5. *Dobbertin H.* The first two rounds of md4 are not one-way // LNCS. 1998. V. 1372. P. 284–292.
6. *De D., Kumarasubramanian A., and Venkatesan R.* Inversion attacks on secure hash functions using SAT solvers // LNCS. 2007. V. 4501. P. 377–382.
7. *Gribanova I., Zaikin O., Otpuschennikov I., and Semenov A.* Using parallel SAT solving algorithms to study the inversion of MD4 hash function // Параллельные вычислительные технологии. XI Междунар. конф. ПаВТ'2017, г. Казань, 3–7 апреля 2017 г. Короткие статьи и описания плакатов. Челябинск: Издательский центр ЮУрГУ, 2017. С. 100–109.
8. *Otpuschennikov I., Semenov A., Gribanova I., et al.* Encoding cryptographic functions to SAT using TRANSALG system // ECAI 2016 — 22nd European Conference on Artificial Intelligence. Frontiers in Artificial Intelligence and Applications. 2016. V. 285. P. 1594–1595.
9. *Biere A.* Lingeling essentials. A tutorial on design and implementation aspects of the the SAT solver lingeling // Proc. Fifth Pragmatics of SAT Workshop. 2014. V. 27. P. 88.
10. `http://hpc.icc.ru` — Иркутский суперкомпьютерный центр СО РАН. Иркутск: ИДСТУ СО РАН.

УДК 519.14+519.25 DOI 10.17223/2226308X/10/62

РАНЖИРОВАНИЕ ПОКАЗАТЕЛЕЙ, ФОРМИРУЮЩИХ КЛАСТЕРНОЕ РАЗБИЕНИЕ, НА ОСНОВЕ КОЭФФИЦИЕНТОВ ОТНОСИТЕЛЬНОГО СХОДСТВА

С. В. Дронов, Е. А. Евдокимов

Рассматривается задача установления относительной информационной ценности числовых показателей, по близости значений которых производится разбиение конечного множества объектов на кластеры. Вводится коэффициент для оценки относительной силы влияния на вид кластерного разбиения каждого из показателей по сравнению с одним или произвольной совокупностью остальных, а также два коэффициента, позволяющих с разных сторон оценить степень связи двух показателей по отношению к этой структуре (кластерная связь). Предложен новый алгоритм сокращения размерности данных на основе этих коэффициентов, в наибольшей степени оставляющий неизменной кластерную структуру исходного множества объектов. Степень искажения оценивается с использованием кластерной метрики, ранее предложенной одним из авторов. Путём реализации этого алгоритма может быть достигнуто более уверенное распознавание угроз компьютерной безопасности при общем снижении нагрузки на систему.

Ключевые слова: *кластерное разбиение, сокращение размерности, кластерная связь, коэффициент силы связи.*

Рассмотрим задачу разбиения конечного множества объектов на кластеры по степени близости совокупностей показателей, которые в этом контексте будем называть формирующими. Нас будет интересовать только результат разбиения, причём договоримся считать, что по совокупности всех рассматриваемых показателей кластеризация объектов производится абсолютно правильно. Мы хотим определить сравнительную силу формирующих показателей по степени их влияния на кластеры. Кроме этого, некоторые из показателей могут быть схожи между собой до такой степени, что использование их вместе совсем не требуется. Такую схожесть показателей для кластерного анализа данных назовём кластерной связью. Силу этой связи тоже можно оценивать с помощью определённых числовых коэффициентов.

Подобные разновидности задачи сокращения размерности данных, по сути являющиеся вариантами post-hoc анализа кластерных разбиений, могут находить примене-

ние для более надёжной классификации компьютерных угроз и за счёт уменьшения количества определяющих показателей увеличивать скорость их распознавания, а также в задачах медицинской диагностики. Исследования были начаты одним из авторов в [1, 2], но в этих работах сравнение различных кластерных разбиений производилось не относительно друг друга, а относительно некоторых предельных разбиений, которые в практических задачах никогда не встречаются. К тому же в этих работах понятие кластерной связи не вводилось вовсе.

Далее предполагается, что все без исключения формирующие показатели обязательно были учтены при построении правильного кластерного разбиения и никакие процедуры их взвешивания или исключения не применялись. Так бывает, например, при использовании иерархических кластерных алгоритмов. Выразимся точнее:

Основное предположение. *Два объекта признаются близкими по совокупности нескольких показателей тогда и только тогда, когда они признаются близкими и по каждому из показателей, участвующих в совокупности.*

Пусть множество изучаемых объектов состоит из n элементов. Рассмотрим два показателя X, Y и три кластерных разбиения $\mathcal{A}, \mathcal{B}, \mathcal{C}$. При этом первое из них построено с учётом исключительно близости значений показателя X, второе — близости значений Y, а для построения третьего использована близость совокупностей значений этих показателей. Пусть количества кластеров, составляющих каждое из разбиений, равны k, m, f соответственно:

$$\mathcal{A} = \{A_1, \ldots, A_k\}, \quad \mathcal{B} = \{B_1, \ldots, B_m\}, \quad \mathcal{C} = \{C_1, \ldots, C_f\}.$$

