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

Домашнее задание

По курсу
«Методы машинного обучения в АСОИУ»

Выполнил:

ИУ5-22М Бабин А.С.

14.04.2024

Проверил:

Гапанюк Ю.Е.

2024 г.

Введение

Анализ формальных понятий АФП (от англ. Formal Concept Analysis – FCA) это теория анализа данных, которая идентифицирует концептуальные структуры в наборах данных. Она была введена Рудольфом Вилле в 1982 году и с тех пор быстро развивалась [5].

В течение первых 10 лет АФП разрабатывалась в основном небольшой группой исследователей и студентов Вилле в Германии [2]. Из-за математической природы большинства публикаций того времени знания о АФП оставались узко специализированными. Однако за последнее десятилетие формальный анализ понятий превратился в широко известный метод. Области его исследования и дальнейшего развития продолжают расширяться и в настоящее время.

Данный алгоритм тесно связан со многими направлениями исследований в области информационных технологий и математики:

- теория решёток (решётки Галуа);
- представление знаний;
- интеллектуальный анализ данных;
- машинное обучение;
- семантические сети и др.

Математическая мощь АФП проистекает из формализации концептуальной решетки, в которой каждый элемент отражает формальную концепцию, в то время как вся структура представляет собой иерархию, которая предлагает концептуальную кластеризацию, просмотр и анализ правил ассоциации. Сильной стороной формального концептуального анализа является также его способность создавать графическую визуализацию присущих данным структур.

На основе классического алгоритма АФП было разработано множество модификаций и усовершенствований, каждое из которых имеет свою специфику, определённую область применения, ограничения.

Классический АФП

Основная идея АФП заключается в том, чтобы выявить скрытые связи между объектами и признаками в данных, используя понятия из теории решеток, а также алгебры логики.

Основные математические концепции, используемые в АФП, включают в себя следующие ключевые понятия:

- формальный контекст;
- операторы Галуа;
- формальное понятие;
- решётка понятий;

Формальный контекст (Formal Context) – это математическая структура, используемая в методе анализа формальных понятий для представления данных. Формальный контекст представляет собой тройку K = (G, M, I), где:

- G - множество объектов;
- M - множество признаков;
- I - бинарное отношение между объектами и признаками.

Рассмотрим эту тройку подробнее.

Множество объектов (G) представляет собой набор элементов или множество объектов в нашем датасете, который мы собираемся анализировать.

Множество признаков (M) представляет собой набор характеристик или свойств, которые могут быть присущи объектам из множества G.

Бинарное отношение (I) определяет, обладает ли объект из множества G определенным признаком из множества M. То есть, каждому объекту сопоставляется набор признаков, которыми он обладает. Это отношение может быть представлено в виде таблицы или матрицы, где строки соответствуют объектам, столбцы - признакам, а элементы матрицы указывают наличие (1) или отсутствие (0) признака у объекта.

Математически формальный контекст K = (G, M, I) можно представить следующим образом:

- G = g1, g2, ..., gn - множество объектов;

- M = m1, m2, ..., mk - множество признаков;

- I - бинарное отношение между объектами и признаками, представленное в виде матрицы размерности n x k, где элемент i,j равен 1, если объект gi обладает признаком mj, и 0 в противном случае.

Операторы Галуа (соответствие Галуа) – операторы, которые позволяют устанавливать связи между формальными понятиями и отношениями между объектами и признаками в контексте.

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

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

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

Если оператор Галуа применить дважды, то получится оператор замыкания. С его помощью можно определить замкнутые множества объектов: множество объектов $A \subseteq G$ называется замкнутым, если выполняется условие $A'' = A$, Для множеств признаков M замкнутость определяется аналогичным образом.

Формальное понятие (Formal Concept) – это основной элемент анализа формальных понятий, который представляет собой пару (A, B), где:

- A - множество объектов, обладающих определенным набором признаков. Называется множество объёмом формального понятия (или extension);

- B - множество признаков, которыми обладают все объекты из множества A. Называется множество содержанием формального понятия (или intention).

Строгое математическое определение:

Пара множеств (A, B), таких, что $A \subseteq G$, $B \subseteq M$, $A' = B$ и $B' = A$, называется формальным понятием (или концептом) [3].

Формальное понятие может быть интерпретировано следующим образом:

- Все объекты из множества A обладают всеми признаками из множества B;

- Никакой другой объект, не входящий в множество A, не обладает всеми признаками из множества B;

- Никакой другой признак, не входящий в множество B, не имеется одновременно у всех объектов из множества A.

Для множества объектов A рассмотрим множество A'. Оно будет представлять группу их общих признаков и служить для описания сходства объектов из этого множества A.

Если применить оператор Галуа ещё раз, то получим замкнутое множество A'', которое является кластером сходных с множеством общих признаков A' объектов. При этом отношение “быть более общим понятием” задается следующим образом:

$(A, B) \geq (C, D)$ если $A \supseteq C$ и $B \subseteq D$.

Решётка понятий (Concept lattice) – частично упорядоченное множество всех формальных понятий, где операция пересечения и объединения формальных понятий определяет структуру решетки.

Согласно основной теореме анализа формальных понятий, частично упорядоченное по вложению объёмов множество формальных понятий контекста K образует математический объект – решётку, которая называется решёткой понятий [4].

Решетка понятий строится путем упорядочивания всех формальных понятий в контексте по включению. Вершина верхнего уровня решетки представляет все объекты и признаки контекста. Каждая вершина нижнего уровня представляет собой некоторое формальное понятие (A, B), а ребра соединяют данные формальные понятия в соответствии с отношением включения.

Решетка понятий, построенная на формальном контексте, является инструментом для представления и извлечения знаний из данных контекста. В роли знаний выступают понятия, организованные иерархично. При этом граф решетки понятий не является деревом, что характерно для графов многих концептуальных моделей, а имеет более обобщённую структуру. Это позволяет представлять знания, выражающиеся понятиями, которые характеризуются меньшей и большей общностью, а также меньшими и большими объемом и содержанием.

При наличии множества импликаций решетки понятий имеется возможность сформировать в ней определённую систему навигации, позволяющую находить частные и общие понятия для заданного входа – узла решетки. В этом состоит большое преимущество решеток понятий как концептуальных моделей [4].

Решётка понятий позволяет визуализировать структуру зависимостей между формальными понятиями и выявлять иерархические отношения между ними. Однако построение хорошей диаграммы во многих ситуациях, особенно на больших контекстах, является нетривиальной задачей и требует значительных ресурсов [1].

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.

В связи с этим применение классического АФП ограничено из-за сложности использования и недостаточной гибкости. Следовательно, возникла необходимость разработки модификации метода АФП, которая позволила бы упростить процесс анализа данных и расширить на практике область его применения.

Для работы со сложными данными, содержащими, например, множества последовательностей или графы, была разработана модификация классического анализа формальных понятий. Она получила название узорных структур.

Узорная структура – это тройка (G, (D, ∏), δ), где:

- G – множество объектов,
- (D, ∏) – полная полурешётка всевозможных описаний,
- δ:G→D – функция, которая сопоставляет каждому объекту из множества G его описание из D.

Полурешёточная операция ∏ соответствует операции сходства между двумя описаниями.

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

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

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

где:

- $\sqsubseteq$ – 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.

Пример преобразования формального контекста в узорную структуру приведён на рис. 1.

	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)

$\Rightarrow$

	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 1 – 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 [6].