Виды признаков и их роль в дифференцировании классов при оценке не полностью описанного объекта[1]

[1, 2] *В.Н. Юдин <yudin@ispras.ru>*
[1, 3] *Л.Е. Карпов <mak@ispras.ru>*
[4] *В.Ю. Абрамов <v_abramov@list.ru>*
[1] *Институт системного программирования РАН,*
Россия, 109004, г. Москва, ул. А. Солженицына, д. 25.
[2]*Московский областной научно-исследовательский клинический институт им. М.Ф. Владимирского, Россия, 129110, г. Москва, ул. Щепкина, д. 61/2*
[3]*Московский государственный университет имени М.В. Ломоносова,*
Россия, 119991, г. Москва, Ленинские горы, д. 1
[4]*НИИ скорой помощи им. Н.В. Склифосовского,*
Россия, 129010, г. Москва, Большая Сухаревская площадь, д. 3.

Аннотация. Разработанный метод в рамках прецедентного подхода к принятию решений позволяет решить проблему выбора наиболее подходящих прецедентов в условиях, когда объект исследования не полностью описан и оценивается неоднозначно. Особенность предлагаемого подхода – в том, что он ориентирован на работу в условиях нефиксированного набора признаков (атрибутов). Это актуально для многих приложений, особенно, при поддержке врачебных решений, когда на процесс принятия решений накладываются ограничения по времени и ресурсам. Чтобы добиться успеха, необходимо дифференцировать возможную принадлежность объекта, расширив его признаковое пространство. Эта задача, в свою очередь, сводится к изучению роли признаков и их сочетаний (по аналогии с дифференциальной диагностикой и семиотикой в медицине). Для выбора порядка, извлечения недостающих признаков, используются введенные понятия: ранг, устойчивые сочетания признаков, частота появления, доступность признака и категории объектов.

Ключевые слова: добыча данных; вывод на основе прецедентов; база прецедентов; дифференциальный ряд; мера близости; устойчивые сочетания признаков.

DOI: 10.15514/ISPRAS-2016-28(3)-14

[1] Работа поддержана грантами Российского фонда фундаментальных исследований № 15-01-02362 и № 15-07-02355.

Для цитирования: Юдин В.Н., Карпов Л.Е., Абрамов В.Ю. Виды признаков и их роль в дифференцировании классов при оценке не полностью описанного объекта. Труды ИСП РАН, том 28, вып. 3, 2016 г., стр. 231-240. DOI: 10.15514/ISPRAS-2016-28(3)-14

1. Введение

Вывод на основе прецедентов – метод принятия решений, в котором используются знания о предыдущих ситуациях или случаях (прецедентах). В такой терминологии прецедент рассматривается как *объект*, включающий в себя описание проблемы, описание решения проблемы и результат применения решения (исход). Накопленная совокупность прецедентов, наполняемая как смоделированными типовыми случаями, так и случаями из практики, образует так называемую базу прецедентов. При рассмотрении новой проблемы (текущего случая) в базе прецедентов находится похожий прецедент. Можно попытаться использовать ранее принятое для него решение, возможно, адаптировав к текущему случаю, вместо того, чтобы искать решение каждый раз сначала. После того, как обработка текущего случая завершится, новый прецедент должен быть внесен в базу прецедентов вместе со своим решением для возможного последующего использования.

Однако чтобы найти наиболее подходящий прецедент, нужно иметь способ измерения близости прецедента и текущего случая. Часто используемым методом в выборе наиболее подходящих прецедентов является *метод ближайшего соседа*. В его основе лежит тот или иной способ измерения степени близости прецедента и текущего случая. В качестве основы измерений в пространстве всех признаков можно ввести какую-либо метрику, определив в этом пространстве точку, соответствующую текущему случаю. На основе выбранной метрики можно отыскивать ближайшую точку, которая и представит прецедент. К сожалению, во многих случаях ввести метрику не удается. Тогда вместо метрики используется так называемая мера близости. Это означает, что вместо метрического пространства используется топологическое.

