ОБУЧЕНИЕ ПО ПРЕЦЕДЕНТАМ НА ОСНОВЕ АНАЛИЗА СВОЙСТВ ПРИЗНАКОВ

В.В. КРАСНОПРОШИН[1], В.Г. РОДЧЕНКО[2]

[1]*Белорусский государственный университет, Республика Беларусь*

[2]*Гродненский государственный университет имени Янки Купалы, Республика Беларусь*

Поступила в редакцию 10 июня 2017

Аннотация. В работе исследуется задача обучения по прецедентам (по примерам). Предложен метод обучения, основанный на анализе свойств сочетаний признаков и построении признаковых подпространств, в которых классы не пересекаются. В рамках метода разработан алгоритм, который допускает распараллеливание процесса обучения и проведение его в автоматическом режиме. Показана возможность использования метода для автоматического обнаружения и интерпретации скрытых, заранее неизвестных закономерностей в задачах интеллектуального анализа данных.

Ключевые слова: обучение по прецедентам, интеллектуальный анализ данных, обучающая выборка.

Abstract. The issue of learning by precedents is studied. A method of learning based on the analysis of the properties of feature combinations and building feature subspaces where classes do not intersect is proposed. An algorithm that allows paralleling of the learning process and performing it in the automatic mode has been developed in this approach. A possibility to use this method for automatic detection and interpretation of hidden, previously unknown patterns in data mining tasks is presented.

Keywords: learning by precedents, data mining, training sample.

Doklady BGUIR. 2017, Vol. 108, No. 6, pp. 35-41
Learning by precedents based on the analysis of the features properties
V.V. Krasnoproshin, V.G. Rodchanka

Введение

Широкое применение информационных технологий в различных сферах человеческой деятельности обеспечило возможность детального протоколирования и представления в электронном виде огромных массивов данных. Необходимость эффективного их использования привела к развитию во второй половине XX столетия целого раздела искусственного интеллекта, называемого машинным обучением и ориентированного на извлечение знаний из данных [1, 2]. Первоначально в машинном обучении выделяли два типа обучения: индуктивное (по прецедентам), которое основано на выявлении закономерностей в эмпирических данных, и дедуктивное, которое предполагает формализацию знаний экспертов и представление их в виде базы знаний. Позже дедуктивное обучение перешло в область исследования экспертных систем, и машинное обучение фактически стало ассоциироваться с обучением по прецедентам [3].

Наиболее распространенным способом формализации прецедентов является их описание в виде набора признаков. Вначале задается априорный словарь признаков, затем измеряются их значения (для всех прецедентов) и формируется обучающая выборка. Целью решения задачи обучения является построение алгоритма, который приближал бы неизвестную целевую зависимость как на объектах обучающей выборки, так и на всем исходном множестве [4].

В статье предлагается метод обучения, который базируется на исследовании свойств сочетаний признаков с целью выявления признаковых подпространств, в которых классы не пересекаются.

Постановка задачи обучения по прецедентам

Обучение по прецедентам представляет собой процесс, при котором интеллектуальной системе предъявляется конечный набор известных (положительных и отрицательных) примеров, связанных с некоторой заранее неизвестной закономерностью. Такие наборы в литературе называют обучающей выборкой. В результате анализа обучающей выборки должен быть построен алгоритм, который мог бы разделять заданные примеры. Проверка качества алгоритма обычно выполняется на объектах так называемой экзаменационной выборки [5].

Формально постановку такой задачи обучения можно записать следующим образом.

Пусть X и Y – соответственно множество описаний объектов и множество допустимых ответов по их классификации. Существует неизвестная целевая зависимость $y^* : X \to Y$, значения которой известны только для объектов обучающей выборки $X^m = \{(x_1,y_1),\ldots,(x_m,y_m)\}$.

Требуется построить алгоритм $a : X \to Y$, который приближал бы эту целевую зависимость не только на объектах конечной выборки, но и на всем множестве X.

В настоящее время наиболее распространенным способом решения такой задачи является метод минимизации эмпирического риска.

Вначале фиксируется некоторая модель алгоритмов $A = \{a : X \to Y\}$ и вводятся:

– функция потерь $L(y,y')$ – величина отклонения работы алгоритма $(y = a(x))$ для произвольного $x \in X$ от правильного значения $(y' = y^*(x))$;

– функционал качества $Q(a, X^m) = \frac{1}{m}\sum_{i=1}^{m} L(a(x_i), y^*(x_i))$ – средняя ошибка (эмпирический риск) алгоритма на объектах произвольной выборки X^m.

