Известия Гомельского государственного университета
имени Ф. Скорины, №5(50), 2008

УДК 004.02:004.9

Об одном методе формирования пространства решений при построении систем распознавания образов

В. Г. Родченко

Введение

Изучая закономерности функционирования сложных систем и явлений реального мира, исследователи часто сталкиваются с отсутствием математического аппарата, позволяющего составлять уравнения, которые адекватно и с удовлетворительной степенью точности описывают соответствующий процесс. При этом возникает ситуация, когда удается реализовать измерение большого числа разнообразных показателей, так или иначе характеризующих этот процесс. Наблюдаемые данные в скрытой форме представляют закономерности функционирования сложной системы, а потому эффективными в данном случае могут оказаться подходы, основанные на использовании методов прикладной статистики, в частности, аппарата математической теории распознавания образов [1, 2]. Действительно, измеряя (наблюдая) значения разнообразных показателей при различных состояниях сложной системы удается построить исходную классифицированную выборку, которая в дальнейшем используется в качестве обучающей выборки. Проводя компаративный анализ данных различных состояний сложной системы, можно формировать такую подсистему признаков, которая будет обеспечивать разделение в многомерном признаковом пространстве соответствующих эталонов этих состояний.

Задача, которую изначально приходится решать, связана, во-первых, с формированием исходного словаря наблюдаемых признаков, образующих пространство наблюдений, и, во-вторых, с реализацией процедуры преобразования этого словаря в априорный словарь признаков (построение пространства признаков) [3, 4]. Многочисленный опыт построения систем распознавания (диагностирования) образов свидетельствует о том, что исследователи стремятся найти решение в пространстве таких признаков, которые относительно легко измеряются и априори обеспечивают достаточно четкое разделение эталонов классов в соответствующем признаковом пространстве [5, 6]. Гипотеза о том, что практически в любом априорном признаков пространстве, сформированном на основе экспертных оценок специалистов, удается построить четко выраженные эталоны классов, не выдерживает критики.

Традиционно алгоритм функционирования системы распознавания рассматривается как последовательное выполнение двух процедур: *процедуры обучения* и *процедуры принятия решения*. Следует четко понимать, что результатом первой процедуры часто является заключение о невозможности построения эталонов классов на основе исходного априорного словаря признаков. Либо может быть получено заключение о необходимости сепарирования исходных признаков по степени их информативности и исключению малоинформативных признаков из дальнейших исследований [7]. Только качественное проведение процедуры обучения будет гарантировать и качественное функционирование всей системы распознавания. Отметим, что даже разделение кластеров эталонов классов в многомерном признаковом пространстве не является гарантией достоверного выполнения заключительной фазы распознавания. Причем, в случае, когда удается в результате обучения получить компактные и разделенные в многомерном признаковом пространстве решений эталоны исследуемых классов, реализация непосредственно процедуры принятия решения является чисто техническим вопросом и принципиальных сложностей не вызывает [8].

В данной статье представлен метод формирования пространства решений, которым предусматривается сепарирование признаков из исходного априорного словаря по степени

их информативности. Процедура сепарирования признаков выполняется путем проверки значимости различий в законах распределений выборок с использованием непараметрического знакового критерия. Эта проверка осуществляется на основе компаративного анализа выборок значений признаков, представленных в исходной классифицированной обучающей выборке. Компактность и разделение образов эталонов классов обеспечивается за счет отбора таких признаков, значения которых, с одной стороны, слабо варьируют внутри каждого отдельного класса, а, с другой стороны, демонстрируют существенную неоднородность при межклассовом парном сравнении. Отобранные в результате признаки и используются для построения искомого пространства решений. Отметим, что полученный таким образом уточненный (рабочий) словарь признаков будут гарантировать устойчивое разделение эталонов классов в пространстве решений и в конечном итоге обеспечит более высокую достоверность всей процедуры распознавания.