2. Структуризация базы прецедентов

Один из способов определения меры близости – структуризация множества прецедентов, например, разбиение базы прецедентов на классы эквивалентности, при котором все случаи одного и того же класса считаются равными. В основу такого разбиения кладутся знания о предметной области (фоновое знание), полученные с помощью методов добычи данных (Data Mining) – классификации и кластеризации [1, 2, 3].

В задачах распознавания образов обычно предполагается, что в основе описаний объектов лежит набор признаков, общий для объектов всех классов (за основу принято признаковое описание случая, когда он описывается набором своих характеристик). Иными словами, классы и исследуемые объекты располагаются в едином признаковом пространстве. В реальных приложениях это условие часто не выполняется. Как сами окружающие объекты, так и

описания классов, могут иметь собственные пространства признаков. Например, в медицине каждое заболевание характеризуется своим набором существенных признаков. И, наконец, исследуемый случай может иметь набор показателей, не совпадающий с наборами показателей введенных в систему классов (в медицине – заболеваний), часто из-за дефицита времени, ресурсов, а иногда и квалификации исследователя.

Ряд признаков, которыми обладает исследуемый случай, может не входить в общее признаковое пространство имеющихся классов, а некоторые признаки могут оказываться несущественными для данного конкретного случая. Такие признаки в дальнейшем не будут нами рассматриваться. С другой стороны, у исследуемого случая по разным причинам могут отсутствовать признаки (например, не проделаны важные измерения, не завершены важные исследования), которые являются существенными по отношению к некоторым классам.

3. Дифференцирование классов выбором разделяющего признака

Уже довольно давно в мире развивается технология, во многом опирающаяся на методы рассуждения по прецедентам [4-7]. В ИСП РАН на базе разработанного исследовательского программного комплекса «Универсальный Анализатор» и системы поддержки врачебных решений «Спутник врача», создаваемой на клинической базе МОНИКИ им. М. Ф. Владимирского, также проводится разработка новой технологии [2, 8-9]. В развиваемом подходе база прецедентов состоит из совокупности прецедентов и описаний классов, каждое из которых включает в себя перечень существенных признаков (причем классы – это структура, накладываемая на совокупность прецедентов сверху). Оценить случай – значит выявить его принадлежность тому или иному классу. Отношения между текущим случаем и классами выявляются в проекциях классов на пространство признаков объекта – текущего случая. Недостаточно полно описанный объект может попасть в проекцию класса, к которому он на самом деле не принадлежит, только потому, что у него не хватает признака, который дифференцировал бы его от этого класса. Проиллюстрируем это на простом примере.

Два непересекающихся класса, A и B (рис. 1), описаны в пространстве признаков $\{x_1, x_2\}$. Текущий случай O представлен одним признаком x_1, признак x_2 отсутствует. В пространстве признаков $\{x_1\}$ проекции классов пересекаются, и объект попадает в это пересечение.

Классы нужно дифференцировать, добавляя значения недостающих признаков для текущего случая. В медицине подобная задача носит название *дифференциальная диагностика* [9]. На практике подобное добавление может быть затруднено из-за нехватки средств, времени или оборудования. Но главная

причина заключается в том, что реальные приложения редко укладываются в рамки фиксированного признакового пространства.

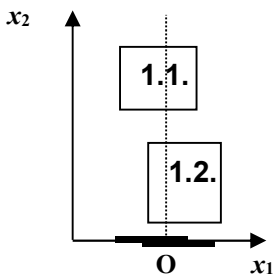

Рис. 1. Одновременное отнесение недостаточно полно описанного объекта к двум классам.

Fig. 1. Not fully described object may be assigned to several different classes.

