Методы программирования в ограничениях в задачах кластеризации с частичным привлечением учителя

Аспирант: Зуенко О.Н.
Научный руководитель: д.т.н. Олейник А.Г.

Машинное обучение

- Классификация
- Регрессия
- Кластеризация
- Поиск ассоциативных правил
- …

Актуальность работы: Традиционный подход к решению задач Constrainted Clustering состоит в модификации методов локального поиска с учетом пользовательских ограничений, но данный подход позволяет находить лишь локальный оптимум. Актуальность темы диссертации обусловлена потребностью в новых методах систематического и гибридного поиска, позволяющих отыскивать глобальный оптимум в пространствах большой размерности. Анализ прототипов показал, что в рамках технологии программирования в ограничениях имеются предпосылки для разработки подобных методов.

Цель исследования: состоит в разработке эффективных гибридных методов и методов систематического поиска для точного решения задачи кластеризации с частичным привлечением учителя (Constrainted Clustering).

Новизна: представление условий задачи с помощью специализированных матрицеподобных структур (табличных ограничений) и их обработка в рамках парадигмы программирования в ограничениях.

Задачи кластерного анализа

Пусть требуется разбить n объектов

$O=\{o_1, \ldots, o_n\}$, для которых задана матрица расстояний, на k кластеров. Полученное разбиение должно удовлетворять некоторому функционалу F.

Методы:

Вероятностные методы

Методы на основе ИИ

Теоретико-графовые методы

Логические методы

Иерархические методы

…

Задача удовлетворения ограничений (Constraint Satisfaction Problem - CSP)

- Множество переменных x_1, x_2, … x_n
- Множество ограничений c_1, c_2, … c_n
- Домен переменных D_i

Методы:

Методы распространения ограничений – arc-consistency, node-consistency, forward checking, looking ahead, …

Методы систематического поиска – conflict directed backtracking, backjumping, dynamic backtracking, chronological backtracking, …

Предлагаемый подход

1 шаг. Оценить диапазон значений, в который должен попадать искомый оптимальный диаметр разбиения. Для нахождения первоначального разбиения предлагается использовать метод FPF (*Furthest Point First*). Данный приближенный метод позволяет найти оценку для оптимального диаметра разбиения $D \in [d/2, d]$. На основе полученной оценки генерируются ограничения *cannot-link* для тех пар кластеров, для которых $d_{ij} > d$.

2 шаг. Выполнить конкретизацию верхней границы интервала $D \in [d/2, d]$. Для этого осуществляется процедура иерархической кластеризации мультимножеств. Существенная модификация данной процедуры заключается в том, что в ходе кластеризации анализируются ограничения *cannot-link*. Применение данного метода повышает эффективность вычислительных процедур и позволяет сократить перебор вариантов объединения кластеров. В результате данного шага получаем новый интервал для оценки *D*.

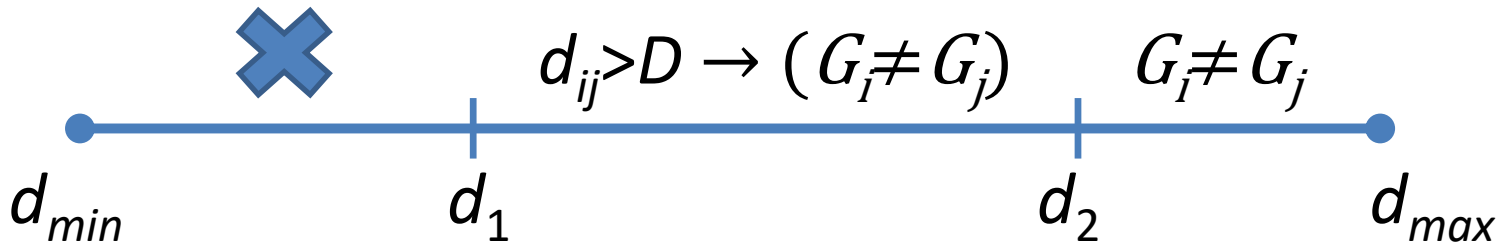
3 шаг. Сгенерировать ограничения для систематического решения задачи CSP. Предыдущие два этапа позволяют генерировать ограничения не для всех пар кластеризуемых объектов, как было описано ранее. Ограничения представляются с помощью табличных ограничений, а именно *smart*-таблиц *D*-типа. Обработка данных ограничений производится с помощью высокоэффективных авторских методов удовлетворения нечисловых ограничений.

