ИНТЕЛЛЕКТУАЛЬНЫЕ ИНФОРМАЦИОННЫЕ СИСТЕМЫ И ТЕХНОЛОГИИ

УДК 004.724

ПРИМЕНЕНИЕ МЕТОДОВ КЛАСТЕРИЗАЦИИ ДЛЯ АНАЛИЗА СВОЙСТВ МАТЕРИАЛОВ

В. П. Корячко, д.т.н., профессор, заведующий кафедрой САПР ВС РГРТУ, Рязань, Россия;
orcid.org/0000-0003-0272-673X, e-mail: koryachko.v.p@rsreu.ru
С. Д. Викулин, аспирант РГРТУ, Рязань, Россия;
orcid.org/0009-0002-9932-1113, e-mail: vikulin97@gmail.ru
А. В. Волков, инженер ключевых проектов, ООО «Техкомпания Хуавэй», Москва, Россия;
orcid.org/0009-0008-1162-3816, e-mail: vic-volk@yandex.ru

Рассматривается задача разработки методов кластерного анализа для изучения базовых характеристик материалов в целях дальнейшей разработки и реализации интеллектуальной поисковой системы в области материаловедения. ***Целью данной работы*** *является развитие методов кластерного анализа для изучения базовых характеристик материалов для дальнейшей разработки и реализации интеллектуальной поисковой системы в области материаловедения. Кластеризация проводится с использованием алгоритма k-средних, а валидация результатов осуществляется с помощью внутрикластерного и межкластерного анализа. Определение оптимальных начальных параметров метода кластерного анализа производилась с помощью методов локтя и силуэтов. Проведено успешное разложение коллекции материалов на кластеры, а визуализация иерархической структуры данных проведена с помощью дендрограммы, которая подтвердила эффективность предложенной методики.*

Ключевые слова: *машинное обучение, кластерный анализ, материаловедение, k-средних, метод локтя, метод силуэтов, дендрограмма.*

DOI: 10.21667/1995-4565-2024-89-77-84

Введение

Методы машинного обучения часто применяются в различных областях науки. Некоторые задачи машинного обучения подразумевают анализ различных неструктурированных данных. При этом сложная неструктурированная информация лучше распознается с помощью паттернов. Для исследования и обнаружения подобных закономерностей широко используется кластерный анализ [1]. Механизм работы кластеризации подразумевает разбиение множества объектов на подмножества (кластеры) по заданному критерию. При этом каждый отдельный кластер должен включать в себя максимально схожие по определенным признакам объекты. Применение кластеризации широко используется в особо проблемных областях науки, тесно связанных с деятельностью человека [2]. Кластерный анализ может успешно применяться в динамических моделях оптимизации топологий для решения проблем при проектировании различных конструкций [3]. С помощью кластерного анализа успешно находили паттерны при постановке диагноза различных заболеваний [4].

В сфере материаловедения кластерный анализ также может играть ключевую роль. Этот метод может помочь выявить основные закономерности, сходства и различия в больших данных о свойствах материалов, тем самым заложить основание для других методов машинного обучения. Методы кластеризации очень тесно связаны со сложной темой классифика-

ции [5]. Например, в [6] кластеризация используется для подготовки изображений к классификации материалов. В работе [7] описан подход к заполнению недостающих значений при неполном объеме данных о материалах. Достижения в области анализа больших данных, в том числе кластерного анализа и других методов машинного обучения, открывают инновационные возможности для обнаружения новых передовых материалов в различных областях промышленности, электроники и энергетики [8].

Таким образом, кластерный анализ служит мощным систематическим инструментом для анализа материалов на основе их свойств и характеристик, тем самым облегчает исследование зависимостей между структурой и свойствами материалов, открытие новых классов и свойств материалов и оптимизацию использования определенных материалов в той или иной сфере деятельности человека.

Целью данной работы является развитие методов кластерного анализа для изучения базовых характеристик материалов для дальнейшей разработки и реализации интеллектуальной поисковой системы в области материаловедения.

