УДК 004.855.5 DOI: 10.17587/it.24.642-648

В. А. Харахинов, аспирант, e-mail: tes4obse@mail.ru,
Иркутский национальный исследовательский технический университет

Генетический алгоритм как альтернатива обучения слоя Кохонена

Предложено применение генетического алгоритма в качестве альтернативного метода обучения слоя Кохонена. Приведено сравнение оценок качества кластерного анализа, полученных следующими методами: K-means; слой Кохонена, обученный стандартным алгоритмом; слой Кохонена, обученный с помощью генетического алгоритма.

Качество кластерного анализа определялось с помощью двух индексов: индекса Рэнда и отрегулированного индекса Рэнда.

Приведено сравнение временных затрат при обучении слоя Кохонена стандартным алгоритмом и генетическим алгоритмом.

Анализ был проведен на наборе экземпляров банковских банкнот, описываемых набором числовых признаков, полученных из образов двух типов карт: реальных и фальшивых.

Ключевые слова: *кластерный анализ, K-means, слой Кохонена, генетический алгоритм, индекс Рэнда*

Введение

В настоящее время существует множество алгоритмов, реализующих задачу кластерного анализа. Технику кластеризации можно применять в самых различных прикладных областях, в том числе и для контроля подлинности банковских банкнот.

Существуют классические методы кластерного анализа, давно ставшие популярными в анализе данных — это использование метода K-means или слоя Кохонена. Однако необходимо расширять набор алгоритмов для повышения качества кластеризации и сокращения временных затрат.

Автором было предложено использование генетического алгоритма для обучения сети, а именно настройку весов для слоя Кохонена с возможностью в дальнейшем решать задачу кластерного анализа.

Программная реализация проводилась в среде разработки MATLAB, были использованы такие расширения, как Neural Network Toolbox, Global Optimization Toolbox, а также неофициальный Exploratory Data Analysis Toolbox [1].

Кластерный анализ

Кластерный анализ (кластеризация) предназначен для разбиения совокупности объектов на однородные группы (кластеры). В результате применения различных методов кластеризации могут быть получены неодинаковые результаты.

Общая постановка задачи кластеризации выглядит следующим образом.

Пусть X — множество объектов, Y — множество номеров кластеров. Задана функция расстояния между объектами $p(x, x')$. Имеется конечная обучающая выборка объектов $X^m = \{x_1, ..., x_m\} \in X$. Требуется разбить выборку на непересекающиеся подмножества, называемые кластерами, так, чтобы каждый кластер состоял из объектов, близких по метрике p, а объекты разных кластеров существенно (по выбранной метрике) отличались. При этом каждому объекту $x_i \in X^m$ приписывается номер кластера y_j [2].

Слой Кохонена и K-means

Алгоритм работы слоя Кохонена схож с известным алгоритмом K-means, их объединяет следующее.

1. Необходимо заранее определить и задать число кластеров.

2. Алгоритмы сходятся, если:

2.1) выполняется один из основных критериев останова алгоритма обучения сети — не

произошло значимого изменения весовых коэффициентов в пределах заданной точности на протяжении последней эпохи обучения;

2.2) в случае K-means алгоритм завершается, когда координаты центров кластеров на текущей итерации не отличаются от соответствующих координат на предыдущей итерации алгоритма.

Отличительной чертой процесса обучения данной сети от процессов обучения многих других видов сетей является то, что необходимо настроить веса синапсов нейронов, а не минимизировать ошибку обучения.

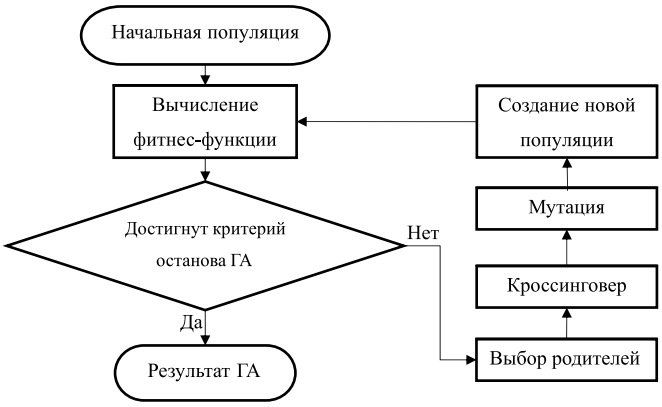

Рис. 1. Блок-схема ГА

Генетический алгоритм

Генетический алгоритм (ГА) — это один из методов решения оптимизационных задач. Как известно, оптимизационные задачи заключаются в нахождении минимума (максимума) заданной функции. Такую функцию называют функцией приспособленности (фитнес-функция).