Далее в модели $A = \{a : X \to Y\}$ строится алгоритм, который минимизирует среднюю ошибку на всей выборке X^m, $a = \arg\min_{a \in A} Q(a, X^m)$.

Опыт практического использования данного метода для решения задачи обучения по прецедентам свидетельствует о том, что:

– выбор модели алгоритмов $A = \{a : X \to Y\}$ является далеко нетривиальной задачей, в этом случае говорят не только о науке, но и об искусстве построения алгоритмов, которые извлекают знания из данных [6];

– использование обучающей выборки X^m означает, что решение ищется только в признаковом пространстве описания объектов, однако, открытым остается вопрос о существовании подпространств, в которых задача могла бы решаться более эффективно;

– даже в случае удачного построения алгоритма, приближающего неизвестную целевую зависимость, не удается провести хотя бы минимально приемлемую интерпретацию полученных результатов (в рамках предметной области), поскольку этот алгоритм фактически представляет собой «черный ящик»;

– в силу необходимости выбора модели алгоритмов и последующей настройки ее параметров можно говорить только об автоматизированном, но не автоматическом режиме процесса обучения.

Таким образом, если принимать во внимание, что в описанном варианте задачи обучения в неявном виде присутствует словарь признаков, на основе которого сформирована обучающая выборка, то можно предложить альтернативную цель задачи. Вместо построения алгоритма, приближающего неизвестную целевую зависимость, цель задачи – объявить поиск признаковых подпространств, в которых классы не пересекаются.

В таком случае постановку задачи обучения можно переформулировать следующим образом. Пусть X и Y – соответственно множество описаний и множество допустимых ответов. Существует неизвестная целевая зависимость – отображение $y^* : X \to Y$, значения которой известны на объектах обучающей выборки $X^m = \{(x_1,y_1),\ldots,(x_m,y_m)\}$.

Требуется найти признаковые подпространства, в которых классы не пересекаются.

Общее описание метода обучения

Пусть задано множество описаний объектов X, множество допустимых ответов Y, а также обучающая выборка $X^m = \{(x_1,y_1),\ldots,(x_m,y_m)\}$, которая сформирована на основе словаря признаков $F = \{f_1,\ldots,f_n\}$. Обозначим через $V = \{v_1,\ldots,v_q\}$ множество всех непустых

подмножеств – всевозможных сочетаний признаков из F (очевидно, V содержит $q = \sum_{i=1}^{n} C_n^i$ таких подмножеств).

Возможна ситуация, когда для некоторого признака f_k интервалы изменения его значений $[min_k, max_k]$ внутри разных классов таковы, что для них выполняется условие $\bigcup_{i=1}^{s}([min_i, max_i] \bigcap [min_j, max_j]) = \varnothing, \forall i \neq j : j = \overline{1,s}$. То есть интервалы изменения значений признака f_k для всех возможных пар классов не пересекаются. Очевидно, что признак f_k, (с точки зрения разделения классов) обладает высокой степенью информативности.

В общем случае для поиска сочетаний признаков, обеспечивающих разделение классов в соответствующем признаковом пространстве, недостаточно анализировать только значения признаков. Необходимо также учитывать и отношения между признаками внутри различных сочетаний. Это можно легко продемонстрировать с помощью геометрической интерпретации классов (рис. 1).

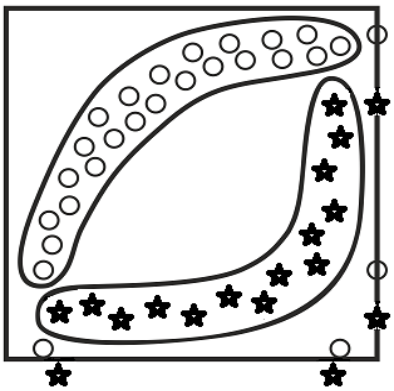

Рис. 1. Пример распределения объектов двух классов

Нетрудно видеть, что интервалы изменения значений признаков в классах «кружки» и «звездочки» или практически совпадают (по оси абсцисс), или на три четверти пересекаются (по оси ординат). Следовательно, можно сделать вывод, что с помощью анализа только значений признаков не всегда удается провести разделение классов. В то же время геометрически очевидное разделение в данном случае обеспечивается путем учета (для конкретных признаков) пар «значение–отношение».