Постановка задачи

Исследовать методы локтя и силуэта в задачах кластерного анализа на основе k-средних для анализа коллекции однородных материалов и сплавов. Содержимое коллекции: легированная сталь, нержавеющая сталь, сплавы алюминия, меди, титана и других цветных металлов. В качестве основных параметров материалов выбраны следующие свойства: плотность, модуль упругости, теплопроводность, удельная теплоемкость, электрическое сопротивление. Эти свойства являются основными для металлических материалов и позволяют грамотно охарактеризовать каждый элемент коллекции для дальнейшего анализа

Теоретические исследования

В качестве метода кластеризации выбран метод k-средних из-за его универсальности и простоты реализации [9]. Основная идея заключается в определении k-центроидов (центров кластера) по одному для каждого кластера. Эти центроиды стоит задать в изначальной выборке на максимальном расстоянии друг от друга для получения надежных результатов. Следующим шагом является связка каждой точки, принадлежащей коллекции с ближайшим центроидом. Далее происходит обновление центроидов. Возникает цикличное вычисление центров кластеров. Центроиды меняют свое местоположение шаг за шагом, пока не произойдет конечное распределение элементов коллекции в кластеры. Алгоритм завершает работу, когда не происходит изменений в кластерах.

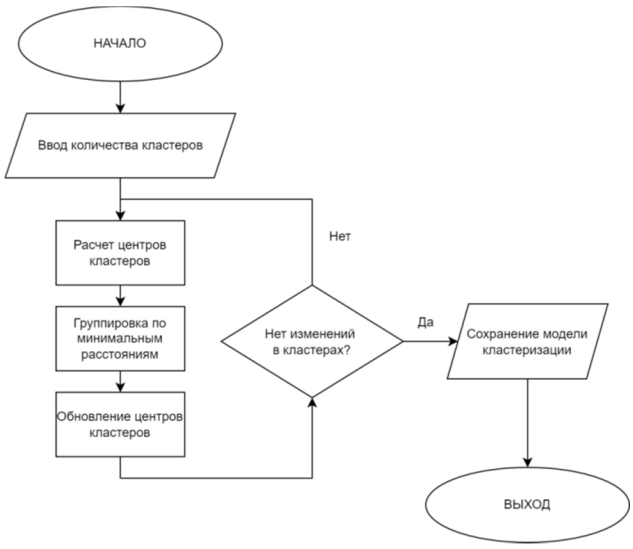

Рисунок 1 – Алгоритм k-средних
Figure 1 – K-means algorithm

Таким образом, алгоритм можно свести к минимизации направленной функции:

$$W(\mathrm{S},\mathrm{C}) = \sum_{k=1}^{K} \sum_{i \in S_k} \left\| y_i - c_k \right\|^z$$

где $S - k -$ кластерное разбиение сущностей, представленное векторами $y_i (i \in I)$ в многомерном пространстве свойств, состоящее из непустых непересекающихся кластеров S_k, каждый из которых имеет центроид $c_k (k = 1, 2...K)$.

Количество кластеров будет определяться с помощью метода локтя и метода силуэтов.

Метод локтя является самым старым способом определения истинного количества кластеров [10]. Он заключается в том, чтобы последовательно получать кластерное разбиение для разного начального количества кластеров. Принято начинать с количества кластеров равного 2, и продолжать увеличивать это число на 1 до тех пор, пока сумма квадратов внутрикластерных расстояний, именуемая инерцией, не перестанет существенно уменьшаться.