Предложенный Джоном Холландом в 1975 г. генетический алгоритм основан на принципах естественного отбора Ч. Дарвина и является эвристическим; он в ряде случаев дает результаты более эффективно, чем классические методы оптимизации.

В генетическом алгоритме используется как механизм генетического наследования, так и аналог естественного отбора. При этом сохраняется биологическая терминология в упрощенном виде и применяются основные понятия линейной алгебры.

Одно пробное решение называется особью, а набор всех пробных решений — популяцией.

Как известно, принцип естественного отбора заключается в том, что в конкурентной борьбе выживает наиболее приспособленный. В нашем случае приспособленность особи определяется фитнес-функцией: чем меньше значение функции, тем более приспособленной является особь, т. е. пробное решение, использовавшееся в качестве аргумента фитнес-функции.

В генетическом алгоритме создание новой популяции реализуется путем применения селекции и генетических операторов (оператора скрещивания и оператора мутации).

Процесс размножения состоит в селекции родительских пар для скрещивания таким образом, чтобы решения (особи) в новой популяции были ближе к искомому глобальному минимуму функции приспособленности. Следующим шагом в работе генетического алгоритма являются мутации, т. е. случайные изменения полученных в результате скрещивания хромосом. Мутации способны улучшить или ухудшить приспособленность особи-потомка.

Основные принципы работы генетического алгоритма (ГА) заключены в следующей схеме (рис. 1) [3].

Из приведенной блок-схемы видно, что вычисление функции приспособленности выполняется для каждой генерируемой популяции, начиная с начальной. От полученных значений функции приспособленности зависит дальнейшая работа генетического алгоритма: если критерий останова удовлетворен, то алгоритм прекращает работу; если критерий не был достигнут, то на основе значений функции приспособленности создается новая популяция и цикл повторяется.

В данной работе формирование начальной популяции осуществлялось с помощью псевдослучайного выбора заданного числа особей.

В проводимом исследовании стояла задача минимизации функции приспособленности.

Определение критерия останова ГА зависит от его конкретного применения. Основные критерии:

1) достижение предельного значения функции приспособленности;

2) формирование новой популяции не приводит к значимому улучшению значения функции приспособленности;

3) достигнут лимит на время выполнения ГА, либо лимит на число поколений.

В данном исследовании использовалось два критерия — 1 и 3.

Процесс селекции родителей реализован в соответствии с алгоритмом стохастической универсальной селекции Бэкера, так как для достижения высокого уровня стабильности результатов ГА предпочтительнее использовать именно эту селекцию [4].

Скрещивание родительских особей происходило согласно алгоритму *scatter*-скрещивания при $P_{sc} = 0{,}8$, где P_{sc} — вероятность скрещивания.

Процесс мутации был организован по принципу гауссовской мутации, согласно которому случайно выбранное с помощью распределения Гаусса число добавляется к каждому элементу родительского вектора [4].

Объединение ГА и нейросетевого подхода при решении задачи кластерного анализа

Объединение генетических алгоритмов и нейронных сетей известно в литературе под аббревиатурой COGANN. Объединение может быть независимым, вспомогательным, либо равноправным. При независимом объединении генетические алгоритмы и нейронные сети используют по отдельности для решения той или иной задачи (никак не взаимодействуют друг с другом). Вспомогательное объединение этих двух методов означает, что они применяются последовательно один за другим, дополняя друг друга. Например, ГА может провести анализ обученной сети, либо выполнить поиск оптимального набора параметров для обучения сети. В случаях равноправного применения ГА можно использовать для выбора топологии сети, обучения сети [5].

