УДК 519.7:004.93

А.А. Олейник, С.А. Субботин

МУЛЬТИАГЕНТНАЯ КЛАСТЕРИЗАЦИЯ С ПРЯМОЙ СВЯЗЬЮ МЕЖДУ АГЕНТАМИ

Введение

При построении моделей для принятия решений в задачах технической и биомедицинской диагностики актуальной является задача кластерного анализа. Данная задача заключается в разделении входной выборки данных на кластеры – компактные, непересекающиеся области в пространстве признаков. Известны методы кластерного анализа [1,2], основным недостатком которых является необходимость предварительного задания количества выделяемых кластеров, что затрудняет применение этих методов при обработке данных в случае, когда заранее сложно предположить необходимое количество кластеров.

Поэтому актуальной является разработка методов кластеризации, свободных от указанных недостатков, и обеспечивающих необходимую точность получаемых решений. Часто с такой целью применяются методы, основанные на случайном поиске, поскольку они характеризуются невысокой итеративностью, и, в случае разработки правильных схем работы, достигают необходимой точности оптимизации. Такими методами, в частности, являются мультиагентные методы интеллектуальной оптимизации, имеющие бионическую природу. К ним относятся: метод муравьиных колоний [3, 4], метод пчелиной колонии [5–7], метод оптимизации на основе моделирования перемещения бактерий [8] и т.д.

Метод муравьиных колоний уже успешно применялся для решения задачи кластерного анализа, а метод оптимизации на основе моделирования перемещения бактерий находится ещё на этапе своего становления, поскольку его математические модели ещё дорабатываются и области возможных применений известны недостаточно широко. В то же время, метод пчелиной колонии является достаточно известным и успешно применялся для решения различных задач оптимизации [5–7, 9–11]. Поэтому в данной работе предлагается решать задачу кластерного анализа на основе метода пчелиной колонии.

Метод пчелиной колонии является эвристическим итеративным методом случайного поиска, основанным на моделировании перемещения пчёл. При этом связь между программными агентами, моделирующими поведение пчёл, является прямой. Таким образом, метод пчелиной колонии является мультиагентным метод оптимизации с прямой связью между агентами.

Целью данной работы является разработка метода кластерного анализа, основанного на применении мультиагентного подхода с прямой связью между агентами, который позволяет исключить необходимость за-

© А.А. Олейник, С.А. Субботин , 2008

 ISSN 1562-9945

дания количества кластеров и позволяет сократить требования к вычислительным ресурсам при выполнении кластерного анализа.

Для достижения поставленной цели необходимо выполнить следующие задачи:

– анализ методов кластерного анализа с целью выявления их преимуществ и недостатков;

– разработка мультиагентного метода кластерного анализа с прямой связью между агентами;

– программная реализация предложенного мультиагентного метода кластеризации;

– сравнение разработанного подхода с существующими методами кластерного анализа путём проведения при помощи созданного программного обеспечения экспериментов по решению тестовых задач и анализа полученных результатов.

Постановка задачи

Пусть задано множество объектов O, каждый из которых характеризуется множеством значений признаков X. Тогда задача кластерного анализа заключается в том, чтобы на основе значений признаков X, разбить множество объектов O на m (m – целое) кластеров (подмножеств) $C_1, C_2, \ldots, C_m$, так, чтобы каждый объект O_i принадлежал одному и только одному подмножеству разбиения и чтобы объекты, принадлежащие одному и тому же кластеру, были сходными, в то время, как объекты, принадлежащие разным кластерам были разнородными.

Кластерный анализ

Кластерный анализ заключается в разбиении данных на группы схожих объектов. Каждая группа, называемая кластером, состоит из объектов, которые схожи между собой, и которые при этом отличные от объектов других групп.

Существует несколько видов методов кластерного анализа, отличающихся между собой допущениями о форме кластеров, видом результирующего разбиения и параметрами, которые должны быть предварительно установлены (например, количеством кластеров).

Исключающая кластеризация: данные группируются путём исключения единиц данных. Если определённый объект принадлежит к одному кластеру, то он не может быть включён в другой кластер (например, метод k-средних). Основными недостатками такого подхода являются необходимость указания количества кластеров, на которые необходимо разбить входную выборку, а также то, что выполняется поиск кластеров только заданной формы.

