УДК 004.8

**Н. Р. Кондратенко, к. т. н., доц.; О. А. Манаева**

# НЕЧЕТКАЯ КЛАСТЕРИЗАЦИЯ АБОНЕНТОВ ИНТЕРНЕТ-ПРОВАЙДЕРА

*Предложен генетический алгоритм нечеткой кластеризации на основе неоднородных хромосом с начальной инициализацией координат центров кластеров. Он был исследован на сходимость и его функционирование, проиллюстрирован компьютерным экспериментом.*

***Ключевые слова:*** *нечеткая кластеризация, провайдер, генетический алгоритм, неоднородная хромосома, тестовые функции, средневзвешенное отклонение, степени принадлежности*.

## Введение

Интернет – глобальная информационная сеть, объединяющая большое количество региональных сетей и в то же время миллионы компьютеров во всех концах планеты с целью обмена данными и доступа к информационным и технологическим ресурсам [1]. Предоставлением услуг доступа к сети Интернет и других услуг, связанных с ней, занимаются интернет-провайдеры. Такого рода организации владеют огромными объемами информации о своих пользователях; эту информацию необходимо определенным образом систематизировать, структурировать, обобщать и пр. Эти задачи тесно связаны с задачей кластерного анализа [2].

Существует большое количество методов кластеризации, которые можно классифицировать на четкие и нечеткие. Четкие методы кластеризации разбивают исходное множество объектов $\boldsymbol{X}$ на несколько непересекающихся подмножеств. При этом любой объект из $\boldsymbol{X}$ принадлежит лишь одному кластеру. Нечеткие методы кластеризации позволяют одному и тому же объекту принадлежать одновременно к нескольким (или даже ко всем) кластерам, но с разной степенью принадлежности. Единственным отличием является то, что при нечетком разбиении степень принадлежности объекта к кластеру принимает значение из интервала [0, 1], а при четком – из двухэлементного множества {0, 1}. Нечеткая кластеризация во многих ситуациях является более "естественной", чем четкая, например, для объектов, расположенных на границе кластеров [3, 4].

Для решения поставленной задачи предложим подход, основанный именно на нечетком разбиении пространства объектов на кластеры. Для задачи распределения на группы абонентов провайдера интернет-услуг именно такой подход имеет практическое значение. Это связано с одним из основных требований к тарифным планам организации – их гибкостью. Для выполнения этого требования при построении разбиения необходимо допускать некоторую неопределенность относительно принадлежности абонента к определенной группе.

Пусть каждый из абонентов выступает как объект, который характеризуется определенными значениями заданных показателей (скорость доступа, объем использованного входящего и исходящего трафика и т. п.). Соответственно они могут быть представлены как точки в многомерном пространстве. Практический смысл такого понимания сходства означает, что абоненты считаются тем более похожими, чем меньшее расхождение между одноименными показателями, с помощью которых они описываются [4].

При таких условиях целесообразно решить такую задачу в масштабах отдельного провайдера интернет-услуг. В нашем случае как объекты рассматриваются абоненты упомянутого провайдера, представленные набором параметров. Задача заключается в разбиении совокупности абонентов, заданных таким образом, на однородные нечеткие множества.

**Целью** представленного **исследования** является математическое моделирование поведения абонентов относительно провайдера телекоммуникационных услуг и разделение их на однородные группы с возможностью последующего анализа полученных результатов.

### Постановка задачи

Поставим задачу разбиения множества абонентов интернет-провайдера на нечеткие однородные подмножества согласно заданному набору показателей. При этом каждый абонент может содержаться в определенном кластере с некоторой степенью принадлежности в пределах от 0 до 1. Необходимо определить все степени принадлежности $\mu_{ij}$ абонента $j$ к кластеру $i$, а также места расположения центров кластеров $c_i, i = \overline{1,m}$.

Для решения поставленной задачи предложим генетический алгоритм, осуществляющий нечеткую кластеризацию абонентов интернет-провайдера согласно указанным показателям, и исследуем его на сходимость на ряде тестовых функций.

### Математическая модель

Пусть имеется набор абонентов $I=\{I_1, I_2, ..., I_n\}$ некоторого провайдера интернет-услуг. Каждый из $n$ абонентов характеризуется множеством признаков (измерений) $X_i=\{x_1, x_2, ..., x_p\}$, среди которых скорость передачи данных, а также объемы входящего и исходящего трафика за заданный период времени. Задача нечеткой кластеризации заключается в том, чтобы на основании данных, содержащихся в множестве $I$, разбить множество абонентов $I$ на $1<m<n$ кластеров, т. е. определить степени принадлежности $\mu_{ij}$ каждого абонента к каждому из $m$ кластеров, которые задаются центрами $c_i, i = \overline{1,m}$.

