УДК 004.93'12

МЕТОД КЛАСТЕРИЗАЦИИ ДАННЫХ, ОСНОВАННЫЙ НА ПАРАЛЛЕЛЬНОМ ГЕНЕТИЧЕСКОМ АЛГОРИТМЕ

И. В. Шевченко, Е. С. Король, И. В. Тимощук

Кременчугский национальный университет имени Михаила Остроградского

Вул. Первомайская, 20, г. Кременчуг, 39600, Украина. E-mail: athome050@yandex.ru

Предложен метод кластеризации данных, в котором используется параллельный генетический алгоритм для отбора оптимального результата кластеризации по одному из известных критериев. За счет использования нескольких популяций появляется возможность провести декомпозицию пространства признаков. В качестве базового используется модифицированный алгоритм формирования кластеров из «ближайших соседей», что позволяет упростить структуру хромосомы и повысить скорость отбора лучших решений. Преимущества предлагаемого метода по трем критериям продемонстрированы на тестовых задачах. Дополнительным преимуществом метода является гарантированное отсутствие перекрытия для всех кластеров и отсутствие необходимости априори задавать количество кластеров.

Ключевые слова: кластеризация, параллельный генетический алгоритм, хромосома, тестирование.

МЕТОД КЛАСТЕРИЗАЦІЇ ДАНИХ, БАЗОВАНИЙ НА ПАРАЛЕЛЬНОМУ ГЕНЕТИЧНОМУ АЛГОРИТМІ

І. В. Шевченко, Є. С. Король, І. В. Тимощук

Кременчуцький національний університет імені Михайла Остроградського

вул. Першотравнева, 20, м. Кременчук, 39600, Україна. E-mail: athome050@yandex.ru

Запропоновано метод кластеризації даних, в якому використовується паралельний генетичний алгоритм для відбору оптимального результату кластеризації за одним з відомих критеріїв. За рахунок використання декількох популяцій з'являється можливість провести декомпозицію простору ознак. Як базовий використовується модифікований алгоритм формування кластерів з «найближчих сусідів», що дозволяє спростити структуру хромосоми і підвищити швидкість відбору кращих рішень. Переваги пропонованого методу за трьома критеріями продемонстровані на тестових задачах. Додатковою перевагою методу є гарантована відсутність перекриття для всіх кластерів і відсутність необхідності апріорі задавати кількість кластерів.

Ключові слова: кластеризація, паралельний генетичний алгоритм, хромосома, тестування.

АКТУАЛЬНОСТЬ РАБОТЫ. Методы кластерного анализа позволяют разделить изучаемую совокупность объектов на группы “схожих” объектов, называемых кластерами. Кластеризация используется на начальных этапах исследования, когда о данных мало что известно. Если кластеры обнаружены, естественно использовать другие методы Data Mining, чтобы попытаться установить причины разбиения на кластеры или использовать кластеры для классификации и распознавания.

Задача кластеризации заключается в определении групп объектов, которые являются наиболее близкими один к другому по некоторому критерию. При этом никаких предварительных предположений об их структуре, как правило, не делается [1]. Большинство методов кластеризации базируется на анализе матрицы коэффициентов сходства, в качестве которых выступают расстояние, сопряженность, корреляция и др. [1]. Если критерием или метрикой выступает расстояние, то кластером называют группу точек Ω такую, что средний квадрат внутригруппового расстояния до центра группы меньше среднего квадрата расстояния s^2 (дисперсии) до общего центра в исходном множестве объектов мощности N [2], т.е.

$$\bar{d}_\Omega^2 < s^2, \qquad (1)$$

где $$\bar{d}_\Omega^2 = \frac{1}{N}\sum_{x_i\in\Omega}\left(x_i - \bar{x}_\Omega\right)^2, \ \bar{x}_\Omega = \frac{1}{N}\sum_{x_i\in\Omega} x_i. \qquad (2)$$

Задаче кластеризации сопутствуют две проблемы: определение оптимального количества кластеров и получение их центров и границ. Исходными данными для задачи кластеризации являются значения параметров (признаков) объектов исследования. Обычно определение оптимального количества кластеров является прерогативой исследователя. Что касается границ кластеров в пространстве признаков, то их определение можно автоматизировать, используя различные методы и алгоритмы.