Перекрывающаяся кластеризация: данные могут входить в два или более кластеров в зависимости от значения функции принадлежности. К таким методам относится метод нечётких С-средних. Основным недостатков этих методов является необходимость указания количества кластеров.

Иерархическая кластеризация: в начале кластеризации каждый объект рассматривается как отдельный кластер, после чего два ближайших кластера объединяются в один и так далее. Метод заканчивает свою работу, когда все данные объединены в один кластер либо если выполнилось условие окончания работы. Основным недостатком такого подхода является существенная вычислительная сложность, которая особенно заметна при обработке многомерных выборок большого объёма.

Вероятностная кластеризация: имеет две разновидности: методы, основанные на смеси многомерных нормальных распределений, и методы интеллектуальной оптимизации, основанные на моделировании коллективного интеллекта общественных живых существ. Поскольку данный подход основан на вероятностном подходе, то есть возможность несходимости к оптимальному решению.

Как видно из приведенной классификации каждый из рассмотренных методов обладает определёнными недостатками, основными из которых являются: необходимость задания количества формируемых кластеров, допущение о форме кластеров, большая вычислительная сложность. В свою очередь, кластеризация, основанная на применении мультиагентных методов такими недостатками не обладает, однако ей характерен другой недостаток – возможность несхождения к оптимальному решению. В связи с этим можно сделать вывод, что применение мультиагентных методов для кластерного анализа является перспективным, однако такие подходы необходимо применять с учётом указанного недостатка так, чтобы сходимость к допустимому оптимальному решению достигалась.

Кластеризация на основе мультиагентного подхода

Рассматриваемый мультиагентный метод направлен на отыскание глобального оптимума некоторой функции, которая зависит от определённых, как правило, непрерывных аргументов. Кластерный анализ относится к задачам дискретной оптимизации, поэтому базовый мультиагентный метод оптимизации с прямой связью между агентами для решения задачи кластеризации не может полностью подходить.

В связи с этим предлагается использовать мультиагентный метод оптимизации, основанный на моделировании поведения пчёл [5–7] при изучении внешней среды и происходящем при этом опылении, за счёт чего возникают отдельные области объектов, обладающих подобными характеристиками. Такие области при работе предлагаемого метода и будут рассматриваться как кластеры.

В общем виде предлагаемый метод мультиагентной оптимизации для кластеризации состоит из следующих основных этапов.

1. Инициализация: создаётся пространство поиска, в котором случайным образом размещаются агенты и объекты входной выборки.

2. Перемещение агентов и выбор ими объектов, которые они будут распространять в пространстве поиска.

3. Перемещение агентов и дублирование выбранных ими объектов.

4. Моделирование обменом информации между агентами об объектах, которые они распространяют. За счёт такого моделирования обеспечивается прямая связь между агентами.

5. Исключение и сокращение количества объектов в точках пространства поиска и выделение, таким образом, кластеров.

Работу мультиагентного метода оптимизации с прямой связью между агентами для выполнения кластеризации можно представить в виде последовательности следующих шагов.

Шаг 1. Инициализация. Создать пространство поиска $m \times m$. При этом $m^2 = 4n$, где n – количество экземпляров во входной выборке. Экземпляры выборки случайным образом распределить по пространству поиска. Создать агентов в количестве $k = n/3$. Агенты размещаются в свободные ячейки пространства поиска случайным образом:

$$x_i^k = \text{rand}(m), i = \overline{1,m}, \forall p \neq l : X^p \neq X^l, \tag{1}$$

где x_i^k – i-ая координата размещения k-го агента в пространстве поиска; $\text{rand}(m)$ – случайное число, выбранное в диапазоне от 1 до m; X^p и X^l – позиции p-го и l-го агентов, соответственно.

Шаг 2. Установить счётчик итераций: $t = 1$.

Шаг 3 . Установить: $i = 1$.

Шаг 4. Установить: $j = 1$.

Шаг 5. Если j-ый агент не выбрал объект, который он распространяет в рабочем пространстве, то агент проверяет соседние ячейки пространства на предмет выбора объекта для его распространения. В случае, если j-ый агент уже выбрал объект для распространения, то выполнить переход к шагу 6.