Описание метода для формирования пространства решений

Построение системы распознавания (*диагностика – diagnōstikos, способный распознавать*, [9]), ориентированной на решение конкретной прикладной проблемы, начинается с формирования исходного пространства наблюдений (словаря наблюдаемых характеристик исследуемых объектов). Решение данной задачи в первую очередь возлагается на специалистов, являющихся экспертами в соответствующей области знаний. Универсального механизма, позволяющего достаточно строго формализовать эту процедуру, на сегодняшний день не существует. А потому приходится использовать накопленный практический опыт и, кроме того, привлекать аналитиков, которые являются специалистами в области использования компьютерных методов анализа данных [10].

Объективность результатов разрабатываемой системы распознавания будет тем выше, чем наиболее представительней будет исходная выборка наблюдаемых характеристик, формируемая на основе соответствующего *генерального словаря*. В конечном итоге, исходя из соображений здравого смысла, с учетом стоимостных, временных и других соответствующих ограничений экспертам и аналитикам удается реализовать вариант решения задачи, сформировав при этом исходное пространство наблюдений и преобразовав его в исходное пространство признаков, которое принято называть *априорным словарем признаков* (АСП).

Параллельно с проблемой построения априорного словаря признаков экспертами во взаимодействии с аналитиками решается и задача определения алфавита классов. Часто при построении конкретных систем распознавания выработка этого алфавита происходит практически автоматически, поскольку непосредственно связано с содержанием исходной задачи. Однако бывают случаи, когда формирование исходного алфавита класса является нетривиальной задачей и требует проведения скрупулезных предварительных исследований, которые реализуются совместными усилиями экспертов и аналитиков.

В конечном итоге в результате выполнения первого предварительного этапа, связанного с выполнением процедуры обучения, получаются алфавит классов $A=\{A_1,A_2,...,A_k\}$ и априорный словарь признаков $P=\{P_1,P_2,...,P_n\}$.

Непосредственно процедура обучения реализуемой системы распознавания будет проводиться на основе данных, размещаемых в классифицированной обучающей выборке. Построение этой выборки осуществляется на основе использования ранее полученных алфавита классов $A=\{A_1,A_2,...,A_k\}$ и априорного словаря признаков $P=\{P_1,P_2,...,P_n\}$. При этом формальное описание j-го экземпляра i-го класса задается в виде вектора-столбца $(x^i_{1j}, \dots, x^i_{nj})^T$.

Объединение всех m_i векторов i-го класса образует матрицу размерности n x m_i:

$$X^i_{n\ x\ m_i} = \begin{pmatrix} x^i_{11} & x^i_{12} & \dots & x^i_{1m_i} \\ x^i_{21} & x^i_{22} & \dots & x^i_{2m_i} \\ \dots & \dots & \dots & \dots \\ x^i_{n1} & x^i_{n2} & \dots & x^i_{nm_i} \end{pmatrix}, \text{ где } i=\overline{1,k}$$

Таким образом, матрица $X^i_{n\ x\ m_i}$ представляет собой формальное описание отдельного i-го класса в соответствующем многомерном априорном признаковом пространстве.

Классифицированная обучающая выборка $X_{n\ x\ m}$ получается путем объединения всех матриц $X^i_{n\ x\ m_i}$, т.е. $X_{n\ x\ m} = \bigcup_{i=1}^{k} X^i_{n\ x\ m_i}$ (где n – количество признаков априорного словаря, а $m=m_1+m_2+\ \ldots+m_k$).

Следующий этап исследований ориентирован на проведение анализа признаков из априорного словаря $P=\{P_1,P_2,...,P_n\}$ по степени их информативности с точки зрения компактности и разделения формальных образов эталонов классов в многомерном признаковом пространстве решений. В данном случае интерес будут представлять только такие признаки, значения которых, с одной стороны, относительно слабо варьируют внутри каждого отдельного класса возле среднего арифметического, а, с другой стороны, демонстрируют неоднородность для всевозможных пар при соответствующем межклассовом сравнении. За счет слабого варьирования значений признака внутри класса обеспечивается компактность эталонных образов класса. А неоднородность при межклассовом сравнении обеспечивает разделение этих образов в соответствующем признаковом пространстве – пространстве решений.