Если увеличить размерность пространства до трех, то несложно представить ситуацию, когда объекты двух классов четко разделены и распределены в районе вершин диагонали куба, в то время как при любом проецировании на грани куба (т. е. в пространство размерности 2) или на ребра куба (т. е. в пространство размерности 1) разделение объектов классов отсутствует. Отсюда можно сделать вывод, что если исходное пространство состоит из n признаков, то пространство наименьшей размерности, в котором классы могут оказаться разделенными, будет иметь размерность p ($p \leq n$).

Для построения признаковых подпространств, в которых на объектах обучающей выборки классы не пересекаются, предлагается выполнить следующие операции:

– в множестве V выбираем произвольный элемент v_i (где $i = \overline{1,q}$) и на основе признаков f_j (где $j = \overline{1,n}$), входящих в v_i, и обучающей выборки путем исключения данных о признаках, которые не входят в состав v_i, получаем обучающую выборку $Z^m = \{(z_1,y_1),\ldots,(z_m,y_m)\}$;

– на основе $Z^m = \bigcup_{i=1}^{k} Z_i^{m_i}$, где k – количество классов, $Z_i^{m_i}$ – выборка объектов i-го класса, m_i – количество объектов i-го класса, проверяем условие $\bigcup_{i=1}^{k}(Z_i^{m_i} \bigcap Z_j^{m_j}) = \varnothing, \forall i \neq j : j = \overline{1,k}$. Если оно выполняется, то в подпространстве, образованном на основе признаков из v_i, классы не пересекаются. Подмножество признаков v_i включаем в результирующее множество V^*.

Нетрудно заметить, что после анализа элементов $V = \{v_1,\ldots,v_q\}$ будет построено множество

V^*. Если оно оказывается пустым, то необходимо либо искать решение в рамках классического подхода, либо переформатировать априорный словарь признаков и искать решение в рамках нового варианта этого словаря.

Если же множество V^* – не пустое, то на основе признаков v_i, попавших в множество V^*, формулируем ранее неизвестную и только что выявленную закономерность: «в пространстве признаков подмножества v_i классы не пересекаются».

Необходимо отметить, что, во-первых, в рамках конкретной прикладной задачи каждый признак из v_i является интерпретируемым, а, значит, можно интерпретировать в терминах предметной области и выявленную закономерность. Во-вторых, на основе выявленных закономерностей можно, например, по правилу «ближайшего соседа» осуществлять классификацию объектов.

Алгоритм обучения на основе анализа сочетаний признаков

Пусть X – множество описаний объектов, $Y = \{y_1,\ldots,y_k\}$ – алфавит классов. Пусть также имеется словарь признаков $F = \{f_1,\ldots,f_n\}$. Признаком в данном случае является результат измерения некоторой характеристики объекта. Формально признак – отображение $f : X \to D_f$, где D_f – множество допустимых значений признака. Вектор $(f_1(x),\ldots,f_n(x))$ таким образом задает признаковое описание объекта $x \in X = D_{f_1} \times \ldots \times D_{f_n}$.

Совокупность признаковых описаний всех объектов обучающей выборки $X^m = (x_1,\ldots,x_m)$, представленную в виде матрицы $Z = \begin{pmatrix} f_1(x_1),\ldots, f_n(x_1) \\ \cdots \\ f_1(x_m),\ldots, f_n(x_m) \end{pmatrix}$ размерности $m \times n$, называют матрицей «объект–свойство».

Обозначим через $Z_i^{m_i}$ матрицу размерности $m_i \times n$, образованную на основе всех объектов i-го класса (m_i – количество объектов i-го класса, $i = \overline{1,k}$, k – количество классов, n – количество признаков, $Z = \bigcup_{i=1}^{k} Z_i^{m_i}$), а через $V = \{v_1,\ldots,v_q\}$ – множество всевозможных сочетаний признаков, полученных на основе словаря F, где $q = \sum_{i=1}^{n} C_n^i$.

Алгоритм поиска признаковых подпространств формально можно описать в виде последовательности трех основных шагов:

Шаг 1. Выбираем очередное сочетание признаков v_i (где $i = \overline{1,q}$);

Шаг 2. Из матриц $Z_l^{m_l}$ (где $l = \overline{1,k}$) исключаем все столбцы со значениями признаков, не входящими в v_i. В итоге получаем множество матриц $W_1^{m_1},\ldots,W_k^{m_k}$.

Шаг 3. Проверяем выполнение условия $\bigcup_{i=1}^{k}(W_i^{m_i} \bigcap W_j^{m_j}) = \varnothing, \forall i \neq j : j = \overline{1,k}$.

