

Научная статья

УДК 681.3.06

EDN VVJQSW

DOI 10.17150/2713-1734.2024.6(2).251-262

**А.В. Родионов***Байкальский государственный университет,
г. Иркутск, Российская Федерация***К.Л. Ищенко***Байкальский государственный университет,
г. Иркутск, Российская Федерация*

Исследование влияния параметров алгоритма К-ближайших соседей на метрики качества моделей

Аннотация. В данной статье было проведено исследование влияния параметра k , размера обучающей выборки и ряда популярных метрик расстояний для метода k -ближайших соседей (kNN – k -nearest neighbors) на качество моделей с использованием метрик RMSE и R^2 . Алгоритм К-ближайших соседей является одним из самых популярных среди ML-моделей для решения задач классификации и регрессии. Тщательная настройка параметров — фундаментальный аспект для достижения баланса между точностью и эффективностью модели. Важность правильного выбора параметров k и метрики расстояния является ключевым фактором для создания модели с высокой точностью. В результате исследования были получены оптимальные значения параметра k алгоритма, которые применимы для решения большинства прикладных задач. Наиболее часто используемые метрики, такие как Евклидова и Манхэттенская, показали сопоставимую эффективность по сравнению с метриками Чебышева и Махаланобиса. Практическая применимость оптимальных характеристик алгоритма делает его решения востребованными в разнообразных прикладных задачах классификации и регрессии.

Ключевые слова. Метод k -ближайших соседей, регрессия, метрики расстояний, Евклидово расстояние, расстояние городских кварталов (манхэттенское расстояние), расстояние Махаланобиса, расстояние Чебышева, расстояние Минковского.

Информация о статье. Дата поступления: 1 декабря 2023 г.; дата принятия к публикации: 23 мая 2024 г.; дата онлайн-размещения: 19 июня 2024 г.

Original article

A.V. Rodionov*Baikal State University,
Irkutsk, Russian Federation***K.L. Ishchenko***Baikal State University,
Irkutsk, Russian Federation*

Investigation the Influence of K-Nearest Neighbors Algorithm Parameters on Model Quality Metrics

Abstract. This article investigates the influence of the k parameter, the size of the training set, and various popular distance metrics on the quality of models using the

k-nearest neighbors (kNN) method, evaluated with RMSE and R2 metrics. The k-nearest neighbors algorithm is among the most popular machine learning models for classification and regression tasks. Careful parameter tuning is a fundamental aspect in achieving a balance between model accuracy and efficiency. The importance of selecting the right k parameter and distance metric is a key factor in creating a high-precision model. The research results provide optimal values for the k parameter applicable to a wide range of practical tasks. Commonly used metrics, such as Euclidean and Manhattan distances, demonstrated comparable effectiveness compared to Chebyshev and Mahalanobis metrics. The practical applicability of the algorithm's optimal characteristics makes it a sought-after solution for diverse applications in classification and regression tasks.

Keywords. k-nearest neighbors algorithm, kNN, classifiers, regression, distance metrics, Euclidian distance, Manhattan distance, Mahalanobis distance, Chebyshev distance.

Article info. Received 1 December, 2023; Accepted 23 May, 2024; Available online 19 June, 2024.

Введение

Метрические методы классификации объектов хорошо раскрывает фраза «Скажи мне, кто твой друг и я скажу, кто ты». Эти алгоритмы фактически не имеют фазы обучения и являются непараметрическими, так не делают выводов о общих правилах, которым подчиняются данные, а вместо этого просто запоминают обучающую выборку и на этапе предсказания ищут похожий объект [1]. То есть метрические методы работают с допущениями, что информация об объекте может быть получена исходя из представления о его соседях.

Одним из популярных метрических алгоритмов является алгоритм k-ближайших соседей. Хотя в чистом виде он уже редко применяется для решения прикладных задач, но является составной частью многих других методов [2].

