УДК 519.86 JEL C02

МЕТОДЫ МАШИННОГО ОБУЧЕНИЯ ДЛЯ РЕШЕНИЯ ЗАДАЧИ КЛАСТЕРИЗАЦИИ

И. С. Чергейко, Е. И. Васенкова

Белорусский государственный университет, г. Минск, Республика Беларусь

В статье рассматривается использование методов машинного обучения для решения задачи кластеризации. На основании оценки эффективности различных методов кластеризации (k-средних, иерархической кластеризации и DBSCAN) был выбран оптимальный метод и определенны основные характеристики каждого выделенного кластера.

Ключевые слова: кластеризация, машинное обучение, кластер, метод k- средних, метрика качества.

MACHINE LEARNING METHODS FOR SOLVING THE CLUSTERING PROBLEM

I. S. Charheika, E. I. Vasenkova

Belarusian State University, Minsk, Republic of Belarus

The paper considers the use of machine learning methods for solving the clustering problem. Based on the evaluation of the effectiveness of various clustering methods (k-means, hierarchical clustering and DBSCAN) the optimal method was chosen and the main characteristics of each allocated cluster were determined.

Key words: clustering, machine learning, cluster, k-means method, quality metrics.

Использование машинного обучения для оптимизации банковских процессов является активно развивающейся областью. Машинное обучение позволяет банкам анализировать данные о клиентах, их предпочтениях, поведении и разделить их несколько соответствующих групп, а затем предлагать им персонализированные продукты и услуги. Кластеризация позволяет получить однородные группы клиентов в соответствии с выявленными характеристиками, а алгоритмы машинного обучения могут определить наиболее подходящие предложения для каждого клиента и помочь банку улучшить взаимодействие с ним. Кластеризация, как метод машинного обучения, решает задачу разбиения заданной выборки объектов на непересекающиеся подмножества, называемые кластерами,

так, чтобы каждый кластер состоял из схожих объектов, а объекты разных кластеров существенно отличались.

Решение задачи кластеризации неоднозначно, так как не существует однозначно наилучшего критерия качества кластеризации; число кластеров, как правило, неизвестно заранее и устанавливается в соответствии с некоторыми субъективными критериями; результат кластеризации существенно зависит от метрики, выбор которой также субъективен и определяется экспертом.

Самым известным и наиболее используемым методом кластеризации является метод k-средних. Метод k-средних используется для кластеризации данных на основе алгоритма разбиения векторного пространства на заранее определенное число кластеров k [2]. Алгоритм представляет собой итерационную процедуру. Остановка алгоритма производится тогда, когда границы кластеров перестанут изменяться от итерации к итерации, т.е. на каждой итерации в каждом кластере будет оставаться один и тот же набор наблюдений. Преимуществом алгоритма являются скорость и простота реализации. К недостаткам можно отнести неопределенность выбора начальных центров кластеров и числа кластеров, что может потребовать некоторой априорной информации об исходных данных.

Иерархическая кластеризация – совокупность алгоритмов упорядочивания данных, направленных на создание иерархии (дерева) вложенных кластеров [3]. Результаты иерархической кластеризации обычно представляются в виде дендограммы, которая позволяет изобразить взаимные связи между объектами из заданного множества. Преимущество иерархической кластеризации в том, что можно попытаться определить нужное количество кластеров, исследуя свойства получившегося дерева. В качестве недостатка метода можно отметить факт, что алгоритм относительно медленный, так как количество вычислений быстро растет по мере увеличения размера набора данных.

Еще одним распространенным алгоритмом кластеризации на основе плотности является DBSCAN (Density Based Spatial Clustering of Application with Noise) [4]. Кластеры представляют собой плотные области некоторых объектов в пространстве данных, разделенных между собой объектами, плотность которых значительно ниже. Расположение точек в одном кластере обусловлено их соединением или некоторой связью между собой. DBSCAN можно использовать для обнаружения кластеров странной или неправильной формы. Он устойчив к выбросам, способен их обнаруживать и полностью исключать из кластеров. Еще одно преимущество DBSCAN заключается в том, что он может автоматически определять количество кластеров. DBSCAN обычно медленнее, чем K-средних, но быстрее, чем иерархическая

кластеризация. В данном методе сложно включать категориальные признаки, большинство признаков должны быть числовыми.

Для оценки качества кластеризации используют внешние и внутренние метрики качества. Внешние используют информацию об истинном разбиении на кластеры, в то время как внутренние метрики не используют никакой внешней информации и оценивают качество кластеризации, основываясь только на наборе данных. Наиболее часто в качестве метрик качества используют следующие показатели:

1) Adjusted Rand Index (ARI), который выражает выражает схожесть двух разных кластеризаций одной и той же выборки. ARI принимает значения в диапазоне [-1;1]. Отрицательные значения соответствуют "независимым" разбиениям на кластеры, значения, близкие к нулю, — случайным разбиениям, и положительные значения говорят о том, что два разбиения схожи.