Выбор j-ым агентом объекта для распространения выполняется следующим образом:

$$o^j = \begin{cases} \text{rand}(o^{l,p}), \text{ если } |o^{l,p}| = 2; \\ o^{l,p}_{worst}, \text{ если } |o^{l,p}| > 2; \\ o^{l,p}, \text{ если } |o^{l,p}| = 1, \end{cases} \tag{2}$$

где $|o^{l,p}|$ – количество объектов в ячейке с координатами $(l;p)$; $o^{l,p}$ – множество объектов, находящихся в ячейке с координатами $(l;p)$; $\text{rand}(o^{l,p})$ – один объект, случайным образом выбранный из множества $o^{l,p}$; $o^{l,p}_{worst}$– объект, для которого условия нахождения в ячейке $(l;p)$ худшие, который выбирается следующим образом:

$$o^{l,p}_{worst} = \arg\max \left[D_n(C^{l,p}, o^{l,p}_r) \right], \tag{3}$$

где $D_n(C^{l,p}, o^{l,p}_r)$ – нормированная разница между r-ым объектом ячейки $(l;p)$ и центром этой ячейки $C^{l,p}$. Центр определяется как среднее значение для каждой характеристики всех объектов, входящих в ячейку (l,p). Нормированная разница определяется на основе расстояния $D(C^{l,p}, o^{l,p}_r)$, которая может рассчитываться, например, как эвклидово расстояние:

$$D(C^{l,p}, o_r^{l,p}) = \sqrt{\sum_{q=1}^{N} \left[C^{l,p}(q) - o_r^{l,p}(q)\right]^2}, \quad (4)$$

где $C^{l,p}(q), o_r^{l,p}(q)$ – значение q-ой характеристики объекта $o_r^{l,p}$ и центра $C^{l,p}$, соответственно.

Если агент выбрал объект для распространения, то он перемещается в ячейку $(l; p)$ и берёт выбранный объект для его дальнейшего распространения.

Если агент, изучив все соседние ячейки, не выбрал объект для распространения, то он случайным образом перемещается в одну из соседний ячеек.

Шаг 6. Если j-ый агент обладает объектом, который распространяет в рабочем пространстве, то он изучает соседние ячейки и решает, где можно продублировать объект, который он распространяет. В случае если j-ый агент не обладает объектом для распространения, тогда агент случайным образом переместить его в одну из соседних ячеек, и выполнить переход к шагу 7.

Если рассматриваемая объектом ячейка не содержит объектов вовсе, то агент не делает ничего и рассматривает следующую соседнюю ячейку.

Если рассматриваемая объектом ячейка содержит только один объект, то агент с вероятностью 0.5 дублирует объект, который распространяет:

$$\text{Если } rand > 0,5, \text{то } o^{l,p} = \{o^{l,p}, o^j\}, \quad (5)$$

где $rand$ – случайное число в интервале [0; 1].

Если ячейка содержит более одного объекта, то возможны следующие случаи:

1) Рассматриваемая ячейка содержит объект, условия для которого хуже, чем для объекта, который распространяет объект. В этом случае агент выполняет следующие действия:

1.1) Объектом для распространения становится объект, для которого условия нахождения в данной ячейки худшие: $o^j = o_{worst}^{l,p}$.

1.2) Агент перемещается в данную ячейку. Переход к шагу 7.

2) Условия в рассматриваемой ячейке для распространяемого объекта лучше, чем в его исходной ячейке. В таком случае агент выполняет следующие действия:

2.1) Объект дублируется в данной ячейке: $o^{l,p} = \{o^{l,p}, o^j\}$.

2.2) Агент перемещается в данную ячейку. Переход к шагу 7.

3) Если ни один из предыдущих двух случаев не произошёл, то агент рассматривает следующую соседнюю ячейку.

В случае, когда после рассмотрения всех соседних ячеек агент не переместился ни в одну из них, агент перемещается в соседнюю ячейку, выбранную случайным образом, и выполняется переход к шагу 7.

Шаг 7. Установить: $j = j + 1$.

Шаг 8. Установить: $i = i + 1$.

Шаг 9. Если $i < N_{move}$, то выполнить переход на шаг 4, в противном случае – переход к шагу 10.

Шаг 10. Моделирование обмена информации между агентами.

В результате обмена информацией одни агенты должны сообщить остальным про ячейки, существенное влияние в которых имеют объекты, распространяемые соответствующими агентами. Таким образом, агенты разделяются на две группы: агенты, которые сообщают информацию о ячейке, к которой относится распространяемый ими объект, и агенты, которые анализируют информацию, сообщаемую другими агентами.

