КОМПЛЕКС ПРОГРАММ ДЛЯ ГРУППИРОВКИ ВЕЩЕСТВЕННЫХ ДАННЫХ ОЦЕНКИ КАЧЕТВА КЛАСТЕРИЗАЦИИ

Гатин Руслан Ришатович[1]
Российская Федерация, г. Казань, КНИТУ-КАИ им. А.Н. Туполева.

Аннотация: Статья посвящена разработке компьютерной программы кластеризации данных и определения качества проведенной кластеризации. Предложена программа с функцией кластеризации данных на максимальное количество кластеров для произвольного входного набора данных. На каждый вариант кластеризации реализована её оценка с применением различных методов.

Ключевые слова: методы кластеризации, оценка качества кластеризации, индекс Дэвиса-Болдина, C#.

COMPLEX OF PROGRAMS FOR GROUP PROPERTIES OF REAL DATA FOR EVALUATION OF THE QUALITY OF CLUSTERIZATION

Gatin R.R.
Russian Federation, Kazan, KNRTU-KAI named after A.N. Tupolev.

Abstract: The article is devoted to the development of programs for clustering data and determining the quality of clustering. A program is proposed with the allocation of data clustering for the maximum number of clusters for the input data set. For each clustering option, its assessment is used using different methods.

Key words: clustering methods, clustering estimates, the Davis-Boldin index, C#.

Введение

Некоторые методы прогнозирования требуют предварительной обработки данных, такой как сортировка, нормировка и кластеризация [2]. При группировании данных возникает вопрос: сколько кластеров необходимо использовать, и какая кластеризация будет лучшей. Для получения информации о качестве объединения данных используются методы оценки кластеризации, такие как оценка силуэта и индекс Дэвиса-Болдуина. Одним из способов оценки является проверка кластеров на компактность и отделимость, алгоритмически реализуемая при помощи такого подхода, как индекс Дэвиса-Болдуина (DBI) [3]. Другим способом служит проверка, насколько объект похож на свой кластер по сравнению с другими кластерами, реализуемая при помощи такого подхода, как метод оценки силуэта (SWC) [4].

Ход исследования

Для кластеризации входных данных используется метод «*k*-средних значений» («*k*-means clustering») [1]. В основе алгоритма лежит процедура минимизации суммарного квадратичного отклонения точек кластеров от центров этих кластеров. Метод *k*-средних разделяет *m* векторов на *k* кластеров, где $k \leq m$.

---

[1]Аспирант 3 года обучения кафедры прикладной математики и информатики КНИТУ-КАИ им. А.Н. Туполева. Научный руководитель: Новикова С.В., доктор технических наук, доцент, профессор кафедры прикладной математики и информатики КНИТУ-КАИ им. А.Н. Туполева.

Вначале определяется k случайных векторов из кластеризуемой выборки. Они становятся центроидами для кластеров S. Определяется евклидово расстояние между векторами входных данных и центроидами по формуле:

$$\rho(x, \mu_k) = \sqrt{\sum_{i=1}^{n} (x_i - \mu_{i_k})^2}, \quad (1)$$

где x_i – i-й элемент входного вектора, μ_{i_k} – i-й элемент центроида k-го кластера, n – количество элементов вектора. Вектор записывается в кластер с наименьшим значением функции расстояния. После определения каждого вектора в кластер, обновляется значение центроидов кластеров:

$$\mu_k = \frac{1}{s_i} \sum_{x^{(j)} \in S_i} x^{(j)}, \quad (2)$$

где j – номер вектора из кластера S_i.

Таким образом, алгоритм «k-средних значений» заключается в пересчете центроида на каждом шаге для каждого кластера, полученного на предыдущем шаге. Алгоритм останавливается, когда значения μ_k начинают меняться незначительно:

$$|\mu_k^{шаг\ t} - \mu_k^{шаг\ t-1}| \leq \varepsilon, \quad (3)$$

где $\varepsilon \approx 0{,}0001$ [5].

Окно программы, реализующей описанный алгоритм, представлено на рисунке 1. С помощью кнопки «Загрузить данные» реализуется загрузка множества входных векторов из excel-файла, представляющих собой четырёхкомпонентные вектора вещественных значений. По нажатию кнопки «Кластеризация k-means» выполняется кластеризация загруженных данных. Кнопка «Подсчёт оценки качества кластеризации для отдельного файла» позволяет провести оценку кластеризации для уже сгруппированной выборки и выводит результат в поле «Состояние программы».

