НЕЙРОІНФОРМАТИКА ТА ІНТЕЛЕКТУАЛЬНІ СИСТЕМИ

НЕЙРОИНФОРМАТИКА И ИНТЕЛЛЕКТУАЛЬНЫЕ СИСТЕМЫ

NEUROINFORMATICS AND INTELLIGENT SYSTEMS

УДК 004.023

Е. В. Бодянский, В. В. Волкова, К. В. Коваль

АВТОМАТИЧЕСКАЯ КЛАСТЕРИЗАЦИЯ ТЕКСТОВЫХ ДОКУМЕНТОВ НА ОСНОВЕ ГЕНЕТИЧЕСКОГО АЛГОРИТМА С ИСКУССТВЕННЫМ ОТБОРОМ

Предложен новый генетический алгоритм с искусственным отбором, в основе которого лежит синтез обычного эволюционного генетического подхода с идеями последовательного комплекс-метода отыскания экстремума произвольных функций многих переменных. Алгоритм используется для кластеризации больших объемов текстовых документов в режиме последовательной обработки.

ВВЕДЕНИЕ

В общей проблеме интеллектуального анализа данных – Data Mining, Exploratory Data Analysis, Text Mining и, особенно, Web Mining важное место занимает задача поиска и классификации информации, содержащейся в текстовых документах, количество которых в Web практически неограниченно и постоянно увеличивается, достигая астрономических значений [1]. Фактически речь идет об очень больших и непрерывно растущих во времени базах данных и знаний, образованных зачастую не связанными между собой текстами самого различного содержания и происхождения, поиск и анализ в которых должен производиться в on-line режиме, обеспечивая эффективное автоматическое разбиение обрабатываемых документов на обычно заранее неизвестное число классов [2].

К настоящему времени сложилось достаточно много подходов к решению этой задачи, однако, большинство из них связано с использованием человеческих интеллекта и труда, которые весьма дороги. В связи с этим сегодня упор делается на средства искусственного интеллекта, позволяющие решать задачу в автоматическом режиме при минимальном участии человека. Среди таких средств достаточно высокую эффективность продемонстрировали как ставшие уже традиционными методы кластеризации, так и более современные методы вычислительного интеллекта такие, как искусственные нейронные сети (самоорганизующиеся карты Кохонена, BSB-нейромодели, сети теории адаптивного резонанса и т. д.) и нечеткие системы (fuzzy C-means Бездека, алгоритмы Густафсона – Кесселя, Ягера – Филева, Хёппнера – Клавонна – Крузе и т. д.).

Применение математических методов кластеризации требует предварительного преобразования анали-

зируемых документов в приемлемые для дальнейшей обработки компактные формы такие, как концепты, число которых также может быть велико [1], либо векторы, образованные частотами появления отдельных термов в тексте, размерность которых может быть неприемлемой для эффективного использования как нейронных сетей, так и четких, и нечетких методов кластеризации. Так или иначе, для решения задачи классификации документ должен быть сначала преобразован в векторную форму с числовыми компонентами.

Дальнейшее решение связано с проблемой оптимизации (критерий самообучения нейронной сети, целевая функция кластеризации) функции многих аргументов, число которых очень велико, а сама эта функция, описывающая либо корреляцию между концептами [3–4], либо меру семантического подобия текстов [3], как правило, многоэкстремальна [2] и может иметь сколь угодно сложную форму [1].

Именно это обстоятельство привело в последние годы к использованию в проблеме кластеризации текстов методов третьего основного направления вычислительного интеллекта – эволюционных вычислений и, прежде всего, генетических алгоритмов [1–7] и дало начало направлению, получившему название Genetic Mining [4]. Генетические алгоритмы в общем случае представляют собой методы эвристической оптимизации, чьи механизмы подобны биологической эволюции на основах принципов естественного отбора [8–12]. Такие их свойства, как адаптивность, робастность, возможность распараллеливания вычислений и отыскание глобального экстремума принятой функции приспособленности (fitness function), обеспечили их эффективное использование для решения задач кластеризации в пространствах высокой размерности [3]. При этом каждый документ, подлежащий обработке, кодируется в форме хромосомы (string) с бинарными компонентами, множество документов образует популяцию с изменяющимся числом особей, а в качестве функции приспособленности используется, как правило, та или иная мера семантического подобия анализируемых текстов.