Для решения задачи кластеризации статистическими методами необходимо располагать условными многомерными плотностями распределения признаков для каждого класса [2]. Тогда задача заключается в отыскании способа принятия оптимального решения о принадлежности проверяемого экземпляра к тому или иному классу в условиях неопределенности, т.е. в условиях действия случайных факторов, маскирующих связь между признаками и классом экземпляра.

Классические статистические методы [1, 2] дают оптимальное решение задачи кластеризации. Однако практическое применение этих методов возможно, если проведен специальный эксперимент по сбору и такой обработке статистических данных о прогнозируемом параметре и признаках, в результате которой найдены подходящие аналитические модели условных многомерных плотностей распределения прогнозируемого параметра и признаков. Но в ре-

альных задачах исследователь сталкивается здесь с рядом проблем, поэтому реализовать классические статистические методы не всегда возможно.

Во-первых, для реальных задач даже при известной совокупности информативных признаков (выявление которых представляет весьма трудоемкую задачу) не всегда доступны для изучения многомерные условные плотности распределения признаков.

Во-вторых, получение аналитических моделей этих условных плотностей распределения представляет трудоемкий процесс и может быть поставлено только отдельной самостоятельной задачей для каждого класса объектов и для определённых условий решения данной задачи.

В-третьих, даже если такие аналитические модели получены, необходимые при этих методах кластеризации аналитические преобразования достаточно сложны. Задача относительно легко решается аналитически, если многомерные условные плотности подчиняются нормальному закону, что в действительности имеет место далеко не всегда.

В связи со сказанным выше представляет интерес применение методов решения задач кластеризации, основанных на эвристических алгоритмах. Смысл понятия "эвристический алгоритм" состоит в том, что в этом случае алгоритм кластеризации не вытекает из строгих положений теории, а в значительной степени основан на интуиции и опыте исследователя. Такие методы могут давать удовлетворительные результаты и при ограниченной исходной информации о вероятностных характеристиках признаков и прогнозируемого параметра. Так, для применения этих методов для кластеризации по признакам достаточно иметь набор признаков, сильно коррелированных с прогнозируемым параметром, и необязательно знать вид их условных плотностей распределения.

Следует сказать, что методы кластеризации, основанные на использовании эвристических алгоритмов, не всегда приводят к оптимальным решениям. Однако для их применения на практике достаточно, чтобы ошибка кластеризации не превышала допустимого значения, а этого можно добиться, например, подбором более информативных признаков, применением соответствующих способов улучшения оператора кластеризации [3].

Таким образом, разработка эффективных эвристических методов кластеризации была и остается актуальной задачей.

Анализ проблемы и постановка задачи. Кластеризацию можно рассмотреть как задачу построения оптимального разбиения объектов на группы. При этом оптимальность может быть определена как требование максимизации плотности кластеров или минимизации среднеквадратичного расстояния между центром кластера и всеми его объектами:

$$F_1 = \sum_{l=1}^{k} \sum_{i \in S_1} d^2(X_i, \overline{X}_l) \qquad (3)$$

где l – номер кластера ($l = 1,2,...k$); $\overline{X}$ – центр l-го кластера; X_i – вектор значений переменных для i-го объекта, входящего в l-й кластер; $d(X_i, \overline{X}_l)$ – расстояние между i-м объектом и центром l-го кластера.

Решая задачу оптимизации, приходится выбирать метод её решения. В рассматриваемой задаче целевая функция (3) является мультимодальной (многоэкстремальной), поэтому предпочтительно выбрать адаптивный метод случайного поиска, каковым является генетический алгоритм (ГА) [4, 5].