Если условие не выполняется, то переходим к шагу 1, в противном случае сочетание v_i включаем в результирующее множество V^* и также переходим к шагу 1.

Предложенный алгоритм является переборным и циклическим (с количеством шагов $q = \sum_{i=1}^{n} C_n^i$, где n – количество признаков). Нетрудно показать, что алгоритм легко можно распараллелить на различных уровнях его выполнения:

– для каждого отдельного сочетания признаков v_i (см. шаг 1);

– при формировании каждой отдельной матрицы $W_l^{m_l}$ (см. шаг 2);

– при проверке условия $(W_i^{m_i} \bigcap W_j^{m_j}) = \varnothing, \forall i \neq j : j = \overline{1,k}$ (см. шаг 3).

Применение метода в задачах интеллектуального анализа данных

В настоящее время под термином интеллектуальный анализ данных (Data Mining) подразумевают совокупность методов обнаружения в данных ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации знаний, необходимых для принятия решений в различных сферах человеческой деятельности [7].

Задачи, решаемые методами Data Mining, принято разделять на описательные и предсказательные. Первые нацелены на выявление и наглядное представление имеющихся внутри данных скрытых закономерностей. Для вторых (в предсказательных задачах) главным является поиск ответа на вопрос о возможности предсказания закономерностей по данным, которые могут появиться в будущем. Например, к описательным задачам относится проблема поиска паттернов, а к предсказательным – задача классификации объектов [8].

Традиционный подход к решению этих задач методами интеллектуального анализа данных предусматривает выполнение следующих этапов:

1. Постановка задачи анализа.
2. Сбор данных.
3. Подготовка данных (фильтрация, дополнение, кодирование).
4. Выбор модели (алгоритма анализа данных).
5. Подбор параметров модели и алгоритма обучения.
6. Обучение модели (автоматический поиск остальных параметров модели).
7. Анализ качества обучения, если неудовлетворительный переход на п. 5 или п. 4.
8. Анализ выявленных закономерностей, если неудовлетворительный переход на п. 1, 4 или 5.

В результате выполнения этапов 1–3 строится обучающая выборка. Следующие четыре этапа (с 4 по 7 включительно) на практике, как правило, реализуются в автоматизированном режиме с участием квалифицированного в области интеллектуального анализа данных специалиста.

Вместо этапов 4–7 предлагается воспользоваться описанным выше методом обучения (на основе анализа свойств сочетаний признаков) и в автоматическом режиме, во-первых, сформировать множество V^*, и, во-вторых, для каждого сочетания признаков v_i, попавшего в V^*, построить паттерны классов в соответствующем признаковом пространстве и сформулировать ранее скрытую и выявленную в процессе анализа закономерность: «в пространстве признаков подмножества v_i классы не пересекаются».

Паттерны классов будут формально описывать выявленные закономерности в соответствующем сочетанию v_i признаковом пространстве. Они, в частности, могут быть использованы в задачах классификации для принятия решений по правилу «ближайшего соседа».

Отметим, что возможна ситуация, когда не будет выявлено ни одного признакового подпространства, в котором классы не пересекаются (т.е. $V^* = \varnothing$). В этом случае, как отмечалось ранее, предлагается либо искать решение в рамках классического подхода, либо сформировать другой вариант априорного словаря и искать решение на основе нового набора признаков.

Поскольку метод обучения на основе анализа свойств сочетаний признаков предусматривает получение оценок по взаимному размещению классов, то для решения задачи классификации объектов можно выбирать подпространства, в которых классы максимально разделены, а затем решать задачу в рамках традиционного подхода.

Заключение

В статье предложен метод обучения по прецедентам, который базируется на анализе свойств сочетаний признаков, сформированных на основе априорного словаря признаков. Исследование свойств сочетаний позволяет выявить подпространства, в которых классы не пересекаются, и проводить интерпретацию закономерностей, выявляемых на основе сочетаний признаков.

Описан алгоритм реализации предложенного метода и показано, что он может быть распараллелен на уровне каждого отдельного сочетания признаков, внутри сочетания признаков – на уровне формализации классов и на уровне получения оценки взаимного размещения классов.

Показана возможность использования метода для автоматического извлечения и представления ранее неизвестных, нетривиальных, практически полезных и доступных интерпретации закономерностей.

