Секция 2. ГЕОМЕТРИЯ И АНАЛИЗ

УДК 519.25

Визуализация иерархических кластерных алгоритмов

В.Н. Андреева, С.В. Дронов
АлтГУ, Барнаул

Иерархией $\mathcal{F}$ на множестве $X = \{X_1,...,X_n\}$ называется система подмножеств $\{S : S \subset X\}$, такая что:

1. $X \in \mathcal{F}$;
2. $\{X_i\} \in \mathcal{F}$, $i = 1,...,n$;
3. Если S и S' из $\mathcal{F}$ имеют непустое пересечение, то $S' \subset S$ либо $S \subset S'$.

Из третьего свойства следует, что каждый элемент иерархии, кроме $\{X_i\}$, получается объединением каких классы k-то других («более простых») ее элементов.

Иерархии, которые рассматриваются в работе, предполагаются порожденными некоторым алгоритмом слияния подмножеств X. Он начинается с ситуации, когда каждый элемент X образует отдельное множество. Как правило, алгоритм этот является бинарным, то есть, на каждом его шаге сливаются (объединяются) два каких-то множества, превращаясь в одно. Тем самым, на каждом шаге алгоритма число рассматриваемых подмножеств уменьшается (для бинарного алгоритма на 1).

Этот алгоритм будем далее называть агломерацией. Его действие можно себе представить следующим образом: сначала мы «различаем» каждый из элементов X. Затем «порог различимости» понемногу уменьшается. При этом некоторые, наиболее близкие друг к другу элементы, перестают различаться и сливаются в один объект. Процесс продолжается до момента, когда все элементы X сольются в один объект и полностью прекратят различаться.

Уровень иерархии $\mathcal{F}$ – очередной этап слияния двух ее элементов в новый. Более точно: агломерация начинается с ситуации, когда каждый элемент X находится в собственном подмножестве (нулевой уровень иерархии). После совершения k-го шага агломерации мы получим набор подмножеств, образующих ее k-й уровень. Этот уровень

обязательно будет кластерным разбиением исходного множества. Таким образом, каждый уровень иерархии получает свой естественный номер. При этом, если k-й уровень обозначить через $\mathcal{F}_k$, то, очевидно, $\mathcal{F}$ представляет собой объединение всех $\mathcal{F}_k$.

Индексация агломерации – это отображение $\vartheta : \mathcal{F} \to R$ ставящее в соответствие множеству $S \in \mathcal{F}$ неотрицательное число $\vartheta(S)$ таким образом, что:

1) $\vartheta(S) = 0$ тогда и только тогда, когда S состоит из одного элемента;

2) $\vartheta(S') \le \vartheta(S)$ для каждой пары (S, S') из $\mathcal{F}$ такой, что $S' \subset S$

Неформально можно воспринимать значение индексации на множестве S как время, прошедшее до возникновения S с момента начала действия агломерации. Еще не более двух десятилетий назад считалось, что достаточно полагать время, затрачиваемое на каждый шаг алгоритма, равным одной и той же величине, например, единице. Получаемую таким образом индексацию назовем традиционной пошаговой. Но такая индексация не может нас устроить, поскольку ясно, что степень понижения порога различимости, приводящая к образованию нового уровня иерархии, на разных шагах может различаться очень сильно.

Ранее в работах [1–2] рассматривались некоторые способы построения индексаций, реализующие подобные идеи. Но даже самый удачный из них (см. [2, с. 261]) на наш взгляд является неестественным, поскольку при его использовании иногда алгоритм объединяет не те множества, которые интуитивно кажутся наиболее близкими.

Новый метод индексации 1

Пусть $\mathcal{F}_0$ – нулевой уровень иерархии, то есть кластерное разбиение, в котором каждый элемент X образует собственный кластер. В работах [3–4] предложен коэффициент кластерных различий K, который оценивает степень различия двух кластерных разбиений одного и того же множества из n элементов. Он оказывается тем больше, чем меньше это разбиение похоже на $\mathcal{F}_0$. Почти очевидно, что если два кластера сливаются в один, а остальные не меняются, то вновь образующийся уровень иерархии будет сильнее предыдущего отличаться от $\mathcal{F}_0$ с точки зрения величины этого коэффициента. Строго это обосновывается с помощью приведенной ниже теоремы 1.

Перейдем к деталям. Пусть есть кластерные разбиения одного и того же множества X, которые мы обозначим $\mathcal{A}$ и $\mathcal{B}$.