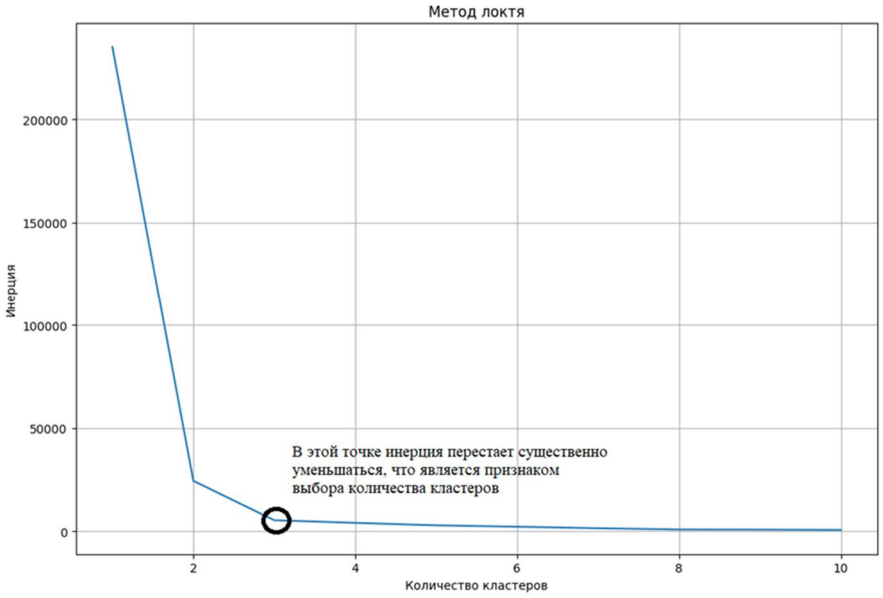

Рисунок 2 – Визуализация метода локтя
Figure 2 – Elbow method visualization

Хорошо сбалансированный коэффициент, такой как ширина силуэта, показал отличную производительность в экспериментах и был введен в [11]. Понятие ширины силуэта включает в себя разницу между плотностью внутри кластера и отрывом его от остальных. Данный коэффициент *S(i)* для каждого кластера *(i ∈ I)* можно представить в следующем виде:

$$S(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}, \quad (2)$$

где *a(i)* – среднее расстояние между *i* и всеми остальными сущностями кластера, к которому принадлежит *i*, а *b(i)* – минимум среднего расстояния между *i* и всеми сущностями в каждом другом кластере. Значения ширины силуэта лежат в диапазоне от -1 до 1. Если значение ширины силуэта для сущности около нуля, это означает, что данная сущность может быть отнесена и к другому кластеру. Если значение ширины силуэта близко к -1, это означает, что сущность неправильно классифицирована. Если все значения ширины силуэта близки к 1, это означает, что коллекция хорошо кластеризована.

Экспериментальные исследования

Метод локтя был использован для исходной коллекции с целью определения оптимального числа кластеров путем анализа суммарного внутрикластерного рассеяния (инерции) при

увеличении количества кластеров. График зависимости инерции от числа кластеров представлен на рисунке 3.

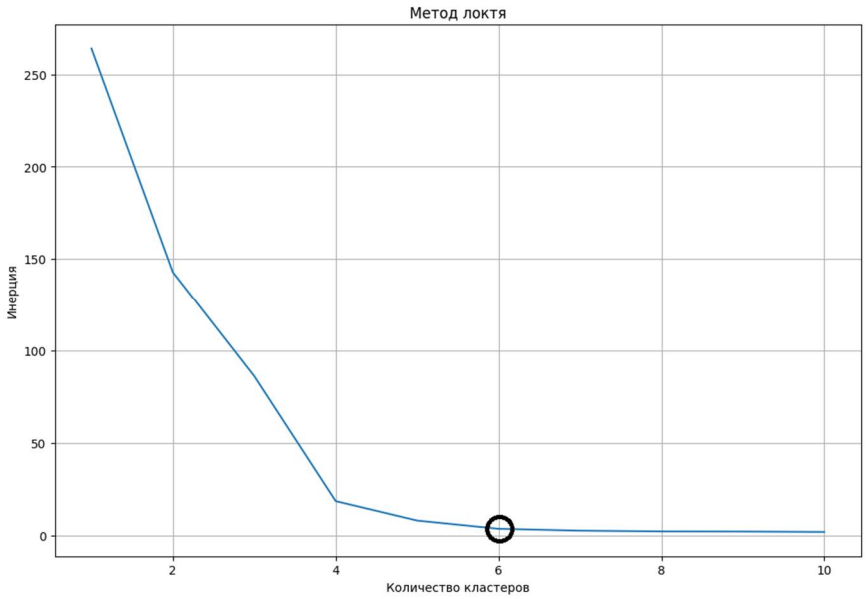

Рисунок 3 – Определение количества кластеров методом локтя
Figure 3 – Determining the number of clusters using the elbow method

На графике метода локтя (рисунок 3) наблюдается стремительное изменение инерции до отметки k = 6, где хорошо заметен характерный «локоть» (помечено черным кругом), что указывает на оптимальное число кластеров. После этого значение инерции снижается на много медленнее, что свидетельствует о том, что последующее добавление кластеров не приводит к значительному улучшению модели, но при этом требует большего количества ресурсов для вычислений. Стоит также отметить, что метод локтя является визуальным методом, поэтому он плохо подходит для автоматизации процесса подбора количества кластеров. Таким образом, проведем анализ повторно, но с помощью метода силуэтов для оценки качества кластеризации при различных значениях k. Средний коэффициент ширины силуэта для каждого значения k был рассчитан и представлен на рисунке 4.