Вместе с тем, холландовские алгоритмы, основанные на идеях естественного отбора, характеризуются низкой скоростью сходимости, не позволяющей им отыскивать решение за приемлемое время. В связи с этим представляется целесообразным ввести в процедуры генетической оптимизации элементы искусственного отбора, отличного от общепринятых стратегий элитизма (пчелиная семья, модель островов и т. д.) и имеющие под собой более строгое математическое обоснование.

ГЕНЕТИЧЕСКИЙ АЛГОРИТМ С ИСКУССТВЕННЫМ ОТБОРОМ

В основе предлагаемого алгоритма лежит синтез обычного эволюционного генетического подхода с идеями адаптационной оптимизации [13] и, прежде всего, последовательного комплекс-метода отыскания экстремума функций многих переменных [13–15]. При этом в каждый момент времени текущая популяция отождествляется с «облаком» – комплексом точек в пространстве переменных-факторов, а кроме традиционных генетических операторов мутации, кроссовера и инверсии дополнительно вводятся операторы комплекс-поиска такие, как отражение, растяжение и сжатие. При этом в отличие от традиционного комплекс-метода нами предлагается отражать не одну наихудшую вершину комплекса, а целое множество наихудших особей популяции.

В общем случае процедура оптимизации на основе обычного последовательного комплекс-метода выглядит следующим образом: требуется отыскать минимум некоторой функции

$$E(x) \to \min_{x \in R^n}$$

достаточно общего вида, при этом о характере этой функции не делается практически никаких априорных предположений. Работа алгоритма начинается с формирования начального комплекса

$$x_i(0) = (x_{i1}(0), x_{i2}(0), \ldots, x_{ij}(0), \ldots, x_{in}(0))^T,$$
$$i = 1, 2, \ldots, N \geq n + 1,$$

представляющего собой «облако» (популяцию) точек (векторов), достаточно произвольно расположенных в n-мерном пространстве факторов. Среди множества этих точек находится «наихудшая» $x_i(1)$, в которой значение функции $E(x_H(0))$ максимально, после чего эта точка отражается через центр тяжести всех остальных вершин-точек, формируя новый комплекс $x_i(1)$, $i = 1, 2, \ldots, N$. Такое отражение вместе с растяжением и сжатием обеспечивают движение комплекса к экстремуму функции $E(x)$, при этом, благодаря достаточно случайному распределению точек «облака», поиск имеет глобальный характер.

С формальной точки зрения рассмотрим процесс оптимизации на k-й итерации поиска, когда сформирован комплекс $x_i(k)$, $i = 1, 2, \ldots, N$. Среди множества точек $x_i(k)$ находится «наихудшая» такая, что

$$E(x_H(k)) = \max_i \{E(x_1(k)), \ldots, E(x_H(k))\},$$

после чего определяется центр тяжести «облака» без наихудшей точки:

$$x_C(k) = \frac{1}{N-1}\left(\sum_{i=1}^{N} x_i(k) - x_H(k)\right).$$

Далее $x_H(k)$ отражается через центр тяжести $x_C(k)$, формируя новую вершину комплекса $x_R(k)$, которая теоретически расположена ближе к экстремуму чем $x_H(k)$ и $x_C(k)$, т. е.

$$E(x_R(k)) < E(x_C(k)) < E(x_H(k)).$$

Операция отражения формально имеет следующий вид:

$$x_R(k) = x_C(k) + \eta_R(x_C(k) - x_H(k)) =$$
$$= \frac{1}{N-1}x_1(k) + \ldots + \frac{1}{N-1}x_{N-1}(k) + \frac{\eta_R}{N-1}x_1(k) + \ldots +$$
$$+ \frac{\eta_R}{N-1}x_{N-1}(k) - \eta_R x_H(k) = X(k)R,$$

где η_R – параметр шага отражения, часто полагаемый равным единице, $X(k) = (x_H(k), x_1(k), \ldots, x_{N-1}(k))$ – $(n \times N)$-матрица координат вершин комплекса, $R = \left(-\eta_R, \frac{1+\eta}{N-1}, \ldots, \frac{1+\eta_R}{N-1}\right)^T$ – $(N \times 1)$-вектор.