Формально сущность предложенного метода оценки не полностью описанных объектов сводится к следующему:

- описание объекта (случая) – набор признаков.
- описание класса – многомерный параллелепипед в пространстве признаков, минимально объемлющий прецеденты класса.
- оценка объекта – сравнение случая с проекциями классов на пространство своих признаков.
- *дифференциальный ряд* случая – набор классов, в пересечение проекций которых объект попал.
- если объект попал в область пересечения проекций классов, то наиболее близкими к нему считаются прецеденты этих классов, находящиеся в той же области пересечения. В этом заключается смысл искомой *меры близости* [2, 3, 8] отражающей сходство текущего объекта (случая) и выбранного прецедента.

В зависимости от сложности пересечения, все прецеденты делятся на группы. Находящиеся в общей с текущим случаем области пересечения, естественно считать более близкими к нему, чем те, что находятся только в одном из классов. В конечном счете, прецеденты самого высокого ранга близости находятся в области пересечения всех классов, образующих дифференциальный ряд текущего случая (рис. 2).

Первоначальный отбор прецедентов может не дать ощутимого результата. Например, наличие в текущем случае всего лишь одного признака «высокая температура» (в медицине это носит название лихорадка неясного генеза) даст обилие пересекающихся классов. Тогда нужно либо согласиться, что с таким набором признаков проблему не решить, либо наращивать набор исследуемых признаков.

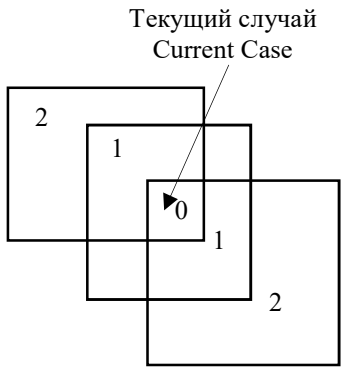

Рис. 2. Степени близости прецедентов (цифрами обозначены расстояния между текущим случаем и прецедентами).

Fig. 2. Similarity degrees (digits designate the degree of similarity between the new case and existing cases).

4. Признаки и их роль в дифференцировании классов

Задача исследователя – расширить признаковое пространство текущего случая так, чтобы однозначно оценить его принадлежность тому или иному классу. Когда случай находится в области пересечения классов, в условиях нехватки времени или ресурсов выявлять все оставшиеся признаки нет возможности. Нужно сформулировать критерий извлечения новых признаков для разделения классов, *определяя приоритеты этого извлечения.* Для этого используются несколько новых понятий: устойчивые сочетания признаков, ранг признака в классе по его степени информативности. Предлагается ввести еще ряд дополнительных критериев отбора: по частоте использования признака в приложении, по категории объекта, по доступности признака.

Не все признаки, которые присутствуют в описании класса, одинаково информативны. Так, в медицине, есть так называемые патогномоничные симптомы, которые имеют абсолютное диагностическое значение (в частности, маркеры рака, гепатита, инфаркта) и позволяют установить наличие заболевания. В общем случае, отвлекаясь от медицины и ее приложений, мы называем признаки класса, имеющие наибольшую информативность, *идеальными*. Они однозначно идентифицируют свой класс, а в других классах не встречаются. Но в той же медицине патогномоничные симптомы не всегда обнаруживаются при соответствующих болезнях, либо обнаруживаются не во всех стадиях или не при всех формах течения. При отсутствии подобных симптомов необходимо принимать во внимание ряд других признаков, относительно более характерных, чем другие, для сопоставляемого заболевания. Есть признаки, появляющиеся в классе с вероятностью, в

несколько раз превышающей вероятность их появления в прецедентах других классов. Назовем такие признаки *детерминирующими* (билирубин, печеночные ферменты при гепатитах). Конечно, для окончательной оценки не следует ориентироваться на один такой признак, следует обязательно учитывать его сочетание с другими.