Для метрических методов очень важно эффективно и быстро находить ближайшие объекты. Поэтому вопрос эффективности и скорости алгоритма является насущной задачей. В научной работе проведены измерения метрик качества алгоритма kNN в задачах регрессии с использованием разных параметров алгоритма, в первую очередь, зависимость от количества k соседей и используемой метрики расстояний.

Метод k-ближайших соседей

Метод k-ближайших соседей (kNN — k-nearest neighbors) представляет собой один из наиболее широко известных метрических алгоритмов, который может использоваться как для задачи классификации объектов, так и в качестве метода регрессии. Этот метод отличается своей простотой, легкостью в понимании логики работы на каждом этапе и прозрачностью в интерпретации полученных результатов¹.

¹ Классификация данных методом k-ближайших соседей // Wiki Loginom. URL: <https://loginom.ru/blog/knn>.

Если метод используется для задач классификации данных, то объект будет присвоен тому классу, который наиболее распространен среди k -соседей данного объекта, классы которых уже известны (рис. 1). Если метод используется для задачи регрессии, то объекту будет присвоено среднее значение по k -ближайшим к нему объектам, значения которых уже известны (рис. 2).

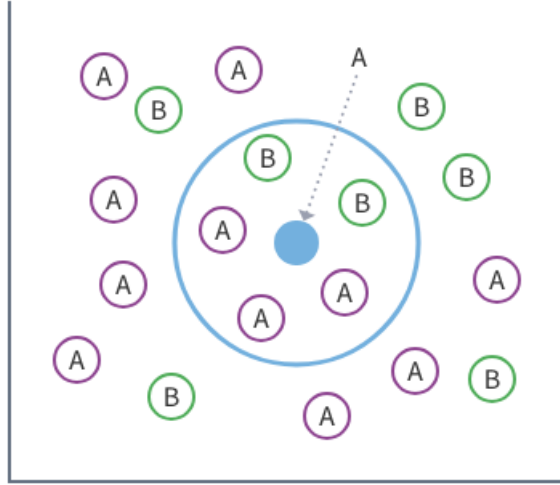


Рис. 1. Определение принадлежности объекта классу в задачах классификации

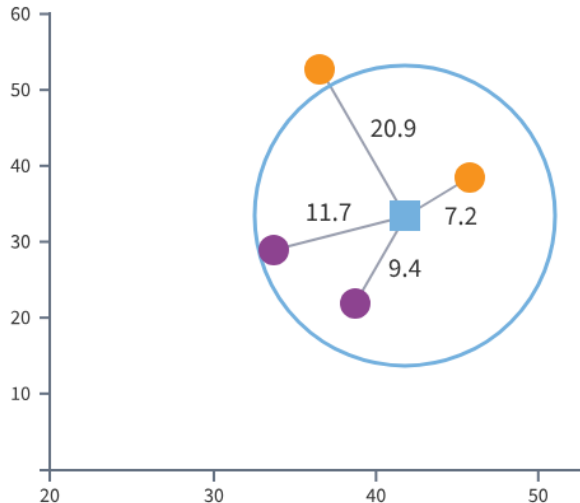


Рис. 2. Пример определение принадлежности классу в задаче регрессии

Алгоритм эффективен в работе с многомерными данными, особенно в случаях, где присутствует большое количество атрибутов. Тем не менее, прежде чем приступить к использованию, важно определить

подходящую функцию расстояния. Сталкиваясь с необходимостью этого выбора, возникает вопрос: какую именно функцию расстояния следует предпочесть? Рассмотрим несколько популярных метрик расстояний, которые можно использовать для работы kNN-алгоритма.

