http://dx.doi.org/10.35596/1729-7648-2019-124-6-20-25

Оригинальная статья
Original paper

УДК 004.855.5

КЛАССИФИКАЦИЯ НА ОСНОВЕ ПРОСТРАНСТВ РЕШЕНИЙ

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

[1]*Белорусский государственный университет, Республика Беларусь*

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

Поступила в редакцию 18 июня 2018

Аннотация. В работе предложен метод классификации, основанный на анализе свойств признаковых подпространств. Описана процедура автоматического выявления пространств, в которых классы не пересекаются, и показана возможность их использования для автоматического построения классификаторов.

Ключевые слова: машинное обучение, интеллектуальный анализ данных, обучение по прецедентам, классификация.

Конфликт интересов. Авторы заявляют об отсутствии конфликта интересов.

Для цитирования. Краснопрошин В.В., Родченко В.Г. Классификация на основе пространств решений. Доклады БГУИР. 2019; 6(124): 20-25.

CLASSIFICATION BASED ON DECISION SPACES

KRASNOPROSHIN V.V.[1], RODCHANKA V.G.[2]

[1]*Belarusian State University, Republic of Belarus*

[2]*Grodno State University named after Yanka Kupala, Republic of Belarus*

Submitted 18 June 2018

Abstract. It's proposed a classification method based on the analysis of the feature subspaces properties. A procedure of automatic space identification where classes do not intersect is described. It's presented a possibility to use these spaces for automatic construction of classifiers.

Keywords: machine learning, data mining, learning by precedents, classification.

Conflict of interests. The authors declare no conflict of interests.

For citation. Krasnoproshin V.V., Rodchanka V.G. Classification based on decision spaces. Doklady BGUIR. 2019; 6(124): 20-25.

Введение

Развитие моделей, методов и технологий эффективного использования больших массивов накопленных, вновь появляющихся структурированных и слабоструктурированных данных является одной из центральных задач информатики. Наиболее перспективными в области обработки данных являются разделы искусственного интеллекта, называемые машинным обучением (Machine Learning) и интеллектуальным анализом данных (Data Mining) [1, 2].

В настоящее время Machine Learning ассоциируется с индуктивным обучением, фактически представляющим собой обучение по прецедентам. В результате обучения строится алгоритм, приближающий неизвестную целевую зависимость для объектов как обучающей выборки, так и всего исходного множества [3, 4]. Методы и средства интеллектуального анализа данных подразумевают обнаружение ранее неизвестных, практически полезных и доступных интерпретации закономерностей, которые необходимы для принятия решений в прикладных задачах [5, 6].

Определяющим этапом машинного обучения и интеллектуального анализа данных является обучение по прецедентам, реализуемое на основе анализа содержимого обучающей выборки. Однако если в Machine Learning процесс обучения направлен на построение алгоритма, то в Data Mining такой подход будет являться методологической ошибкой, поскольку искомый алгоритм представляет собой «черный ящик», который практически не поддается интерпретации.

В статье для интеллектуального анализа данных обучение по прецедентам предложено реализовать в рамках поиска признаковых подпространств – пространств решений, в которых классы не пересекаются. Описан оригинальный метод, который базируется на исследовании свойств сочетаний признаков, выявлении и использовании пространств решений для автоматического построения классификаторов.

Метод выявления пространств решений

Машинное обучение – раздел искусственного интеллекта, нацеленный на изучение алгоритмов, способных обучаться. Обучение основано на анализе частных эмпирических данных, задаваемых прецедентами, и на практике оно выполняется в рамках решения задачи классификации.

Постановка задачи классификации следующая.

Пусть X – множество описаний объектов, Y – множество номеров (наименований) классов. Существует неизвестная целевая зависимость $y^*: X \to Y$, значения которой известны только на объектах конечной обучающей выборки $X^m = \{(x_1,y_1),\dots,(x_m,y_m)\}$. Требуется построить алгоритм $a: X \to Y$, который приближал бы эту целевую зависимость для любого объекта из исходного множества X [7].

Традиционный подход к решению задачи предусматривает следующую последовательность действий. Вначале выбирается некоторая модель алгоритмов $A = \{a: X \to Y\}$; вводится функция потерь $L(y,y')$ – измерение отклонения алгоритма ($y = a(x)$) для произвольного $x \in X$ от правильного значения ($y' = y^*(x)$); вводится функционал качества $Q(a, X^m) = \frac{1}{m}\sum_{i=1}^{m} L(a(x_i), y^*(x_i))$ – величина средней ошибки алгоритма a на объектах выборки X^m. Затем в модели A строится алгоритм, обеспечивающий минимальное значение средней ошибки на всей выборке X^m, $a = \arg\min_{a \in A} Q(a, X^m)$.