К агентам, информирующим других агентов о ячейке, к которой относится распространяемый ими объект, относятся следующие агенты:

1. Агенты, объект которых находится от центра соответствующей ячейки не дальше, чем $\Delta(D_n(C^{l,p}, o_r^{l,p}) < \Delta)$, при условии, что в ячейке находится 3 и более объектов. При этом Δ выбирается экспериментально и зависит от конкретной практической задачи. Из таких агентов случайным образом отбирается половина, и они информируют других агентов о соответствующей ячейке.

2. Агенты, объект которых относится к ячейке, в которой данный объект является единственным ($|o^{l,p}| = 1$). Из таких агентов также случайным образом отбирается половина для информирования о распространяемых ими объектах.

Все агенты, которые не вошли в группу агентов, выполняющих информирование, автоматически входят в группу агентов, анализирующих информацию от остальных агентов.

После разделения на группы для каждого агента, который анализирует информацию, рассчитывается расстояние между объектом, который он распространяет, и между объектами, которые распространяют агенты, относящиеся к информирующей группе агентов. Если минимальная из полученных разниц меньше ΔD, то объект, который распространяет информируемый агент, дублируется в ячейке с объектом, который распространяет соответствующий информирующий агент.

Шаг 11. Естественный отбор.

Поскольку один объект может находиться в нескольких ячейках одновременно, то нужно произвести отбор и оставить каждый объект только в одной ячейке. Для этого необходимо выполнить процедуру отбора. Предлагается выполнять жёсткий отбор, в соответствии с которым для каждого объекта необходимо учитывать, насколько он близок к каждому из центров ячеек $D(C^{l,p}, o_r^{l,p})$, взвешенное на нормализированное значение данного расстояния для текущей ячейки. Таким образом, необходимо объект оставить в той ячейке, в которой данное взвешенное расстояние наименьшее:

$$(q, w) = \arg\min_{l,p} \left[D(C^{l,p}, o_r^{\rangle} \cdot \left(1 - D_n(C^{l,p}, o_r^{\rangle}\right) \right], \forall l, p = \overline{1, m}, \qquad (6)$$

где (q, w) – ячейка, в которой следует оставить объект r-го агента o_r.

Шаг 12. Установить: $t = t + 1$.

Шаг 13. Если $t < t_{\max}$, то выполнить переход к шагу 3, в противном случае – переход к шагу 14.

Шаг 14. Рассчитать окончательные центры кластеров. Каждая отдельная ячейка считается кластером. На основании объектов, находящихся в ячейках, рассчитать центры кластеров:

$$x_i^c = \frac{1}{N^c} \cdot \sum_{j \in O^c}^{x_i^j}, \tag{7}$$

где x_i^c – i-ая координата ячейки c; N^c – количество объектов в ячейке c; O^c – объекты, находящиеся в ячейке c.

Шаг 15. Останов.

При разработке предложенного метода учитывались недостатки, связанные с возможностью несходимости к оптимальному решению. В связи с этим данный метод обладает следующими особенностями:

1. Прямая связь между агентами обеспечивается путём обмена информацией между агентами, за счёт чего одни агенты могут получить информацию об областях пространства поиска, в которых они не находились и от которых находятся далеко. Таким образом, достигается лучшее изучение пространства поиска, что положительно влияет на сходимость к оптимальному решению.

2. Введение процедуры естественного отбора позволяет исключить объекты из кластеров, условия нахождения для которых являются неудовлетворительными. Для этого вводится мера, характеризующая условия нахождения объекта в кластере, как расстояние объекта до центра кластера, взвешенное нормализированным значение расстояния, за счёт чего учитывается как абсолютное значение расстояние, так и относительное влияние данного объекта на кластер в целом.

3. Также для лучшего изучения пространства поиска предлагается выполнять шаг 6 несколько раз, что позволит каждому агенту изучать область, в которой он находится, более детально.

Эксперименты и результаты

Предложенный метод кластерного анализа на основе мультиагентного подхода с прямой связью между агентами был программно реализован в среде пакета Matlab 7.0.

При помощи разработанного программного обеспечения и встроенных средств пакета Matlab 7.0 проводились эксперименты, которые заключались в разбиении на кластеры искусственно сформированных выборок при помощи разработанного метода, а также при помощи методов кластеризации: K-средних и агломеративного иерархического метода.