Метрики расстояний

1. Евклидова метрика представляет собой измерение расстояния между двумя точками в n -мерном пространстве, основанное на теореме Пифагора. Сущность данной метрики заключается в определении кратчайшего расстояния между указанными точками, выраженного в виде длины прямой, соединяющей их. В общем случае для n -мерного пространства [3]:

$$p(x, y) = \sqrt{\sum_{i=1}^n (x_i - y_i)^2}.$$

2. Манхэттенское расстояние, также известное как расстояние городских кварталов, представляет собой метрику, в которой расстояние между двумя точками определяется как сумма модулей разностей их координат [2].

$$p(x, y) = \sum_{i=1}^n |x_i - y_i|.$$

3. Расстояние Чебышева представляет собой метрику в векторном пространстве, определяемую как максимум модуля разности компонент векторов, составляющих это пространство [3].

$$p(x, y) = \max(|x_i - y_i|), \text{ где } (i = 1, \dots, n).$$

Все три вышеописанных расстояния являются частными случаями метрики Минковского:

$$p(x, y) = \left(\sum_{i=1}^n |x_i - y_i|^p \right)^{1/p},$$

где, при $p = 1$ метрика обращается в расстояние городских кварталов, при $p = 2$ это евклидова метрика, а при $p = \infty$ метрика обращается в расстояние Чебышева [3].

4. Расстояние Махаланобиса представляет собой показатель несходства между двумя случайными векторами $\vec{x}$ и $\vec{y}$, принадлежащими одному распределению вероятностей с матрицей ковариации S . Эта метрика широко применяется в задачах классификации и описывается формулой:

$$p(\vec{x}, \vec{y}) = \sqrt{(\vec{x} - \vec{y})^T S^{-1} (\vec{x} - \vec{y})},$$

где S — матрица ковариации. Если матрица ковариации является единичной, то расстояние Махаланобиса совпадает с расстоянием Евклида. В случае, когда матрица ковариации является диагональной (но не обязательно единичной), получившаяся метрика называется нормализованным расстоянием Евклида [3]:

$$p(\vec{x}, \vec{y}) = \sqrt{\sum_{i=1}^n \frac{(x_i - y_i)^2}{\sigma_i^2}}.$$

Для эффективного использования расстояния Махаланобиса в задаче классификации, ориентированной на определение принадлежности объекта к одному из классов, требуется предварительное вычисление матриц ковариации для каждого класса на обучающей выборке. В этом контексте вероятнее всего объект будет отнесен к классу, для которого значение расстояния Махаланобиса минимально. Это эквивалентно применению метода максимального правдоподобия, где более маленькое расстояние свидетельствует о более вероятной принадлежности объекта к определенному классу.

Набор данных

В качестве тестовых данных для задачи регрессии в исследовании данной статьи использовался набор по физико-химическим характеристикам португальским красным винам «Vinho Verde»². В качестве целевой переменной используется экспертная оценка от 1 до 10. Список признаков: Fixed acidity, Volatile acidity, Citric acid, Residual sugar, Chlorides, Free sulfur dioxide, Total sulfur dioxide, Density, pH, Sulphates, Alcohol, Quality (фиксированная кислотность, ароматы уксусной кислоты или этилацетата, количество лимонной кислоты, остаточный сахар, хлориды, свободный SO₂, общий SO₂, плотность, уровень pH, сульфаты, уровень алкоголя, целевой признак (качество, экспертная оценка от 1 до 10)).

Постановка эксперимента

Исследование проведено на наборе данных «Vinho Verde» используя разные параметры (режимы) работы алгоритма kNN:

1. Количество k-соседей ($k = 3, 5, 7, 10, 20, 30$).
2. Метрика расстояний (Евклидова метрика, расстояние городских кварталов, расстояние Чебышева, расстояние Махаланобиса).

² Red Wine Quality // Kaggle URL: <https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009>.

3. Размер обучающей выборки (размеры: 0.5, 0.33, 0.2 и кросс-валидация k-fold с $k = 10$ подвыборок).

В методе kNN основополагающим и ключевым параметром является k и его выбор играет решающую роль в формировании качественной модели. При выборе слишком маленьких значений k возникает явление переобучения. В таких случаях принятие решения о принадлежности объекта к классу основывается на ограниченном числе примеров, что приводит к низкой статистической значимости. При установке $k=1$ алгоритм просто присваивает новому наблюдению метку класса ближайшего объекта. Использование небольших значений k также увеличивает воздействие шумов на результаты, что проявляется в чувствительности модели к малым изменениям данных, вызывая значительные изменения в результатах регрессии³.

