Study of Supervised K-Nearest Neighbors Algorithm

Prof. Londhe D.R.[1], Monika khot[2], Mansi Shitole[3], Payal Palve[4], Rutika Patil[5]
[1](Professor, SRCOE, Department of Computer Engineering Pune)
[2],[3,4,5](Student, SRCOE, Department of Computer Engineering Pune)

Abstract:*The K-Nearest Neighbors (KNN) algorithm is a straightforward, flexible method used for both classification and regression tasks in machine learning. It works by predicting the label or value of a data point based on the labels or values of its closest neighbors in the data.In classification, KNN looks at the labels of the k nearest neighbors of a given point and assigns the most common label among them. In regression, it predicts the value by averaging the values of the k nearest neighbors. One of the key features of KNN is that it doesn't make assumptions about the underlying data (this is called "non-parametric"), and it doesn't build an explicit model during training ("lazy learning"). Instead, it simply stores all the data and makes predictions when it needs to, by comparing new points to the stored data. However, KNN's performance depends on a few factors: the number of neighbors (k) chosen, the distance metric used to measure similarity (such as Euclidean distance), and the scale of the data features. While KNN can be slow and computationally expensive, especially with large datasets, it's easy to implement and works well for problems where the decision boundaries are complex or not linear. It's a useful algorithm for many practical applications despite its computational drawbacks.*

Key Word*: KNN algorithm, Supervised learning, AI (Artificial Intelligence), Ecommerce, Recommendation, ML (Machine Learning).*

I. Introduction

The K-Nearest Neighbors (KNN) algorithm is a simple, easy-to-understand method used in machine learning to make predictions for both classification and regression problems. It's a supervised learning technique, meaning it learns from examples (labeled data) and then uses that knowledge to predict the outcome for new, unseen data. Here's how it works: When you give the algorithm a new data point, it looks for the k closest data points from its training set (the "neighbors"). It then uses the information from these nearest neighbors to make a decision. In classification, KNN assigns a label to the new data point based on what the majority of its nearest neighbors are labeled. For example, if most of the nearest neighbors are labeled as "A," the new data point will be classified as "A." In regression, instead of a label, KNN predicts a value by averaging the values of its nearest neighbors.

II. Literature Review

Zhe Wang "An Improved Multilabel k-Nearest Neighbor Algorithm Based on Value and Weight" This paper introduces a novel algorithm, VWML-kNN, designed to address the challenge of classifying imbalanced multilabel data. It separates labels into minority and majority categories and applies different strategies to each. The algorithm calculates the Mean Average Precision (MAP) and uses distance-based weights from the nearest neighbors to classify minority labels, which often have fewer examples in the dataset.The results from experiments on various datasets show that VWML-kNN performs well, especially on datasets with high Total Class Strength (TCS) and Mean Inverse Rank (MeanIR), indicating its effectiveness in imbalanced classification tasks. The algorithm is suitable for applications like drug molecule identification, building identification, and text categorization. However, the algorithm has some limitations, such as its reliance on the Euclidean distance metric and its limited feature integration. Future research will focus on improving distance calculation methods and exploring label relationships to enhance the algorithm's performance.[1]

Shichao Zhang "Challenges in KNN Classification" This paper provides a comprehensive review of the latest research on KNN classification, specifically focusing on its four most challenging issues. The paper highlights the key findings from recent work within the authors' own research group, distinguishing their approach from other existing methods. It emphasizes the introduction of new research directions aimed at advancing KNN classification.[2]

Gongde Guo “KNN Model-Based Approach in Classification” This paper introduces a new approach to address the limitations of the KNN algorithm, particularly its inefficiency and reliance on the choice of k. To overcome these issues, the proposed method selects a subset of representative data points from the training dataset, using additional information to better capture the characteristics of the entire dataset. Each representative is chosen based on an optimal value of k, which is determined automatically by the dataset itself, eliminating the need for user intervention. Experimental results on six public datasets show that this KNN Model is highly competitive in classification tasks. Its average classification accuracy is comparable to that of C5.0 and traditional KNN. Additionally, the model significantly reduces the size of the dataset used for classification, with an average reduction rate of 90.41% in the number of data points. This makes the kNNModel a promising alternative to traditional kNN, especially for applications like dynamic web mining in large data repositories.[3]

Shichao Zhang “Efficient kNN Classification With Different Numbers of Nearest Neighbors” In this paper, we introduce two new kNN classification algorithms: kTree and k*Tree. These methods aim to select the optimal k-value for each test sample, making kNN classification more efficient and effective. The main idea behind our approach is to design a training stage that reduces the computational cost during the test stage, while also enhancing classification performance. We conducted two sets of experiments to compare our methods with other existing approaches. The results show that kTree and k*Tree outperformed the competing methods in both classification accuracy and running cost. Looking ahead, we plan to focus on improving the performance of these algorithms when dealing with high-dimensional data.[4]

III. K-Nearest Neighbors Archietecture

K-NEAREST NEIGHBOR(KNN) :-

The K-Nearest Neighbors (KNN) algorithm is unique because it doesn't build a separate model for training. Instead, it simply stores the entire dataset and uses it directly to make predictions. There is no explicit "training phase" in KNN, as the algorithm doesn’t learn any parameters in advance.When predicting for a new instance (let’s say xxx), KNN searches through the entire dataset to find the k most similar data points (or neighbors). It then uses these neighbors to predict the outcome. For classification tasks, the prediction is based on the majority class among the k nearest neighbors. For regression, the prediction is usually the average value of the output variable of these neighbors.