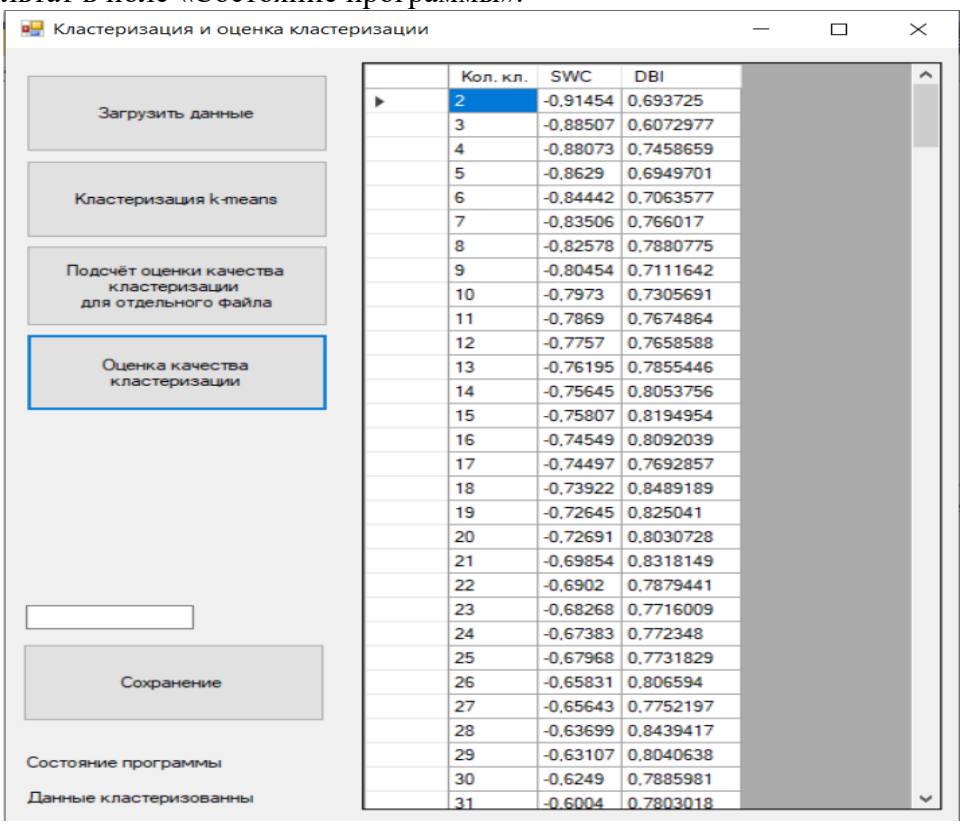

Рисунок 1 – Окно программы «Кластеризация и оценка кластеризации»

Кнопка «Оценка качества кластеризации» запускает процесс оценки процедур кластеризации на разное количество кластеров. Оценка производится по алгоритмам методов SWC и DBI.

Блок-схема кластеризации представлена на рисунке 2.

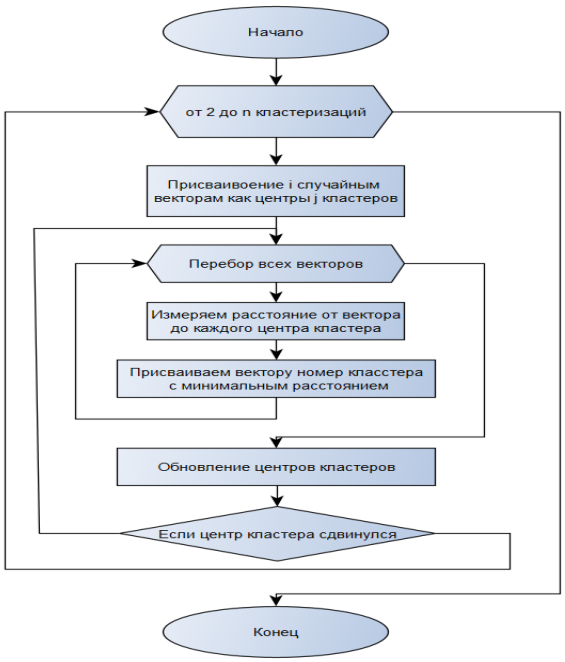

Рисунок 2 – Блок-схема алгоритма кластеризации

Алгоритм работы метода SWC:

1. Измеряется расстояние от вектора X до векторов в своём кластере (a_i);
2. измеряется расстояние от вектора X до векторов из других кластеров (b_i);
3. определяется силуэт по формуле

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{4}$$

4. SWC рассчитывается как среднее значение всех силуэтов:

$$SWC = \frac{1}{n}\sum_{i=1}^{n} s(i) \tag{5}$$

Алгоритм работы метода DBI:

1) Определение центров кластеров (μ_k) по формуле (2);

2) измерение евклидова расстояния от каждого вектора до центра его кластера по всем кластерам (ρ) по формуле (1);

3) расчёт среднего расстояния между векторами до их кластера (Мера компактности кластера);

$$S(c_k) = \frac{1}{|c_k|}\sum_{x_i \in c_k} \|x_i - \bar{c_k}\| \tag{6}$$

4) измеряется расстояние между центрами кластеров (Мера отделимости кластеров);

$$d_{ij} = \|\mu_i - \mu_j\| \tag{7}$$

5) расчёт меры качества схемы кластеризации;

$$R_{ij} = \frac{S_i + S_j}{d_{ij}} \tag{8}$$

6) расчёт индекса DBI – сумма максимальных Ri, делённая на количество кластеров.

$$DBI = \frac{1}{n}\sum_{i=1}^{n} \max(Ri). \quad (9)$$

Полученные результаты и выводы (Заключение)

Разработанная программа участвовала в испытаниях безопасного уровня металла в воде с учётом особенностей местности применительно к четырём возрастным категориям граждан: 0 – 1 год (условная группа «младенцы»), 1-9 лет (условная группа «дети»), 9-14 лет (условная группа «подростки»), 14 лет и старше (условная группа «взрослые»).

Список использованных источников

1) J. Hartigan. Clustering Algorithms. John Wiley & Sons, 1975.

2) Гатин Р.Р. Комплекс программ для моделирования накопления в биосредах организма токсикантов, поступающих с питьевой водой // XXV Туполевские Чтения (школа молодых ученых). 2021. Т – V. С. 164-168.

3) Гатин Р.Р., Новикова С.В., Моисеев Г.В. Исследование применимости моделей различной структуры для решения обратных задач определения пороговых концентраций металлов в питьевой воде, безопасных для населения //Вестник Казанского Государственного Энергетического Университета. 2022. № 2., Т – 14. С. 71-81.

4) Сивоголовко Е.В. Методы оценки качества чёткой кластеризации // Компьютерные инструменты в образовании, 2011г. № 4. - С.14-31.

5) Ширшова Д.В., Гатин Р.Р. Метод каскадного расширения выборки на основе 2-3d-линейной интерполяции для модели прогнозирования уровня металла в крови человека// Вестник технологического университета. 2023. Т. 26. № 1 С.106-112.

АНАЛИЗ ИНФЛЯЦИОННЫХ ПРОЦЕССОВ НА ПОТРЕБИТЕЛЬСКОМ РЫНКЕ САМАРСКОЙ ОБЛАСТИ В 2020 – 2022 ГГ.

Градова Анна Евгеньевна[1]

Российская Федерация, г. Самара, Самарский университет.

Аннотация: в работе предпринята попытка осветить основные тенденции развития потребительского рынка в Самарской области, в частности, инфляционных процессов за истекшие три года. В статье также исследуются наибольшие внутригрупповые колебания и их причины по продовольственным и непродовольственным товарам, а также в сфере платных услуг.

Ключевые слова: продовольственные товары, инфляционные процессы, услуги, индекс потребительских цен, непродовольственные товары, изменения.

ANALYSIS OF INFLATIONARY PROCESSES IN THE CONSUMER MARKET OF THE SAMARA REGION IN 2020 – 2022

Gradova A.E.

Russian Federation, Samara, Samara University.

[1]Студент 1 курса магистратуры Института экономики и управления Самарского университета. Научный руководитель: Миронова Е. А., доктор экономических наук, профессор кафедры экономики инноваций Самарского университета.