В противоположность этому, при установке высоких значений параметра k в процессе классификации участвует множество объектов, принадлежащих разным классам. В результате вывод оказывается грубым и недостаточно отражает локальные особенности данных. При увеличении значения параметра k происходит расширение участия объектов в классификации [4]. Таким образом, оптимальный выбор значения параметра k представляет собой баланс между достижением точности и сохранением обобщающей способности модели.

Для экспериментальной части работы использовался язык Python в среде Google Colab и библиотека машинного обучения Scikit-learn⁴. Для предобработки данных, очистки и нормализации набора использовалось средство визуального анализа данных Orange⁵.

Задачей поставленной в научной работе не являлось создание максимально правдоподобной модели. Задачей являлось исследование влияния параметров алгоритма kNN на качество модели регрессии на тестовом наборе данных.

Метрики качества модели регрессии

Оценку качества моделей мы будем проводить на основе двух показателей.

³ Классификация данных методом k-ближайших соседей // Wiki Loginom. URL: <https://loginom.ru/blog/knn>.

⁴ SciKit Learn. Nearest Neighbors. // Scikit-learn. Machine Learning in Python URL: <https://scikit-learn.org/stable/modules/neighbors.html>.; Mean squared error regression loss // Scikit-learn. Machine Learning in Python. URL: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_squared_error.html.; kNN // Orange Data Mining. Open source machine learning and data visualization. URL: <https://orangedatamining.com/widget-catalog/model/knn>.

⁵ Red Wine Quality // Kaggle URL: <https://www.kaggle.com/datasets/uciml/red-wine-quality-cortez-et-al-2009>.

MSE (Mean Square Error) представляет собой одну из наиболее распространенных метрик в задаче регрессии, используемую в качестве функции потерь. Иногда для согласования размерности показателя эффективности MSE с исходными данными, измерениям этой метрики придают такую же размерность, вычисляя квадратный корень из MSE, что приводит к созданию метрики RMSE (Root Mean Square Error). Эффективность модели в соответствии с набором данных оценивается по значению RMSE, и чем ниже это значение, тем лучше соответствие модели исходным данным⁶.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}.$$

MSE не имеет ограничения сверху, что затрудняет оценку того, насколько конкретное значение является «хорошим» или «плохим». Для установления ориентиров в оценке эффективности модели важно обратить внимание на R2 — коэффициент детерминации. R2 представляет собой универсальную меру, характеризующую степень зависимости одной случайной величины от множества других. Этот показатель служит ориентиром, обеспечивающим более осмысленное понимание качества модели в контексте оценки ошибки.

$$R^2 = 1 - \frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}.$$

Результаты эксперимента

Полученные результаты RMSE и R², для разного количества обучающей выборки, количества k-соседей, и метрик сведем в общую таблицу и построим графики зависимости RMSE от варьируемых параметров.

Результаты эксперимента замера метрик RMSE и R²

	Train data 0,5		Train data 0,33		Train data 0,2		CV (10 folds)	
	RMSE	R2	RMSE	R2	RMSE	R2	RMSE	R2
Euclidean (n = 3)	0,723	0,216	0,692	0,281	0,726	0,263	0,683	0,284
Manhattan (n = 3)	0,692	0,280	0,686	0,294	0,718	0,279	0,676	0,299

⁶ Метрики классификации и регрессии // Yandex Educational. URL: <https://education.yandex.ru/handbook/ml/article/metriki-klassifikacii-i-regressii>.