Генетический алгоритм – это эвристический алгоритм поиска, используемый для решения задач оптимизации и моделирования путем последовательного подбора, комбинирования и вариации искомых параметров с использованием механизмов, напоминающих биологическую эволюцию. Является разновидностью эволюционных вычислений. Отличительной особенностью генетического алгоритма является акцент на использование оператора «скрещивания», который производит операцию рекомбинации решений-кандидатов, роль которой аналогична роли скрещивания в живой природе.

Общая схема ГА в контексте задачи кластеризации выглядит следующим образом:

1. Выбрать начальную, случайную популяцию множества решений и получить оценку качества для каждого решения, например, по критерию (3).

2. Создать и оценить следующую популяцию решений, используя эволюционные операторы:

Оператор выбора предков – с большей вероятностью предпочитаются хорошие решения

Оператор рекомбинации (обычно это кроссовер) – создает новое решение на основе рекомбинации из существующих.

Оператор мутации – создает новое решение на основе случайного незначительного изменения одного из существующих решений.

3. Повторять шаг 2 до получения приемлемого (квазиоптимального) результата

Главным достоинством ГА в данном применении является то, что они с большой долей вероятности находят глобальное оптимальное решение. Операторы рекомбинации и мутации позволяют получить решения, сильно не похожие на исходные – таким образом осуществляется глобальный поиск. Большинство популярных алгоритмов кластеризации выбирают начальное решение, которое затем изменяется в ту или иную сторону. Таким образом, получается хорошее разбиение, но не всегда – оптимальное.

Недостатками классического ГА являются:

– склонность к стагнации;

– эффективная работа с задачами малой размерности.

Стагнация генетического алгоритма – это такое состояние алгоритма, при котором на протяжении большого числа поколений не было изменения лучшего значения функции приспособленности у популяции, но текущее решение сильно отличается от глобального минимума. Адаптивный генетический алгоритм исключает периоды стагнации либо сводит их длительность к минимуму за счет увеличения

разнообразия популяции.

Оба указанных недостатка классического ГА преодолеваются при использовании квазипараллельного ГА (ПГА), в котором используются несколько субпопуляций (СП) [5]. Обмен генетической информацией между популяциями создает хорошие условия для обеспечения изменчивости даже без применения оператора мутаций. В то же время удается снизить размерность задачи путем её декомпозиции – в подпространствах небольшой размерности ведут поиск разные популяции, после чего их решения анализируются и объединяются. Даже для однопроцессорного компьютера реализация параллельного генетического алгоритма в виде псевдопараллельних вычислений дает лучшие результаты (алгоритм, чаще находит глобальный оптимум или использует для этого меньшее количество вычислений целевой функции, чем классический ГА.

Целью работы является разработка эвристического метода кластеризации, основанного на применении ПГА и отвечающего следующим требованиям:

1. Метод должен работать с метрическим пространством признаков достаточно большой размерности.

2. Алгоритм, построенный по данному методу, должен быть достаточно простым в реализации и потреблять минимум ресурсов.

3. Участие пользователя в процессе кластеризации должно сводится к минимуму.

МАТЕРИАЛ И РЕЗУЛЬТАТ ИССЛЕДОВАНИЙ.

Выбор базового метода кластеризации с учётом использования ПГА. Поскольку каждая особь в каждой СП является решением задачи кластеризации, необходимо выбрать базовый метод кластеризации и структуру хромосомы таким образом, чтобы по возможности сократить сложность алгоритма проверки каждой особи на приспособленность по целевой функции (3). Для выбора были рассмотрены два метода кластеризации – алгоритм *k*-средних и модифицированный алгоритм «ближайших соседей», предложенный в работе [Минаш*].

Алгоритм k-средних (k-means). Алгоритм k-средних строит k кластеров, расположенных на возможно больших расстояниях друг от друга. Основной тип задач, которые решает алгоритм k-средних, - наличие предположений (гипотез) относительно числа кластеров, при этом они должны быть различны настолько, насколько это возможно. Выбор числа k может базироваться на результатах предшествующих исследований, теоретических соображениях или интуиции. Общая идея алгоритма: заданное фиксированное число k кластеров наблюдения сопоставляются кластерам так, что средние в кластере (для всех переменных) максимально возможно отличаются друг от друга. Ограничения: небольшой объем данных. Достоинства: простота использования; быстрота использования; понятность и прозрачность алгоритма. Недостатки: алгоритм слишком чувствителен к выбросам, которые могут искажать среднее; медленная работа на больших базах данных; необходимо задавать количество кластеров.