4 шаг. Решить сгенерированную на предыдущем шаге задачу *Constrained Clustering* с помощью описанных далее эвристик для поиска переменной и её значения. Предлагаемый метод систематического поиска опирается на следующие эвристики выбора переменной на текущем шаге поиска: выбирается переменная, домен которой содержит наименьшее количество значений. При выборе значения переменной руководствуемся следующим правилом: поскольку переменная представляет один из кластеризуемых объектов, а её значение – номер кластера, то присваиваем переменной номер того кластера, который ближе к рассматриваемому объекту (рассчитываются расстояния между соответствующими мультимножествами).

Сокращение количества ограничений и упрощение их вида

$$d_{ij} > D \rightarrow (G_i \neq G_j)$$

$$D \in [d_1, d_2]$$



Пример

Представление информации о кластеризуемых объектах

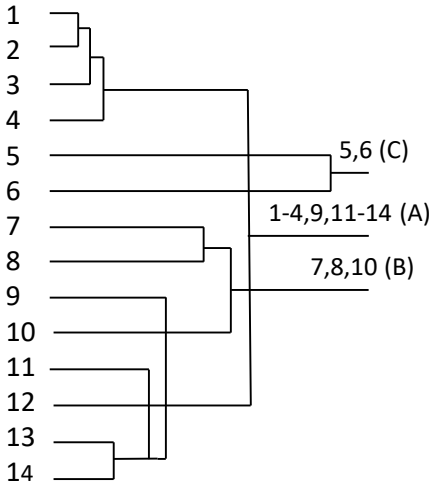
N	*R*1	*R*2	*OP*	*OPv*	*RT*	*VP*	*RT/VP*	*W*	*WB*	*LB*	*N*
1	2 0	2 0	2 0	2 0	2 0	0 2	2 0	2 0	0 2	2 0	2
2	0 2	2 0	2 0	2 0	2 0	0 2	2 0	2 0	0 2	2 0	2
3	6 0	6 0	6 0	6 0	6 0	0 6	6 0	6 0	0 6	6 0	6
…											

Матрица расстояний $d_{11}(o_i, o_j) = \sum_{l=1}^{n} \left| k_{Ai}(x^l) - k_{Aj}(x^l) \right|$

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	0	4	40	84	1062	1264	446	1088	242	442	104	328	10	66
2	4	0	44	80	1066	1262	450	1084	246	438	108	324	10	62
3	40	44	0	52	1046	1272	438	1104	250	454	112	344	60	58
4	84	80	52	0	1050	1260	480	1100	270	450	140	340	100	72
5	1062	1066	1046	1050	0	832	980	1294	1020	1156	1030	1006	1070	1098
6	1264	1262	1272	1260	832	0	1260	936	1260	1088	1250	1132	1260	1288
7	446	450	438	480	980	1260	0	810	200	104	410	386	450	478
8	1088	1084	1104	1100	1294	936	810	0	930	650	1080	1016	1080	1108
9	242	246	250	270	1020	1260	200	930	0	280	210	270	250	278
10	442	438	454	450	1156	1088	104	650	280	0	430	366	430	458
11	104	108	112	140	1030	1250	410	1080	210	430	0	240	100	114
12	328	324	344	340	1006	1132	386	1016	270	366	240	0	320	320
13	10	10	60	100	1070	1260	450	1080	250	430	100	320	0	70
14	66	62	58	72	1098	1288	478	1108	278	458	114	320	70	0

Пример

1 шаг Оценка, полученная методом *FPF* $D \in [465, 930]$.
Диаметр кластера A – 344, диаметр кластера B – 930, диаметр кластера C – 832. Диаметр разбиения – 930.



2 шаг Уточненная оценка после иерархической кластеризации - $D \in [465, 832]$.
Диаметр кластера A – 344, диаметр кластера B – 810, диаметр кластера C – 832.
Диаметр разбиения – 832.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	0	4	40	84	1062	1264	446	1088	242	442	104	328	10	66
2	4	0	44	80	1066	1262	450	1084	246	438	108	324	10	62
3	40	44	0	52	1046	1272	438	1104	250	454	112	344	60	58
4	84	80	52	0	1050	1260	480	1100	270	450	140	340	100	72
5	1062	1066	1046	1050	0	832	980	1294	1020	1156	1030	1006	1070	1098
6	1264	1262	1272	1260	832	0	1260	936	1260	1088	1250	1132	1260	1288
7	446	450	438	480	980	1260	0	810	200	104	410	386	450	478
8	1088	1084	1104	1100	1294	936	810	0	930	650	1080	1016	1080	1108
9	242	246	250	270	1020	1260	200	930	0	280	210	270	250	278
10	442	438	454	450	1156	1088	104	650	280	0	430	366	430	458
11	104	108	112	140	1030	1250	410	1080	210	430	0	240	100	114
12	328	324	344	340	1006	1132	386	1016	270	366	240	0	320	320
13	10	10	60	100	1070	1260	450	1080	250	430	100	320	0	70
14	66	62	58	72	1098	1288	478	1108	278	458	114	320	70	0

