

РАСПОЗНАВАНИЕ НА ОСНОВЕ СИНТЕЗА ПАТТЕРНОВ КЛАССОВ

В.В. Краснопрошин¹, А.В. Карканица², В.Г. Родченко³

¹Белорусский государственный университет, пр. Независимости, 4,
220030, г. Минск, Беларусь, krasnoproshin@bsu.by

²Гродненский государственный университет имени Янки Купалы, ул. Ожешко, 22,
230023, г. Гродно, Беларусь, a.karkanica@grsu.by

³Гродненский государственный университет имени Янки Купалы, ул. Ожешко, 22,
230023, г. Гродно, Беларусь, rovar@grsu.by
Контактный автор: a.karkanica@grsu.by

Рассмотрен альтернативный подход к распознаванию, основанный на выявлении различительных свойств классов. Вначале предлагается в признаковых пространствах наблюдаемых объектов формировать области определения классов (паттерны). Затем на основе оценок взаимного размещения паттернов выявлять признаковые подпространства, в которых классы не пересекаются.

Ключевые слова: Распознавание образов; интеллектуальный анализ данных; машинное обучение.

RECOGNITION BASED ON THE SYNTHESIS OF CLASS PATTERNS

V.V. Krasnoproshin^a, A.V. Karkanitsa^b, V.G. Rodchenko^c

^aBelarussian State University, Nezavisimosti Avenue, 4,
220030, Minsk, Belarus, krasnoproshin@bsu.by

^bYanka Kupala State University of Grodno, Ozheshko str, 22,
230023, Grodno, Belarus, a.karkanica@grsu.by

^cYanka Kupala State University of Grodno, Ozheshko str, 22,
230023, Grodno, Belarus, rovar@grsu.by
Corresponding author: a.karkanica@grsu.by

An alternative approach to pattern recognition based on identifying the distinctive properties of classes is considered. Initially, it is proposed to form domains of classes (patterns) in the feature spaces of observed objects. Then, based on estimates of patterns mutual placement, identify feature subspaces where classes do not intersect.

Keywords: Pattern recognition; data mining; machine learning.

Введение

Одной из приоритетных задач информатики является разработка новых подходов (методов и технологий) анализа больших данных, направленных на увеличение объема извлекаемой из них полезной информации [1, 2].

Известно, что современные тренды развития в области информационных систем связаны в первую очередь с внедрением технологий искусственного интеллекта. Эти технологии наиболее ярко представлены методами машинного обучения (*Machine Learning*) и интеллектуального анализа данных (*Data Mining*) [3, 4].

Различают два вида обучения: *дедуктивное* и *индуктивное*. Дедуктивное обучение предусматривает формализацию знаний экспертов и размещение их в базе знаний. Индуктивное (обучение по прецедентам) – направлено на выявление в данных эмпирических закономерностей [5, 6].

Машинное обучение и интеллектуальный анализ данных ассоциируются с индуктивным обучением. В результате машинного обучения строится алгоритм, приближающий неизвестную целевую зависимость, как на объектах обучающей выборки, так и на всем исходном множестве [7, 8]. В то же время методы интеллектуального анализа данных предусматривают обнаружение ранее неизвестных, практически полезных и доступных интерпретации закономерностей для принятия решений в различных сферах человеческой деятельности [9, 10].

В настоящее время традиционно обучение по прецедентам сводится к решению задачи оптимизации. В данном случае требуется построить алгоритм, который наилучшим образом приближал бы неизвестную целевую зависимость, как на элементах обучающей выборки, так и на всем множестве [11]. Здесь необходимо указать на ряд серьезных проблемных моментов, связанных с использованием такого подхода. Во-первых, процесс обучения удастся реализовать только в автоматизированном режиме, поскольку требуется участие специалиста в области компьютерного анализа данных. Во-вторых, единственным результатом обучения является алгоритм классификации, неявным образом отражающий выявленную скрытую закономерность. Построенный алгоритм фактически представляет собой некий «black box», который не поддается интерпретации. В-третьих, затратив существенные ресурсы на формирование и обработку обучающей выборки, в итоге не удастся получить никаких новых знаний о предметной области [12].