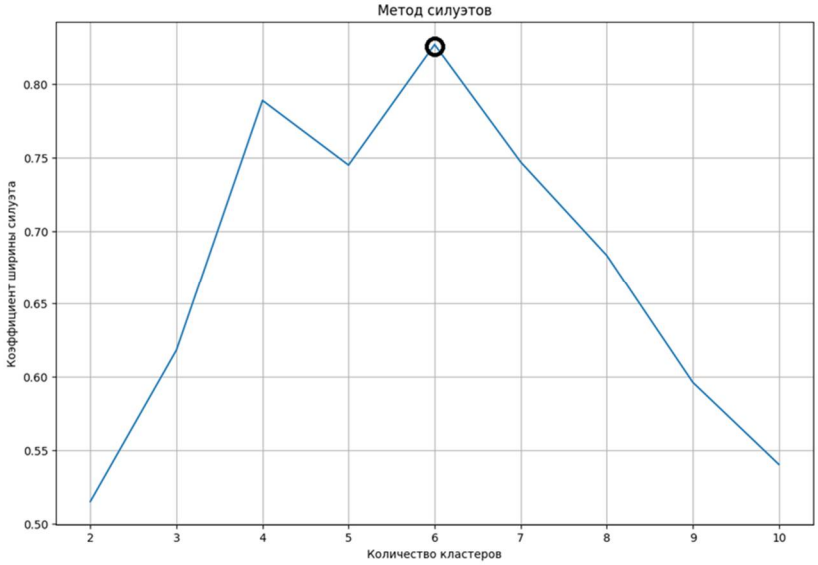

Рисунок 4 – Определение количества кластеров методом силуэтов
Figure 4 – Determining the number of clusters using the silhouette method

На рисунке 4 видно, что оптимальное количество кластеров для нашей коллекции равно 6, что помечено черным кругом. Эти результаты подтверждают результат, полученный методом локтя. При этом данный метод лучше подходит для автоматического определения кластеров, так как в ходе его работы можно получить вектор значений коэффициента ширины силуэта (2) для определенного диапазона количества кластеров и найти максимальный из них. Количество кластеров для максимального значения инерции и будет являться искомым числом кластеров. Таким образом, кластеризация при $k = 6$ должна обеспечить наиболее четкое разделение данных на кластеры. Для этого визуализируем результаты в виде построения силуэтов и разбиения на кластеры (рисунок 5).

