Министерство науки и высшего образования Российской Федерации Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)» (МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ Информатика и системы управления

КАФЕДРА Системы обработки информации и управления

Отчет по рубежному контролю №2

По курсу «Методы машинного обучения в АСОИУ»

Выполнил:

ИУ5-22М Барышников М.И.

14.02.2024

Проверил:

Балашов А.М.

2024 г.

Концепция метода АФП

С точки зрения философии науки, формальное понятие является естественным способом описания группы объектов через их характерные свойства, которое определяется объемом – величиной всех объектов, принадлежащих понятию и содержанием – совокупностью всех признаков, общих для объектов [1-4].

Первая задача АФП заключается в формировании формального контекста в виде тройки из множеств объектов, признаков и отношений их инцидентности. Обычно формальный контекст представляют в виде объектно-признаковой таблицы, в которой столбцам соответствуют признаки, а строками – объекты. Наличие крестика в таблице на пересечении столбца и строки означает вхождение соответствующей пары «объект, признак» в отношение инцидентности. Воздействуя на контекст оператором Галуа, представленным парой отображений объектов на признаки, и признаков на объекты, получаем (формальные) понятия. Для упорядочивания понятий строится решетка понятий. Вершинами решетки соответствуют (формальные) понятия, с соответствующими им объемами и содержаниями, а ребрами – отношения соседства между самими понятиями.

Отношения между формальными понятиями определяются на основе их свойств. Например, два понятия могут быть связаны отношением включения, если одно понятие является подмножеством другого. Также могут существовать отношения эквивалентности, когда два понятия имеют одинаковые свойства.

Теория АФП успешно применяется для классификации, информационного поиска, построения онтологий, формирования рекомендаций и прочие [5-7].

Формальный контекст – это математическая модель, которая представляет собой таблицу, состоящую из трех частей: множества объектов, множества признаков и отношения между ними. Каждый объект может иметь некоторые признаки, и каждый признак может быть присущ только определенным объектам. Отношение между объектами и признаками может быть двух видов: присутствие и отсутствие.

Контекстом в АФП называют тройку K = (G, M, I), где G — множество объектов, M — множество признаков, а отношение $I \subseteq G \times M$ говорит о том, какие объекты какими признаками обладают.

Для произвольных $A \subseteq G$ и $B \subseteq M$ определены соответствия Галуа:

$$A' = \{m \in M \mid \forall g \in A\ (g\ I\ m)\};$$

$$B' = \{g \in G \mid \forall m \in B\ (g\ I\ m)\};$$

которые задают соответствия между частично упорядоченными множествами объектов и признаков.

Оператор '' (двукратное применение оператора ') является оператором замыкания и отвечает следующим свойствам:

- идемпотентен ($A'''' = A''$),
- монотонен ($A \subseteq B$ влечет $A'' \subseteq B''$),
- экстенсивен ($A \subseteq A''$).

Множество объектов $A \subseteq G$, такое, что $A'' = A$, называется замкнутым. Аналогично для замкнутых множеств признаков — подмножеств множества M. Пара множеств (A, B), таких, что:

- $A \subseteq G$,
- $B \subseteq M$,
- $A' = B$,
- $B' = A$,

называется формальным понятием контекста K. Множества A и B замкнуты и называются объемом и содержанием формального понятия (A, B) соответственно. Для множества объектов A множество их общих признаков A' служит описанием сходства объектов из множества A, а замкнутое множество A'' является кластером сходных объектов (с множеством общих признаков A').

Отношения между объектами и признаками определяются на основе наличия или отсутствия признака у объекта. Если объект обладает определенным признаком, то мы говорим, что между ними существует отношение принадлежности. Если объект не обладает признаком, то между ними существует отношение непринадлежности.

Формальный контекст может быть представлен в виде таблицы (рис. 1), где строки соответствуют объектам, а столбцы - признакам. В каждой ячейке таблицы может быть значение 1 или 0, которое указывает наличие или отсутствие признака у объекта.

	bookable	rentable	driveable	rideable	joinable
hotel	x				
apartment	x	x			
car	x	x	x		
bike	x	x	x	x	
excursion	x				x
trip	x				x

Рисунок 1 – формальный контекст

Формальный контекст может быть использован для построения решетки понятий, которая позволяет визуально представить отношения между понятиями. Понятия формального контекста K = (G, M, I), упорядоченные по вложению объемов образуют решетку B(G, M, I), называемую решеткой понятий.

Также он может быть использован для классификации объектов и понятий, а также для построения систем рекомендаций на основе анализа данных.

Решетка понятий – это графическое представление формального контекста, которое позволяет визуально выявить отношения между понятиями. В решетке понятий каждое понятие представлено вершиной, а отношения между ними – ребрами.

ДСМ-метод

ДСМ-метод — это метод автоматического порождения гипотез, который используется для анализа данных и выявления скрытых закономерностей. Этот метод основан на теории анализа формальных понятий (АФП), которая изучает связи между объектами и их свойствами. Основной идеей ДСМ является представление объектов в виде формальных понятий, которые определяются через их характеристики и отношения между ними. Эти формальные понятия представляют собой абстрактные концепции, которые объединяют объекты с одинаковыми свойствами или отношениями в единое целое.

В задаче бинарной классификации, объекты классификационного контекста разделяются в зависимости от значения целевого признака t [5]:

- положительные примеры: Множество $G_+ \subseteq G_{train}$ объектов, про которые известно, что они обладают целевым признаком t,
- отрицательные примеры: Множество $G_- \subseteq G_{test}$ объектов, про которые известно, что они не обладают целевым признаком t,
- недоопределенные примеры: Множество $G_\tau = G_{test}$ объектов, про которые не известно, обладают ли они целевым признаком или нет.