2) Adjusted Mutual Information (AMI), которая измеряет долю информации, общей для обоих разбиений, т.е. насколько информация об одном из них уменьшает неопределенность относительно другого. AMI принимает значения в диапазоне [0,1]. Значения, близкие к нулю, говорят о независимости разбиений, а близкие к единице – об их схожести.

3) Гомогенность, полнота, V-мера. Первые две метрики измеряют, насколько каждый кластер состоит из объектов одного класса и насколько насколько объекты одного класса относятся к одному кластеру. Метрики принимают значения в диапазоне [0,1] и большие значения соответствуют более точной кластеризации. Для учёта обеих величин используют V-меру, которая показывает, насколько две кластеризации схожи между собой.

4) Силуэт, который показывает, насколько среднее расстояние до объектов своего кластера отличается от среднего расстояния до объектов других кластеров. Данная величина лежит в диапазоне [-1;1]. Чем больше значение силуэта, тем более четко выделены кластеры, и они представляют собой компактные, плотно сгруппированные облака точек.

Для решения задачи кластеризации использованы данные кредитных карт банковских клиентов. Набор данных содержит действия с 8950 кредитных карт клиентов за 6 месяцев, которые представлены следующими показателями: сумма остатка денег на счету держателя; частота обновления баланса; количество сделанных покупок за все время использования карт; максимальная сумма одной покупки; сумма покупок, сделанных в рассрочку; сумма денег на счете для оплаты кредитных процентов; частота совершения покупок; частота совершения разовых покупок; частота совершения покупок в рассрочку; частота снятия наличных денег; количество снятия наличных денег всего; количество переводов с карты на карту; кредитный лимит по кредитной карте; сум-

ма платежей, сделанных пользователем; минимальная сумма платежа; процент, который платит пользователь от полной стоимости продукта; срок обслуживания кредитной карты.

Для нахождения кластеров будем использовать язык программирования Python. Для нормализации количественных данных использован минимаксный метод.

Результаты кластеризации по методу k-средних представлены на рисунке 1.

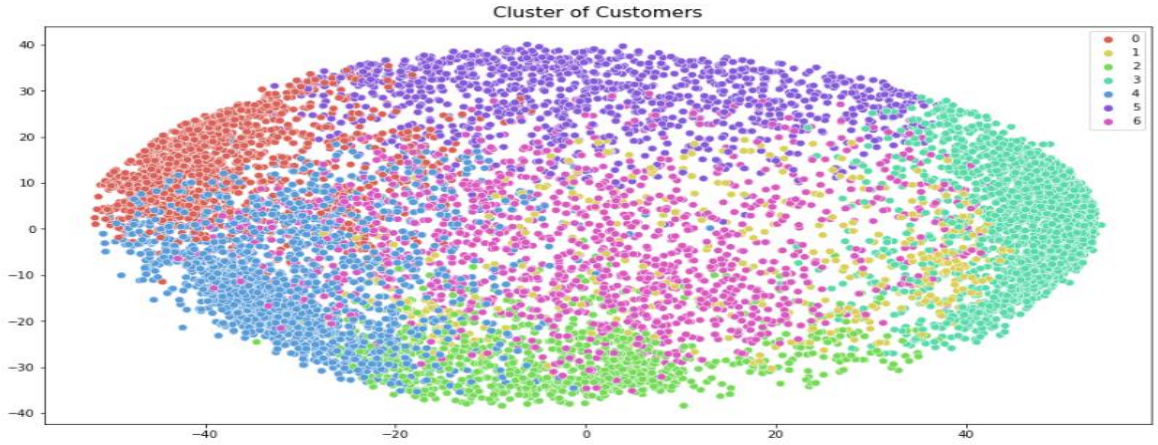

Рис.1. Разбиение данных на кластеры по методу k-средних

Источник: собственная разработка.

Результаты кластеризации по методу DBSCAN представлены на рисунке 2.

Рис. 2. Разбиение данных на кластеры по метод DBSCAN

Источник: собственная разработка

Для нахождения оптимального количества кластеров иерархическим методом используем дендограмму (рисунок 3).