	Train data 0,5		Train data 0,33		Train data 0,2		CV (10 folds)	
Chebyshev (n = 3)	0,711	0,240	0,699	0,267	0,731	0,253	0,700	0,248
Mahalanobis (n = 3)	0,683	0,299	0,687	0,291	0,707	0,301	0,687	0,276
E u c l i d e a n (n = 5)	0,686	0,293	0,677	0,311	0,686	0,342	0,681	0,289
M a n h a t t a n (n = 5)	0,669	0,328	0,674	0,318	0,698	0,320	0,671	0,309
C h e b y s h e v (n = 5)	0,688	0,290	0,685	0,296	0,718	0,280	0,683	0,284
Mahalanobis (n=5)	0,663	0,340	0,672	0,321	0,688	0,338	0,664	0,323
Euclidean (n = 7)	0,675	0,315	0,676	0,315	0,698	0,319	0,674	0,302
Manhattan (n = 7)	0,669	0,328	0,659	0,348	0,681	0,352	0,663	0,325
Chebyshev (n=7)	0,671	0,324	0,677	0,312	0,712	0,291	0,667	0,297
Mahalanobis (n = 7)	0,663	0,340	0,671	0,323	0,691	0,333	0,663	0,327
Euclidean (n = 10)	0,665	0,337	0,668	0,329	0,698	0,319	0,666	0,319
Manhattan (n = 10)	0,668	0,330	0,659	0,347	0,680	0,355	0,658	0,336
Chebyshev (n = 10)	0,671	0,325	0,668	0,330	0,698	0,319	0,674	0,303
Mahalanobis (n = 10)	0,659	0,348	0,665	0,336	0,691	0,333	0,666	0,319
Euclidean (n = 20)	0,662	0,341	0,659	0,348	0,685	0,344	0,660	0,331
Manhattan (n = 20)	0,664	0,338	0,661	0,344	0,684	0,346	0,656	0,339
Chebyshev (n = 20)	0,672	0,321	0,674	0,318	0,699	0,317	0,663	0,326
Mahalanobis (n = 20)	0,661	0,344	0,664	0,338	0,691	0,334	0,659	0,333
Euclidean (n = 30)	0,666	0,335	0,663	0,340	0,688	0,339	0,660	0,331
Manhattan (n = 30)	0,664	0,338	0,657	0,351	0,691	0,333	0,660	0,332
Chebyshev (n = 30)	0,672	0,321	0,673	0,319	0,702	0,312	0,667	0,318
Mahalanobis (n = 30)	0,664	0,329	0,660	0,346	0,700	0,315	0,661	0,330

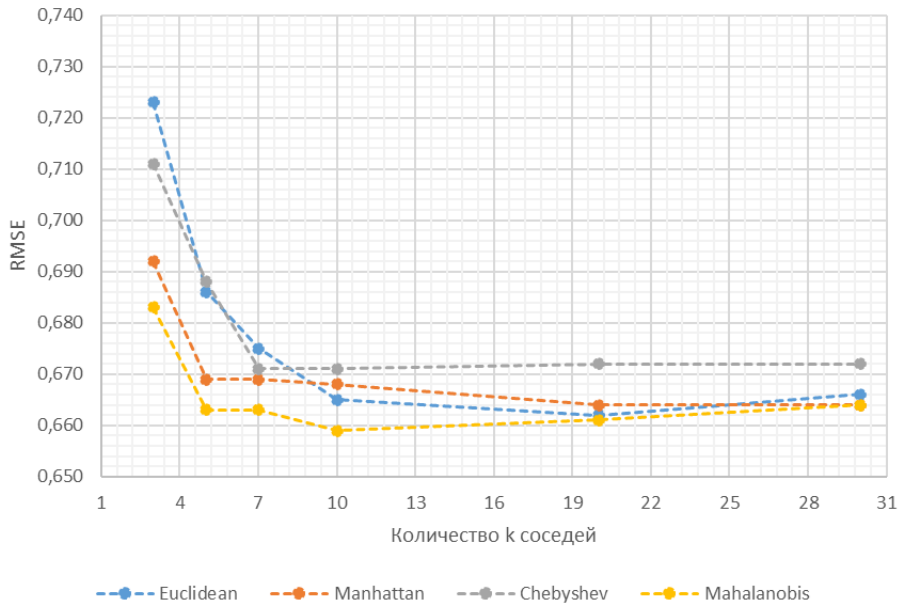


Рис. 3. График зависимости $RMSE$ от количества k -соседей и используемой метрики расстояний для обучающей выборки 0,5