В статье рассмотрен альтернативный подход для реализации процесса обучения по прецедентам. Подход основан на исследовании свойств классов, содержащихся в сочетании различных признаков и обеспечивающих различение классов. Описан метод построения и сравнения паттернов классов в признаковых пространствах, сформированных на основе сочетаний признаков из априорного словаря.

1. Об альтернативном подходе к распознаванию

Распознавание образов относится к числу фундаментальных проблем информатики, которая изучается в рамках теории интеллектуальных систем. Под распознаванием в широком смысле этого термина подразумевается процедура отнесения исходных данных (на основе выделения различительных признаков и свойств) к определенному классу. При этом класс представляет собой некоторое множество объектов, обладающих набором общих свойств.

В процессе построения систем распознавания образов решаются три основные подзадачи. Первая посвящена представлению исходных данных, полученных для распознаваемых объектов в результате измерений. Вторая связана с выделением существенных признаков и свойств из исходных данных, а третья заключается в нахождении оптимальных решающих правил классификации [13].

В настоящее время обучение в системах распознавания традиционно сводится к оптимизационной задаче, в рамках которой строятся решающие правила, обеспечивающие экстремум заранее выбранного критерия. Ее решение начинается с выбора исходной *модели (параметрического семейства алгоритмов, реализующих решающее правило)*. Затем на основе данных обучающей выборки производится многократное выполнение сценария *«тренинг + тестирование»*. По существу, обучение представляет собой итерационный процесс, когда с помощью положительных и отрицательных подкреплений формируются оптимальные значения параметров алгоритма распознавания (*классификатора*) [14].

Фактически в теории распознавания образов сложился и используется классический подход к обучению по прецедентам, который предусматривает построение (на основе данных обучающей выборки) прогностического алгоритма:

задано пространство описания объектов и в этом пространстве необходимо построить поверхность, разделяющую классы (Рисунок 1).

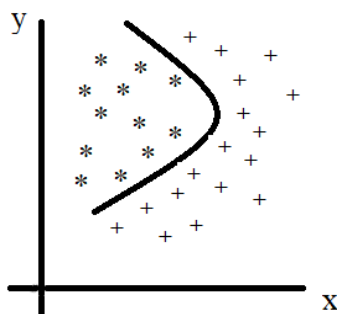


Рисунок 1 – Пример разделения двух классов кривой

В настоящее время на основе классического подхода удалось эффективно решить большое число практически полезных задач, особенно в рамках использования технологий искусственных нейронных сетей [15].

Однако детальный анализ и практический опыт использования данного подхода позволяют указать на наличие у него ряда существенных ограничений и недостатков. Во-первых, выбор модели является нетривиальной задачей, которая выполняется специалистом в области компьютерного анализа данных. И, как следствие, весь процесс обучения удастся реализовать только в автоматизированном режиме. Во-вторых, результатом обучения является практически полезный алгоритм классификации, представляющий собой «*black box*», механизм работы которого неизвестен, то есть невозможно никак проинтерпретировать. В-третьих, построенный алгоритм является *единственно* полезным результатом. То есть, затратив на формирование обучающей выборки серьезные ресурсы, в итоге удастся построить *только* алгоритм классификации, но никак не расширить свои знания ни о свойствах классов, ни о предметной области.

От указанных выше недостатков можно избавиться в случае использования предлагаемого альтернативного подхода. Когда акцент в обучении направлен не на построение алгоритма классификации, а на исследование свойств классов с целью выявления тех из них, которые обеспечивают различение классов.

Процесс обучения предлагается строить на основе гипотезы компактности, которая гласит, что в пространстве признаков образам классов соответствуют компактные множества:

задано пространство описания объектов и необходимо проверить выполняется ли в этом пространстве условие гипотезы компактности [16].

