны значений признака для классов K_1 и K_2 .

вательно, то решения о появлении того или иного класса являются несовместными событиями.

Как и в предыдущем случае, ошибки при принятии системой распознавания могут возникать на участке bc (рис. 3). Здесь $f_1^*(X_i)$ и $f_2^*(X_i)$ — эмпирические плотности распределения значений признака по классам K_1 и K_2 . Вероятность принятия правильных решений на участках ab и cd равна единице. На участке bc

$$f_1^*(X_i) > f_2^*(X_i),$$

следовательно алгоритм принятия решений проголосует за класс K_1 . Однако, существует вероятность того, что на bc появится объект из K_2 , что приведет к ошибке распознавания.

Вероятность появления объекта класса K_1 на участке bc вычисляется следующим образом:

$$P_{bc}^{K_1} = \frac{\int_b^c f_1^*(X_i) dx}{\int_b^c f_1^*(X_i) dx + \int_b^c f_2^*(X_i) dx}.$$

Появление объектов из классов K_1 и K_2 есть несовместные события, которые образуют полную группу, поэтому $P_{bc}^{K_2} = 1 - P_{bc}^{K_1}$.

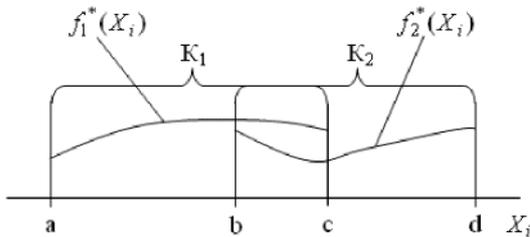


Рис. 3. Плотности распределения значений признака для классов K_1 и K_2 .

Введем обозначения: $l_{K_1} = ab$ — область значений признака $X \in K_1 \setminus K_2$; $l_{K_2} = cd$ — область значений признака $X \in K_2 \setminus K_1$; $l_{K_1 K_2} = bc$ — область значений признака $X \in K_1 \cap K_2$; $L_{K_1 K_2} = ad$ — область значений признака $X \in (K_1 \cap K_2) \cup (K_1 \setminus K_2) \cup (K_2 \setminus K_1)$; $P_n = \max(P_{bc}^{K_1}, P_{bc}^{K_2})$ — максимальная из вероятностей появления объектов классов K_1 и K_2 на участке bc .

Для ситуации, показанной на рис. 3, эффективность обучения будет вычисляться следующим образом

$$K_{\text{ЭО}}^{1,2} = \frac{l_{K_1} + P_n l_{K_1 K_2} + l_{K_2}}{L_{K_1 K_2}}.$$

Для случая одного признака и Y классов:

$$K_{\text{ЭО}}^{i,Y} = \frac{1}{Y} \sum_{j=1}^{Y-1} \frac{l_{K_j} + P_{nj} l_{K_j K_{j+1}} + l_{K_{j+1}}}{L_{K_j K_{j+1}}}. \quad (2)$$

Следует отметить, что величины l_{K_j} являются интервалами признака, соответствующие классам K_j , где вероятность принятия правильного решения равна единице.

Применение выражения (2) представляется удобным для оценки эффективности каждого из признаков, при обосновании рабочего алфавита. Введя пороговое значение $K_{\text{ЭО}}^{i,Y} \leq 1$ можно на этапе обучения системы исключить те признаки, которые ему не удовлетворяют. Таким образом, использование параметра $K_{\text{ЭО}}^{i,Y}$ позволяет гибко оперировать имеющимся множеством признаков при решении различных задач в различных предметных областях.

Распространив выражение (2) на случай N признаков получим:

$$K_{\text{ЭО}}^{N,Y} = \frac{1}{NY} \sum_{i=1}^N \sum_{j=1}^{Y-1} \frac{l_{K_j}^i + P_{nj}^i l_{K_j K_{j+1}}^i + l_{K_{j+1}}^i}{L_{K_j K_{j+1}}^i}. \quad (3)$$

Выражение (3) позволяет оценить коэффициент эффективности обучения системы распознавания использующей множество вещественнозначных вероятностных признаков и алгоритмы принятия решений, основанные на принципе частичной прецедентности.

В настоящее время существует достаточно большое количество предметных областей, в которых с успехом применяются алгоритмы вычисления оценок. Однако, опираясь на результаты исследований полученных в [3], представляется, что наибольший эффект от применения модифицированного алгоритма может быть достигнут при решении таких практических задач, как диагностика заболеваний, распознавание рукописных символов и отпечатков пальцев. Проведенная серия экспериментов с рукописными изображениями геометрических фигур (крест, треугольник, круг, квадрат, звезда), показала эффективность предлагаемого подхода.

Качество обучающей выборки и качество признаков определяют тип решаемой задачи распознавания.

По результатам вычисления $K_{\text{ЭО}}$ любую задачу распознавания можно условно классифицировать следующим образом:

- полностью определенная задача;
- частично определенная задача;
- неопределенная задача.

Полностью определенной является задача, если $K_{\text{эо}} = 1$. Обучающая выборка и качество признаков слабо определенной задачи по количественному или качественному составу могут быть недостаточными для описания всего множества классов. В этом случае классы обучающей выборки могут пересекаться и $0 \leq K_{\text{эо}} < 1$. Неопределенная задача характеризуется полным отсутствием обучающей выборки, непригодностью признаков, комбинацией указанных факторов. В любом случае для неопределенной задачи $K_{\text{эо}} = 0$. Естественно, что в плохо формализованных предметных областях по значению коэффициента эффективности обучения нельзя однозначно судить о качестве работы системы.

ЗАКЛЮЧЕНИЕ

Таким образом, в работе показаны особенности вычисления качества обучения системы

Дурденко Владимир Андреевич — д.т.н., профессор кафедры менеджмента, Воронежский институт инновационных систем. Тел. (4732) 354-898. E-mail: dva_viis@mail.ru

Ильичев Михаил Александрович — к.т.н., доцент кафедры технических систем безопасности Воронежского института МВД РФ. Тел. (4732) 312-412. E-mail: maip@rambler.ru

распознавания, которая использует алгоритмы, основанные на принципе частичной прецедентности.

Предложенный подход позволяет гибко оперировать имеющимся множеством признаков при решении задач распознавания в различных предметных областях, а также оценить коэффициент эффективности обучения системы распознавания, использующей множество вещественнозначных вероятностных признаков.

СПИСОК ЛИТЕРАТУРЫ

1. Дюкова Е. В. Асимптотически оптимальные тестовые алгоритмы в задачах распознавания. — Проблемы кибернетики. М.: Наука, 1982, вып. 39, С. 165—199.
2. Журавлев Ю. И. Избранные научные труды. — М.: Магистр, 1998, 420 с.
3. Журавлев Ю. И., Рязанов В. В., Сенько О. В. «РАСПОЗНАВАНИЕ». Математические методы. Программная система. Практические применения. — М.: ФАЗИС, 2006, 176 с.

Durdenko Vladimir A. — Doctor of Technical Sciences, Professor, Department of Management, Voronezh Institute of Innovation Systems. Tel. (4732) 354-898. E-mail: dva_viis@mail.ru

Pichiov Mikhail A. — Candidate of Technical Sciences, Associate Professor, Department of Security and Communication Technical System, Voronezh Institute of the Russia Ministers Interior. Tel. (4732) 312-412. E-mail: maip@rambler.ru