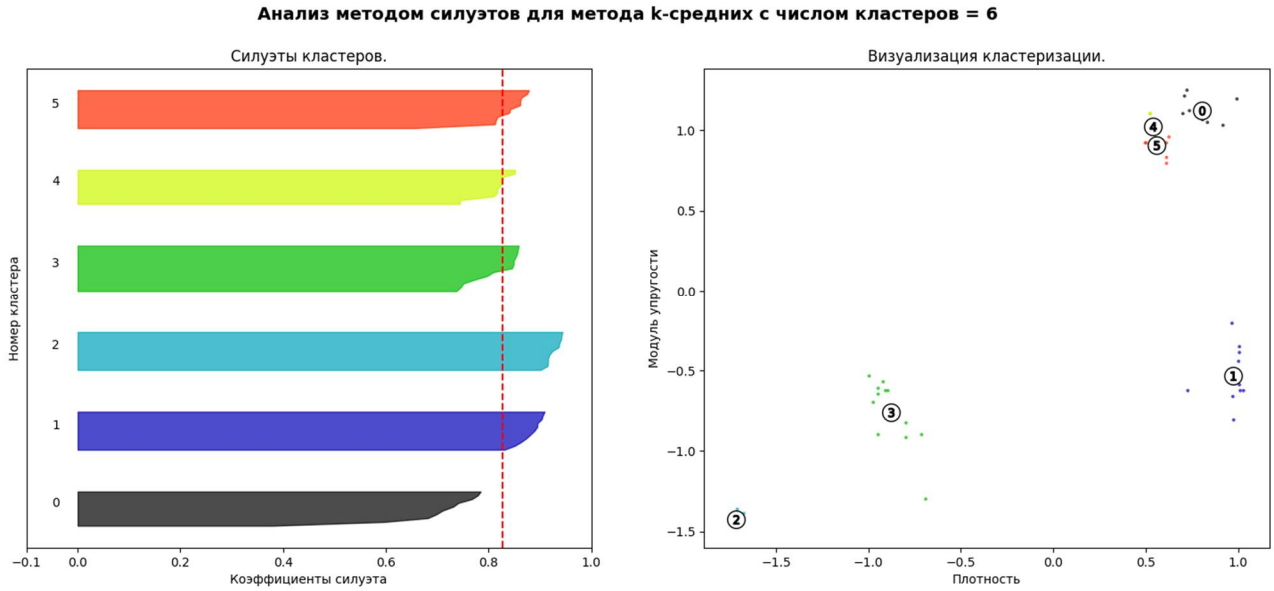

Рисунок 5 – Визуализация работы метода k-средних с выбором начального количества кластеров, равного 6, с помощью метода силуэтов
Figure 5 – Visualization of k-means method with the selection of initial number of clusters equal to 6 using the silhouette method

На рисунке 5 видно, что силуэты достаточно сбалансированы по количеству и внутренней инерции кластеров. Также на визуализации можно увидеть разбиение на кластеры и их центроиды. Заметно, что кластеры 0,4 и 5 находятся достаточно близко в пространстве плотности и модуля упругости, что говорит о необходимости учитывать больше параметров материалов. Остальные кластеры легко различимы, что свидетельствует о качественной кластеризации.

Таким образом, алгоритм позволил автоматически определить количество кластеров с помощью метода силуэтов и выполнил кластерный анализ методом k-средних, получив при этом разбиение коллекции материалов на кластеры. Если провести обратное наложение изначальных классов материалов, то станет заметно, что материалы идеально разбились в свои исходные классы. Продемонстрируем это в виде дендрограммы на части исходной коллекции (рисунок 6).

Дендрограмма на рисунке 6 показывает иерархическое разделение данных, позволяя визуализировать отношения между кластерами и подгруппами данных. На дендрограмме можно видеть, что данные разбиваются на 6 основных ветвей, что согласуется с результатами k-средних. При этом каждый материал оказался в том же кластере, что и материалы из той же группы. Эти результаты еще раз демонстрируют, что методы кластеризации могут быть весьма полезны в задачах классификации в области материаловедения, в частности там, где отсутствует информация о классификации материала, полагаясь исключительно на его свойства.

