

Машинное обучение, теория решеток и анализ формальных понятий

С. О. Кузнецов, Всероссийский Институт Научной и Технической Информации (ВИНИТИ РАН), Москва

Машинное обучение обычно определяют как область искусственного интеллекта, “занятую вопросом о том как построить компьютерную программу, которая бы автоматически улучшала свое поведение с приобретением опыта” (Т. Mitchell). Неудачи чисто автоматического подхода к приобретению знаний в парадигме машинного обучения привели к моделям обнаружения знаний (knowledge discovery), которое есть “человеко-центрированный процесс открытия”. “Преобразование информации в знание поддерживается наилучшим образом когда информация вместе с ее общим смыслом представляется в соответствии с социальными и культурными паттернами понимания в сообществе, члены которого создают знание.” (R. Wille) Несмотря на видимую здесь смену взгляда от чисто автоматического к человеко-центрированному,

В данной работе будет представлен обзор по применению теории решеток в моделях машинного обучения и обнаружения знаний. Повидимому первыми среди подобных работ следует считать работы (J. Reynolds, G.D. Plotkin, 1970) по антиунификации. Антиунификация как операция супремума в решетке термов исследовалась впервые в работе Дж. Рейнолдса.

Пример:

Для $A = P(a, x, f(x))$ и $B = P(y, f(b), f(f(b)))$ имеем
 $\wedge(A, B) = P(z_1, z_2, f(z_2))$.

В работе Г.Д. Плоткина антиунификация использовалась как метод индуктивного обобщения. Позже эта идея использовалась в методах Индуктивного Логического Программирования (ILP) [5].

В конце 1970х гг. идеи теории решеток, легшие в основу Анализа Формальных Понятий (Formal Concept Analysis [8, 2]), также использовались в процедуре интерактивного обучения, называемой исследованием признаков (Attribute exploration).

Вспомним основные определения из АФП [2]. Положим, что у нас есть

- M , множество **признаков**
- G , множество **объектов**
- $I \subseteq G \times M: (g, m) \in I \iff g$ обладает признаком m .

- $\mathbb{K} := (G, M, I)$ называется **формальным контекстом**.

Операторы Галуа (derivation operators) задаются следующим образом:

$$A' \stackrel{\text{def}}{=} \{m \in M \mid gIm \forall g \in A\}, B' \stackrel{\text{def}}{=} \{g \in G \mid gIm \forall m \in B\}$$

Тогда **(формальное) понятие** есть пара (A, B) : $A \subseteq G$, $B \subseteq M$, $A' = B$, и $B' = A$, A называется **(формальным) объемом**, а B - **(формальным) содержанием** данного понятия. Понятия, упорядоченные отношением

$$(A_1, B_1) \geq (A_2, B_2) \iff A_1 \supseteq A_2,$$

образуют полную решетку, называемую **решеткой понятий** $\mathfrak{B}(G, M, I)$.

Импликация $A \rightarrow B$, где $A, B \subseteq M$, имеет место если $A' \subseteq B'$, т.е. каждый объект, обладающий всеми признаками из множества A , также обладает всеми признаками из множества B . Импликации удовлетворяют **правилам Армстронга**, известным из теории функциональных зависимостей. Относительно этих правил существует базис - минимальное множество импликаций, из которых с помощью правил выводимы все импликации. В начале 1980х был предложен "пообъектный" алгоритм **Next Closure** для построения базиса импликаций и Исследование признаков (Attribute exploration) - интерактивная процедура обучения.

В 1980х годах В.К. Финном был предложен ДСМ-метод, которым сходство определялась как операция пересечения на описаниях объектов. В системе CHARADE (J. Ganascia, 1987) ищались базисы импликаций для заданных соответствий Галуа. В системе GRAND (G. D. Oosthuizen, 1988) ищались Замыкание Дедекинда-МакНила порядка общности (отношения "быть более общим чем") и осуществлялось построение базисов импликаций В 1990е годы идея пространства версий была разработана средствами логического программирования в рамках **Индуктивного Логического Программирования (ILP)**, в котором большое значение имеет решетка поглощения. Кроме того, в 1990е годы решетки замкнутых множеств признаков вновь приобрели большое значение из-за установления их связи с базисами ассоциативных правил (обобщение импликаций, когда допускается некоторая доля исключений).