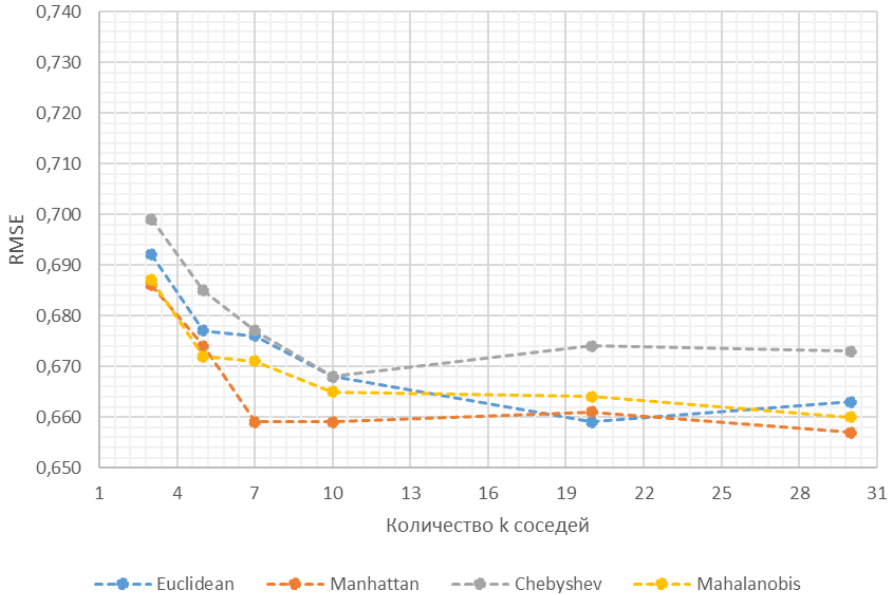


Рис. 4. График зависимости $RMSE$ от количества k -соседей и используемой метрики расстояний для обучающей выборки 0,33

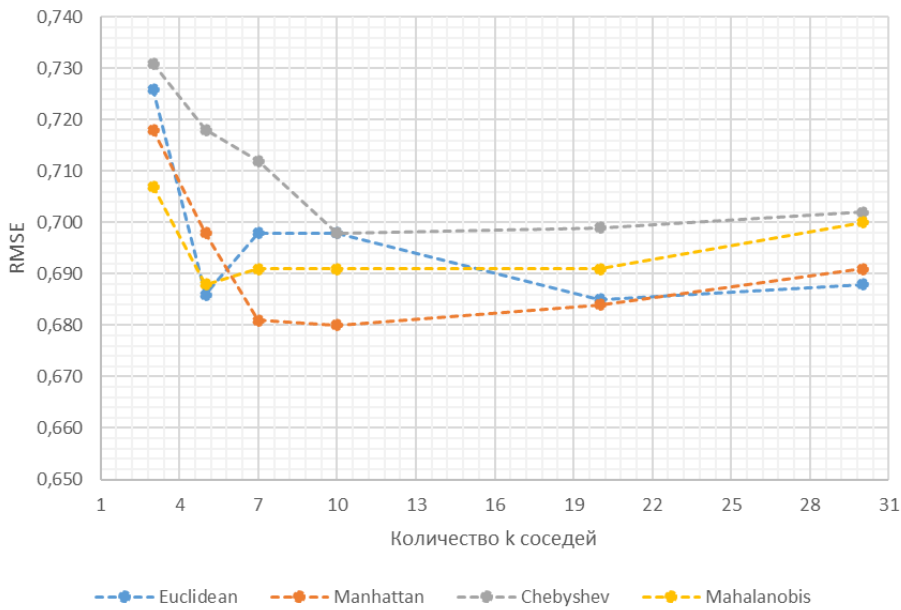


Рис. 5. График зависимости *RMSE* от количества *k*-соседей и используемой метрики расстояний для обучающей выборки 0,2