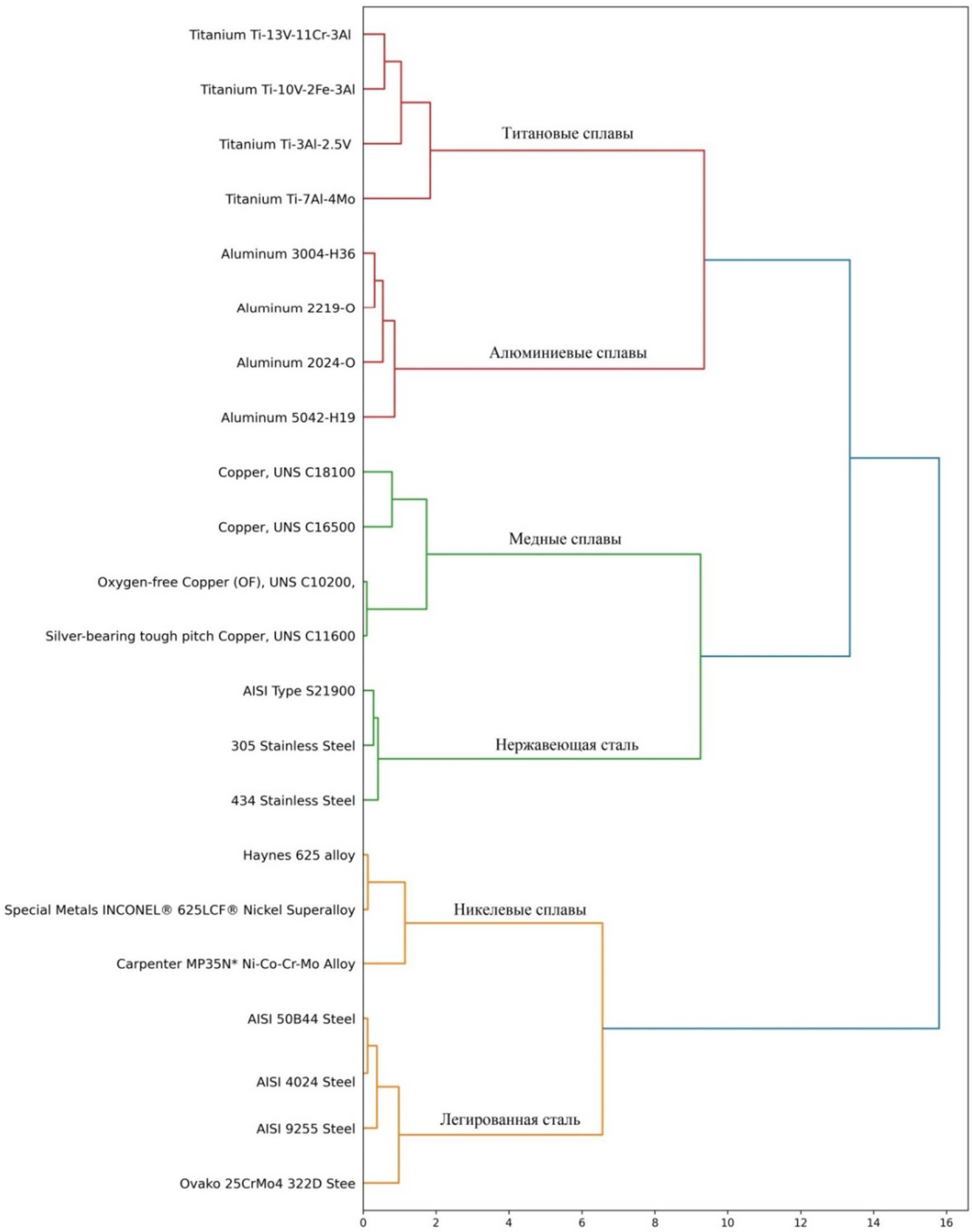

Рисунок 6 – Дендрограмма разбиения исходной коллекции
Figure 6 – Dendrogram of original collection partition

Заключение

В ходе проведенного исследования была реализована методика кластеризации данных в сфере материалов с использованием алгоритма k-средних. Предлагается автоматизировать определение количества кластеров с помощью метода силуэтов, так как он позволил достичь высокого качества кластеризации. Результаты были проверены с помощью дендрограммы.

Методика продемонстрировала свою эффективность в разбиении небольшой коллекции (< 1000 материалов). Каждый из кластеров целиком и полностью повторяет группы классификации данной коллекции. Это разделение позволяет легко идентифицировать группы материалов с похожими характеристиками, что является важным для дальнейшего анализа и применения в различных областях материаловедения.

Данная методика перспективна для использования в поиске по материалам, а также поиска схожих по свойствам материалов, используя лишь информацию о базовых свойствах материалов. Это может упростить процесс поиска аналогов в различных приложениях хранения материалов, особенно там, где может возникнуть потребность в замене материала похожим по определенным характеристикам.

В дальнейшем планируется также изучить взаимосвязь свойств материалов и их химический состав, что может вместе с полученной моделью кластеризации открыть возможность для целенаправленного улучшения свойств материалов путем модификации их свойств и структуры.

Данную методику планируется масштабировать на более крупные и разнообразные коллекции, а также добавить дополнительные характеристики материалов. В перспективе методика может быть дополнена регрессионными моделями и быть полезной при валидации данных и заполнении недостающих свойств материалов в коллекции.

Библиографический список

1**. Landau S., Ster I.C.** Cluster analysis: overview .Á Á. – 2010. – Т. 11. – №. x12. – С. x1p.

2. **Charu A.** An Introduction to Cluster Analysis, 2018

3**. Qiu Z., Li Q., Liu S., Xu R**. Clustering-based concurrent topology optimization with macrostructure, components, and materials. Structural and Multidisciplinary Optimization. 2020.

4. **Vogt W., Nagel D.** Cluster Analysis in Diagnosis. Clinical Chemistry, 1992, 38(2), 182-198.

5. **Blashfield R.K., Aldenderfer M.S.** (1978). The Literature On Cluster Analysis. Multivariate Behavioral Research, 13(3), 271–295.

6**. Bidhendi S.K., Shirazi A.S., Fotoohi N., Ebadzadeh M.M.** Material Classification of Hyperspectral Images Using Unsupervised Fuzzy Clustering Methods. 2007. Third International IEEE Conference on Signal-Image Technologies and Internet-Based System.