И, наконец, часто встречается еще одна группа признаков – *сопутствующие*. Они не являются характерными признаками класса (в медицине, например, это симптомы, которые могут сопутствовать основному заболеванию: лихорадка, скорость оседания эритроцитов и т. д.). Их наличие можно считать необходимым, но не достаточным условием принадлежности к классу. Роль сопутствующих признаков в дифференцировании классов ничтожна. Резюмируя, будем считать, что признаки в описании каждого класса в базе прецедентов ранжированы, а именно, подчинены отношению порядка: идеальные – детерминирующие – сопутствующие. Достоверно идентифицировать состояние объекта может только идеальный признак. При его отсутствии, даже если к рассмотрению привлечен целый ряд детерминирующих признаков, говорить о принадлежности к классу можно лишь условно.

При выявлении любых дополнительных признаков должны учитываться их ранги в каждом из классов дифференциального ряда. Само собой разумеется, что исследование признака, который во всех классах ряда относится к сопутствующим, не даст такого эффекта, как если бы это был признак более высокого ранга.

Для окончательного определения класса, к которому относится текущий случай, не всегда можно ориентироваться на единственный признак, нужно учитывать его возможную связь с другими признаками. Так, в медицине особое диагностическое значение имеет устойчиво наблюдаемая совокупность симптомов, определяемая как синдром (семиотика – направление в медицине, где изучаются симптомы различных заболеваний, в особенности их сочетания и их роль в дифференциальной диагностике, так называемая синдромная специфичность). Здесь видна прямая связь с методом выявления знаний при добыче данных, который носит название *анализ ассоциаций*. Этот метод весьма полезен и часто успешен при обработке описаний классов в базе прецедентов на предмет выявления устойчивых сочетаний и рангов признаков.

Вводимое понятие *Устойчивые сочетания признаков* – дополнительный путь к дифференцированию классов. Взаимосвязь *Признаки - Сочетания признаков* похожа на взаимосвязь *Признаки - Классы*. И в том, и в другом случае – это связь многие-к-многим. Если первая связь изначально была отражена в структуре данных базы прецедентов, то вторая только воплощается на текущем этапе. По аналогии с классами, в базу прецедентов заносятся сочетания и входящие в них признаки. На практике такой подход уже давно используется. В медицине синдром как устойчивый набор признаков может указывать не на одно, а на ряд заболеваний. С другой стороны, заболевание может проявляться

не одним, а рядом синдромов. Оба эти факта указывают на связь *Классы - Сочетания признаков* как на связь многие-к-многим. Эта связь тоже должна быть отражена в базе прецедентов.

Текущий случай при его оценке в базе прецедентов попадает в дифференциальный ряд сочетаний признаков, где в перекрестии находятся признаки случая, а в остальной части лепестков – признаки, пока отсутствующие в описании случая, но которые при наличии смогут образовать с первыми устойчивое сочетание. Естественно, при принятии решения, какой из признаков выявлять в первую очередь, выбирается сочетание, где в наборе признаков случая не хватает только одного признака из известного устойчивого сочетания (или наименьшего числа таких признаков).

Итак, если первая стадия оценки случая – получение дифференциального ряда классов, то вторая – получение дифференциального ряда сочетаний признаков. По связи *Классы - Сочетания признаков* можно выбрать набор классов, который соответствует этим сочетаниям. Третья стадия – к двум наборам применяется операция конъюнкции, в результате которой возникает уменьшенный дифференциальный ряд классов.

Итак, когда текущий случай находится в области пересечения классов, выявлять все его недостающие признаки в условиях нехватки времени или ресурсов нет возможности. В этой ситуации для выбора рекомендуется использовать введенные ранее понятия *Ранг признака* и *Устойчивые сочетания признаков*.