Выборки формировались случайным образом на основе нормального распределения с различными математическими ожиданиями и дисперсиями. Было сформировано четыре двумерные выборки, отличающиеся

 ISSN 1562-9945

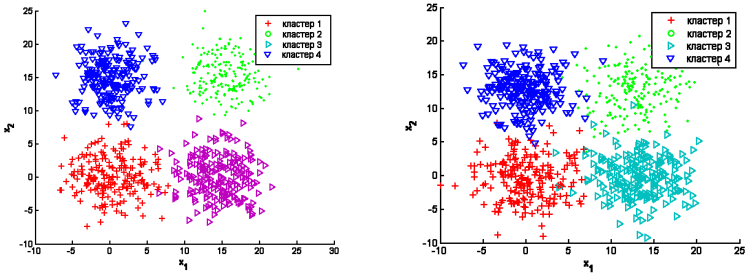

Рис. 1 – Графическое представление первой (а) и второй (б) выборок

между собой степенью пересеченья кластеров. Параметры распределений, на основании которых формировались выборки, приведены в таблице 1. Распределение выборок 1–4 в пространстве переменных представлено на рисунках 1 а), 1 б), 2 а) и 2 б), соответственно. Каждая выборка состояла из четырёх кластеров, каждый из которых, в свою очередь, состоял из 200 экземпляров, характеризующихся двумя признаками. Как можно видеть из таблицы 1 и рисунков 1 и 2 вторая и четвёртая выборки характеризуются большим пересечением кластеров по сравнению с первой и третьей выборками.

В качестве критерия сравнения результатов работы исследуемых методов кластеризации использовалась ошибка классификации:

Таблица 1

Табл. 1 – Парметры распределений выборок

Выборка	Кластер	x_1		x_2	
		$M(x)$	$D(x)$	$M(x)$	$D(x)$
1	1	0	3	0	3
	2	15	3	15	3
	3	15	3	0	3
	4	0	3	15	3
2	1	0	3	0	3
	2	13	3	13	3
	3	13	3	0	3
	4	0	3	13	3
3	1	0	3	0	3
	2	0	3	25	4
	3	16	3	25	4
	4	25	5	0	3
4	1	0	3	0	3
	2	0	3	12	3
	3	12	3	0	3
	4	12	3	12	3

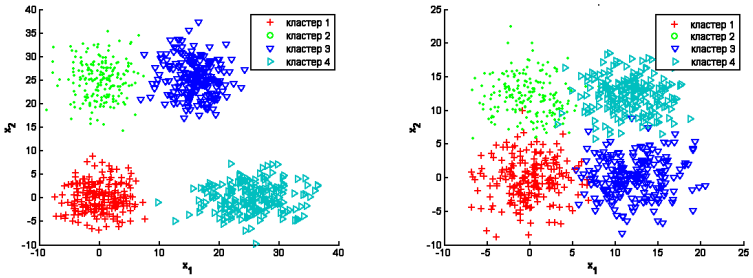

Рис. 2 – Графическое представление третей (а) и четвертой (б) выборок

$$\varepsilon = \frac{1}{N}\sum_{i=1}^{N} res_i, \tag{8}$$

где $res_i = 1$, если $cluster_i^* \neq cluster_i$, в противном случае – $res_i = 0$; $cluster_i^*$ – номер кластера, к которому отнесён i-ый объект при помощи заданного метода кластерного анализа, $cluster_i$ – номер кластера, к которому относится i-ый объект в заданной обучающей выборке.

Результаты работы традиционных методов кластеризации и предложенного метода представлены в табл. 2.

Исходя из результатов экспериментов, представленных в таблице 2, можно видеть, что предложенный метод характеризуется меньшей ошибкой классификации по сравнению с методами: K-средних и иерархическим агломеративным. При этом наибольшая ошибка классификации наблюдалась для всех методов при анализе второй и четвёртой выборок, для которых характерна ощутимая пересекаемость кластеров.

Таблица 2

Табл. 2 – Результаты работы методов кластерного анализа

Метод	Ошибка			
	Выборка 1	Выборка 2	Выборка3	Выборка 4
Метод K-средних	0.0113	0.0288	0.0050	0.0288
Иерархический агломеративный метод	0.0138	0.0325	0.0075	0.0300
Мультиагентный метод с прямой связью между агентами	0.0063	0.0150	0.0037	0.0125

Также важно отметить, что для работы предложенного метода не надо было задавать количество выходных кластеров, в отличие от рассматриваемых традиционных методов. При этом количество кластеров, на которое разбивал входную выборку разработанный метод, было правильным и составило четыре кластера для всех выборок.

