# Кластеризация алгоритмом прыгающих лягушек

***В. С. Ковалев, М. Б. Бардамова, студенты кафедры КИБЭВС Научный руководитель И. А. Ходашинский, профессор кафедры КИБЭВС, д.т.н. г. Томск, ТУСУР, vitaly_979@mail.ru***

***Проект ГПО КИБЭВС-1404 – Нечеткие классификаторы обнаружения вторжений***

## Введение

Кластеризация находит применение в самых разных сферах деятельности, наибольшее применение кластеризация первоначально получила в таких науках как биология, антропология, психология. В информатике кластеризация может применяться для анализа данных, извлечения и поиска информации, группировки и распознавания объектов. Во многих прикладных задачах измерять степень сходства объектов существенно проще, чем формировать признаковые описания. Кластеризация (или кластерный анализ) — это задача разбиения множества объектов на группы, называемые кластерами. Внутри каждой группы должны оказаться схожие по каким-либо характеристикам объекты, а объекты разных группы должны быть как можно более отличны. Главное отличие кластеризации от классификации состоит в том, что перечень групп четко не задан и определяется в процессе работы алгоритма.

Кластеризация является описательной процедурой, она не делает никаких статистических выводов, но дает возможность провести разведочный анализ и изучить "структуру данных".

## Кластеризация

Применение кластерного анализа в общем виде сводится к следующим этапам:

- Выделение вектора характеристик. Свойства, которые характеризуют объекты, могут быть как количественными (координаты, интервалы), так и качественными (цвет, материал) характеристиками. Каждому объекту отождествляется вектор характеристик: $\mathbf{X} = (x_1,\ldots, x_n)$, где $x_i$ - отдельная характеристика объекта, $n$ – количество характеристик, определяет размерность пространства характеристик;
- Определение метрики: расстояние $d(X_i, X_j)$ между объектами $X_i$ и $X_j$ – результат применения выбранной метрики в пространстве характеристик. В работе будет использоваться евклидово расстояние, оно является геометрическим расстоянием в многомерном пространстве, или квадрат евклидова расстояния для того, чтобы придать большие веса более отдаленным друг от друга объектам.
- Разбиение объектов на группы путем использования алгоритмов кластеризации [1].

## Алгоритм k-средних

Одним из самых известных алгоритмов кластеризации является алгоритм k-means (k-средних), он разбивает множество элементов векторного пространства на заданное число кластеров $k$ [2]. Каждый кластер имеет свой центр масс [3]. Главная идея алгоритма заключается в том, что на каждой итерации центр масс вычисляется заново для каждого кластера, полученного на предыдущем шаге, затем векторы разбиваются на кластеры вновь в соответствии с тем, какой из новых центров оказался ближе по выбранной метрике. Алгоритм завершается, когда на последующих итерациях не происходит никакого изменения кластеров.

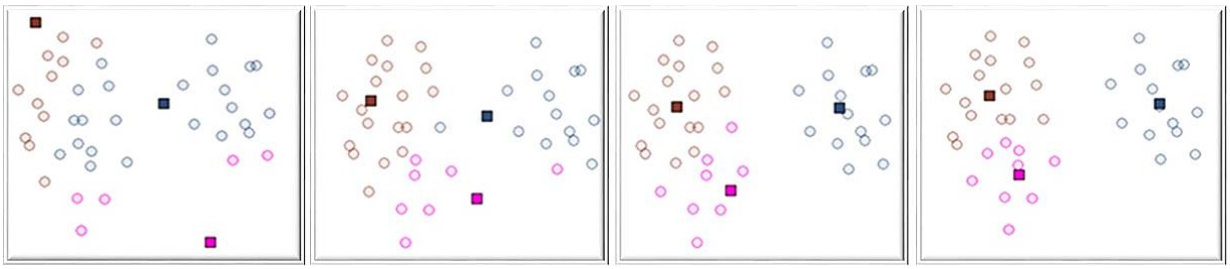

Рисунок 1 – Принцип работы алгоритма кластеризации методом k-средних

Алгоритм k-средних имеет несколько значительных недостатков: необходимо заранее знать или подбирать количество кластеров, и алгоритм крайне чувствителен к начальному выбору центров масс [4]. В связи с этими недостатками, мы предлагаем использовать идею алгоритма k-средних в совокупности с алгоритмом прыгающих лягушек для кластеризации данных.

**Алгоритм прыгающих лягушек**

Frog Leaping Algorithm (алгоритм прыгающих лягушек) является метаэвристическим алгоритмом, основанным на поведении лягушек в процессе поиска безопасной среды обитания. Основой алгоритма является комбинирование локального поиска в пределах каждого из мемеплексов [5] и глобального поиска путем обмена информацией о положениях лучших агентов этих мемеплексов и определения на этой основе глобально лучшего агента.