Как уже говорилось ранее, отличительной чертой процесса обучения слоя Кохонена является то, что необходимо настроить веса синапсов нейронов, а не минимизировать ошибку обучения. Иными словами, для того чтобы получить обученную сеть необходимо найти центры кластеров, координаты которых, в свою очередь, составят матрицу весов синапсов нейронов, тем самым исключается потребность в проведении обучения сети классическим алгоритмом Кохонена.

Как уже говорилось ранее, алгоритм работы слоя Кохонена схож с известным алгоритмом K-means.

В случае K-means распространен критерий — минимизация суммы квадратов расстояний от точек до центров кластеров, к которым они относятся [6]. То есть задачу кластерного анализа методом K-means можно свести к задаче минимизации функции

$$F = \sum_{i=1}^{k} \sum_{x_j \in S_i} (x_j - c_i)^2,$$

где k — число кластеров; S_i — полученные кластеры (подмножества множества всех анализируемых наблюдений); x_j — наблюдения; c_i — центры кластеров [7, 8].

В данном случае целью ГА будет минимизация приведенной выше функции F и нахождение глобальных экстремумов.

На основании приведенного ранее критерия минимизации была сформирована функция приспособленности. Однако сравнение принципов работы алгоритма K-means с принципами работы ГА в качестве алгоритма обучения слоя Кохонена выявляет ряд особенностей (табл. 1).

Таблица 1

Сравнение этапов работы алгоритмов

Особенность	Алгоритм K-means	Сформированная функция приспособленности
Начало работы алгоритма	Определение координат центров кластеров	Формирование начальной популяции (вектор, содержащий координаты центров кластеров)
Проверка останова алгоритма	Координаты центров кластеров на текущей итерации не отличаются от соответствующих координат на предыдущей итерации алгоритма	Формирование новой популяции не приводит к значимому улучшению значения фитнес-функции
Изменение координат центров кластеров	Перерасчет координат происходит по четко определенной формуле	Выполняется ряд операций: селекция, скрещивание, мутация и формирование новой популяции — нового вектора, содержащего координаты центров кластеров

Оценка качества кластерного анализа

Задача оценки качества кластеризации является более сложной по сравнению с задачей оценки качества классификации. Во-первых, такие оценки не должны зависеть от номеров кластеров, а зависеть только от самого разби-

ения выборки. Во-вторых, не всегда известны истинные номера кластеров объектов, поэтому также нужны оценки, позволяющие оценить качество кластеризации, используя только неразмеченную выборку.

Наиболее популярными метриками качества являются: индекс Рэнда — Rand Index (RI); отрегулированный индекс Рэнда — Adjusted Rand Index (ARI). В обоих случаях предполагается, что известны истинные номера кластеров объектов. ARI не зависит от самих номеров кластеров, а зависит только от разбиения выборки на кластеры.

Пусть n — число объектов в выборке; A — вектор истинных номеров кластеров каждого объекта; B — вектор номеров кластеров, полученный алгоритмом кластеризации. Обозначим a — число пар объектов, имеющих одинаковые номера кластеров в векторах A и B и находящихся в одном кластере; b — число пар объектов, имеющих различные метки и находящихся в разных кластерах. Тогда Rand Index (RI)

$$RI = \frac{a+b}{\binom{n}{2}} = \frac{2(a+b)}{n(n-1)}.$$

То есть это доля объектов, для которых эти разбиения (исходное и полученное в результате кластеризации) "согласованы" [9].

В отличие от RI, ARI не зависит от самих значений номеров кластеров и перестановок этих номеров, а зависит только от разбиения выборки на кластеры.

Для вычисления ARI необходимо построить таблицу сопряженности (табл. 2). В таблице $X = \{X_1, X_2, ..., X_r\}$ и $Y = \{Y_1, Y_2, ..., Y_s\}$ — результаты кластерного анализа, либо результат анализа и целевое разбиение; r — число кластеров, полученное методом X; s — число кластеров, полученное методом Y (либо число s заранее известно, так как известно целевое разбиение).