Для $x \in X$ через A_x мы обозначаем тот кластер в $\mathcal{A}$, в который входит x, а через B_x – соответствующий кластер в $\mathcal{B}$. Положим

$$Q(\mathcal{A},\mathcal{B}) = \sum_{x\in X} |A_x \Delta B_x|, \quad Q_k = Q(\mathcal{F}_k, \mathcal{F}_0),$$

где $|A_x \Delta B_x|$ – число элементов симметрической разности множеств A_x и B_x.

Теорема 1. *Для произвольного k справедливо неравенство $Q_k < Q_{k+1}$, а следовательно, отображение, которое каждому из множеств $\mathcal{F}_k$ ставит в соответствие число Q_k, является индексацией.*

Отметим, что $Q_k = n(n-1)K(\mathcal{F}_k, \mathcal{F}_0)$, где K – коэффициент кластерных различий. Ясно также, что все числа Q_k являются целыми. А это значит, что такая индексация близка к традиционной пошаговой.

Новый метод индексации 2

Второй метод индексации, предлагаемый нами, похож на метод Айвазяна – Жамбю [2], но представляется более естественным.

Полагая для одноэлементных множеств S $V^*(S) = 0$, введем для множеств следующих за нулевым уровней иерархии

$$V^*(S_1 \cup S_2) = \sum_S \|X - Z\|^2,$$

если $S = S_1 \cup S_2$, а Z – центр множества S. Индукцией по уровню индексации несложно доказывается.

Теорема 2. $V^*(S_1 \cup S_2) > V^*(S_1) + V^*(S_2)$, *то есть V^* является индексацией.*

В полном тексте работы представлены также простые формулы, позволяющие находить значения обеих предложенных индексаций. Они имеют итеративный характер.

Библиографический список

1. Жамбю М. Иерархический кластер-анализ и соответствия. – М.: Финансы и статистика, 1988.

2. Айвазян С.А., Бухштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Классификация и снижение размерности. – М.: Финансы и статистика, 1989.

3. Дронов С.В. Одна кластерная метрика и устойчивость кластерных алгоритмов // Известия АлтГУ. – 2011. – Вып. 1. – С. 32–35.

4. Dronov S.V., Dementjeva E.A. A new approach to post-hoc problem in cluster analysis // Model Assisted Statistics and Applications. – 2012. – Vol. 7, № 1. – P. 49–55.

УДК 510.223

Единственность сегмента класса N, исчерпываемого всеми своими подсегментами

С.В. Дронов
АлтГУ, г. Барнаул

Работа выполнена в аксиоматике альтернативной теории множеств (AST), подробное изложение основных положений которой можно найти в [1]. В этой теории вслед за множествами, образующими нулевой уровень сложности объектов и теоретико-множественными классами (*Sd*-классы, первый уровень) следуют так называемые σ- и π-классы. Приведем два этих определения. Пусть, как обычно, через **FN** обозначен класс всех конечных натуральных чисел. Если $X_n, n \in \mathbf{FN}$ – какая-то цепочка теоретико-множественных классов, то класс $X = \bigcup\{X_n, n \in \mathbf{FN}\}$ называют σ-классом, а $Y = \bigcap\{X_n, n \in \mathbf{FN}\}$, соответственно, π-классом. Такие классы подробно изучены в [1] и на них основаны многие фундаментальные конструкции AST.

Будем пользоваться также понятием сегмента (начального отрезка) класса натуральных чисел **N.** Это понятие неоднократно вводилось и обсуждалось в работах автора. Известно, что если сегмент теоретико-множественно определим (*Sd*-сегмент), то он, либо совпадает со всем классом **N**, либо является натуральным числом.

В настоящей работе делается попытка заменить класс **FN** в определениях σ- и π-классов на некоторый более широкий сегмент C, который, таким образом, играет роль более далекого горизонта, чем ближайший, ограничиваемый в AST конечными натуральными числами. Конечно же, любой разумный горизонт не может быть четким, поэтому для всех рассматриваемых ниже сегментов C мы потребуем, чтобы они являлись последовательными, то есть

$$(\forall n \in \mathbf{N})\,(n \in C) \Rightarrow (n+1 \in C).$$

Следуя [2], сегмент A класса натуральных чисел **N** далее будем называть C-исчерпываемым, если найдется неубывающая последова-