Описание алгоритма [6]:

Этап 1. Первоначальное распределение объектов по кластерам. Выбирается число k, и на первом шаге эти точки считаются "центрами" кластеров. Каждому кластеру соответствует один центр. Выбор начальных центров может осуществляться следующим образом:

– выбор k-наблюдений для максимизации начального расстояния;
– случайный выбор k-наблюдений;
– выбор первых k-наблюдений.

В результате каждый объект назначен определенному кластеру.

Этап 2. Вычисляются центры кластеров, которыми затем и далее считаются покоординатные средние кластеров. Объекты опять перераспределяются. Процесс вычисления центров и перераспределения объектов продолжается до тех пор, пока не выполнено одно из условий: кластерные центры стабилизировались, т.е. все наблюдения принадлежат кластеру, которому принадлежали до текущей итерации; число итераций равно максимальному числу итераций. Выбор числа кластеров является сложным вопросом. Если нет предположений относительно этого числа, рекомендуют создать два кластера, затем три, четыре, пять и т.д., сравнивая полученные результаты. В структуре хромосомы должны содержаться координаты центров кластеров.

Второй рассматриваемый метод – однопроходной метод, основанный на присоединении очередной точки к рассматриваемому кластеру, в случае, если расстояние между новой точкой и предыдущей рассматриваемой точкой меньше заданного порога [Мин*]. Для каждой оси пространства признаков выполняются следующие шаги:

1. Выполняется нормирование пространства признаков.

2. Выполняется расчет плотности расположения точек по каждой координате пространства признаков.

3. Выполняется расчет среднего расстояния r_i между двумя соседними точками на выбранной оси координат, где i – номер оси:

$$r_i = \frac{1}{m-1}\sum_{t=2}^{m} d_t \quad , \qquad (4)$$

где $d_t = (x_t - x_{t\text{-}1})$; t – порядковый номер точки на оси i по возрастанию значения координаты, $t=1...m$.

4. Выполняется расчет степени неравномерности распределения точек на каждой i-й оси (дисперсия плотности):

$$s_i^2 = \frac{1}{m}\sum_{k=2}^{m}\left(d_{ik} - r_i\right), \qquad (5)$$

5. Выполняется ранжирование осей координат по возрастанию значения s_i^2.

6. Выполняется процедура распределения точек по кластерам для каждой i-й оси. При описании данной процедуры индекс оси (i) будем опускать.

Обозначим кластеры G_j, где j – порядковый номер кластера, $j=1...m$. Точку x_1 сразу отнесём к кластеру G_1. После этого необходимо циклически выполнить m-1 следующих итераций:

6.1. Рассчитывается расстояние $d_t = x_t - x_{t-1}$ между точками x_t и x_{t-1}, ($t = 2...m$, так как точка x_1 уже находится в кластере G_1).

6.2. Если $d_t < \alpha_i r_i$, (α_i коэффициент, определяющий порог плотности кластеров), то точка x_t относится к той же группе G_j, что и точка x_{t-1} иначе кластер G_j считается укомплектованным. Состав этого кластера фиксируется, и точка x_t становится первой точкой в новом кластере G_{j+1}, после чего t увеличивается на 1 и выполняется следующая итерация.

Как результат обработки данных на оси i получаем некоторое количество кластеров, в которых плотность превышает среднюю плотность по оси.

Далее выполняем шаг 6 для оси i+1. При этом последовательно обрабатываются точки, попавшие в кластеры, определенные на предыдущей оси. Каждый из этих кластеров, после прохождения шага 6 может быть разбит еще на некоторое количество более мелких кластеров. Таким образом, повторяем шаг 6 для всех координат.