Попытаемся описать остальные критерии отбора:

- Новый признак хотя бы в одном из классов дифференциального ряда должен иметь высокий ранг (идеальный или детерминирующий, но не сопутствующий).
- Выявляются в первую очередь признаки, которые имеют бóльшую частоту появления (на уровне всей базы прецедентов). Эту величину можно получить приблизительно, поддерживая в описании базы прецедентов при каждом признаке счетчик использований данного признака, значение которого делится на значение счетчика использований всех признаков базы. В базе вводится еще один дополнительный тег для признака.
- Учитывается категория исследуемого объекта. Выбираются только признаки, соответствующие данной категории. Для примера можно опять обратиться к медицине: конкретное лечебное учреждение во многом определяет контингент больных, находящихся там, их заболевания, характерные симптомы. В базу прецедентов вводится сущность *Категория*, и отношение *Категория - Признак* вида многие-к-многим.
- Выбираются наиболее доступные признаки. Это довольно широкий термин, под которым понимается ряд параметров: стоимость

выявления признака, наличие аппаратуры для его выявления, целый ряд параметров предпочтения (в медицине известен термин неинвазивность) и ряд других. В зависимости от приложения, это один или несколько дополнительных тегов признака на уровне базы.

5. Заключение

Подход к оценке не полностью описанных объектов востребован, особенно в такой области, как медицина, хотя медицинскими приложениями этот подход не ограничивается. Понятие дифференциального ряда и мера близости в оценке объектов являются оригинальными, они разработаны и подробно описаны в более ранних работах авторов.

Список литературы

[1]. I. Watson and F. Marir. Case-based reasoning: A review. The Knowledge Engineering Review, 9(4):327-354, 1994.

[2]. Л. Е. Карпов, В. Н. Юдин. Интеграция методов добычи данных и вывода по прецедентам в медицинской диагностике и выборе лечения. Сборник докладов 13-й Всероссийской конференции "Математические методы распознавания образов (ММРО-13)", октябрь, 2007, стр. 589-591.

[3]. Valery. Yudin, Leonid Karpov. The Case-Based Software System for Physician's Decision Support. Sami Khari, Lenka Lhotska, Nadia Pisanti (eds.), "Information Technology in Bio- and Medical Informatics, ITBAM 2010", Proceedings of the First International Conference, Bilbao, Spain. Lecture Notes in Computer Science Sublibrary: SL 3, Springer Verlag, Berlin, Heidelberg, 2010, pp. 78-85.

[4]. Agnar Aamodt, Enric Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. AI Communications, 7(1). pp. 39-59, 1994.

[5]. I. Watson and F. Marir. Case-based reasoning: A review. The Knowledge Engineering Review, 9(4):327-354, 1994.

[6]. Mario Lenz, Brigitte Bartsch-Spörl, Hans-Dieter Burkhard. Case-Based Reasoning Technology: From Foundations to Applications. Springer, 2003.

[7]. Lorraine McGinty, David C. Wilson, Case-Based Reasoning Research and Development: 8th International Conference on Case-Based Reasoning, ICCBR 2009 Seattle, WA, USA, July 20-23, 2009 Proceedings (Lecture Notes in Artificial Intelligence).

[8]. В.Н. Юдин. Мера близости в системе вывода на основе прецедентов. Доклады 12-й Всероссийской конференции Математические Методы Распознавания Образов (ММРО-12), МАКС Пресс, Москва 2005, стр. 241-244 .

[9]. В. Н. Юдин, Л. Е. Карпов, А. В. Ватазин. Методы интеллектуального анализа данных и вывода по прецедентам в программной системе поддержки врачебных решений, М., Альманах клинической медицины, № 17 в двух частях, Москва 2008, ч. 1, стр. 266-269.

Feature's types and their role in differentiating classes for estimation of not fully described object

[1,2] *V.N. Yudin <yudin@ispras.ru>*
[1,3] *L.E. Karpov <mak@ispras.ru>*
[4] *V.Y. Abramov <v_abramov@list.ru>*
[1]*Institute for System Programming of the Russian Academy of Sciences, 25, Alexander Solzhenitsyn st., Moscow, 109004, Russia*
[2]*Moscow Regional Research and Clinical Institute n.a. M.F.Vladimirsky 61/2, Shepkina, Moscow, 129110, Russia*
[3]*Lomonosov Moscow State University, GSP-1, Leninskie Gory, Moscow, 119991, Russia*
[4]*N.V. Sklifosovsky Research Institute for Emergency Medicine of Moscow Healthcare Department, 3, Bolshaya Sukharevskaya Square, Moscow, 129010, Russia*