Возникают три подконтекста: $\mathrm{K}_\varepsilon := (\mathrm{G}_\varepsilon, \mathrm{M}, \mathrm{I}_\varepsilon), \varepsilon \in \{-, +, \tau\}$.

Формальное содержание $H \subseteq M$ контекста K_+ есть положительная гипотеза, если H не является подмножеством содержания ни одного отрицательного примера $g \in G_-$:

$$H^{++} = H, \qquad \forall g \in G_- \quad H \nsubseteq g^-.$$

Отрицательные гипотезы определяются симметрично (с заменой + на –). Формальное содержание H ⊆ M контекста K_- есть отрицательная гипотеза, если H не является подмножеством содержания ни одного положительного примера $g \in G_+$:

$$H^{--} = H, \qquad \forall g \in G_+ \quad H \nsubseteq g^+.$$

Классификация недоопределенного примера g^τ:

– Если g_τ^τ содержит в качестве подмножества положительную гипотезу и не содержит ни одной отрицательной гипотезы, то g^τ классифицируется положительно (предсказывается наличие целевого признака ω).

– Если g_τ^τ содержит в качестве подмножества отрицательную гипотезу и не содержит ни одной положительной гипотезы, то g^τ классифицируется отрицательно (предсказывается отсутствие целевого признака ω).

– Если g_τ^τ содержит в качестве подмножеств гипотезы обоих знаков или если g_τ^τ вообще не содержит в качестве подмножеств ни положительных ни отрицательных гипотез, то классификация объекта, соответственно, противоречива или недоопределена.

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

Триадический анализ формальных понятий

Триадический подход к анализу формальных понятий основан на формализации триадического отношения, соединяющего объекты, атрибуты и условия. Реальные ситуации могут быть проанализированы только в ограниченном контексте, и триадический концептуальный анализ основан на формальных понятиях триадических контекстов. Формальные условия могут формализовать различные отношения, посредничество, репрезентации, интерпретации, доказательства, оценки и другие аспекты.

Подход, предлагаемый в стандартном АФП, не рассматривает возможного наличия условий, накладываемых на связи (отношения) объект – признак.

Триадический формальный контекст K = (G, M, B, Y) состоит из множеств G (объекты), M (признаки), B (условий) и тернарного отношения $Y \subseteq G \times M \times B$. Запись $(g, m, b) \in Y$ означает, что g имеет признак m при условии b.

Трикластером называется тройка вида:

$$T = (g', m', b'), \text{где } (g, m, b) \in Y.$$

Точно так же, как двоичные контексты часто описываются двумерными перекрестными таблицами, триадические контексты могут быть представлены трехмерными перекрестными таблицами [9, 10].

Трипонятием контекста K называется такая тройка вида (X, Y, Z), которая максимальна по вложению своих компонент, такая что $X \subseteq G$, $Y \subseteq M$, $Z \subseteq B$ и $X \times Y \times Z \subseteq Y$.

Для описания вывода операторов используются альтернативные обозначения базового триадического контекста. Для построения триадических концепций требуются дополнительные операторы вывода.

Метод трикластеризации является эффективным средством сокращения количества формальных трипонятий (время работы алгоритма и размер выхода полиномиальны), порождаемых для входных контекстов и может быть использован для решения аналогичных задач анализа данных.

Узорные структуры

АФП преобразует формальный контекст, представленный как бинарное отношение, в решётку формальных понятий, но во многих случаях исследуемые «объекты» могут иметь более сложное описание, чем множество некоторых наперед заданных признаков.

Узорные структуры расширяют АФП для работы со сложными данными, такими как численные значения, множества последовательностей или графов. Узорная структура представляет собой тройку (G, (D, ∏), δ), где:

- G - множество объектов,
- (D, ∏) - полная полурешётка описаний,
- δ:G→D - функция, сопоставляющая каждому объекту его описание из D [8].

Соответствие Галуа между множествами объектов и множеством описаний перепишется для узорной структуры следующим образом:

$A^{\Xi} \coloneqq \prod_{g \in A} \delta(g)$ для $A \subseteq G,$

$d^{\Xi} \coloneqq \{g \in G | d \sqsubseteq \delta(g)\}$ для $d \subseteq D,$

где:

- ⊑ – это отношение поглощения.

Узорное понятие узорной структуры (G, (D,∏), δ) – это пара (A, d), где:

- A ⊆ G – подмножество множества объектов,
- d ∈ D – одно из описаний из полурешётки, такие что A^{Ξ} = d и d^{Ξ} = A, A называется объёмом понятия, а d – узорным содержанием.

Соответствие между объектами и множествами признаков является функцией δ, а множество всех подмножеств множества признаков с операцией пересечения множеств является полной полурешёткой.

Пример преобразования формального контекста в узорную структуру приведён на рис. 3.

	m_1	m_2	m_3	m_4
g_1	x			x
g_2			x	x
g_3		x		
g_4			x	x

Контекст (G, M, I)

	Description	Name
g_1	$\{m_1, m_4\}$	d_1
g_2	$\{m_3, m_4\}$	d_2
g_3	$\{m_2\}$	d_3
g_4	$\{m_3, m_4\}$	d_2

Узорная Структура $(G, (\wp(M), \cap), \delta)$.

Рисунок 3 – преобразование формального контекста в узорную структуру.

Стоит обратить внимание, что размер решётки узорных понятий может быть крайне велик. Это приводит к вычислительно затратным операциям и, следовательно, время построения такой решётки будет занимать существенное время.

Чтобы оптимизировать работу данного алгоритма были введены понятия проекций узорных структур. Проекция может быть представлена как некоторый фильтр полурешётки описания с определенным перечнем математических свойств, благодаря которым в спроецированной решётке для любого выбранного понятия гарантированно найдётся соответствующее понятие из исходной решётки [8].