На величины $\mu_{ij}$ накладываются такие ограничения:

$$1.\ 0 \le \mu_{ij} \le 1; \qquad 2.\ \sum_{i=1}^{m} \mu_{ij} = 1 \text{ для всех } j. \tag{1}$$

Для оценки качества нечеткого разбиения используется средневзвешенное отклонение точек-абонентов от центров кластеров:

$$E = \sum_{i=1}^{m}\sum_{j=1}^{n} \mu_{ij}^{m} \left\| x_j - c_i \right\|^2,$$

где $m \ge 1$ – экспоненциальный вес, определяющий нечеткость, рассеянность кластеров.

Необходимо найти такое размещение центров кластеров $c_i$ и величины $\{\mu_{ij}\}$, при котором величина этого критерия была бы минимальной при одновременном соблюдении условий ограничения (1) [5]. Для решения этой задачи предложим генетический алгоритм оптимизации. Классический генетический алгоритм представляет собой итерационный процесс, на каждой итерации которого популяция последовательно подвергается операциям отбора, скрещивания и мутации. Остановив итерационный процесс в определенный момент и выбрав лучшую особь из популяции, можно получить вполне приемлемое решение задачи.

Предложим неоднородную хромосому как формализованное представление решения. Хромосомный набор состоит из двух качественно различных частей. Первая из них представляет собой прямоугольную матрицу, содержащую степени принадлежности точек-абонентов к соответствующим кластерам. Ее элементы определяют степень связи соответствующего абонента с определенным кластером. Вторая определяет координаты центров кластеров в пространстве признаков. Для каждой такой хромосомы вычисляется некоторое значение целевой функции.

На подготовительном этапе происходит начальная инициализация координат центров кластеров. Она опирается на геометрическое представление абонентов в виде точек в пространстве. Для этого каждая из осей разбивается на $N = 2 + E(3{,}322 \lg n)$ интервалов. Таким образом, все пространство признаков разделяется на $N^d$ равных по объему кубов, где $d$ –

количество признаков (измерений). За начальные центры кластеров принимаются геометрические центры кубов, внутрь которых попадается наибольшее количество точек.

В дальнейшем над начальной популяцией выполняются операции скрещивания и мутации в такой последовательности:

– одноточечное скрещивание: две хромосомы разрезаются в случайно выбранной точке и обмениваются полученными частями. Операция производится по одинаковой схеме над обеими компонентами хромосомы;

– двухточечное скрещивание: хромосомы расцениваются как циклы, образованные соединением концов линейной хромосомы. Для замены сегмента одного цикла сегментом другого цикла выбираются две точки разреза;

– равномерное скрещивание: каждый ген потомка создается копированием соответствующего гена от одного или второго из родителей в соответствии со случайно сгенерированной маской. Если в соответствующей позиции маски стоит 1, то ген копируется из первой родительской хромосомы, если 0 – то из второй. Процесс повторяется с родителями, которых обменяли, для создания второго потомка. Для каждой пары родителей случайно генерируется новая маска;

– мутация: генерирование новых степеней принадлежности для одной, случайно выбранной точки, а также случайное изменение положения центра каждого кластера по одному измерению в пространстве признаков.

Особи, полностью идентичные по хромосомному набору, из популяции исключаются и заменяются мутантами, образованными по приведенной выше схеме.

В результате выполнения таких действий получаем $7n$ генетически уникальных потомков, для каждого из которых подсчитываем целевую функцию, после чего в пределах популяции реализуется механизм естественного отбора на основе стратегии элитизма. При этом решения с низшим значением целевой функции гарантированно переходят в популяцию следующего поколения, что способствует более скорой сходимости генетического алгоритма. В целом за счет скрещивания происходит обработка наиболее перспективных вариантов решения, в то время как мутации реализуют механизм выхода оптимизационного процесса из локальных минимумов. Как результат, алгоритм с высокой вероятностью сходится к решению, максимально близкому к оптимальному.

## Компьютерный эксперимент

Решим задачу кластеризации абонентов провайдера телекоммуникационных услуг по трем признакам: скорость передачи данных и объемы входящего и исходящего трафика за фиксированный период времени. Объем исследуемой выборки составляет 100 пользователей.

Результат кластеризации абонентов по данным показателям приведен в таблицах 1 и 2.