На рисунке 2 показано, что на основе наблюдаемых прецедентов двух классов (* и +) и пары признаков (x, y) построены области определения классов (*паттерны*), которые не пересекаются. Отметим, что пара признаков (x, y) обладает свойством различения классов, а построение классификатора на основе выявленного свойства пары превращается в тривиальную процедуру.

Если априорный словарь включает n признаков, то из них можно сформировать $2^n - 1$ различных сочетаний. Тогда, казалось бы, в процессе обучения необходимо проверить выполнение гипотезы компактности в каждом из $2^n - 1$ соответствующих подпространств описания объектов. Однако при детальном рассмотрении свойств отдельных сочетаний очевидным является факт, что если, например, пара признаков (x, y) обладает

свойством различения классов, то это свойство сохраняется и у всех остальных сочетаний, содержащих эту пару.

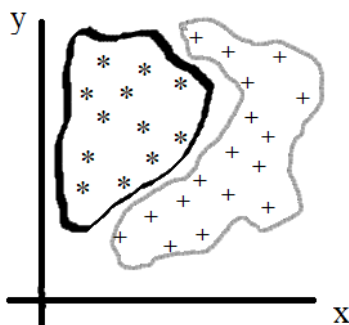


Рисунок 2 – Пример взаимного размещения паттернов двух классов

2. Алгоритм обучения

В основе предлагаемого подхода лежит идея гипотезы компактности. В пространстве описания наблюдаемых объектов предлагается строить области определения классов и на основе оценки их взаимного размещения проверить условие выполнения гипотезы компактности.

Отметим, что построение на основе наблюдаемых прецедентов области определения класса представляет собой фактически *синтез паттерна класса*, а проверка гипотезы компактности выполняется на основе *сравнения паттернов классов*.

В данном случае постановка задачи распознавания формально выглядит следующим образом.

Пусть заданы:

множество описаний объектов X , априорный словарь признаков $F=\{f_1, \dots, f_n\}$ и алфавит классов $Y=\{y_1, \dots, y_k\}$. И пусть обучающая выборка $X^m=\{(x_1, y_1), \dots, (x_m, y_m)\}$ сформирована на основе словаря F , а $V=\{v_1, \dots, v_q\}$ (где $q=2^n-1$) – множество всевозможных сочетаний признаков из F .

Требуется найти такие сочетания признаков из V , для которых в соответствующем признаковом подпространстве паттерны классов не пересекаются.

Алгоритм нахождения сочетаний признаков из $V=\{v_1, \dots, v_q\}$, для которых в соответствующем признаковом подпространстве паттерны классов не пересекаются, следующий:

Шаг 1. Выбираем из V подмножество $V^+=\{v_1^+, \dots, v_n^+\}$, где v_i^+ содержит только один признак.

Шаг 2. Для каждого v_i^+ строим паттерны классов и проводим оценку их взаимного размещения [17].

Шаг 3. Включаем v_i^+ в результирующее множество $V^* = \{v_1^*, \dots, v_k^*\}$ в случае, если паттерны классов не пересекаются.

Шаг 4. Исключаем из множества $V = \{v_1, \dots, v_q\}$ подмножество $V^+ = \{v_1^+, \dots, v_n^+\}$ и получаем $V^\wedge = \{v_1^\wedge, \dots, v_p^\wedge\}$.

Шаг 5. Исключаем из $V^\wedge$ все сочетания $v_i^\wedge$, которые содержат любое сочетание из $V^* = \{v_1^*, \dots, v_k^*\}$.

Шаг 6. Выбираем очередное сочетание $v_i^\wedge$ из $V^\wedge$ и на его основе строим признаковое пространство.

Шаг 7. В этом признаковом пространстве строим паттерны классов и проводим оценку их взаимного размещения.

Шаг 8. Если паттерны классов не пересекаются, то сочетание $v_i^\wedge$ включаем в результирующее множество V^* , а из $V^\wedge$ исключаем все сочетания, содержащие $v_i^\wedge$;

Шаг 9. Повторяем шаги 6 – 8 до тех пор, пока $V^\wedge$ не станет пустым.