С практической точки зрения слабыми местами такого подхода, в частности, являются следующие:

1) проблема выбора модели алгоритмов, функции потерь и функционала качества относится к разряду нетривиальных. Причем реально такой выбор выполняется в ручном режиме;

2) результатом решения задачи является алгоритм, но интерпретируемых знаний (в аспекте интеллектуального анализа данных) извлечь не получается.

Интеллектуальный анализ данных представляет собой процесс обнаружения в наблюдаемых эмпирических данных ранее неизвестных, практически полезных и доступных интерпретации знаний, необходимых для принятия обоснованных решений [8]. Фактическим результатом такого процесса будут являться, во-первых, выявленные знания, обладающие указанными выше свойствами, и, во-вторых, алгоритмы принятия обоснованных решений, которые строятся на основе выявленных знаний.

Процедура обнаружения знаний представляет собой «обучение с учителем» и реализуется путем решения задачи в следующей постановке.

Пусть X – множество описаний, Y – 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 – пространства решений, в которых классы не пересекаются [9].

Предположим, что имеются множество описаний объектов X, множество допустимых ответов Y, обучающая выборка $X^m = \{(x_1,y_1),\ldots,(x_m,y_m)\}$, сформированная на основе словаря признаков $F = \{f_1,\ldots,f_n\}$. Обозначим через $V = \{v_1,\ldots,v_q\}$ множество всех непустых подмножеств, образованных всевозможными сочетаниями признаков из F. Очевидно, что множество V состоит из $q = \sum_{i=1}^{n} C_n^i = 2^n - 1$ подмножеств.

Выявление среди множества $V = \{v_1,\ldots,v_q\}$ признаковых подпространств – пространств решений, в которых образы классов, построенные на объектах обучающей выборки, не пересекаются, предлагается провести следующим образом:

выбираем очередное сочетание v_i (где $i = \overline{1, q}$) и на основе всех его признаков определяем соответствующее признаковое подпространство;

в этом признаковом подпространстве строим образы классов и проводим оценку их взаимного размещения;

сочетание признаков v_i включаем в результирующее множество V^* только в том случае, когда образы классов не пересекаются.

В результате анализа всех элементов $V = \{v_1,\ldots,v_q\}$ будет построено множество $V^* = \{v_1^*,\ldots,v_t^*\}$, где $0 \le t \le q$. Если множество V^* оказывается пустым, то необходимо сформировать новый вариант априорного словаря признаков и искать решение в рамках этого словаря.

Для каждого отдельного подмножества признаков $v_i^* \in V^*$ формулируем ранее неизвестную и выявленную эмпирическим путем закономерность: «в пространстве признаков подмножества v_i^* классы не пересекаются». Отметим, что в рамках конкретной прикладной задачи каждое сочетание признаков v_i^* из V^* может быть проинтерпретировано, а значит, и все выявленные закономерности могут быть проинтерпретированы.

Каждое сочетание признаков $v_i^* \in V^*$ определяет пространство решений, в котором классы не пересекаются, т. е. для паттернов классов внутри таких пространств подтверждается гипотеза компактности, а потому построение классификаторов (алгоритмов классификации) принципиальных затруднений не вызывает.

Построение классификатора на основе пространств решений

Пусть X – множество описаний объектов, $Y = \{y_1,\ldots,y_k\}$ – алфавит классов, $F = \{f_1,\ldots,f_n\}$– словарь признаков, где признаком является результат измерения некоторой характеристики объекта.

Признак представляет собой отображение f: $X \to D_f$, где D_f – множество допустимых значений признака. Вектор $(f_1(x),\ldots,f_n(x))$ – описание объекта $x \in X = D_{f1} \times D_{fn}$. Совокупность признаковых описаний всех объектов образует обучающую выборку $X^m = (x_1,\ldots,x_m)$, которая

представляет собой матрицу $Z = \begin{pmatrix} f_1(x_1), \ldots, f_n(x_n) \\ \ldots \\ f_1(x_m), \ldots, f_n(x_m) \end{pmatrix}$ размерности $m \times n$.

Обозначим через $Z_i^{m_i}$ матрицу размерности m_i x 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 $q = \sum_{i=1}^{n} C_n^i = 2^n - 1$.