Последовательно анализу подвергаются все признаки из АСП $P=\{P_1,P_2,...,P_n\}$. При этом в качестве критерия однородности при выполнении процедуры сравнительного анализа предлагается использовать непараметрический знаковый критерий.

Отметим, что, в результате выполнения предложенной процедуры исследования информативности признаков с точки зрения разделения эталонов образов классов не всегда удается из априорного словаря сформировать непустое множество подходящих признаков. В этом случае необходимо возвращаться и проводить переформирование АСП, а затем повторять процедуру анализа обновленного набора признаков. Если же из априорного словаря удается выделить подмножество соответствующих признаков, то именно на их основе и формируется пространство решений.

Алгоритм метода

Алгоритм метода для формирования пространства решений при построении системы распознавания образов предполагает выполнение следующей последовательности шагов:

1 эксперты формируют соответствующий целям исследований алфавит классов $A=\{A_1,A_2,...,A_k\}$. Совместно со специалистами в области компьютерного анализа данных они определяют исходное пространство наблюдений и на его основе строят априорный словарь признаков $P=\{P_1,P_2,...,P_n\}$. Здесь выполняется процедура преобразования пространства наблюдений в пространство признаков. Эта процедура носит ярко выраженный проблемно-ориентированный характер, а потому невозможно предложить универсальный подход для ее выполнения. Однако опыт построения систем распознавания показывает, что при решении конкретных задач реализация указанного преобразования принципиальных сложностей не вызывает;

2 формальное представление каждого отдельного класса A_i (где $i=\overline{1,k}$) в априорном признаковом пространстве изначально определяется совокупностью соответствующих многомерных объектов. Каждый из объектов на основе признаков из АСП описывается в многомерном признаковом пространстве в виде вектора-столбца $x=(x_1,x_2,...,x_n)^T$, где x_i – значение i-го признака. В конечном итоге объединение всех соответствующих векторов-столбцов i-ого класса образует матрицу $X^i_{n\ x\ m_i}$, где m_i – число объектов i-го класса;

3 классифицированная обучающая выборка получается в результате объединения всех соответствующих матриц классов. Она представляет собой прямоугольную матрицу, состоящую из n строк и m столбцов (где $m=m_1+\ m_2+\ \ldots+m_k$, а m_i – количество объектов i-го класса);

4 каждый из признаков априорного словаря анализируется и относится к одному из двух видов, то есть словарь $P=\{P_1,P_2,...,P_n\}$ разбивается на два словаря: $P^{(1)}=\{P_1^{(1)},P_2^{(1)},...,P_{n1}^{(1)}\}$ и $P^{(2)}=\{P_1^{(2)},P_2^{(2)},...,P_{n2}^{(2)}\}$, где $P=P^{(1)}\cup P^{(2)}$ и $n=n_1+n_2$. Очередной анализируемый P_i (где $i=\overline{1,n}$) при-

знак будет отнесен к словарю $P^{(2)}$ при условии что, во-первых, значение коэффициента вариации внутри каждого класса не превысило допустимое пороговое значение V^*, и, во-вторых, для всех пар классов непараметрический знаковый критерий однородности показал существенное различие между выборками значений этого признака для двух сравниваемых классов;

5 полученный словарь $P^{(2)}=\{P_1^{(2)},P_2^{(2)},...,P_{n2}^{(2)}\}$ включает в себя признаки, на основе которых и формируется пространство решений. Из классифицированной обучающей выборки исключаются все строки, содержащие значения признаков, попавших в словарь $P^{(1)} = \{P_1^{(1)}, P_2^{(1)}, ..., P_{n1}^{(1)}\}$. Затем осуществляется нормирование значений оставшихся признаков и строятся эталоны классов в полученном многомерном пространстве решений.