В случае, если отраженная вершина $x_R(k)$ окажется «наилучшей» среди всех остальных точек комплекса, т. е.

$$E(x_R(k)) < E(x_i(k)) < E(x_H(k)), \quad i = 1, 2, \ldots, N-1,$$

производится операция растяжения комплекса в направлении от центра тяжести $x_C(k)$ до $x_R(k)$ согласно выражению

$$x_E(k) = x_C(k) + \eta_E(x_R(k) - x_C(k)) = X(k)E,$$

где η_E – параметр шага растяжения, часто полагаемый равным двум,

$$E = \left(-\eta_E\eta_R, \frac{1-\eta_E(1-\eta_R)}{N-1}, \ldots, \frac{1-\eta_E(1-\eta_R)}{N-1}\right)^T.$$

Если же $x_R(k)$ окажется наихудшей среди всех $x_i(k)$, комплекс сжимается согласно соотношению

$$x_S(k) = x_C(k) + \eta_S(x_R(k) - x_C(k)) = X(k)S,$$

где η_S – параметр шага сжатия, обычно полагаемый равным 0,5,

$$S = \left(-\eta_S\eta_R, \frac{1-\eta_S(1-\eta_R)}{N-1}, \ldots, \frac{1-\eta(1-\eta)}{N-1}\right)^T.$$

При $\eta_S = 1$, $\eta_E = 2$, $\eta_S = 0,5$ приходим к простым выражениям

$$R = \left(-1, \frac{2}{N-1}, \ldots, \frac{2}{N-1}\right)^T,$$
$$E = \left(-2, \frac{1}{N-1}, \ldots, \frac{1}{N-1}\right)^T,$$
$$S = \left(-0,5, \frac{1}{N-1}, \ldots, \frac{1}{N-1}\right)^T.$$

Таким образом, в процессе своего движения к экстремуму оптимизируемой функции комплекс на каждой итерации теряет одну наихудшую вершину и приобретает одну новую точку так, что на $(k+1)$-й итерации новый комплекс также имеет N точек-вершин.

В отличие от комплекс-метода, в генетических алгоритмах в результате селекции из популяции одновременно исключаются несколько особей с наихудшими (максимальными) значениями функции приспособленности. В связи с этим представляется целесообразным ввести алгоритм комплекс-метода с отражением, растяжением и сжатием сразу нескольких вершин.

Итак, пусть на k-й итерации процесса оптимизации имеется комплекс $x_i(k)$, $i = 1, 2, \ldots, N$ с $P < N$ наихудшими вершинами $x_{H_p}(k)$, $p = 1, 2, \ldots, P$. Тогда координаты центра тяжести комплекса без вершин $x_{H_p}(k)$ задаются выражением

$$x_C(k) = \frac{1}{N-P}\left(\sum_{i=1}^{N} x_i(k) - \sum_{p=1}^{P} x_{H_p}(k)\right),$$

а процедура отражения описывается системой уравнений

$$\begin{cases} x_{R_1}(k) = x_C(k) + \eta_R(x_C(k) - x_{H_1}(k)), \\ \vdots \\ x_{R_p}(k) = x_C(k) + \eta_R(x_C(k) - x_{H_p}(k)), \end{cases}$$

или

$$\begin{cases} x_{R_1}(k) = (1+\eta_R)x_C(k) - \eta_R x_{H_1}(k), \\ \vdots \\ x_{R_p}(k) = (1+\eta_R)x_C(k) - \eta_R x_{H_p}(k). \end{cases}$$

В матричной форме эти системы уравнений могут быть записаны более компактно

$$X_R(k) = X(k)R_P,$$

где $X(k) = (\underbrace{x_{H_1}(k), \ldots, x_{H_p}(k)}_{(n \times P)}, \underbrace{x_1(k), \ldots, x_{N-P}(k)}_{(n \times (N-P))})$ – $(n \times N)$-матрица, $X_R(k) = (x_{R_1}(k), \ldots, x_{R_p}(k))$ – $(n \times P)$-

матрица, $R_P = \begin{matrix} P \\ N-P \end{matrix}\left\{\underbrace{\begin{pmatrix} -\eta_R I_P \\ \hline \frac{1+\eta_R}{N-P} I_{N-P,P} \end{pmatrix}}_{P}\right.$ – $(N \times P)$-матрица, I_p – $(P \times p)$-единичная матрица, $I_{N-P,P}$ – $((N-P) \times P)$-матрица, образованная единицами. В случае, если среди отраженных вершин оказывается $Q \le P$ наилучших, комплекс растягивается в их направлении согласно уравнениям

$$\begin{cases} x_{E_1}(k) = x_C(k) + \eta_E(x_{R_1}(k) - x_C(k)), \\ \vdots \\ x_{E_Q}(k) = x_C(k) + \eta_E(x_{R_Q}(k) - x_C(k)), \end{cases}$$