Таблица 1

**Результаты кластеризации: степени принадлежности**

| Код абонента | Кластер 1 | Кластер 2 | Кластер 3 | Кластер 4 |
|---|---|---|---|---|
| 0 | 0,004049983 | 0,058683567 | 0,268992839 | 0,668273611 |
| 1 | 0,233707476 | 0,039341806 | 0,080751559 | 0,646199159 |
| 2 | 0,083366999 | 0,115781932 | 0,751763956 | 0,049087113 |
| 3 | 0,189849775 | 0,73390408 | 0,010152989 | 0,066093155 |
| 4 | 0,997684019 | 0,000579759 | 0,001224275 | 0,000511947 |
| 5 | 0,010728999 | 0,784860715 | 0,19198237 | 0,012427916 |
| … | … | | | |
| 99 | 0,059285226 | 0,903040624 | 0,002809315 | 0,034864835 |

Таблица 2

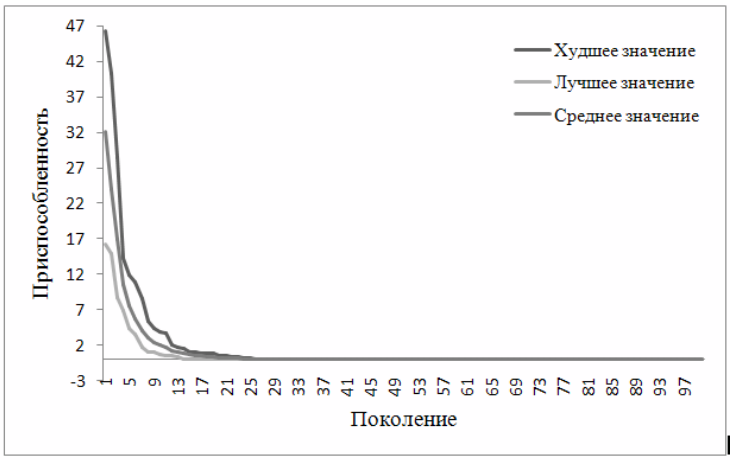
**Результаты кластеризации: положения центров кластеров**

| Измерение пространства признаков | Кластер 1 | Кластер 2 | Кластер 3 | Кластер 4 |
|---|---|---|---|---|
| Скорость передачи данных | 3208,12 | 430,24 | 484,15 | 819,71 |
| Объем входящего трафика | 4133,44 | 1613,21 | 887,96 | 701,37 |
| Объем исходящего трафика | 514,09 | 2503,17 | 105,46 | 88,01 |

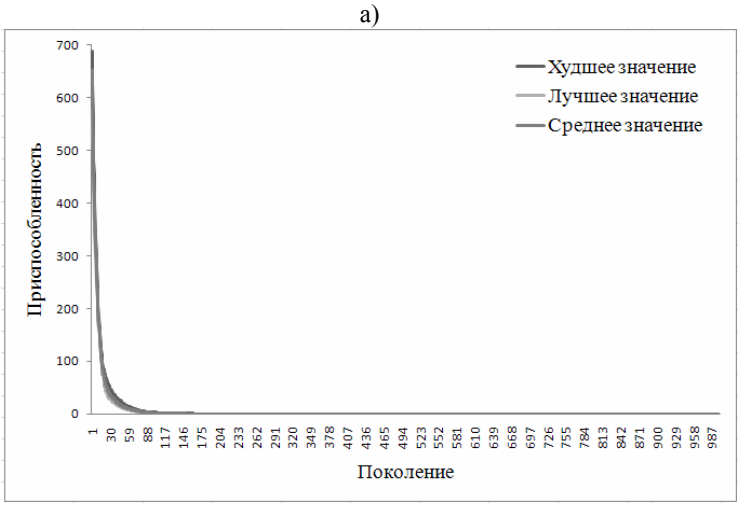
Из расположения центров кластеров видим, что в первом кластере находятся абоненты, имеющие наивысшую скорость доступа, второй представляет собой небольшой сегмент пользователей, у которых объем исходящего трафика за рассматриваемые сутки приближается к входящему или превышает его. К третьему кластеру были отнесены абоненты, скорость передачи данных для которых преимущественно невысока, а входящий трафик значительно преобладает над исходящим. Четвертый кластер отличается от третьего более высокой скоростью доступа, соотношение же входящего и исходящего трафика приблизительно такое же, как и в третьем.