Таблица 2

Таблица сопряженности

X \ Y	Y_1	Y_2	...	Y_s	*Sums*
X_1	n_{11}	n_{12}	...	n_{1s}	a_1
X_2	n_{21}	n_{22}	...	n_{2s}	a_2
⋮	⋮	⋮	⋱	⋮	⋮
X_r	n_{r1}	n_{r2}	...	n_{rs}	a_r
Sums	b_1	b_2	...	b_s	

Значения n_{ij} — число объектов, пересекающихся во множествах X_i и Y_j: $n_{ij} = |X_i \cap Y_j|$. Значение $a_r = \sum_{i=1}^{s} n_{ri}$, значение $b_s = \sum_{j=1}^{r} n_{js}$.

ARI вычисляется по формуле

$$ARI = \frac{\sum_{ij}\binom{n_{ij}}{2} - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] / \binom{n}{2}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \left[\sum_i \binom{a_i}{2}\sum_j \binom{b_j}{2}\right] / \binom{n}{2}}.$$

Данный индекс является мерой расстояния между различными разбиениями выборки. ARI принимает значения в диапазоне [–1, 1]. Отрицательные значения соответствуют "независимым" разбиениям на кластеры, значения, близкие к нулю, — случайным разбиениям, и положительные значения говорят о том, что два разбиения схожи (совпадают при ARI = 1) [10].

Приведенный ниже пример наглядно демонстрирует вычисление ARI.

Пусть вектора X и Y имеют следующие значения:

X	1	2	3	3	2	1	1	3	3	1	2	2
Y	3	2	3	2	2	1	1	2	3	1	3	1

Все анализируемые объекты разделены на три кластера.

Строится таблица сопряженности (табл. 3)

Имея все значения n_{ij}, a_i, b_j, легко рассчитать значение отрегулированного индекса Рэнда.

В данном случае

$$ARI = \frac{6 - [18 \times 18] / \binom{12}{2}}{\frac{1}{2}[18+18] - [18 \times 18] / \binom{12}{2}} = 0,08333.$$

Чем больше значения RI и ARI, тем больше соответствие между полученными и истинными кластерами.

Таблица 3

Таблица сопряженности

X \ Y	Y_1	Y_2	Y_3	*Sums*
X_1	3	0	1	4
X_2	1	2	1	4
X_3	0	2	2	4
Sums	4	4	4	

Исходные данные

Анализируемые данные были получены из работы [11]. Исходная выборка содержит 1372 объекта, разделенных на две группы — подлинные и фальшивые купюры. Путем использования различных методов вейвлет-преобразований создатели данной выборки из множества изображений купюр получили матрицу, содержащую информацию об исследуемых банкнотах. В этой матрице каждый объект описывается четырьмя признаками: дисперсия изображения; коэффициент асимметрии изображения; коэффициент эксцесса изображения; энтропия изображения.

Достоинства данной выборки заключаются в следующем.

1. Относительно большой объем анализируемых данных при малом числе кластеров, что делает процесс работы алгоритмов на данной выборке не столь затратным по объему производимых вычислений, сохраняя при этом достаточное число объектов для обучения сети.

2. Наличие целевого вектора (вектора, содержащего истинные номера кластеров объектов).

3. Поскольку создатели выборки получали данные из реальных источников (подлинные и фальшивые купюры), то результаты работы алгоритмов, реализующих кластерный анализ, будут представлять научный интерес.

Результаты кластерного анализа

Как уже говорилось ранее, генетический алгоритм является эвристическим алгоритмом, исходя из этого более корректно будет резюмировать результаты кластерного анализа из ряда проведенных экспериментов. В данной работе было проведено 25 экспериментов, в каждом из которых кластерный анализ проводился несколькими методами.

На рис. 2 отображены значения индексов Рэнда, вычисленные по результатам экспериментов.

Аналогично, на рис. 3 отображены значения отрегулированных индексов Рэнда.

Чем больше значения индексов, тем лучше качество кластерного анализа. Из рис. 2, 3 можно сделать вывод, что худший результат по перечисленным индексам у слоя Кохонена, обученного стандартным алгоритмом.

Слой Кохонена, обученный с помощью генетического алгоритма, имеет широкий диапазон значений индексов. На приведенных рис. 2, 3 в 8 из 25 экспериментов значения обоих индексов были выше (в трех экспериментах индексы имеют значения, значительно превосходящие средние значения индексов, полученных другими методами), чем в остальных рассматриваемых методах. В четырех экспериментах индексы были ниже, чем индексы во всех других методах.