или

$$\begin{cases} x_{E_1}(k) = (1-\eta_E)x_C(k) + \eta_E x_{R_1}(k), \\ \vdots \\ x_{E_Q}(k) = (1-\eta_E)x_C(k) + \eta_E x_{R_Q}(k), \end{cases}$$

или

$$X_E(k) = X(k)E_Q,$$

где $X_E(k) = (x_{E_1}(k), \ldots, x_{E_Q}(k))$ – $(n \times Q)$-матрица, $E_Q =$

$= \begin{matrix} Q \\ N-Q \end{matrix}\left\{\underbrace{\begin{pmatrix} -\eta_E\eta_R I_Q \\ \hline (1-\frac{\eta_E(1-\eta_R)}{N-P} I_{N-Q,Q} \end{pmatrix}}_{Q}\right.$ – $(N \times Q)$-матрица.

Если, далее, среди отражаемых вершин окажется $U \le P$ наихудших, комплекс сжимается в их направлении согласно уравнениям

$$\begin{cases} x_{S_1}(k) = x_c(k) + \eta_S(x_{R_1}(k) - x_C(k)), \\ \vdots \\ x_{S_U}(k) = x_C(k) + \eta_S(x_{R_U}(k) - x_C(k)), \end{cases}$$

или

$$\begin{cases} x_{S_1}(k) = (1+\eta_S)x_C(k) - \eta_R x_{R_1}(k), \\ \vdots \\ x_{S_U}(k) = (1+\eta_S)x_C(k) - \eta_S x_{R_U}(k), \end{cases}$$

или

$$X_S(k) = X(k)S_U,$$

где $X_S(k) = (x_{S_1}(k), \ldots, x_{S_U}(k))$ – $(n \times U)$-матрица, $S_U =$

$= \begin{matrix} U \\ N-U \end{matrix}\left\{\underbrace{\begin{pmatrix} -\eta_S\eta_R I_U \\ \hline (1-\frac{\eta_S(1-\eta_R)}{N-U} I_{N-U,U} \end{pmatrix}}_{Q}\right.$ – $(N \times U)$-матрица.

Таким образом, комплекс-метод приобретает черты генетического алгоритма, у которого в результате селекции на каждой итерации из популяции удаляется несколько наихудших особей.

Объединяя введенную модификацию комплекс-метода с холландовской генетической процедурой, приходим к алгоритму, реализующему идею искусственного отбора, состоящую в данном случае в том, что из популяции не только удаляются наихудшие особи, но и одновременно создаются их «антиподы», обладающие улучшенными свойствами.

Работа такого алгоритма образована последовательностью следующих шагов:

– создание начальной популяции, образованной $P(0)$ особями хромосомами – вершинами комплекса;

– операция кроссовера с увеличением популяции $P_{CR}(0) > P(0)$;

– операция мутации $P_M(0) > P_{CR}(0)$;

– операция инверсии $P_I(0) > P_M(0)$;

– первая селекция (определение наихудших особей) без сокращения популяции $P_{SEL1}(0) = P_I(0)$;

– операция отражения с удалением P наихудших особей $P_R(0) < P_{SEL1}(0)$;

– операция растяжения без увеличения популяции $P_E(0) = P_R(0)$;

– операция сжатия без увеличения популяции $P_I(0) = P_E(0)$;

– вторая селекция с удалением $P_W(0)$ наихудших особей $P_{SEL2}(0) = P_I(0) - P_W(0) = P(1)$ и формирование популяции $P(1)$ для следующей итерации алгоритма. Предлагаемый алгоритм может быть представлен схемой, приведенной на рис. 1.

РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТАЛЬНЫХ ИССЛЕДОВАНИЙ

Целью исследования было оценить качество кластеризации стандартными генетическими алгоритмами и разработанным в ходе исследования генетическим алгоритмом с искусственным отбором. Работа предложенного алгоритма оценивалась на выборке новостных статей Reuters-21578, которая является одной из наиболее часто используемых тестовых выборок в Text Mining и информационном поиске. Для эксперимента было выбрано 200 документов из топиков coffee, crude, sugar и trade выборки Reuters-21578.

Эксперимент показал, что предложенный генетический алгоритм с искусственным отбором работает быстрее и дает более точные результаты (в среднем 6–8 %) по сравнению со стандартными генетическими алгоритмами и может быть использован для работы с большими массивами текстовых документов.

Таким образом, было установлено, что в задаче кластеризации документов предложенный генетический алгоритм с искусственным отбором дает более точные результаты, чем стандартные генетические алгоритмы.