Алгоритм поиска пространств решений должен предусматривать выполнение последовательности следующих трех шагов:

Шаг 1. Берем очередное сочетание признаков $v_i \in V$ (где $i = \overline{1,q}$).

Шаг 2. В матрицах $Z_l^{m_l}$ (где $l = \overline{1,k}$) исключаем все столбцы с данными о тех признаках, которые не входят в сочетание v_i, и 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 выполняется, то сочетание v_i включаем в множество V^* и переходим к шагу 1, а иначе просто возвращаемся к шагу 1.

Результатом работы алгоритма является множество $V^* = \{v_1^*,\ldots,v_t^*\}$, где $0 \le t \le q$. Если это множество оказывается непустым, то переходим к построению классификатора, а иначе необходимо перейти к формированию нового варианта априорного словаря признаков и далее искать решение в рамках этого словаря.

Классификатор будет представлять собой алгоритм, предусматривающий выполнение следующей последовательности шагов:

Шаг 1. Из непустого множества $V^* = \{v_1^*,\ldots,v_t^*\}$ последовательно, начиная с $i = 1$, выбираем сочетание признаков v_i^*.

Шаг 2. В матрицах $Z_l^{m_l}$ (где $l = \overline{1,k}$) исключаем все столбцы с данными о признаках, не входящих в сочетание v_i^*, и получаем матрицы $W_1^{*m_1},\ldots,W_k^{*m_k}$.

Шаг 3. В признаковом пространстве, сформированном на основе v_i^*, строим кластерные структуры $C_1^{v_i},\ldots,C_k^{v_i}$ [10].

Шаг 4. Если $i < t$, то переходим к шагу 1, а иначе – к следующему.

Шаг 5. Берем классифицируемый объект.

Шаг 6. Последовательно, начиная с $i = 1$, выбираем очередное сочетание признаков v_i^* и в рамках соответствующего признакового пространства находим и присваиваем d_i либо номер кластерной структуры из множества $C_1^{v_i},\ldots,C_k^{v_i}$, в которую попал объект, либо присваиваем d_i ноль, если объект не попал ни в одну из кластерных структур. Итогом выполнения этого шага будет d_i – номер класса, к которому был отнесен классифицируемый объект.

Шаг 7. Если $i < t$, то переходим к шагу 1, а иначе – к следующему.

Шаг 8. Из сформированного в результате выполнения шагов 6 и 7 множества $D = \{d_1,\ldots,d_t\}$ по правилу большинства голосов находим номер класса y^* (где $0 \le y^* \le k$) классифицируемого объекта.

Отметим, что если $y^* = 0$, то это означает, что классифицируемый объект не является представителем ни одного из заявленных классов, и говорят, что он относится к джокер-классу.

Применение метода для решения задачи классификации

В настоящее время машинное обучение и интеллектуальный анализ данных относятся к числу наиболее перспективных подразделов искусственного интеллекта, связанных с изучением методов построения алгоритмов, обладающих способностью обучаться.

Решаемые методами Machine Learning и Data Mining задачи принято разделять на описательные и предсказательные. Целью описательных задач является выявление и наглядное представление ранее неизвестных скрытых внутри данных закономерностей, например, проблема поиска паттернов. В предсказательных задачах центральным моментом является проблема поиска ответа на вопрос о возможности предсказания закономерностей на основе данных, которые появятся позже. Ярким примером в данном случае являются задачи классификации.

В отличие от классического подхода для решения задачи классификации предлагается воспользоваться описанным выше методом выявления пространств решений. Сначала на основе данных обучающей выборки фактически решить описательную задачу, т. е. отыскать пространства решений, в которых паттерны классов не пересекаются. А затем построение классификатора реализовать на основе выявленных пространств решений.

Отличительной чертой предложенного метода решения задачи классификации является возможность автоматической классификации, когда на вход подаются данные обучающей выборки и без всяких внешних вмешательств на выходе формируется результат классификации исследуемого объекта.

Заключение

В статье предложен метод классификации, который базируется на исследовании и использовании свойств сочетаний признаков исходного словаря. На основе анализа данных обучающей выборки выявляются пространства решений, в которых классы не пересекаются. Показано, что на основе свойств этих пространств удается не только выявлять скрытые закономерности и проводить их интерпретацию в терминах предметной области, но и проводить автоматическое построение классификаторов.

Описаны метод выявления пространств решений, алгоритм поиска таких пространств и алгоритм автоматического построения классификаторов. Показана возможность использования разработанных метода и алгоритмов для решения задач классификации.