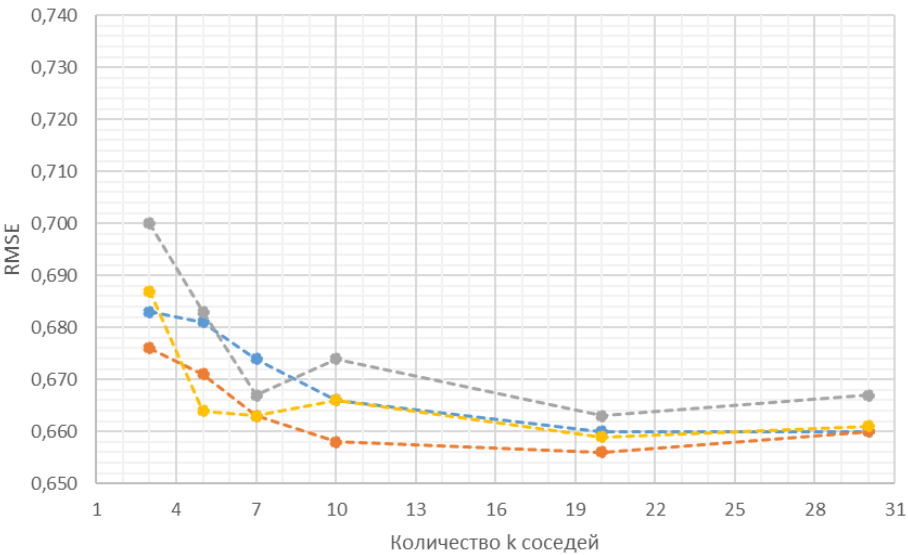


Рис. 6. График зависимости *RMSE* от количества *k*-соседей и используемой метрики расстояний для кроссвалидации из 10 частей

Выводы

Для задач классификации данных настройки параметров алгоритмов имеют решающее значение для построения качественной модели данных. Хотя алгоритм kNN обладает рядом достоинств, та-

кими как: легкость реализации и интерпретации результатов, нечувствительность к выбросам данных — так же у него есть минусы. Прежде всего, стоит отметить значительные вычислительные издержки, поскольку при определении принадлежности объекта к классу используются все имеющиеся данные. Это особенно ощутимо в случае обширных объемов данных, что приводит к высокой трудоемкости из-за необходимости вычисления расстояний до всех примеров. Сам процесс расчета расстояний представляет собой отдельную задачу, сопряженную с дополнительными вычислительными затратами.

Поэтому, выбор оптимального значения параметра k -соседей и выбор функции расстояния способствует как высокому качеству классификации алгоритма, так и экономии времени при метрическом расчете.

В исследовании, на примере реальных данных с варьированием параметров алгоритма показано, что не имеет смысла большое увеличение значение k -соседей, так как это не ведет к более качественному предсказанию. Оптимальным значением k являются значение 5, 7.

Что касается метрических расстояний, то стандартные используемые метрики в алгоритме kNN (евклидово расстояние и расстояние городских кварталов) по своим результатам значительно не отличаются от остальных проверяемых метрик (расстояние Чебышева, расстояние Махаланобиса). Хотя расстояние Махаланобиса во многих случаях показывает лучшие результаты при небольших значениях k -соседей, но следует учитывать, что метрика Махаланобиса очень чувствительна к выбросам данных, особенно если выбросы присущи только определенным классам в наборе данных.

Результаты исследования применимы к большинству других наборов данных из области прикладных расчетов в экономике и науке при использовании метрического алгоритма kNN . Использование метрик отличных от евклидовой и манхэттенской (то есть метрики Минковского с параметрами $p = \{1, 2\}$), а также большое значение параметра k может быть характерно только для специализированных расчетов там, где обусловлено самой предметной областью данных.