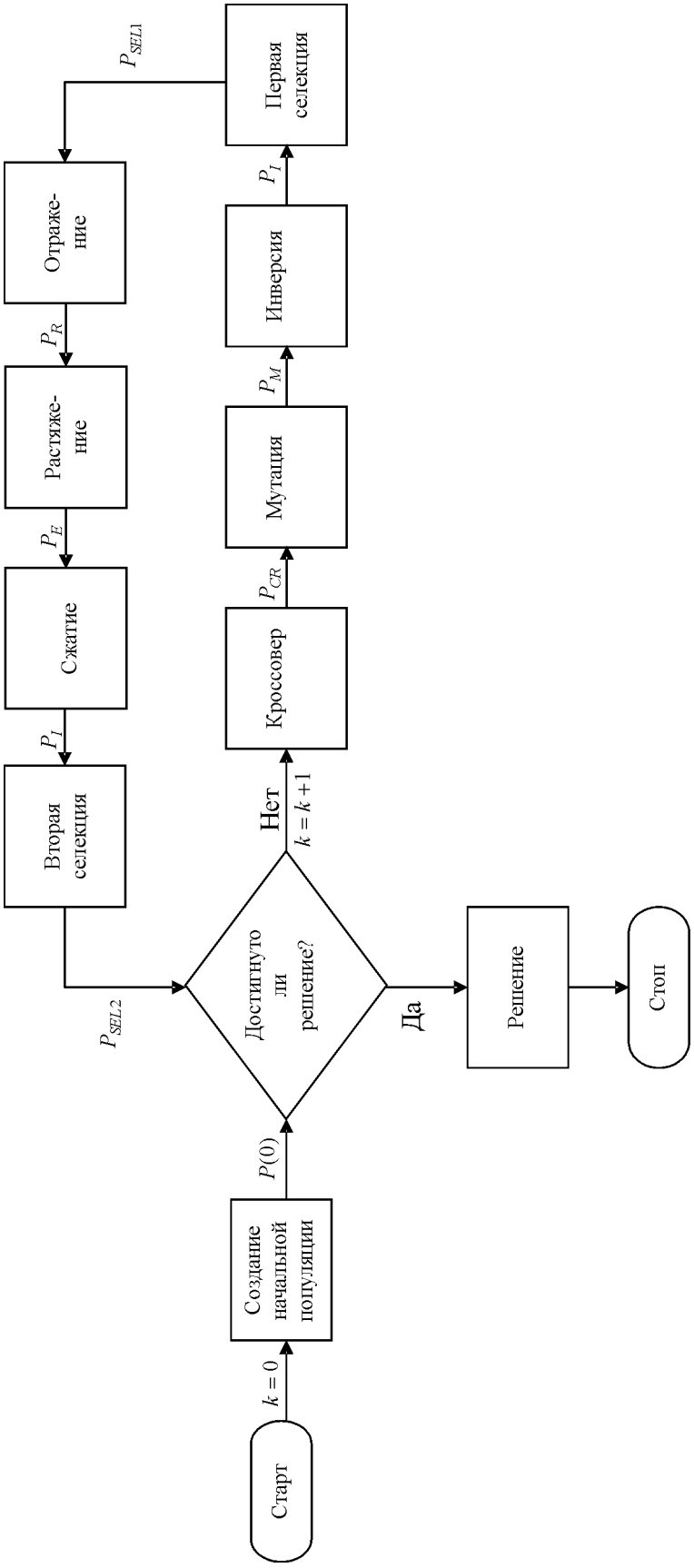

Рисунок 1 – Генетический алгоритм с искусственным отбором на основе последовательного комплекс-метода

ВЫВОДЫ

В работе предложен новый генетический алгоритм с искусственным отбором на основе комплекс-метода адаптационной оптимизации, предназначенный для отыскания экстремума произвольных функций большого числа аргументов в условиях существенной неопределенности о характере этих функций. Алгоритм обладает улучшенными характеристиками по сравнению с традиционными генетическими процедурами, прост в реализации и предназначен для использования в Genetic Mining больших массивов текстовых документов в режиме последовательной обработки.