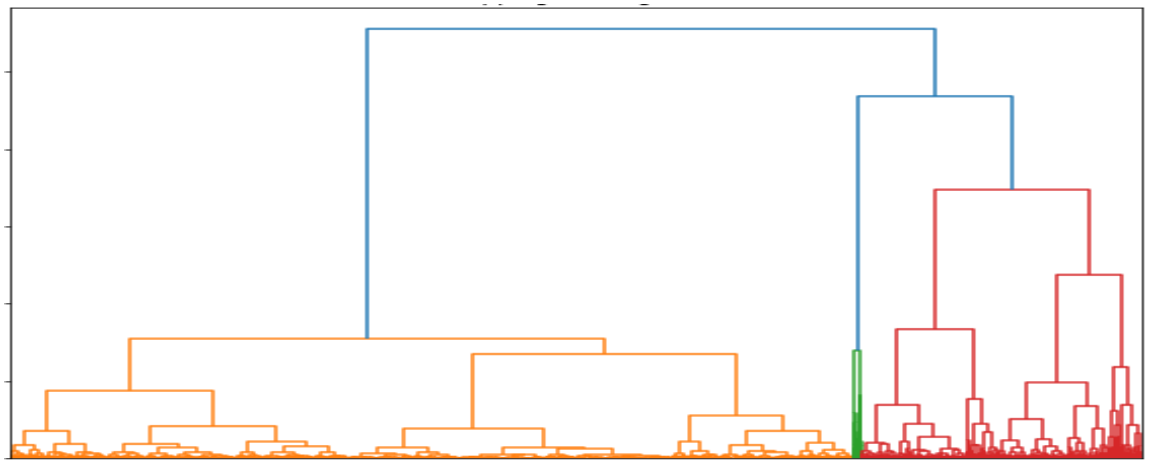

Рис. 3. Дендограмма для исходных данных

Источник: собственная разработка

Исходя из 7 отдельных блоков на дендограмме, исходные данные были разделены на соответствующее число кластеров (рисунок 4).

Рис. 4. Разбиение данных на кластеры по иерархическому методу

Источник: собственная разработка

Метрики оценки качества построенных кластеров для каждого из методов кластеризации представлены в таблице.

По данным таблицы видно, что наиболее оптимальным методом кластеризации является метод k-средних, так как все метрики оценки качества данного метода имеют максимальное значение.

На основе анализа выделенных кластеров были определенны основные характеристики каждого кластера, что в будущем поможет банку заранее относить людей к какому-либо из кластеров, предлагать ему нужные банковские продукты.

Результаты расчетов метрик оценки качества кластеров

	ARI	*AMI*	*Гомогенность*	*Полнота*	*V-мера*	*Силуэт*
K-средних	0.765	0.849	0.833	0.825	0.829	0.183
DBSCAN	0.176	0.334	0.963	0.486	0.634	0.115
Agglomerative	0.635	0.723	0.734	0.745	0.739	0.174

Примечание. Собственная разработка.

Кластер 0: Люди со средним или высоким кредитным лимитом, совершающие все виды покупок.

Кластер 1: Люди, которые чаще всего просто снимают наличные деньги.

Кластер 2: Люди, которые меньше тратят деньги со средними или высокими кредитными лимитами, покупают в основном в рассрочку.

Кластер 3: Люди, которые почти не пользуются картой, платежи совершаются только на малые нужды.

Кластер 4: Высокие траты с высоким кредитным лимитом, совершающие дорогие покупки.

Кластер 5: Люди, которые не тратят много денег и имеют кредитный лимит от низкого до среднего.

Кластер 6: Люди с очень высоким кредитным лимитом.

Таким образом, используя машинное обучение, была решена задача кластеризации клиентов, что поможет банку в будущем проводить более таргетированную маркетинговую политику, создавать другие модели, с уже заранее разбитой на кластеры выборкой, прогнозировать, предлагать клиентам подходящие им решения на основе их кредитной истории.

Библиографические ссылки

1. Обзор алгоритмов кластеризации числовых пространств данных”, 30 декабря 2012. [Электронный ресурс] - Режим доступа: https://habr.com/ru/post/164417/ (дата доступа: 20.02.2023).

2. The Most Comprehensive Guide to K-Means Clustering You'll Ever Need", by Pulkit Sharma, . [Электронный ресурс]: Analytics Vidhya, August 19, 2019 — Режим доступа: https://ijcset.net/docs/Volumes/volume6issue4/ijcset2016060404.pdf. (дата доступа: 14.02.2023).

3. Performance Evaluation of Clustering Algorithm Using Different Datasets". [Электронный ресурс]: International Journal of Advance Research in Computer Science and Management Studies, Volume 3, Issue 1 (Январь 2015) – Режим доступа: http://www.ijarcsms.com/docs/paper/volume3/issue1/V3I1-0058.pdf. (дата доступа: 18.02.2023).

4. Kelvin Salton do Prado, статья "How DBSCAN works and why should we use it?", 2 апреля 2017. [Электронный ресурс]: Сайт "Towards Data Science". – Режим доступа: https://towardsdatascience.com/howdbscan-works-and-why-should-i-use-it-443b4a191c80. (дата доступа: 20.02.2023).