После выполнения шага 6 для координаты n получим некоторое количество самых малых по размеру кластеров, которые и будут являться результатом кластеризации. Этот результат и его оценка (3) будут зависеть от того, какие значения принимают пороговые коэффициенты α_i.

Преимуществами данного метода является то, что предварительного задания количества кластеров не требуется, если в этом нет необходимости, а кроме того, всегда выполняется свойство сепарабельности, то есть, кластеры не перекрываются.

Хромосома особи должна содержать в закодированном виде (или в виде массива чисел с плавающей точкой) значения пороговых коэффициентов. Число аллелей соответствует числу координатных осей в пространстве признаков.

Анализ двух упомянутых методов привел к выводу, что для использования в ПГА целесообразно выбрать модифицированный метод «ближайших соседей». Для сравнения качества кластеризации был использован метод *k-means*.

Далее рассмотрим стратегию использования ПГА в задаче кластеризации.

Стратегию работы ПГА во многом определяет миграционная модель, отражающая способ обмена генетической информацией между СП. Миграционная модель делит популяцию на несколько СП. Эти СП эволюционируют независимо друг от друга в течение определенного периода времени (т.н. периода изоляции). После этого несколько особей перемещаются между популяциями (мигрируют). Число особей, подвергающихся миграции (степень миграции), метод селекции особей для миграции и схема миграции определяют, насколько значимым будет генетическое разнообразие в СП и обмен информацией между ними.

Выбор особей для миграции может происходить:

– случайным образом (выбираются особи по случайному закону);

– на основе целевой функции (выбираются именно приспособленные особи).

– Существует большое разнообразие структур миграции особей между СП. Например, миграция может происходить:

– между всеми СП (топология полного графа);

– по кольцу;

– между соседними СП.

Гипотеза данного исследования заключается в том, что увеличить скорость поиска глобального экстремума можно за счет усиления изменчивости популяции с одной стороны и интенсификации процесса отбора лидеров с другой стороны. Усиление изменчивости можно реализовать двумя путями:

1) Увеличение вероятности мутации одновременно во всех СП.

2) Специализация СП, то есть реализация в разных СП разных способов генерации новых поколений.

Первый способ фактически уравнивает возможности изменчивости во всех популяциях и тем самым отрицательно влияет на скорость поиска глобального экстремума.

Специализация элемента любой системы всегда способствует улучшению того параметра, на котором она сфокусирована. В таком случае важно только, чтобы на системном уровне осуществлялось «правильное» руководство элементами. В данном случае необходимо для каждой СП назначить:

– способ генерации новых поколений;

– способ отбора лидеров (один или несколько лидеров);

Остальные параметры алгоритма (число СП, число особей в популяциях, длительность периода изоляции в каждой СП) одинаковы для всех СП.

Для всех популяций необходимо назначить способ передачи генетической информации. В данной работе был выбран случайный способ.

Поскольку число СП в наших экспериментах могло изменяться от 2 до 6, то всегда имеется одна СП, в которой используется только случайный способ генерации особей (ССГ).

Алгоритм ПГА включает следующие шаги:

Шаг 1. Инициализация СП случайными значениями фенотипических координат особей.

Шаг 2. Выполнение в каждой СП заданного количества эпох эволюции. Выделение лидеров популяций. Копирование координат лидеров в отдельный массив.

Шаг 3. Анализ лидеров на идентичность. Не исключено, что в группе лидеров имеются точки, принадлежащие области одного и того же экстремума. Для выделения таких лидеров достаточно сравнить значения целевой функции по условию $\left| f^{L1} - f^{L2} \right| \le \varepsilon$, где ε – малое число. Сравниваются также координаты лидеров. Если идентичные лидеры найдены, то один из них заменяется на особь со

случайным значением координат.

Шаг 4. Уточнение координат лидеров.

4.1. Для каждого лидера:

4.1.1. Определяются границы поиска уточненного значения экстремума:

$a' = x_i^* - A(b-a)/2,$

$b' = x_i^* + A(b-a)/2,$

где a', b' – новые границы поиска; a, b – исходные границы поиска; x_i^* – координата найденного допустимого решения; A – размер границы для уточненного поиска экстремума.