Список использованной литературы

1. Malkov Y.A. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs / Y.A. Malkov, D.A. Yashunin // IEEE Transactions on Pattern Analysis and Machine Intelligence. — 2018. — Vol. 42, no. 4. — P. 824–836.
2. Di Franco G. Machine learning, artificial neural networks and social research / G. Di Franco, M. Santurro // Qual Quant. — 2021. — No. 5. — P. 1007–1025.
3. Deza M.M. Encyclopedia of Distances / M.M. Deza, E. Deza. — 4 изд. — Berlin Heidelberg : Springer-Verlag, 2016. — 650 p.
4. Лебедев И.С. Адаптивное применение моделей машинного обучения на отдельных сегментах выборки в задачах регрессии и классификации / И.С. Лебедев. — DOI 10.31799/1684-8853-2022-3-20-30. — EDN ZOEVFC // Информационно-управляющие системы. — 2022. — № 3. — С. 20–30.

References

1. Malkov Y.A., Yashunin D.A. Efficient and Robust Approximate Nearest Neighbor Search Using Hierarchical Navigable Small World Graphs. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018, vol. 42, no. 4, pp. 824–836.
2. Di Franco G., Santurro M. Machine Learning, Artificial Neural Networks and Social Research. *Qual Quant*, 2021, no. 5, pp. 1007–1025.
3. Deza M.M., Deza E. *Encyclopedia of Distances*. 4th ed. Berlin Heidelberg, Springer-Verlag, 2016. 650 p.
4. Lebedev I.S. Adaptive Application of Machine Learning Models on Separate Segments of a Data Sample in Regression and Classification Problems. *Informatsionno-upravlyayushchie sistemy = Information management systems*, 2022, no. 3, pp. 20–30. (In Russian). EDN: ZOEVFC. DOI: 10.31799/1684-8853-2022-3-20-30.

Информация об авторах

Родионов Алексей Владимирович — кандидат технических наук, доцент, кафедра математических методов и цифровых технологий, Байкальский государственный университет, г. Иркутск, Российская Федерация, e-mail: RodionovAV@bgu.ru.

Ищенко Константин Леонидович — магистрант, Институт культуры, социальных коммуникаций и информационных технологий, Байкальский государственный университет, г. Иркутск, Российская Федерация, e-mail: is.konstantin@gmail.com.

Information about the Authors

Aleksei V. Rodionov – PhD in Technical Sciences, Associate Professor, Department of Mathematical Methods and Digital Technologies, Baikal State University, Irkutsk, Russian Federation, e-mail: RodionovAV@bgu.ru.

Konstantin L. Ishchenko – Master's Degree Student, Institute of Culture, Social Communications and Information Technologies, Baikal State University, Irkutsk, Russian Federation, e-mail: is.konstantin@gmail.com.

Вклад авторов

Все авторы сделали эквивалентный вклад в подготовку публикации. Авторы заявляют об отсутствии конфликта интересов.

Contribution of the Authors

The authors contributed equally to this article. The authors declare no conflicts of interests.

Для цитирования

Родионов А.В. Исследование влияния параметров алгоритма k-ближайших соседей на метрики качества моделей / А.В. Родионов, К.Л. Ищенко. — DOI 10.17150/2713-1734.2024.6(2).251-262. — EDN VVJQSW // *System Analysis & Mathematical Modeling*. — 2024. — Т. 6, № 2. — С. 251–262.

For Citation

Rodionov A.V., Ishchenko K.L. Investigation the Influence of K-Nearest Neighbors Algorithm Parameters on Model Quality Metrics. *System Analysis & Mathematical Modeling*, 2024, vol. 6, no. 2, pp. 251–262. (In Russian). EDN: VVJQSW. DOI: 10.17150/2713-1734.2024.6(2).251-262.