1 Гипотезы на основе замкнутых множеств признаков

В ДСМ-методе гипотезы относительно причины явления ищутся среди пересечений описаний положительных примеров явления. На пересечения могут быть наложены различные дополнительные условия. Логическим средством ДСМ-метода является многозначное многосортное расширение логики предикатов первого порядка с помощью кванторов по кортежам переменной длины (слабая логика предикатов второго порядка). В предикатах ДСМ-метода задается система замыканий относительно операции "сходства" объектов. В терминах АФП описание метода таково: Помимо признаков из множества M имеется **целевой признак** $w \notin M$, относительно которого все объекты разделяются следующим образом:

- **положительные примеры:** Множество $G_+ \subseteq G$ объектов, про которые известно, что они обладают целевым признаком w ,
- **отрицательные примеры:** Множество $G_- \subseteq G$ объектов, про которые известно, что они не обладают целевым признаком w ,
- **недоопределенные примеры:** Множество $G_\tau \subseteq G$ объектов, про которые неизвестно, обладают ли они целевым признаком или нет.

Возникают три подконтекста: $\mathbb{K}_\varepsilon := (G_\varepsilon, M, I_\varepsilon)$, $\varepsilon \in \{-, +, \tau\}$.

Формальное содержание $H \subseteq M$ контекста \mathbb{K}_+ есть **положительная гипотеза** с запретом на контрпример если H не является подмножеством содержания ни одного отрицательного примера $g \in G_-$:

$$\forall g \in G_- \quad H \not\subseteq g^-$$

Пример обучающей выборки

G \ M	цвет	жесткий	гладкий	форма	фрукт
яблоко	желтое	нет	да	круглое	+
грейпфрут	желтый	нет	нет	круглый	+
киви	зеленое	нет	нет	овальное	+
слива	синяя	нет	да	овальная	+
кубик	зеленый	да	да	кубический	-
яйцо	белое	да	да	овальное	-
теннисный мяч	белый	нет	нет	круглый	-

Естественное шкалирование выборки:

G \ M	w	y	g	b	f	f̄	s	s̄	r	r̄	фрукт
яблоко		x				x	x		x		+
грейпфрут		x				x		x	x		+
киви			x			x		x		x	+
слива				x		x	x			x	+
кубик			x		x		x			x	-
яйцо	x				x		x			x	-
теннисный мяч	x					x		x	x		-

Используемые сокращения:

“g” - зеленый, “y” - желтый, “w” - белый, “f” - твердый, “f̄” - нетвердый,
 “s” - гладкий, “s̄” - негладкий, “r” - круглый,
 “r̄” - некруглый.

Пример правила классификации

- Если недоопределенный пример g'_τ содержит в качестве подмножества положительную гипотезу и не содержит ни одной отрицательной гипотезы, то g'_τ **классифицируется положительно** (предсказывается наличие целевого признака w).

- Если g'_τ содержит в качестве подмножества отрицательную гипотезу и не содержит ни одной положительной гипотезы, то g_τ **классифицируется отрицательно** (предсказывается отсутствие целевого признака w).
- Если g'_τ содержит в качестве подмножеств гипотезы обоих знаков или если g'_τ вообще не содержит в качестве подмножеств ни положительных ни отрицательных гипотез, то классификация объекта, соответственно, **противоречива** или **недоопределенна**.

Как следует из определения, для классификации достаточно иметь множество всех **минимальных** (относительно \subseteq) гипотез.

Классификация недоопределенного примера манго

	$G \setminus M$	w	y	g	b	f	\bar{f}	s	\bar{s}	r	\bar{r}	фрукт
1	яблоко		x			x		x		x		+
2	грейпфрут		x			x			x	x		+
3	киви			x		x			x		x	+
4	слива				x	x		x			x	+
5	кубик			x		x		x			x	-
6	яйцо	x				x		x			x	-
7	теннисный мяч	x					x		x	x		-
8	манго		x			x		x			x	τ

Объект **манго** классифицируется положительно, поскольку:

- $\{\bar{r}, \bar{f}\}$ есть (+)-гипотеза,
 $\{\bar{r}, \bar{f}\} \subseteq \text{манго}' = \{y, \bar{f}, s, \bar{r}\}$;
- для (-)-гипотез $\{w\}$ и $\{f, s, \bar{r}\}$:
 $\{w\} \not\subseteq \text{манго}'$,
 $\{\bar{f}, s, \bar{r}\} \not\subseteq \text{манго}'$.

Вариации модели обучения

- разрешение $\alpha\%$ контрпримеров (для гипотез и/или классификаций),
- проверка других логических условий (например, в духе “Метода различий” Д.С. Милля): **“решетка методов” В.К. Финна**,
- несимметричная классификация: использование условий различной логической силы для положительных и для отрицательных гипотез.