ПЕРЕЧЕНЬ ССЫЛОК

1. *Desjardins G.* A genetic algorithm for text mining / G. Desjardins, G. R. Godin, R. Proulx // Sixth International Conference on Data Mining, Text Mining and their Business Applications. – 2005. – P. 133–142.
2. *Zhang C.* Self-adaptive GA, quantitative semantic similarity measures and ontology-based text clustering / C. Zhang, W. Song, C. Li, W. Yu // 2008. – http://eprints.rclis.org/14909/ (15.12.2008).
3. *Othman R. M.* Incorporating semantic similarity measure in genetic algorithm: an approach for searching the gene ontology terms / R. M. Othman, S. Deris, R. M. Illias, H. T. Alashwal, R. Hassan, F. Mohamed // International Journal of Computational Intelligence. – 2006. – № 3. – P. 257–266.
4. *Khalessizadeh S. M.* Genetic mining: using genetic algorithm for topic based on concept distribution / S. M. Khalessizadeh, R. Zaefarian, S. H. Nasseri, E. Ardil // Proceedings of World Academy of Science, Engineering and Technology – 2006. – 13. – P. 144–147
5. *Mani I.* Advances in Automatic Text Summarization / I. Mani, M. T. Maybury // Cambridge : MIT Press, 1999. – 442 p.
6. *Othman R. M.* Automatic clustering of gene ontology by genetic algorithm / R. M. Othman, S. Deris, R. M. Illias, Z. Zakaria, S. M. Mohamad // International Journal of Information Technology. – 2006 – 3.–№ 1. – P.37–46.
7. *Rocha F. E. L.* A new approach to meaningful learning assessment using concept maps: ontologies and genetic algorithms / F. E. L. Rocha, J. V. da Costa, E. L. Favero // 2004. – http://cmc.ihmc.us/papers/cmc2004-238.pdf (15.12.2008)
8. *Holland J. H.* Genetic algorithms and the optimal allocations of trails // SIAM Journal of Computing. – 1973. – 2. – P. 88–105.
9. *Holland J. H.* Adaptation in Natural and Artificial Systems. An Introductory Analysis with Application to Biology, Control and Artificial Intelligence. – London : Bradford Book Edition, 1994. – 211 p.
10. *Батищев Д. И.* Генетические алгоритмы решения экстремальных задач. – Воронеж : Воронежский государственный технический университет, 1995 – 69 с.
11. *Курейчик В. М.* Генетические алгоритмы. Состояние. Проблемы. Перспективы // Известия РАН. Теория и системы управления. – 1999. – 1. – С. 144–160.
12. *Рутковская Д.* Нейронные сети, генетические алгоритмы и нечеткие системы / Рутковская Д., Пилиньский М., Рутковский Л. – Москва : Горячая линия – Телеком, 2004. – 452 с.
13. *Горский В. Г.* Планирование промышленных экспериментов / Горский В. Г., Адлер Ю. П. – Москва : Металлургия, 1974. – 264 с.
14. *Химмельблау Д. М.* Прикладное нелинейное программирование / Химмельблау Д. М. – Москва : Мир, 1975. – 534 с.
15. *Реклейтис Г.* Оптимизация в технике : кн. 1 / Реклейтис Г., Рейвиндран А., Рэгсдел К. – Москва : Мир, 1986. – 349 с.

Надійшла 16.03.2009

Запропоновано новий генетичний алгоритм зі штучним відбором, в основі якого лежить синтез звичайного еволюційного генетичного підходу з ідеями послідовного комплекс-методу пошуку екстремуму довільних функцій багатьох змінних. Алгоритм використовується для кластеризації великих обсягів текстових документів у режимі послідовної обробки.

The new genetic algorithm with artificial selection is proposed. The algorithm is based on the synthesis of ordinary evolutional genetic approach with the ideas of sequential complex-method for extremum searching arbitrary multivariable functions. The algorithm is used for a clusterization of large data collection in a data-processing mode.

УДК 519.7:004.93

О. О. Олійник, С. О. Субботін

ОПТИМІЗАЦІЯ НА ОСНОВІ КОЛЕКТИВНОГО ІНТЕЛЕКТУ РОЮ ЧАСТОК З КЕРУВАННЯМ ЗМІНОЮ ЇХНЬОЇ ШВИДКОСТІ

Досліджено метод оптимізації на основі моделювання поведінки рою часток. Розроблено модифікацію дослідженого методу з керуванням зміною швидкості часток. Проведено експерименти зі знаходження глобального оптимуму багатовимірної функції на основі запропонованої модифікації.

ВСТУП

Градієнтні методи безумовної оптимізації, що традиційно застосовуються при синтезі моделей складних об'єктів і систем, є високоітеративними та накладають певні вимоги (наприклад, унімодальність, безперервність, монотонність, диференційованість та інш.)