В табл. 4 подведены итоги полученных результатов, отображенных графически на рис. 2, 3.

Среднее значение индексов Рэнда, вычисленных по результатам, полученным от слоя

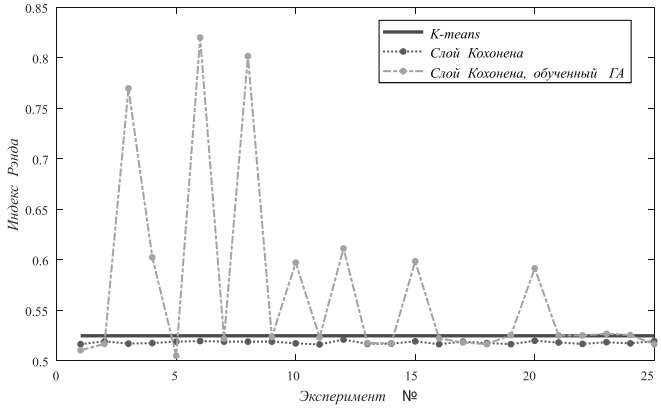

Рис. 2. Индексы Рэнда

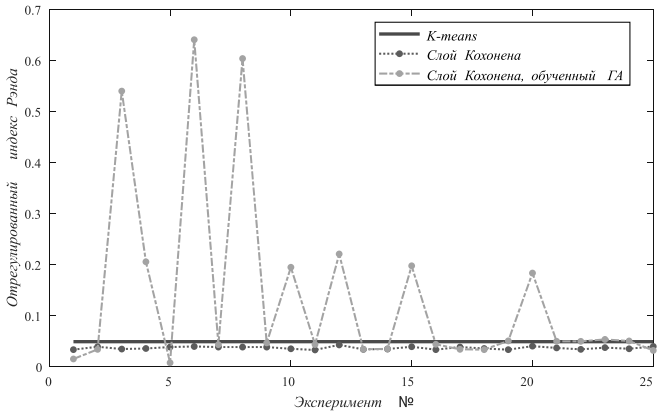

Рис. 3. Отрегулированные индексы Рэнда

Таблица 4

Индексы Рэнда

Метод кластерного анализа	Среднее значение RI по экспериментам	Среднее значение ARI по экспериментам
K-means	0,5249	0,0485
Слой Кохонена (стандартный алгоритм обучения)	0,5182	0,0362
Слой Кохонена (использован ГА для обучения)	**0,5692**	**0,1372**

Кохонена (ГА) за 25 экспериментов, выше соответствующих значений: K-means (RI на 0,0444; ARI на 0,0887), слой Кохонена обученный стандартным правилом Кохонена (RI на 0,0511; ARI на 0,1009).

Для процессов обучения нейронных сетей важным аспектом является время, затраченное на обучение. На рис. 4 приведены затраты времени на обучение сети Кохонена двумя методами — стандартным алгоритмом и генетическим алгоритмом. Табл. 5 содержит полученные результаты, отображенные графически на рис. 4.

На рис. 5 отображены значения функции приспособленности в каждом проведенном эксперименте.