Список литературы

1. Плас Дж.В. Python для сложных задач: наука о данных и машинное обучение. СПб.: Питер, 2018. 576 с.
2. Силен В., Мейсман А., Али М. Основы Data Science и Big Data. Python и наука о данных. СПб.: Питер, 2017. 336 с.
3. 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.
4. Абламейко С.В., Краснопрошин В.В., Образцов В.А. Модели и технологии распознавания образов с приложением в интеллектуальном анализе данных // Вестник БГУ. Сер. 1. № 3. 2011. С. 62–72.
5. Технологии анализа данных: Data Mining. Text Mining, Visual Mining, OLAP / А.А. Барсегян [и др.]. СПб.: БХВ-Петербург, 2007. 384 с.
6. Флах П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных. М.: ДМК Пресс, 2015. 402 с.
7. Машинное обучение / Профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных [Электронный ресурс]. URL: http://www.machinelearning.ru/wiki/index.php?title=Machine_Learning (дата обращения: 05.04.2018).
8. Интеллектуальный анализ данных / Профессиональный информационно-аналитический ресурс, посвященный машинному обучению, распознаванию образов и интеллектуальному анализу данных [Электронный ресурс]. URL: http://www.machinelearning.ru/wiki/index.php?title=Интеллектуальный_анализ_данных (дата обращения: 05.04.2018).
9. Краснопрошин В.В., Родченко В.Г. Обучение по прецедентам на основе анализа свойств признаков // Доклады БГУИР. 2017. № 6 (108). С. 35–41.
10. Краснопрошин В.В., Родченко В.Г. Кластерные структуры и их применение в интеллектуальном анализе данных // Информатика. № 2. 2016. С. 71–77.

References

1. Plas Dzh.V. Python dlja slozhnyh zadach: nauka o dannyh i mashinnoe obuchenie. SPb.: Piter, 2018. 576 s. (in Russ.)
2. Silen V., Mejsman A., Ali M. Osnovy Data Science i Big Data. Python i nauka o dannyh. SPb.: Piter, 2017. 336 s. (in Russ.)
3. Krasnoproshin V.V., Obrazcov V.A. Problema prinjatija reshenij po precendentnosti: razreshimost' i vybor algoritmov // Vybr. navuk. pracy Belarus. dzjarzh. un-ta. 2001. T. 6. Matjematyka. S. 285–311. (in Russ.)
4. Ablamejko S.V., Krasnoproshin V.V., Obrazcov V.A. Modeli i tehnologii raspoznavanija obrazov s prilozheniem v intellektual'nom analize dannyh // Vestnik BGU. Ser. 1. № 3. 2011. S. 62–72. (in Russ.)
5. Tehnologii analiza dannyh: Data Mining. Text Mining, Visual Mining, OLAP / A.A. Barsegjan [i dr.]. SPb.: BHV-Peterburg, 2007. 384 s. (in Russ.)
6. Flah P. Mashinnoe obuchenie. Nauka i iskusstvo postroenija algoritmov, kotorye izvlekajut znanija iz dannyh. M.: DMK Press, 2015. 402 s. (in Russ.)
7. Mashinnoe obuchenie / Professional'nyj informacionno-analiticheskij resurs, posvjashhennyj mashinnomu obucheniju, raspoznavaniju obrazov i intellektual'nomu analizu dannyh [Electronic resource]. URL: http://www.machinelearning.ru/wiki/index.php?title=Machine_Learning (date of access: 05.04.2018). (in Russ.)
8. Intellektual'nyj analiz dannyh / Professional'nyj informacionno-analiticheskij resurs, posvjashhennyj mashinnomu obucheniju, raspoznavaniju obrazov i intellektual'nomu analizu dannyh [Electronic resource]. URL: http://www.machinelearning.ru/wiki/index.php?title= Intellektual'nyj_analiz_dannyh (date of access: 05.04.2018). (in Russ.)
9. Krasnoproshin V.V., Rodchenko V.G. Obuchenie po precedentam na osnove analiza svojstv priznakov // Doklady BGUIR. 2017. № 6 (108). S. 35–41. (in Russ.)
10. Krasnoproshin V.V., Rodchenko V.G. Klasternye struktury i ih primenenie v intellektual'nom analize dannyh // Informatika. № 2. 2016. S. 71–77. (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, 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 of Belarusian State University.

Rodchanka V.G., PhD, associate professor, associate professor of the Modern Programming Technologies Department of Grodno State University named after Yanka Kupala.

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