Пример

3 шаг

$D \in [465,\ 832]$

$$\begin{cases} d_{4,7} = 480 \\ d_{5,6} = 832 \\ d_{7,8} = 810 \\ d_{7,14} = 478 \\ d_{8,10} = 650 \end{cases}$$

$$\begin{array}{l} (480 > D) \to (G_4 \neq G_7) \\ (832 > D) \to (G_5 \neq G_6) \\ (810 > D) \to (G_7 \neq G_8) \\ (478 > D) \to (G_7 \neq G_{14}) \\ (650 > D) \to (G_8 \neq G_{10}) \end{array} \Leftrightarrow$$

D	G_4G_7	G_5G_6	G_7G_8	G_7G_{14}	G_8G_{10}
≥ 480	$\neq$	$\varnothing$	$\varnothing$	$\varnothing$	$\varnothing$
≥ 832	$\varnothing$	$\neq$	$\varnothing$	$\varnothing$	$\varnothing$
≥ 810	$\varnothing$	$\varnothing$	$\neq$	$\varnothing$	$\varnothing$
≥ 478	$\varnothing$	$\varnothing$	$\varnothing$	$\neq$	$\varnothing$
≥ 650	$\varnothing$	$\varnothing$	$\varnothing$	$\varnothing$	$\neq$

4 шаг

***Утверждение* 1**. Если строка *smart*-таблицы *D*-типа пуста, то таблица пуста.

***Утверждение* 2**. Если все компоненты некоторого атрибута пусты, то данный атрибут можно удалить из *smart*-таблицы *D*-типа.

***Утверждение* 3**. Если в *smart*-таблице *D*-типа есть строка, содержащая лишь одну непустую компоненту, то все кванты, не входящие в эту компоненту, удаляются из соответствующего домена.

Утверждение 4. Если строка *smart*-таблицы *D*-типа содержит хотя бы одну полную компоненту, то она удаляется.

Утверждение 5. Если компонента атрибута *smart*-таблицы *D*-типа содержит квант, не принадлежащий соответствующему домену, то квант удаляется из компоненты.

***Утверждение* 6**. Если в *smart*-таблице *D*-типа усечён один или несколько доменов простых атрибутов, которые формируют некоторый составной атрибут, то: из домена составного атрибута исключаются кванты, которые обращаются в пустое множество при новых доменах соответствующих простых атрибутов.

Утверждение 7. В случае конкретизации домена сложного атрибута должны быть конкретизированы и домены соответствующих простых атрибутов.