В результате выполнения алгоритма будет построено множество $V^* = \{v_1^*, \dots, v_t^*\}$, где $0 \leq t \leq q$. Для каждого сочетания $v_i^* \in V^*$ формулируем выявленную закономерность: «в пространстве сочетания признаков v_i^* классы не пересекаются».

Поскольку в рамках конкретной прикладной задачи каждое сочетание v_i^* можно проинтерпретировать, то и все выявленные закономерности могут быть проинтерпретированы.

Каждое сочетание признаков из V^* определяет пространство решений, в котором паттерны классов не пересекаются, а потому построение алгоритмов классификации принципиальных затруднений не вызывает.

Заключение

В статье рассмотрен альтернативный подход к проблеме распознавания образов. Предложено в процессе обучения отказаться от общепринятого целеполагания на построение алгоритма классификации в пользу исследования свойств классов: выявление среди них тех, которые обеспечивают различение классов.

Описан алгоритм обучения, который в автоматическом режиме на основе признаков априорного словаря и данных обучающей выборки выявляет сочетания признаков, которые обеспечивают различение классов. В результате удастся не только построить классификатор, но и обнару-

жить ранее неизвестные и интерпретируемые закономерности в предметной области.

Библиографические ссылки

1. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. Новосибирск: ИМ СО РАН, 1999. 270 с.
2. Краснопрошин В.В., Родченко В.Г. Обучение по прецедентам на основе анализа свойств признаков // Доклады БГУИР. 2017. № 6. С. 35–41.
3. Плас Дж.В. Python для сложных задач: наука о данных и машинное обучение. СПб.: Питер, 2018. 576 с.
4. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Технологии анализа данных: Data Mining. Text Mining, Visual Mining, OLAP. СПб.: БХВ-Петербург, 2007. 384 с.
5. Аркадьев А.Г., Браверман Э.М. Обучение машины распознаванию образов. М.: Наука, 1964. 112 с.
6. Краснопрошин В.В., Образцов В.А. Проблема принятия решений по прецедентности: разрешимость и выбор алгоритмов // Выбр. навуц. працы Беларус. дзярж. ун-та. 2001. № 6. Р. 285–311.
7. Машинное обучение [Электронный ресурс]. URL: http://www.machinelearning.ru/wiki/index.php?title=Machine_Learning (дата обращения: 21.08.2022).
8. Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных. М.: ДМК Пресс, 2015. 402 с.
9. Data mining [Электронный ресурс]. URL: https://ru.wikipedia.org/wiki/Data_mining (дата обращения: 21.08.2022).
10. Краснопрошин В.В., Родченко В.Г. Классификация на основе пространств решений // Доклады БГУИР. 2019. № 6. С. 20-25. DOI: 10.35596/1729-7648-2019-124-6-20-25.
11. Рашка С. Python и машинное обучение. М.: ДМК Пресс, 2017. 418 с.
12. Rodchenko V. Pattern Recognition: Supervised Learning on the Bases of Cluster Structures. 13th International Conference, PRIP 2016 Minsk, Belarus, October 3-5, 2016, Communications in Computer and Information Science, Springer. 2017. P. 106-113. DOI: 10.1007/978-3-319-54220-1_11
13. Ту Дж., Гонсалес Р. Принципы распознавания образов. М.: Мир, 1978. 412с.
14. Васильев В.И. Проблема обучения распознаванию образов. К.: Выща шк. Головное изд-во, 1989. 64 с.
15. Хайкин С. Нейронные сети: полный курс, 2-е изд. М.: ООО “И.Д. Вильямс”, 2016. 1104 с.
16. Белозерский Л.А. Современный взгляд на гипотезу компактности // Искусственный интеллект. 2005. № 3. С. 5–12.
17. Краснопрошин В.В., Родченко В.Г. Кластерные структуры и их применение в интеллектуальном анализе данных // Информатика. 2016. № 2. С. 71–77.