Выводы

В статье предложен метод кластерного анализа, основанный на мультиагентном подходе с прямой связью между агентами, позволяющий выполнять кластерный анализ без задания количества выходных признаков.

Предложенный метод кластерного анализа обеспечивает нахождение оптимального решения за счёт использования прямой связи между агентами, а также за счёт выполнения процедуры естественного отбора, исключающей лишние объекты из пространства поиска.

При помощи разработанного программного обеспечения были проведены эксперименты по кластеризации тестовых выборок, сформированных случайно по нормальному распределению. Работа предложенного метода сравнивалась с традиционными методами кластерного анализа: методом k-средних и аггломеративным иерархическим методов. Исходя из полученных результатов экспериментов, можно сделать вывод, что предложенный мультиагентный метод кластерного анализа с прямой связью между агентами обеспечивает более высокую точность классификации. При этом предложенный метод сам в процессе своей работы выделил правильное количество выходных кластеров, в то время как для других рассматриваемых методов количество выходных кластеров необходимо задавать, что часто является невозможным при решении практических задач в виду отсутствия априорных знаний о выборке данных, которую нужно разбить на кластеры.

Таким образом, научная новизна работы заключается в том, что разработан новый метод кластерного анализа, основанный на мультиагентном подходе с прямой связью между агентами, позволяющий выполнять кластерный анализ без необходимости задания количества кластеров.

Практическая ценность работы заключается в том, что разработано программное обеспечение, реализующее предложенный мультиагентный метод кластерного анализа с прямой связью между агентами. Данное программное обеспечение может быть использовано при решении практических задач.

Литература

1. Интеллектуальные средства диагностики и прогнозирования надежности авиадвигателей: Монография / В. И. Дубровин, С. А. Субботин, А. В. Богуслаев, В. К. Яценко. – Запорожье : ОАО “Мотор-Сич”, 2003. – 279 с.
2. Berkhin P. Survey of clustering data mining techniques / P. Berkhin // Technical report. – San Jose : Accrue Software, 2002. – 56 p.
3. Dorigo M. The Ant System: Optimization by a Colony of Cooperating Agents / M. Dorigo, V. Maniezzo, A. Colorni // IEEE Transactions on Systems, Man, and Cybernetics. – 1996. – Part B, 26 (1). – P. 29–41.
4. Олейник Ал. А. Сравнительный анализ методов оптимизации на основе метода муравьиных колоний / Ал. А. Олейник // Комп’ютерне

моделювання та інтелектуальні системи : Збірник наукових праць / За ред. Д.М. Пізи, С.О. Субботіна. – Запоріжжя : ЗНТУ, 2007. – С. 147–159.

5. Quijano N. Honey Bee Social Foraging Algorithms for Resource Allocation: Theory and Application / N. Quijano, K. M. Passino. – Columbus : Publishing house of the Ohio State University, 2007. – 39 p.
6. Sumpter D.J. Formalising the Link between Worker and Society in Honey Bee Colonies / D. J. Sumpter, D. S. Broomhead // Lecture Notes In Computer Science : Proceedings of the First International Workshop on Multi-Agent Systems and Agent-Based Simulation. – MABS '98 LNAI, 1998. – P. 95-110.
7. Seeley T.D. The Wisdom of the Hive / T. D. Seeley. – Cambridge : Harvard University Press, 1995. – 265 p.
8. Passino K.M. Biomimicry of bacterial foraging for distributed optimization and control / K. M. Passino // IEEE Control System Magazine. – 2002. – 3 (22). – P. 52–67.
9. Karaboga D. An Idea Based on Honey Bee Swarm for Numerical Optimization / D. Karaboga // Technical report TR06. – Erciyes : Erciyes University Press, 2005. – 10 p.
10. Camazine S. Model of Collective Nectar Source by Honey Bees: Self-organization Through Simple Rules / S. Camazine, J. A. Sneyd // Journal of Theoretical Biology. – 1991. – 149. – P. 547–571.
11. Chong S. C. A Bee Colony Optimization Algorithm to Job Shop Scheduling / S. C. Chong, M. Y. Low // Proceedings of the 38th conference on Winter simulation. – Monterey : Monterey Press, 2006. – P. 1954–1961.

Получено 27.11.2008