Список литературы

1. Потапов А.С. Распознавание образов и машинное восприятие: общий подход на основе принципа минимальной длины описания. СПб.: Политехника, 2007. 548 с.
2. Журавлёв Ю.И., Рязанов В.В., Сенько О.В. Распознавание. Математические методы. Программная система. Практические применения. М.: Фазис, 2005. 159 с.
3. Люгер, Дж.Ф. Искусственный интеллект: стратегии и методы решения сложных проблем. М.: Издательский дом «Вильямс», 2005. 864 с.
4. Краснопрошин В.В., Образцов В.А. Проблема принятия решений по прецендентности: разрешимость и выбор алгоритмов // Выбр. навук. працы Беларус. дзярж. ун-та. 2001. Т. 6. Матэматыка. С. 285–311.
5. Аверкин А.Н., Гаазе-Рапопорт М.Г., Поспелов Д.А. Толковый словарь по искусственному интеллекту. М.: Радио и связь, 1992. 256 с.
6. Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных. М.: ДМК Пресс, 2015. 402 с.
7. Интеллектуальный анализ данных / Профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных [Электронный ресурс]. – Режим доступа: http://www.machinelearning.ru/wiki/index.php?title= Интеллектуальный_анализ_данных. Дата доступа: 25.05.2017.
8. 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.

References

1. Potapov A.S. Raspoznavanie obrazov i mashinnoe vospriyatie: obschij podkhod na osnove printsipa minimalnoj dliny opisaniya. SPb.: Politechnika, 2007. 548 s. (in Russ.)
2. Zhuravlev Yu.I., Ryazanov V.V., Sen'ko O.V. Raspoznavanie. Matematicheskie metody. Programmnaya sistema. Prakticheskie primeneniya. M.: Fazis, 2005. 159 s. (in Russ.)
3. Luger G.F. Iskusstvennyj intellect: strategii i metody resheniya slozhnykh problem. M.: Izdatel'skij Dom «Vil'yams», 2005. 864 s. (in Russ.)
4. Krasnoproshin V.V., Obraztsov V.A. Problema prinyatiya reshenij po pretsedentnosti: razreshimost' i vybor algoritmov // Vybr. navuk. pracy Belarus. dzjarzh. uni-ta. 2001. T. 6: Matematyka. S. 285–311. (in Russ.)
5. Averkin A.N., Gaaze-Rapoport M.G., Pospelov D.A. Tolkovyj slovar po iskusstvennomu intellektu. M.: Radio i svyaz', 1992. 256 s. (in Russ.)
6. Flach P. Mashinnoe obuchenie. Nauka i iskusstvo postroeniya algoritmov, kotorye izvlekajut znaniya iz dannykh. M.: DMK Press, 2015. 402 s. (in Russ.)
7. Intellektual'nyj analiz dannykh / Professional'nyj informatsionno-analiticheskij resurs, posvyaschennyj mashinnomu obucheniyj, raspoznavaniyj obrazov i intellektual'nomu analizu dannykh [Electronic data]. – Access mode: http://www.machine-learning.ru/wiki/index.php?title= Интеллектуальный_анализ_данных. Date of access: 25.05.2017. (in Russ.)
8. Barsegyan A.A., Kupriyanov M.S., Stepanenko V.V., Kholod I.I. Tekhnologii analiza dannykh: Data Mining. Text Mining, Visual Mining, OLAP. SPb.: BHV-Peterburg, 2007. 384 s. (in Russ.)

Сведения об авторах

Краснопрошин В.В., д.т.н., профессор, заведующий кафедрой информационных систем управления Белорусского государственного университета.

Родченко В.Г., к.т.н., доцент, доцент кафедры современных технологий программирования Гродненского государственного университета имени Янки Купалы.

Адрес для корреспонденции

230023, Республика Беларусь,
г. Гродно, ул. Ожешко, д. 22,
Гродненский государственный
университет имени Янки Купалы
тел. +375-29-786-98-48;
e-mail: rovar@mail.ru
Родченко Вадим Григорьевич

Information about the authors

Krasnoproshin V.V., D. Sci., professor, head of the department of management information systems, Belarusian state university.

Rodchanka V.G., PhD, associate professor, associate professor of the modern programming technologies department, Yanka Kupala state university of Grodno.

Address for correspondence

230023, Republic of Belarus,
Grodno, Ozheshko str., 22,
Grodno state university
named after Yanka Kupala
tel. +375-29-786-98-48;
e-mail: rovar@mail.ru
Rodchanka Vadzim Rygoravich