Инвариантом остается то, что гипотезы ищутся среди положительных и отрицательных содержаний (замкнутых множеств признаков). Среди многочисленных приложений ДСМ-метода можно упомянуть работы в области фармакологии, автоматическом синтезе лекарств, токсикологии, медицинской диагностики, социологии, технической диагностики, анализе исторических

документов и т.д. Помимо объектно-признакового представления, ДСМ-метод допускает представление с помощью объектов, на которых задана произвольная операция сходства \sqcap , обладающая свойствами полурешетки. Примером такой операции является операция сходства на множествах помеченных графов, которая используется в химических приложениях.

Возможным источником происхождения операции \sqcap является частично-упорядоченное множество (P, \leq) описаний объектов, где \leq – отношение типа “быть более общим описанием” (отношением общности), по которому строится (дистрибутивная) решетка порядковых идеалов упорядоченного множества (P, \leq) . Для такого рода конструкций возможно создание семейства аппроксимирующих представлений на основе операции проекции (взятия внутреннейности).

Одним из новых областей применения ДСМ-гипотез является их (комбинированное с другими методами) использование в системах фильтрации нежелательных сообщений (спама).

2 Построение деревьев решений

На вход системы построения деревьев решений поступают описания положительных и отрицательных примеров, заданные множествами значений признаков. Все вершины дерева (за исключением корня и листьев) помечены признаками, а ребра деревьев помечены значениями признаков (например, 0 или 1 в случае бинарных признаков), каждый лист помечен классом + или –: примеры со всеми значениями признаков на пути, ведущем от корня к дереву, принадлежат к определенному классу, + либо –. Системы подобные ID3 [R. Quinlan 86], вычисляют значение функционала *прироста информации (information gain)* (IG) или *негэнтропии* для каждой вершины дерева и каждого признака, еще не выбранного выше по ветви дерева. Алгоритм последовательно продлевает ветви дерева, на каждом шагу выбирая признак с наибольшим приростом информации: он “наиболее сильно разделяет” объекты из классов + и –. Продление ветви прекращается когда последующее значение признака вместе с значениями признаков, выбранными ранее, однозначно классифицируют примеры относительно классов + и –. Часто процедуру заканчивают раньше для того, чтобы избежать переобучения (overfitting). В работе показывается как деревья решений, погруженные в т.н. полупроизведения дихотомических шкал, сопоставляются с ДСМ-гипотезами.

- Гипотезы соответствуют “наиболее осторожным” (наиболее частным) классификаторам, совместным с обучающей выборкой: они являются наименее общими обобщениями описаний положительных примеров.
- Кратчайшие пути решения (для которых ни в одном дереве решений не существует полных путей с подмножеством значений признаков) соответствуют “самым смелым” (или “самым различающим”) классификаторам: будучи кратчайшими возможными правилами, они являются самыми общими обобщениями описаний положительных примеров.

- Нет гарантий того, что для данной обучающей выборки существует дерево решений такое что минимальные гипотезы являются замыканиями путей решений, соответствующих ветвям дерева.

Вопросы общности обобщений естественно рассматривать в терминах пространств версий.

3 Пространства версий.

Пространства версий рассматриваются в следующих терминах:

- **Язык примеров** L_e , который описывает множество примеров E ;
- **Язык классификаторов** L_c , который описывает множество C классификаторов;
- **Предикат соответствия** $M(c, e)$: Имеет место $M(c, e)$ тогда и только тогда когда пример e "подпадает" под классификатор c . Множество классификаторов частично упорядочено **отношением (порядком) поглощения**: для классификаторов $c_1, c_2 \in L_c$,

$$c_1 \leq c_2 : \iff \forall e \in E M(c_1, e) \rightarrow M(c_2, e).$$

- Множества E_+ и E_- **положительных примеров и отрицательных примеров целевого признака**, для которых имеет место $E_+ \cap E_- = \emptyset$.
- **Предикат согласованности** $\text{cons}(c)$:
Имеет место $\text{cons}(c)$ если для каждого примера $e \in E_+$ имеет место $M(c, e)$ и для каждого отрицательного примера $e \in E_-$ имеет место $\neg M(c, e)$.
- **Пространство версий** есть множество всех согласованных предикатов: $\text{VS}(L_c, L_e, M(c, e), E_+, E_-)$.
- **Задача обучения**:
Дано $L_c, L_e, M(c, e), E_+, E_-$.
Найти пространство версий $\text{VS}(L_c, L_e, M(c, e), E_+, E_-)$.
- **Классификация**:
Классификатор $c \in \text{VS}$ **классифицирует** пример e положительно если e подпадает под c , иначе пример e классифицируется отрицательно.
Пример e **$\alpha\%$ -классифицируем** если не менее $\frac{\alpha}{100} \cdot |\text{VS}|$ классификаторов из пространства версий классифицируют его положительно.