4.1.2. Запускается цикл многостадийной мутации и проверки лидеров, для выявления экстремальных значений критерия в заданном узком интервале значений рабочих координат. Как результат в массиве лидеров фиксируются новые значения координат и экстремумов.

4.2. Массив лидеров сортируется по возрастанию значения критерия.

4.3. Первый в списке лидеров претендует на замещение предыдущего найденного значения глобального экстремума на данном этапе поиска.

4.4. Конец процедуры уточнения.

Шаг 5. Если выполняется условие останова, то конец, иначе переход к Шагу 2.

Программная реализация предложенного метода позволила провести его сравнительные испытания на задачах малой и средней размерности. Список кластеров, их плотность и их границы по всем осям координат выводятся для пользователя в табличном виде.

Для сравнения результатов кластеризации по разработанному методу и по методу *k-means* использовались среднеквадратичное расстояние между центром кластера и всеми его объектами (3) (критерий 1), минимальное межкластерное расстояние (6) (критерий 2) и максимальное внутрикластерное расстояние (7) (критерий 3). В сравниваемых алгоритмах использовались одни и те же метрики и одна и та же целевая функция. Для сравнения по двум упомянутым критериям требуется выполнить следующие вычисления:

минимальное межкластерное расстояние:

$$D_K = \min_l \left(\sum_{t,g \in L} d_{tg}^2 \right), \qquad (6)$$

где d_{tg} – расстояние между точками – «ближайшими соседями» из кластеров t и g;

максимальное внутрикластерное расстояние:

$$D_I = \max_l \left(\sum_{i,j \in S_l} d_{ij}^2 \right), \qquad (7)$$

где d_{ij} – расстояние между крайними точками кластера S_l.

Для тестов были выбраны типовые задачи классификации с размерностью от 3 до 6. Пример тестовой задачи – классификация магазинов по трем признакам: X1 – площадь торгового зала, м^2, X2 – товарооборот на одного продавца, ден. ед., X3 – уровень рентабельности, %. Данные по магазинам приведены в табл. 1.

Для расчета расстояний использовалась евклидова метрика. Значения расстояний были стандартизованы по формуле:

$$z_{ij} = \frac{x_{ik} - \bar{x}_k}{\sigma_k}, \qquad (8)$$

где z_{ij}, – стандартизированное значение расстояния между i-м и j-м объектами; k – номер признака.

Таблица 1 – Данные для кластеризации

Номер магазина	X1	X2	X3
1	100	160	25
2	130	200	30
3	80	180	20
4	40	100	22
5	150	90	15
6	85	200	35
7	60	170	28
8	110	150	18
9	55	110	15
10	110	100	12

На диаграмме (рис. 1) представлены усредненные оценки качества кластеризации методом k-means и по предлагаемому методу. Видно, что по всем трем критериям предлагаемый метод имеет преимущества. Следует добавить, что предлагаемый метод всегда обеспечивает сепарабельность кластеров по всем координатам, что немаловажно при решении задач распознавания.

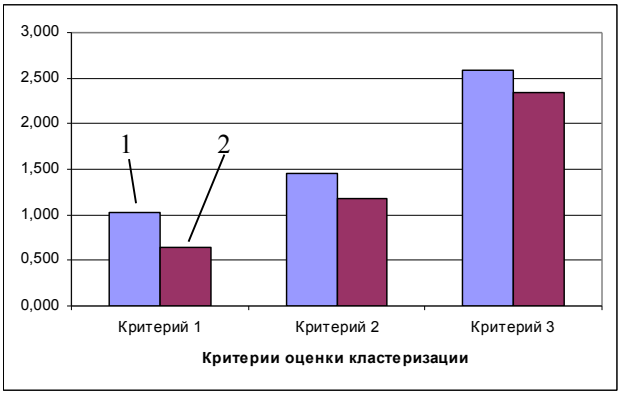

Рисунок 1 – Усредненные оценки качества кластеризации для тестовых задач: 1 – оценки алгоритма *k-means*; 2 – оценки предлагаемого метода

Отметим также, что преимуществом ПГА, как инструмента решения задачи оптимизации, является не только ускорение процесса поиска глобального экстремума, но и возможность избежать стагнации процесса глобального поиска, так как во-первых, за

счет усиленной изменчивости стагнация наблюдается реже, а во-вторых, даже если в отдельной СП происходит стагнация, то это компенсируется развитием других СП.