Abstract. Authors are developing precedent approach to solving the problem of optimal decision making. The method they develop makes it possible to make the most adequate precedent selection in conditions where the object under consideration is not fully described, and cannot be estimated unambiguously. The originality of the approach offered by authors is in its focus on functioning with varying set of features (attributes). It is important for different applications, but it is especially important while supporting physician's decision making, who often has a lack of time and resources. The method presumes the need in differentiating possible object membership that may be done by widening of its feature space. This task may in its turn be reduced to investigating of feature's roles and their combinations (as in differential diagnosis and semiotics in medicine). In order to determine in what way should one retrieve missing features the authors offer to use the following conceptions: range, persistent feature combination, frequency of occurrence, availability of a feature, and object category. This work is supported by Russian Foundation for Basic Research (projects 15-01-02362 and 15-07-02355).

Keywords: data mining; case-based reasoning; case base; differential set; measure of closeness; persistent feature combinations.

DOI: 10.15514/ISPRAS-2016-28(3)-14

For citation: Yudin V.N., Karpov L.E., Abramov V.Y. Feature's types and their role in differentiating classes for estimation of not fully described object. *Trudy ISP RAN /Proc. ISP RAS*, vol. 28, issue 3, 2016, pp. 231-240 (in Russian). DOI: 10.15514/ISPRAS-2016-28(3)-14

References

[1]. Ian H. Witten, Eibe Frank and Mark A. Hall. Data Mining: Practical Machine Learning Tools and Techniques. 3rd Edition. Morgan Kaufmann, 2011. pp. 664.
[2]. L. E. Karpov, V. N. Yudin. Integration of data Mining and Case-Based Reasoning methods in medical diagnostics and treatment choosing. Sbornik dokladov 13-j

Vserossijskoj konferentsii Matematicheskie metody raspoznavaniya obrazov [Proc. of 13-th All-Russian conference Math. methods of pattern recognition], October 2007, MAKS Press, 2007, pp. 589-591 (in Russian).
[3]. Valery. Yudin, Leonid Karpov. The Case-Based Software System for Physician's Decision Support. Sami Khari, Lenka Lhotska, Nadia Pisanti (eds.), "Information Technology in Bio- and Medical Informatics, ITBAM 2010", Proc. of the First International Conference, Bilbao, Spain. Lecture Notes in Computer Science Sublibrary: SL 3, Springer Verlag, Berlin, Heidelberg, 2010, pp. 78-85.
[4]. Agnar Aamodt, Enric Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. AI Communications, 7(1), 1994, pp. 39-59.
[5]. I. Watson and F. Marir. Case-based reasoning: A review. The Knowledge Engineering Review, 9(4):327-354, 1994.
[6]. Mario Lenz, Brigitte Bartsch-Spörl, Hans-Dieter Burkhard. Case-Based Reasoning Technology: From Foundations to Applications. Springer, 2003.
[7]. Lorraine McGinty, David C. Wilson, Case-Based Reasoning Research and Development: 8th International Conference on Case-Based Reasoning, ICCBR 2009 Seattle, WA, USA, July 20-23, 2009 Proceedings (Lecture Notes in Artificial Intelligence).
[8]. Yudin V. N. Measure of closeness in Case-based Reasoning System. Dokladyi 12-y Vserossiyskoy konferentsii Matematicheskie Metodyi Raspoznavaniya Obrazov (MMRO-12) [Proc. of 12-th All-Russia Conf. on Mathematical Methods in Pattern Recognition], MAKS Press, Moscow, 2005, pp. 241-244 (in Russian).
[9]. Yudin V. N., Karpov L. E., Vatazin A. V. Data mining and case-based reasoning methods in physician's decision making support software system. Almanah klinicheskoy meditsinyi [Almanac of Clinical Medicine], Moscow, 2008, v. 17, part 1, pp. 266-269, (in Russian).