7**. Zhao J., Plagge R., Ramos N.M., Simões M.L., Grunewald J.** Application of clustering technique for definition of generic objects in a material database. Journal of Building Physics, 39(2), 124–146.

8. **Goldsmith B.R., Boley M., Vreeken J., Scheffler M., Ghiringhelli L.M.** Uncovering structure-property relationships of materials by subgroup discovery. New Journal of Physics, 19(1), 013031.

9**. Kodinariya T., Makwana P.R.** Review on determining number of Cluster in K-Means Clustering. Int. J. Adv. Res. Comput. Sci.Manag. Stud. 2013, 1, pp. 90–95.

10**. Andrew Ng,** Clustering with the K-Means Algorithm, Machine Learning, 2012.

11. **Kaufman l., Rousseeuw P.,** Finding Groups in Data: An Introduction to Cluster Analysis, New York: J. Wiley & Son, 1990.

UDC 004.724

APPLICATION OF CLASSTERIZATION METHODS TO ANALYZE THE PROPERTIES OF MATERIALS

V. P. Koryachko, Dr. Sc. (Tech.), full professor, CAD department, Head of the Department, RSREU, Ryazan, Russia;
orcid.org/0000-0003-0272-673X, e-mail: koryachko.v.p@rsreu.ru
S. D. Vikulin, post-graduate student, RSREU, Ryazan, Russia;
orcid.org/0009-0002-9932-1113, e-mail: vikulin97@gmail.ru
A. V. Volkov, key project engineer, Huawei Technologies Co., Moscow, Russia;
orcid.org/0009-0008-1162-3816, e-mail: vic-volk@yandex.ru

The problem of developing cluster analysis methods to study basic characteristics of materials for the purpose of further development and implementation of intelligent search system in the field of materials science is considered. ***The aim of this work*** *is to develop cluster analysis methods to study the basic character-*

*istics of materials for further development and implementation of intelligent search system in the field of materials science. Clustering is carried out using the k-means algorithm, and validation of the results is carried out using intra-cluster and inter-cluster analysis. The determination of optimal initial parameters of cluster analysis method was carried out using elbow and silhouette methods. A collection of materials was successfully decomposed into clusters, and d*ata *hierarchical structure was visualized using a dendrogram, which confirmed the effectiveness of the method proposed.*

Keywords: *machine learning, cluster analysis, material science, k-means, elbow method, silhouette method, dendrogram.*

DOI: 10.21667/1995-4565-2024-89-77-84

References

1**. Landau S., Ster I.C.** Cluster analysis: overview. Á Á. 2010, vol. 11, no.12. – C. x1p.

2. **Charu A.** An Introduction to Cluster Analysis. 2018.

3**. Qiu Z., Li Q., Liu S., Xu R**. 2020. Clustering-based concurrent topology optimization with macrostructure, components, and materials. Structural and Multidisciplinary Optimization.

4. **Vogt W., Nagel D.** Cluster Analysis in Diagnosis. Clinical Chemistry.1992, no. 38(2), pp. 182–198.

5. **Blashfield R.K., Aldenderfer M.S.** The Literature On Cluster Analysis. Multivariate Behavioral Research. 1978, no. 13(3), pp. 271–295.

6**. Bidhendi S.K., Shirazi A.S., Fotoohi N., Ebadzadeh M. M.** Material Classification of Hyperspectral Images Using Unsupervised Fuzzy Clustering Methods. *Third International IEEE Conference on Signal-Image Technologies and Internet-Based System*. 2007

7**. Zhao J., Plagge R., Ramos N.M., Simões M.L., Grunewald J.** Application of clustering technique for definition of generic objects in a material database. *Journal of Building Physics*. 2015, no. 39(2), pp. 124–146.

8. **Goldsmith B.R., Boley M., Vreeken J., Scheffler M., Ghiringhelli L.M.** Uncovering structure-property relationships of materials by subgroup discovery. *New Journal of Physics*. 2017, no. 19(1), 013031.

9**. Kodinariya T., Makwana P.R.** Review on determining number of Cluster in K-Means Clustering. Int. J. Adv. Res. Comput. Sci. Manag. Stud. 2013, no. 1, pp. 90–95.

10**. Andrew Ng.** Clustering with the K-Means Algorithm, Machine Learning, 2012.

11. **Kaufman L., Rousseeuw P.,** Finding Groups in Data: An Introduction to Cluster Analysis, New York: J. Wiley & Son, 1990.