ВЫВОДЫ. Предложен метод кластеризации данных, в котором используется параллельный генетический алгоритм для отбора оптимального результата кластеризации по одному из известных критериев. За счет использования нескольких популяций появляется возможность провести декомпозицию пространства признаков. В качестве базового используется модифицированный алгоритм формирования кластеров из «ближайших соседей», что позволяет упростить структуру хромосомы и повысить скорость отбора лучших решений. Преимущества предлагаемого метода по трем критериям продемонстрированы на тестовых задачах. Дополнительным преимуществом метода является гарантированное отсутствие перекрытия для всех кластеров и отсутствие необходимости априори задавать количество кластеров.

ЛИТЕРАТУРА

1. Дуда Р., Харт П. Распознавание образов и анализ сцен. – М.: Мир, 1976. – 511 с.

2. Дидэ Э. Методы анализа данных. – М.: Финансы и статистика, 1985. – 357 с.

3. Мандель И.Д. Кластерный анализ. – М.: Финансы и Статистика, 1988. – 431 с.

4. Генетические алгоритмы: учебн. пособие. / Гладков Л.А., Курейчик В.В., Курейчик В.М. − М.: Физматлит, 2006. – 320 с.

5. Теория и практика эволюционного моделирования / В.В. Емельянов, В.В. Курейчик, В.М. Курейчик. – М.: Физматлит, 2003. – 432 с.

6. Эвристический метод кластеризации в метрическом пространстве признаков / И.В. Шевченко, А.О. Минашкин, Л.Н. Осипчук // Нові технології. – 2009. – № 4 (26). – С. 101–106.

CLUSTERING METHOD OF DATA BASED ON A PARALLEL GENETIC ALGORITHM

I. Shevchenko, E. Korol, I. Timoshuk

Kremenchuk Mykhailo Ostrohradskyi National University

vul. Pershotravneva, 20, Kremenchuk, 39600, Ukraine. E-mail: athome050@yandex.ru

A method of data clustering, which uses a parallel genetic algorithm to select the optimum clustering result on any known criteria. Through the use of multiple populations will be possible to decompose the space of attributes. As the base used a modified algorithm for the formation of clusters of "nearest neighbors". This simplifies the structure of the chromosome and improving the speed of selecting the best solutions. The advantages of the proposed method on three criteria demonstrated on test problems. An additional advantage of this method is guaranteed by the absence of overlap for all clusters, and there is no need to set a priori the number of clusters.

Key words: clustering, parallel genetic algorithm, chromosome, testing.

REFERENCES

1. Duda R., Hart P. (1976), *Raspoznavanie obrazov i analiz scen* [Detection and Scene Analysis], Mir, Moscow, Russia.

2. Dide E. (1985), *Metodi analiza dannih* [Methods of data analysis], Financi i statistica, Moscow, Russia.

3. Mandel I.D. (1988), *Klasternii analiz* [Cluster analysis], Financi i statistica, Moscow, Russia.

4. Gladkov L.A. (2006), *Geneticheskie algoritmu* [Genetic algorithms], Fizmatlit, Moscow, Russia.

5. Emelyanov V.V. (2003), *Teoriya i practica evolucionnogo modelirovaniya*, Fizmatlit, Moscow, Russia.

6. Shevchenko I.V. (2009) “Evristicheskiy metod klasterizacii v metricheskom prostranstve”, *Novi tehnologii,* no. 4 (26), pp. 101–106.

Стаття надійшла 25.10.2014.