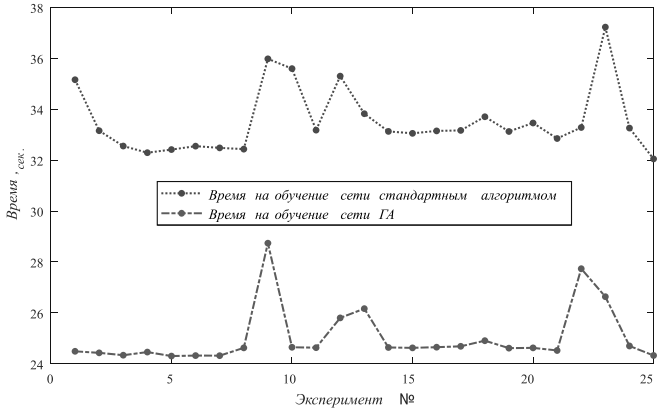

Рис. 4. График временных затрат на обучение

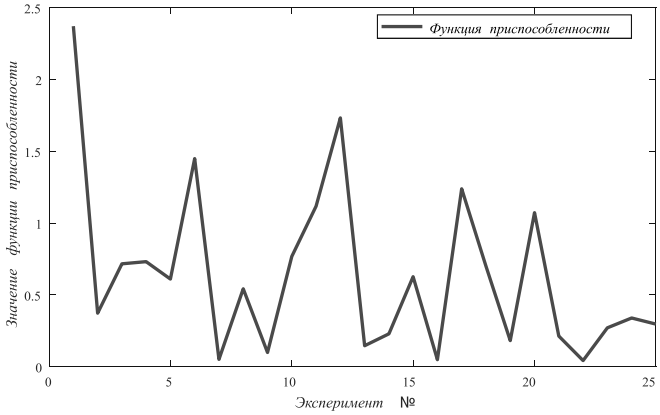

Рис. 5. Значения функции приспособленности

Таблица 5

Временные затраты на обучение сети

Метод кластеризации	Среднее значение, с
Стандартный алгоритм	33,5323
ГА	**25,0388**

Заключение

Обзор результатов исследований, представленных в данной работе, показывает, что классические алгоритмы кластерного анализа не всегда эффективнее (с точки зрения затрат времени на анализ) и не всегда дают более высокие оценки качества проведенного анализа. Качество кластеризации сильно варьируется от анализируемых данных.

Предложенный автором метод кластерного анализа, в котором применяется генетический алгоритм в качестве алгоритма настройки матрицы весов слоя Кохонена, по сравнению с другими, описанными в этой статье алгоритмами, на использованной выборке показал более высокое качество анализа. Однако он имел очень широкий диапазон значений индексов Рэнда и значений функции приспособленности, что говорит о необходимости поиска более оптимального набора параметров генетического алгоритма, в особенности таких, как размер популяции, число поколений, вероятность скрещивания и мутации, а также критерии останова генетического алгоритма.

Тем не менее данный подход представляется интересным и заслуживает дальнейшей разработки.

Список литературы

1. **Exploratory** Data Analysis with MATLAB, 2nd edition. URL: http://pi-sigma.info/EDA.htm.
2. **Мандель И. Д.** Кластерный анализ. М.: Финансы и статистика, 1988. 176 с.
3. **Гладков Л. А., Курейчик В. В., Курейчик В. М.** Генетические алгоритмы. М.: Физматлит, 2006. 320 с.
4. **Goldberg D. E.** Genetic Algorithms in Search, Optimization and Machine Learning. Addison-Wesley Publishing Company, Inc., 1989. 412 p.
5. **Рутковская Д., Пилиньский М., Рутковский Л.** Нейронные сети, генетические алгоритмы и нечеткие системы. М.: Горячая линия-Телеком, 2006. 452 с.
6. **Wierzchon S. T., Klopotek M. A.** Modern Algorithms of Cluster Analysis. Springer, 2018. 421 p.
7. **Миркин Б. Г.** Методы кластер-анализа для поддержки принятия решений: обзор. М.: Высшая школа экономики, 2011. 88 с.
8. **Ujjwal Maulik, Sanghamitra Bandyopadhyay, Anirban Mukhopadhyay.** Multiobjective Genetic Algorithms for Clustering. Berlin—Heidelberg: Springer-Verlag, 2011. 281 p.
9. **Rand W. M.** Objective criteria for the evaluation of clustering methods // Journal of the American Statistical Association. 1971. Vol. 66. P. 846—850.
10. **Santos J. M., Embrechts M.** On the Use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification. Berlin—Heidelberg: Springer-Verlag, 2009. Vol. 2. P. 175—184.
11. **Репозиторий** реальных и модельных задач машинного обучения. URL: https://archive.ics.uci.edu/ml/datasets/banknote + authentication.

V. A. Kharakhinov, Graduate Student, e-mail: tes4obse@mail.ru,
National Research Irkutsk State Technical University

The Genetic Algorithm as the Alternative Method for Training Kohonen Layer

In this article genetic algorithm suggested as the alternative method for training Kohonen layer. This paper considers evaluations of cluster analysis which were produced by follow methods: K-means; Kohonen layer trained with standard kohonen learning algorithm; Kohonen layer trained with genetic algorithm.