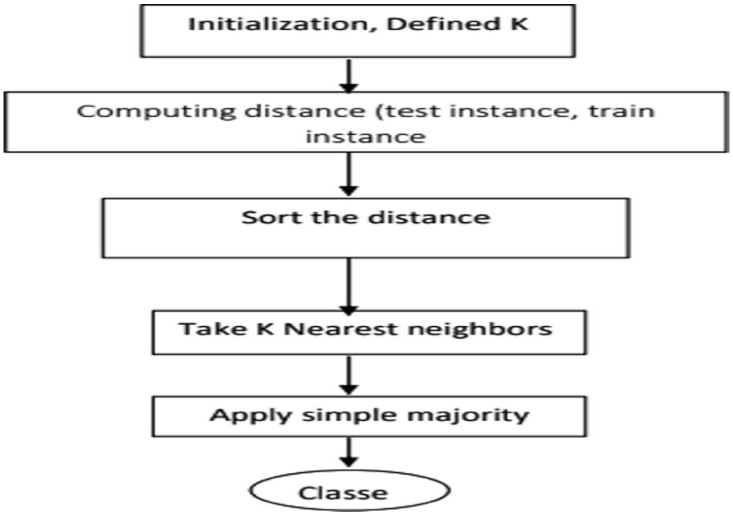
Initialization, Defined K
Computing distance (test instance, train instance
Sort the distance
Take K Nearest neighbors
Apply simple majority
Classe

Fig.1.KNN Architecture

KNN Works:

1. **Data Preprocessing:**
   Before using KNN, it's important to normalize or scale the features (attributes) so that all features contribute equally to the distance calculation. This ensures that no feature dominates due to its larger numerical range.
2. **Choose K:**
   The next step is to decide on the number of nearest neighbors, K, that will be considered when making predictions. This is a key parameter in the algorithm and influences its accuracy.
3. **Distance Metric:**

KNN calculates the distance between data points to measure their similarity. Common distance metrics include Euclidean distance (straight-line distance) or Manhattan distance (sum of absolute differences in coordinates).

4. **Find Neighbors:**
   Once the distance is calculated, KNN identifies the K nearest neighbors (data points that are closest to the query point) based on the chosen distance metric.
5. **Voting (for Classification):**
   In a classification problem, the algorithm assigns a class label to the query point based on the majority class among the K nearest neighbors. The class that appears most frequently among the neighbors becomes the predicted label.
6. **Averaging (for Regression):**
   In a regression task, the algorithm predicts the value of the target variable by averaging the values of the K nearest neighbors.

IV. Advantages

1. **Easy to Understand and Implement:** KNN is simple and intuitive, making it beginner-friendly with no complex formulas. Flexible (Non-Parametric): KNN doesn't assume anything about the data's structure, so it can handle various data types, whether linear or non-linear.
2. **No Training Required:** KNN doesn't require a training phase. It stores the entire dataset and uses it for predictions, making it easy to adapt to new data.
3. **Handles Multi-Class Problems:** KNN can naturally handle problems with multiple categories, unlike algorithms designed for binary classification.
4. **Works Well with Non-Linear Data:** KNN is effective with non-linear decision boundaries, as it relies on similarity rather than linear relationships.

V. Disadvantages

1. **Slow for Large Datasets:** KNN can be slow and computationally expensive, especially with large datasets, as it needs to calculate the distance between the test point and every point in the training set.
2. **Sensitive to the Choice of k:** The algorithm's performance depends on selecting the right number of neighbors (k). A small k may lead to overfitting, while a large k can cause underfitting. Finding the optimal k often requires experimentation or cross-validation.
3. **Affected by Irrelevant Features:** KNN is sensitive to irrelevant features, which can distort distance calculations. This is especially problematic in high-dimensional data due to the "curse of dimensionality."
4. **Requires Feature Scaling:** KNN is affected by the scale of features. If features have different ranges, the algorithm may give more weight to those with larger values. Feature normalization or standardization is usually required.
5. **Memory Intensive:** Since KNN stores the entire training dataset, it can be memory-intensive, making it impractical for very large datasets.