Алгоритм прыгающих лягушек позволит решить проблему центров масс кластера. В классическом варианте алгоритма k-средних используется случайное распределение центров масс кластеров, что очень часто является источником погрешности. Таким образом, результатом работы алгоритма прыгающих лягушек будет обновление расположения на каждой итерации центров масс, что даст возможность получить лучшее решение.

Для определения эффективности работы алгоритма k-средних вместе с алгоритмом прыгающих лягушек будем использовать метод машинного обучения, основанный на нечеткой логике.

**Нечеткий классификатор**

Основная идея нечеткого классификатора состоит в описании предполагаемого кластера нечетким прототипом, размерность которого определена размерностью пространства исследуемых данных [6]. Таким образом, i-й кластер определяется нечетким правилом следующего вида:

$R_{ij}$ : ЕСЛИ $x_1=A_{1i}$ И $x_2=A_{2i}$ И $x_3=A_{3i}$ И … И $x_n=A_{ni}$ ТО class=$c_j$, (1)

где $\mathbf{x} = (x_1, x_2, x_3, \ldots, x_n)$ – вектор признаков классифицируемого объекта; $A_{ki}$ – нечеткий терм, характеризующий $k$-ый признак в $i$-том правиле ($i \in [1, R]$), $R$ – число правил; $c_j$ – идентификатор j-того класса, $j \in [1, m]$, $m$ – количество классов.

В процессе нечеткой классификации объект относится к каждому классу с определенной степенью принадлежности, вычисленной следующим образом:

$$\beta_j(x) = \sum_{R_{ij}} \prod_{k=1}^{n} A_{ki}(x_k), j = 1, 2, \ldots, m. \quad (2)$$

Выходом классификатора является метка класса, определяемая следующим образом:

$$class = c_{j^*}, j^* = \arg \max_{1 \le j \le m} \beta_j. \quad (3)$$

Нечеткий классификатор может быть представлен функцией $c = f(\mathbf{x}, \boldsymbol{\theta})$, где $\boldsymbol{\theta}$ – вектор, описывающий базу правил.

На множестве обучающих данных (таблице наблюдений) $\{(\mathbf{x}_p; c_p), p = 1, \ldots, z\}$ определим единичную функцию

$$delta(p, \boldsymbol{\theta}) = \begin{cases} 1, \text{если } c_p = f(\mathbf{x}_p, \boldsymbol{\theta}) \\ 0, \text{иначе} \end{cases}, p = 1, 2, \ldots, z, \quad (4)$$

тогда численный критерий качества классификации выражается следующим образом:

$$E(\boldsymbol{\theta}) = \frac{1}{z} * \sum_{p=1}^{z} delta(p, \boldsymbol{\theta}). \quad (5)$$

Точность классификации зависит от расположения лингвистических термов, то есть этапа инициализации классификатора. Существуют различные способы построения термов, в том числе на основе алгоритмов кластеризации.

В данной работе для построения термов впервые будет использован алгоритм k-средних с оптимизацией алгоритмом прыгающих лягушек. Результаты совместной работы двух алгоритмов требуется сравнить с работой «чистого» алгоритма кластеризации k-средних, чтобы подтвердить или опровергнуть гипотезу об улучшении результатов кластеризации при использовании оптимизации алгоритмом прыгающих лягушек.

Основными проблемами данной работы является адаптация алгоритма прыгающих лягушек для оптимизации k-средних и подбор оптимального количества кластеров для каждого набора данных. В качестве наборов данных для тестирования нечеткого классификатора будут использоваться наборы данных KEEL (www.keel.es).

**Заключение**

В настоящий момент реализован программный код алгоритма k-средних и алгоритм оптимизации прыгающих лягушек. Получены предварительные результаты тестирования инициализации данными алгоритмами нечеткого классификатора, которые подтверждают гипотезу о улучшении работы алгоритма k-средних алгоритмом оптимизации прыгающих лягушек и позволяют вести дальнейшую работу в области кластеризации.

**Список литературы**

1. Воронцов К.В. Алгоритмы кластеризации и многомерного шкалирования. // Курс лекций. МГУ, 2007.

2. Jain A., Murty M., Flynn P. Data Clustering: A Review. // ACM Computing Surveys. 1999. Vol. 31, no. 3.

3. Navernіouk I. Multiobjective Graph Clustering with Variable Neighbourhood Descent // A thesis submitted in partial fulfilment of the requirements. 2005

4. Котов А., Красильников Н. // Кластеризация данных. 2006.

5. Narimani M.R. A New Modified Shuffle Frog Leaping Algorithm for NonSmooth Economic Dispath // World Applied Sciences Journal. 2011. С. 803–814

6. Ходашинский И.А., Дудин П.А. Идентификация нечетких систем на основе непрерывного алгоритма муравьиной колонии // Автометрия. 2012. Т. 48, № 1. С. 45–71.