Отметим, что предложенный алгоритм ориентирован на построение компактно размещенных и разделенных друг от друга эталонов классов в полученном пространстве решений. Для выполнения заключительной процедуры принятия окончательного решения могут быть использованы различные хорошо отработанные сценарии [11].

Заключение

Применение методов математической теории распознавания образов позволяет проводить исследования, связанные с изучением закономерностей поведения сложных систем и объектов даже в случае, когда классический аппарат математики оказывается пока бессильным. Использование статистических критериев однородности позволяет реализовать новый подход к решению задачи качественного обучения системы.

На основе анализа данных из классифицированной обучающей выборки каждый из признаков исходного априорного словаря в обязательном порядке исследуется по степени информативности его с точки зрения разделения эталонных образов исследуемых классов в многомерном пространстве решений.

Для определения возможности включения признака из исходного АСП в рабочий словарь, на основе которого в дальнейшем строится пространство решений, предложено проводить обязательную процедуру сравнительного статистического анализа выборок значений признаков с использованием непараметрического знакового критерия однородности.

Использование предложенного метода позволяет повысить объективность и достоверность выполнения процедуры распознавания. При проведении процедуры сепарирования по степени информативности фактически параллельно исследуются закономерности поведения всех признаков из априорного исходного словаря. В конечном итоге каждый признак не просто включается или не включается в рабочий словарь для формирования пространства решений, но и вычисляются объективные статистические характеристики, на основе которых принимается соответствующие заключение.

Предложенный алгоритм формирования пространства решений позволяет автоматизировать все шаги, связанные с обучением системы и непосредственным получением окончательного заключения о результатах распознавания.

Abstract. The paper presents the method of designing decision space in which, on the one hand, the standard of one class forms compact set, and, on the other hand, the standards of the classes are divided among themselves.

Литература

1. Журавлев, Ю.И. Распознавание образов и распознавание изображений / Ю.И. Журавлев, И.Б. Гуревич // Распознавание, классификация прогноз. Математические методы и их применение. Вып.2 // Москва: Наука, 1989. – 302 с.

2. Верхаген, К. Распознавание образов: состояние и перспективы / К.Верхаген, Р.Дёйн, Ф.Грун // Москва.: Радио и связь, 1985. – 104 с.

3. Ту, Дж. Принципы распознавания образов / Дж.Ту, Р.Гонсалес // Москва: Мир, 1978. – 412 с.

4. Шестаков, К.М. Теория принятия решений и распознавание образов / К.М. Шестаков // Минск: БГУ, 2005. – 184 с.

5. Загоруйко, Н.Г. Прикладные методы анализа данных и знаний / Н.Г. Загоруйко // Новосибирск: Изд-во Института математики, 1999. – 268 с.

6. Марусенко, М.А. Атрибуция анонимных и псевдоанонимных литературных произведений методами распознавания образов / М.А. Марусенко // Ленинград: Издательство Ленинградского университета, 1990. – 168 с.

7. Родченко, В.Г. Метод реализации стилеметрических исследований на основе применения аппарата математической теории распознавания образов / В.Г. Родченко // Известия Гомельского государственного университета им. Ф.Скорины, 2007. – № 5(44). – С. 58–62.

8. Васильев, В.И. Проблема обучения распознаванию образов / В.И. Васильев // Киев: Выща шк. Головное изд-во, 1989. – 64 с.

9. Словарь иностранных слов / 14-е изд., испр. // Москва: Рус. яз., 1987. – 608 с.

10. Джарратано, Дж. Экспертные системы: принципы разработки и программирование, 4-е издание : Пер. с англ. / Дж. Джарратано, Г. Райли // Москва: ООО “И.Д. Вильямс”, 2007. – 1152 с.

11. Закревский, А.Д. Логика распознавания / А.Д. Закревский // Минск: Наука и техника, 1988. – 118 с.

Гродненский государственный
университет имени Янки Купалы

Поступило 16.05.08