Пространство версий часто рассматривают в терминах граничных множеств (Т. Mitchell 1982)

Если каждая цепь в порядке поглощения имеет минимальный и максимальный элементы, то пространство версий может быть описано с помощью множества $S(VS)$ наиболее частных и множества $G(VS)$ наиболее общих классификаторов:

$$\begin{aligned} G(VS) &:= \text{MIN}(VS) := \{c \in VS \mid \neg \exists c_1 \in VS c_1 < c\}, \\ S(VS) &:= \text{MAX}(VS) := \{c \in VS \mid \neg \exists c_1 \in VS c < c_1\}. \end{aligned}$$

Рассмотри формальный контекст (E, C, I) , где:

- E есть множество примеров, которое содержит непересекающиеся множества положительных и отрицательных примеров целевого признака:
 $E \supseteq E_+ \cup E_-, E_+ \cap E_- = \emptyset;$
- C есть множество классификаторов;
- отношение I задается предикатом соответствия $M(c, e)$: для $c \in C, e \in E$ имеет место отношение eIc тогда и только тогда когда $M(c, e) = 1$;
- \bar{I} есть дополнительное отношение: имеет место $e\bar{I}c$ если $M(c, e) = 0$.

Тогда пространство версий естественно выражаются в терминах соответствий Галуа как

$$VS(E_+, E_-) = E_+^I \cap E_-^{\bar{I}}.$$

Это позволяет с легкостью описывать и вычислять слияние пространств версий, выражать множество всех примеров, классифицируемых всем пространством версий, множество примеров, классифицируемых по крайней мере одним элементом пространства версий.

Если классификаторы, упорядоченные отношением поглощения, образуют полную решетку, то пространство версий есть полная подполурешетка независимо от множеств положительных и отрицательных примеров E_+ and E_- . Это свойство позволяет устанавливать интересные соотношения между гипотезами и пространствами версий, когда максимальным элементом пространства версий является минимальная гипотеза.

Если классификаторы, упорядоченные отношением поглощения, образуют конечную полурешетку по операции пересечения (инфимум), то пространство версий можно вычислить с помощью стандартного алгоритма **NextClosure** из анализа формальных понятий:

Суммируя, можно сказать, что деревья решений и пространства версий естественным образом выражаются на языке соответствий Галуа и формальных понятий. При разумных предположениях пространства версий могут быть вычислены как решетки понятий. Множество классификаторов между (относительно порядка общности или отношения поглощения) минимальными гипотезами и собственными предикторами представляется более интересным и/или более компактным чем конъюнктивно-дизъюнктивное пространство версий, поскольку

оно фактически вводит “ограниченную” дизъюнкцию по минимальным гипотезам. В общем случае, Анализ Формальных Понятий представляет собой удобное средство для формализации символьных моделей машинного обучения, основанных на отношении поглощения (порядка общности).

Список литературы

- [1] V.K. Finn, On Machine-Oriented Formalization of Plausible Reasoning in the Style of F. Backon–J. S. Mill, *Semiotika Informatika*, **20** (1983) 35-101.
- [2] B. Ganter and R. Wille, *Formal Concept Analysis: Mathematical Foundations*, Springer, 1999.
- [3] T. Mitchell, Generalization as Search, *Artificial Intelligence* **18**, no. 2, 1982.
- [4] T. Mitchell, *Machine Learning*, The McGraw-Hill Companies, 1997.
- [5] S.-H. Nienhuys-Cheng and R. de Wolf, *Foundations of Inductive Logic Programming*, Lecture Notes in Artificial Intelligence, vol. 1228, 1997.
- [6] J.R. Quinlan, Induction on Decision Trees, *Machine Learning*, **1**, No. 1, 81-106 (1986).
- [7] G. Stumme, R. Wille, U. Wille, Conceptual Knowledge Discovery in Databases Using Formal Concept Analysis Methods, In: J. Zytkow, M. Quafou, ed., Proc. *2nd European Symposium on PKDD'98. Principles of Data Mining and Knowledge Discovery*, Lecture Notes in Artificial Intelligence, **1510**, Springer, 1998, 450-458.
- [8] R. Wille, Restructuring Lattice Theory: an Approach Based on Hierarchies of Concepts, In: *Ordered Sets* (I. Rival, ed.), Reidel, Dordrecht–Boston, 445-470, 1982.