Согласно основному предположению, каждый из кластеров первого и второго разбиений составлен из кластеров третьего разбиения, как из кирпичиков. Из этого предположения легко выводятся формулы

$$d(\mathcal{A}, \mathcal{C}) = \sum_{i=1}^{k} |A_i|^2 - \sum_{j=1}^{f} |C_j|^2, \; d(\mathcal{B}, \mathcal{C}) = \sum_{i=1}^{m} |B_i|^2 - \sum_{j=1}^{f} |C_j|^2, \tag{1}$$

где d — расстояние между кластерными разбиениями, введённое в [1], а через $|A|$ обозначено число элементов конечного множества A.

Теорема 1. Имеет место равенство

$$d(\mathcal{A}, \mathcal{B}) = d(\mathcal{A}, \mathcal{C}) + d(\mathcal{C}, \mathcal{B}).$$

Если мы договоримся представлять все возможные кластерные разбиения множества объектов точками некоторого метрического пространства с метрикой d, то точка, соответствующая разбиению $\mathcal{C}$ по совокупности показателей, расположена в этом пространстве на отрезке, соединяющем индивидуальные разбиения $\mathcal{A}$ и $\mathcal{B}$. Поэтому можно оценить взаимную силу показателей по её расположению на этом отрезке, для чего определим коэффициент кластерной силы показателя X в паре X, Y формулой

$$Q_{X,Y}(X) = 1 - \frac{d(\mathcal{A}, \mathcal{C})}{d(\mathcal{A}, \mathcal{B})} = \frac{d(\mathcal{B}, \mathcal{C})}{d(\mathcal{A}, \mathcal{B})}, \; d(\mathcal{A}, \mathcal{B}) \neq 0,$$

и равным 1 иначе. Чем ближе этот коэффициент к 1, тем сильнее X по отношению к Y. В случае, когда он равен 1, влиянием второго показателя на вид кластерного разбиения можно полностью пренебречь. Следует отметить, что этот коэффициент после

нормировки совпадает с функцией конкурентного сходства FRiS [3] взаимодействия показателей с X в конкуренции с Y в рассматриваемом частном случае.

Из формулы (1) видно, что расстояние между совместным и индивидуальными разбиениями полностью определяется суммой квадратов количеств элементов в кластерах совместного разбиения. Обозначим её $q = q(f)$. Поскольку сумма самих $|C_j|$ постоянна и равна n, то q может оказаться тем большим, чем меньше число кластеров f. С другой стороны, при фиксированном f величина $q(f)$ оказывается самой большой, когда все кластеры в $\mathcal{C}$, кроме одного, содержат ровно по одному элементу, и самой маленькой, когда все эти кластеры содержат одинаковое число элементов — естественно, это возможно только если n нацело делится на f. Имеет место

Теорема 2. Пусть выполнено основное предположение. Тогда

1) для заданных k, m для числа кластеров совместного разбиения f верно

$$f_{\min} = \max\{k, m\} \leqslant f \leqslant km = f_{\max};$$

2) при фиксированном f выполнено ($[.]$ — целая часть числа)

$$q_{\min} = (2[n/f] + 1)n - f[n/f]([n/f] + 1) \leqslant q \leqslant (n - f + 1) + f - 1 = q_{\max}.$$

При этом неравенства обоих утверждений являются неулучшаемыми, т. е. в них могут достигаться равенства.

Следовательно, можно характеризовать силу кластерной связи, под которой мы понимаем способность показателей замещать друг друга в процессе кластеризации, с разных сторон при помощи двух коэффициентов:

$$K_1(X, Y) = \frac{f_{\max} - f}{f_{\min} - f_{\max}}, \quad K_2(X, Y) = \frac{q - q_{\min}}{q_{\max} - q_{\min}}.$$

Оба этих коэффициента принимают значения от 0 до 1 и тем больше по величине, чем более сильной является кластерная связь между показателями, хотя прямой зависимости между K_1 и K_2 не существует.

В работе обсуждается соотношение между введёнными коэффициентами, возможности их видоизменения, а также предложены некоторые алгоритмы снижения размерности данных для кластерного анализа на основе ранжирования формирующих показателей по величине коэффициентов кластерной силы и степени кластерных связей между ними.

ЛИТЕРАТУРА

1. *Дронов С. В.* Одна кластерная метрика и устойчивость кластерных алгоритмов // Известия АлтГУ. 2011. № 1/2. С. 32–35.
2. *Dronov S. V. and Dementjeva E. A.* A new approach to post-hoc problem in cluster analysis // Model Assisted Statistics and Applications. 2012. No. 1. P. 49–65.
3. *Загоруйко Н. Г., Кутненко О. А.* Цензурирование обучающей выборки // Вестник ТГУ. Управление, вычислительная техника и информатика. 2013. № 1 (22). С. 66–73.