The evaluations of clustering were determined using two indices: the Rand index and Adjusted Rand index.

In addition to this, the article contains the graphical display of the training time for standard kohonen learning algorithm and realized genetic algorithm.

The cluster analysis was accomplished using banknote authentication dataset that described by a set of numerical feature and considers two output flag: authentic and counterfeit.

Keywords: *Cluster analysis, K-means, Kohonen layer, genetic algorithm, Rand index*

DOI: 10.17587/it.24.642-648

References

1. **Exploratory** *Data Analysis with MATLAB*, 2nd Edition. URL: http://pi-sigma.info/EDA.htm.

2. **Mandel' I. D.** *Klasternyj analiz*, Moscow, Finansy i statistika, 1988, 176 p. (in Russian).

3. **Gladkov L. A., Kurejchik V. V., Kurejchik V. M.** *Geneticheskie algoritmy,* Moscow, Fizmatlit, 2006, 320 p. (in Russian).

4. **Goldberg D. E.** *Genetic Algorithms in Search, Optimization and Machine Learning*, Addison-Wesley Publishing Company, Inc., 1989, 412 p.

5. **Rutkovskaja D., Pilin'skij M., Rutkovskij L.** *Nejronnye seti, geneticheskie algoritmy i nechetkie sistemy*, Moscow, Gorjachaja linija-Telekom, 2006, 452 p. (in Russian).

6. **Wierzchon S. T., Klopotek M. A.** *Modern Algorithms of Cluster Analysis*, Springer, 2018. 421 p.

7. **Mirkin B. G.** *Metody klaster-analiza dlja podderzhki prinjatija reshenij: obzor*, Moscow, Vysshaja shkola jekonomiki, 2011. 88 p. (in Russian).

8. **Maulik U., Bandyopadhyay S., Mukhopadhyay A.** *Multiobjective Genetic Algorithms for Clustering,* Berlin—Heidelberg, Springer-Verlag, 2011, 281 p.

9. **Rand W. M.** Objective criteria for the evaluation of clustering methods, *Journal of the American Statistical Association*, 1971, vol. 66, pp. 846—850.

10. **Santos J. M., Embrechts M.** *On the Use of the Adjusted Rand Index as a Metric for Evaluating Supervised Classification*. Berlin—Heidelberg. Springer-Verlag, 2009. Vol. 2, pp. 175—184.

11. **Repozitorij** real'nyh i model'nyh zadach mashinnogo obuchenija. URL: https://archive.ics.uci.edu/ml/datasets/banknote + authentication (in Russian).

УДК 004.434 DOI: 10.17587/it.24.648-656

Л. Н. Лядова, канд. физ.-мат. наук, доц., доц. кафедры информационных технологий в бизнесе, e-mail: LLyadova@hse.ru,
А. О. Сухов, канд. физ.-мат. наук, доц. кафедры информационных технологий в бизнесе, e-mail: ASuhov@hse.ru,
Е. Ю. Медведева, магистрант, e-mail: medvedevaeyu@mail.ru,
Национальный исследовательский университет "Высшая школа экономики", г. Пермь

Алгоритмы синтаксического разбора для текстовых динамически настраиваемых предметно-ориентированных языков

Предложены алгоритмы разбора для текстовых динамически настраиваемых предметно-ориентированных языков и проверки синтаксической корректности написанных с их помощью программ. В процессе своей работы на основе описания расширенной грамматики языка анализатор строит псевдодерево разбора, которое в дальнейшем используется при проверке синтаксической корректности программ пользователя. В основе алгоритма проверки синтаксиса лежит метод леворекурсивного спуска с возвратом.

Ключевые слова: *разбор грамматики, проверка синтаксиса, предметно-ориентированные языки, текстовые языки, языковой инструментарий, метод леворекурсивного спуска, формальные грамматики, дерево разбора*

Введение

В настоящее время все большее число инструментальных программных систем предоставляют в распоряжение пользователей встроенные текстовые языки программирования, позволяющие выполнять создание моделей предметной