VI. Application

1. **Marketing and Advertising:**
   Targeted Campaigns: Social media analytics helps businesses create marketing campaigns that are tailored to specific groups by understanding audience demographics, interests, and behaviors.
   Ad Performance Tracking: Advertisers can track how their ads are performing in real-time, making adjustments to targeting and content to improve results.
   Influencer Marketing: Brands can identify key influencers on social media to collaborate with, helping to expand their reach and build credibility for their campaigns.
   Content Optimization: Insights from social analytics guide marketers in creating content that resonates more with their audience, increasing engagement and shareability.

2. **Customer Service and Support:**
   Sentiment Analysis: Social media analytics can help businesses monitor customer sentiment, identifying potential issues early and allowing them to respond to complaints or feedback quickly.
   Improved Responsiveness: By tracking comments, questions, and complaints on social media, companies can provide timely and effective customer service, improving overall customer satisfaction.
   Crisis Management: Social analytics can alert businesses to a sudden rise in negative sentiment, enabling them to react quickly, manage the situation, and protect their reputation.

3. **Product Development and Innovation:**
   Customer Feedback Analysis: Social media provides direct feedback from customers through comments and reviews, which can guide businesses in improving products or services and even developing new ones.
   Trend Analysis: Analyzing trending topics helps businesses identify emerging consumer needs or interests, enabling them to innovate and stay ahead of the market.

4. **Brand Monitoring and Reputation Management:**
   Brand Sentiment Tracking: Social media tools can track how people feel about a brand, helping businesses detect both positive and negative sentiment, and take action to improve their reputation.
   Competitor Analysis: By monitoring competitors' social media activity, businesses can gain insights into their strategies, strengths, and weaknesses, allowing them to refine their own approach.

5. **Sales and Lead Generation:**
   Identifying Potential Leads: Social analytics helps businesses identify users who show interest in their products or services, allowing them to engage with potential leads more effectively.
   Sales Conversion Optimization: By analyzing user behavior and engagement patterns, sales teams can tailor their outreach efforts to increase the likelihood of turning leads into customers.

6. **Public Relations and Communications:**
   Managing Public Perception: Social media analytics helps PR teams monitor how the public is reacting to announcements, events, or crises, enabling them to adjust messaging and maintain a positive image.
   Campaign Effectiveness: PR teams can measure the success of press releases, events, or social campaigns by analyzing metrics like shares, mentions, and overall engagement to assess their impact.

VII. Conclusion

The K-Nearest Neighbors (KNN) algorithm is a simple yet effective machine learning method used for both classification and regression. Its main strength is its simplicity and flexibility, as it doesn't assume anything about the data and works well for complex, non-linear relationships. KNN is easy to implement and doesn't require a training phase, making it quick to adapt to new data. However, KNN does have some drawbacks. It can be slow for large datasets because it calculates the distance between the test point and every training point. The performance also depends on the choice of k (the number of neighbors), with a small k leading to overfitting and a large k causing underfitting. KNN can also struggle with irrelevant features or high-dimensional data, which can affect its accuracy. Despite these challenges, KNN remains a popular choice for many applications, especially when the data is not too large or high-dimensional. With some optimizations, such as scaling features or using distance-weighted neighbors, KNN can be very effective for solving classification and regression problems.

References

[1]. Wang, Z.; Xu, H.; Zhou, P.; Xiao, G. An Improved Multilabel k-Nearest Neighbor Algorithm Based on Value and Weight. Computation 2023, 11, 32. https://doi.org/10.3390/ computation11020032

[2]. Shichao Zhang "Challenges in KNN Classification" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 34, NO. 10, OCTOBER 2022

[3]. Gongde Guo "KNN Model-Based Approach in Classification" this publication at: https://www.researchgate.net/publication/2948052

[4]. Shichao Zhang "Efficient kNN Classification With Different Numbers of Nearest Neighbors" IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, VOL. 29, NO. 5, MAY 2018

[5]. Labhade, Neelam S., Parul S. Arora, and Y. Sharma. "Study of neural networks in video processing." J Emerg Technol Innov Res (JETIR) 6.3 (2019): 330-335

[6]. Neelam Labhade-Kumar "To Study Different Types of Supervised Learning Algorithm" May 2023, International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), Volume 3, Issue 8, May 2023,PP-25-32, ISSN-2581-9429 DOI: 10.48175/IJARSCT-10256

[7]. Neelam Labhade-Kumar "To Study the Different Types of Face Reorganization Algorithm" May 2023, International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), Volume 3, Issue11 , May 2023,PP-108-114, DOI: 10.48175/IJARSCT-10572

[8]. Neelam Labhade-Kumar "To Study Different Algorithms of Machine Learning to Detect Mobile Botnets" International Journal of Advanced Research in Science, Communication and Technology (IJARSCT), Volume 12, Issue 5 , May 2023,PP-5979-5982, 10.15680/IJIRSET.2023.1205238

[9]. Neelam Labhade-Kumar To Study Different Types Of Machine Learning Algorithm For Gesture Recognition, International Research Journal of Modernization in Engineering Technology and Science, Volume 5, Issue 5 , May 2023,PP- 7751-7754, https://www.doi.org/10.56726/IRJMETS40656