Решения

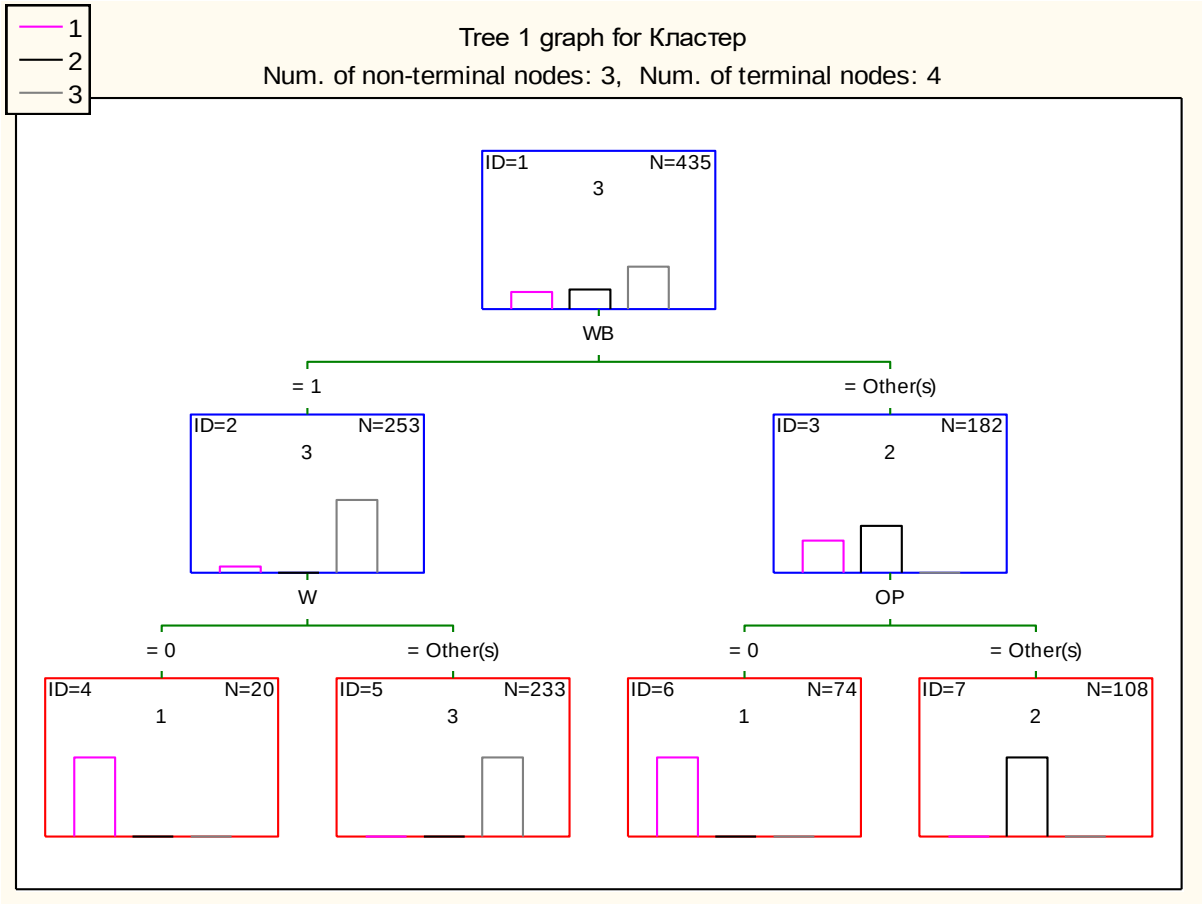
1 решение: в класс A попадают объекты 1, 2, 3, 4, 9, 10, 11, 12, 13, 14; в класс B – объекты 7, 8; в класс C – объекты 5, 6.

2 решение: в класс A попадают объекты 1; 2; 3; 4; 7; 9; 11; 12; 13; 14; в класс B – объекты 8, 10; в класс C – объекты 5, 6.

3 решение: в класс A попадают объекты 1; 2; 3; 4; 9; 11; 12; 13; 14; в класс B – объекты 7, 8, 10; в класс C – объекты 5, 6.

Ячейка	*R*1	*R*2	*OP*	*OPv*	*RT*	*VP*	*RT*/*VP*	*W*	*WB*	*LB*	*N*	Кластер
1	0	0	0	0	0	1	0	0	1	0	2	*A*
2	1	0	0	0	0	1	0	0	1	0	2	*A*
3	0	0	0	0	0	1	0	0	1	0	6	*A*
4	1	0	0	0	0	1	0	0	1	0	10	*A*
5	0	0	0	0	0	0	1	1	1	0	107	*C*
6	1	0	1	1	0	0	1	1	1	0	126	*C*
7	0	0	0	1	1	0	0	1	0	0	45	*A, B*
8	1	0	1	1	1	0	0	1	0	0	108	*B*
9	0	0	0	1	1	0	0	1	0	0	25	*A*
10	1	0	1	1	1	0	0	1	0	0	43	*A, B*
11	0	0	0	1	0	0	1	1	0	1	10	*A*
12	1	0	0	1	0	0	1	1	0	1	32	*A*
13	0	1	0	0	0	1	0	0	0	1	0	*A*
14	1	1	0	0	0	1	0	0	0	1	7	*A*

Правила



1 правило: Если ячейка относится к висячему боку и там имеются выработки, то уровень сейсмической активности оценивается как высокий.

2 правило: Если ячейка относится к висячему боку и там отсутствуют выработки, то уровень сейсмической активности оценивается как низкий.

3 правило: Если ячейка не относится к висячему боку и в ней располагается граница очистного пространства текущего горизонта, то уровень сейсмической активности оценивается как средний.

4 правило: Если ячейка не относится к висячему боку и в ней не располагается граница очистного пространства текущего горизонта, то уровень сейсмической активности оценивается как низкий.

Спасибо за внимание