### Исследование генетического алгоритма на сходимость

Оценим эффективность предложенного генетического алгоритма с помощью тестовых функций. Исследуем его оптимизационные возможности для числа переменных $n$=10 и $n$=100 при предельном числе поколений 100 и 1000 соответственно. Для этого выполним для каждой из приведенных ниже тестовых функций серию из десяти экспериментов.




а)




б)

Рис. 1. Изменения лучшей, худшей и средней приспособленности лучшей особи в популяции для сферической функции при n=10 (а) и n=100 (б)

1. Сферическая функция (первая функция де Джонга) – непрерывная выпуклая унимодальная тестовая функция, считается самой простой для оптимизации:

$$f_1(\mathbf{x}) = \sum_{i=1}^{n} x_i^2, \tag{3}$$

где $-5{,}12 \le x_i \le 5{,}12$, $i=1\ldots n$. Имеет один глобальный минимум, равный 0 в точке, где $x_i=0$, $i=1\ldots n$.

Изменения лучшей, худшей и средней приспособленности лучшей особи в популяции для данной функции в серии из десяти экспериментов показаны на рис. 1.

Аналогичным образом предложенный алгоритм был испытан на других распространенных тестовых функциях. В таблице 3 приведены лучшие, худшие и средние значения приспособленности лучшей особи в популяции в последнем поколении, полученные для разных тестовых функций.

Таблица 3

**Результаты исследования алгоритма на сходимость**

| Тестовая функция | Значение приспособленности лучшей особи в популяции | | | | | |
|---|---|---|---|---|---|---|
| | Лучшее | | Худшее | | Среднее | |
| | $n$=10 | $n$=100 | $n$=10 | $n$=100 | $n$=10 | $n$=100 |
| Сферическая | 0,0005884 | 0,000238 | 0,002541 | 0,000478 | 0,001472 | 0,000324 |
| Шаговая | 0 | 0 | 0 | 0 | 0 | 0 |
| Растригина | 0,088847 | 0,031292 | 0,775811 | 0,053048 | 0,335579 | 0,044519 |
| Швефеля | 0,417962 | 0,20015783 | 3,743808 | 0,468031 | 0,914933 | 0,338687 |
| Гриванка | 0,3411336 | 0,0493624 | 1,001547 | 0,141237 | 0,65774 | 0,092017 |

Из табл. 3 видно, что средняя ошибка нахождения глобального экстремума ни для одной из тестовых функций не превышает порядка первого знака после запятой. Это подтверждает высокую эффективность работы предложенного генетического алгоритма, в том числе и в случае большого числа переменных.

## Выводы

В статье предложен подход для решения задачи нечеткой кластеризации абонентов провайдера интернет-услуг и разработан генетический алгоритм с использованием неоднородных хромосом.

Исследование предложенного генетического алгоритма на сходимость показало, что он является мощным оптимизационным алгоритмом, следовательно может использоваться в задаче кластеризации, для которой характерно наличие большого количества параметров и, как правило, значительного числа локальных экстремумов.

С помощью предложенного алгоритма была проведена кластеризация абонентов интернет-провайдера по показателям, характеризующим особенности использования ими услуг данной организации. В результате множество пользователей разбили на компактные группы, между которыми существуют существенные отличия по данным показателям. Произведенный кластерный анализ позволил сделать вывод, что применение нечеткости, в частности в задачах кластерного анализа, позволяет работать и получать результаты в условиях значительной зашумленности данных. Поэтому дальнейшие исследования в этом направлении являются перспективными.

## СПИСОК ЛИТЕРАТУРЫ

1. Олифер В. Г. Компьютерные сети: принципы, технологии, протоколы / В. Г. Олифер, Н. А. Олифер. – СПб.: Питер, 2006. – 958 с.

2. Муссель К. Предоставление и биллинг услуг связи. Системная интеграция / К. Муссель. – М.: Эко-Трендз, 2003. – 319 с.

3. Дюран Б. Кластерный анализ / Б. Дюран, П. Оделл; Пер. с англ. Е.З. Демиденко. – М.: Статистика, 1977. –

128 с.

4. Мандель И. Д. Кластерный анализ / И. Д. Мандель. – М.: Статистика, 1988. – 176 с.

5. Зайченко Ю. П. Нечеткие модели и методы в интеллектуальных системах / Ю. П. Зайченко. – К.: Издательский дом «Слово», 2008. – 344 с.

***Кондратенко Наталия Романовна*** – к. т. н., доцент, профессор кафедры защиты информации.

***Манаева Ольга Алексеевна*** – магистрант.

Винницкий национальный технический университет.