Proceedings of the 3rd International Conference on Environmental Geoscience and Earth Ecology
DOI: 10.54254/2753-8818/93/2025.22373

A Comparison Between the K-Nearest Neighbors Algorithm and Logistic Regression in the Field of Cell Type Annotation

Lezhou Wen

College of Life Sciences, Sichuan University, Chengdu, China
lezhouwen@stu.scu.edu.cn

Abstract: With the advancement of single-cell sequencing technologies, high-capacity gene expression data have made cell type annotation across diverse cell populations feasible. However, the high-dimensional and complex nature of these datasets poses challenges for algorithm selection, as traditional manual annotation methods have become inadequate. Machine learning algorithms offer a robust alternative, yet choosing the optimal algorithm remains a critical step. This study provides a detailed analysis of two classical machine learning algorithms--k-Nearest Neighbors (KNN) and Logistic Regression and compares their strengths and limitations in cell type annotation from the perspective of algorithmic principles and data characteristics, aiming to offer practical guidance for selecting machine learning approaches. KNN, a distance-based non-parametric method, excels in small-sample and nonlinear scenarios but suffers from the "curse of dimensionality" in high-dimensional spaces, requiring efficiency optimization via dimensionality reduction or locality-sensitive hashing. In contrast, LR, relying on linear assumptions, performs well with large-scale, high-dimensional data through regularization to prevent overfitting, yet its performance declines with small samples or nonlinear distributions. Each algorithm has its own benefits; the choice between algorithms should consider factors such as sample size, feature dimensionality, data quality, interpretability, and the alignment between the true data distribution and the algorithm's inherent assumptions.

Keywords: Cell Type Annotation, k-Nearest Neighbors, Logistic Regression

1. Introduction

With the rapid advancement of single-cell sequencing technologies, feature data for cell type annotation has grown exponentially. Unlike traditional morphology-based parameters (e.g., cell size, nuclear-to-cytoplasmic ratio) limited to specific cell types, gene expression data from single-cell RNA sequencing applies universally across all cell types. These datasets, however, are more complex due to their massive scale and technical noise [1]. To address these challenges, analytical workflows typically involve four critical phases: data preprocessing, upstream analysis, clustering, and ultimately cell type annotation [2]. This study focuses on the final annotation stage, which means predicting cellular categories in new datasets using established feature-classification references.

As genomic data surge, conventional marker gene selection methods prove inefficient for large-scale analyses. Machine learning offers a viable solution through its computational power for processing high-throughput data. There are numerous available algorithms, and selecting optimal

© 2025 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

methods remains challenging due to their heterogeneous strengths and limitations. Previous studies have predominantly focused on algorithm applications and dataset-specific performance comparisons, lacking systematic comparisons at the principle level. To address this gap, this paper elucidates two classical machine learning algorithms: k-nearest neighbors (KNN) and logistic regression, contrasting their strengths and limitations across diverse datasets from theoretical perspectives [3, 4]. By analyzing their data type adaptability, computational efficiency and premise assumption, this research provides guidance for algorithm selection in cell annotation workflows.

2. k-Nearest Neighbors (KNN) Method

2.1. Principles and Algorithm Description of KNN

When classifying a new sample based on its feature data, it is natural to consider which reference sample it is most similar to in terms of features, and thus assume it belongs to the same category. Building on this idea, determining the similarity of features using a computer requires manually defining methods for calculating distances between features, such as Euclidean distance, Manhattan distance, and others. In practice, since different features may have different units, it is necessary to perform standardization, such as Z-score normalization, before calculating distances. At this point, the basic framework of the KNN algorithm has been successfully established.

However, in real-world scenarios, training samples may contain noise, and two different classes may not be completely separable based on features. Relying solely on the single nearest sample in terms of feature distance often results in low prediction accuracy. To address this issue, the algorithm selects a hyperparameter “k”, which means the number of nearest samples and assigns the new sample to the class that the majority of these k samples belong to. This approach helps reduce the impact of noise and improves classification accuracy [5].

2.2. Application of KNN in Single-Cell Type Annotation

To perform cell type annotation, the first step is to collect feature data of the cells. In addition to the data obtained through scRNA-seq, other feature data such as cell morphology can be used to enhance the reliability of the algorithm. For instance, when studying white blood cells, data such as cell perimeter, nuclear ratio, and roundness can be collected from blood smear images to assist in the analysis. This multi-faceted approach allows for a more comprehensive and accurate classification of cell types [6].

In practical applications, using an exhaustive algorithm to calculate the distances between each pair of sample points can result in a prohibitively large computational load, especially when dealing with massive datasets. To reduce computational complexity, various methods such as Kd-trees, dimensionality reduction, and template compression are employed [7, 8]. Among these, the classic Kd-tree method, which recursively partitions the space, reduces the time complexity from $O(n^2)$ to $O(n \log n)$. This efficiency enables the handling of high-dimensional space searches across tens of thousands of cells in single-cell data, making it a powerful tool for managing large-scale datasets [7].

When selecting the value of the hyperparameter k, cross-validation can identify an appropriate k through exhaustive search, but this approach consumes significant computational resources, especially with large datasets. aKNNO addresses this issue by dynamically choosing the k value through statistical analysis of local distance distributions. This method adaptively selects suitable k values for different types of cells, automatically balancing the sensitivity and specificity of rare cell detection, thereby optimizing performance without the computational burden of traditional methods [9].

3. Logistic Regression Method

3.1. Principles and Model Construction of Logistic Regression

Regression methods are typically used to predict continuous numerical values, while Logistic Regression, despite its name, is actually a classic classification algorithm. Its core idea is based on the maximum likelihood method: assuming the data have a linear relationship and is independent, then using the Sigmoid function to map the results of linear regression to the interval (0, 1), it represents the probability that a sample belongs to a certain category. The category with the highest probability is then selected as the predicted class value. Its probability prediction formula (1) and loss function formula (2) are as follows [10, 11]:

$$P(\mathrm{Y=1} \mid X) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n)}} \tag{1}$$

$$J(\beta) = -\frac{1}{m}\sum_{i=1}^{m}[y_i \, ln \, P_i + (1-y_i) \, ln(1-P_i)] \tag{2}$$

Where:
$X = (x_1, x_2, \ldots, x_n)$ is the feature vector,
β_0 is the intercept term,
$\beta_1, \beta_2, \ldots, \beta_n$ are the feature coefficients,
The output value P(Y=1|X) represents the probability of Y=1 given the feature vector X.
y_i is the true label (0 or 1) of the sample,
P_i=P(Y=1|X) is the probability predicted by the model.

The extension of binary classification to multi-class problems can be achieved by employing the One-vs-Rest or Softmax regression approaches. The idea of One-vs-Rest is to transform a multi-class classification problem into multiple binary classification problems, where each time only the probability of a sample belonging to a certain class versus not belonging to that class is considered. On the other hand, the idea of Softmax is to convert the linear combination of inputs into a probability distribution, ensuring that the predicted probability of each class is between 0 and 1, and the sum of the probabilities of all classes is 1. Both methods have their own advantages and disadvantages. One-vs-Rest ignores the relationships between categories and involves a larger computational cost, while Softmax performs poorly when dealing with non-mutually exclusive data. If the number of classes is small and the classes are relatively independent, One-vs-Rest can be chosen. And if the number of categories is large or there is a strong correlation between categories, it is recommended to use Softmax.

3.2. A Meaningful Application of Logistic Regression

In a specific cell population, a robust predictive model can be constructed through the utilization of substantial high-quality data. By compiling models tailored to diverse cell populations into a centralized database, future research efforts can be streamlined, and scientific efficiency significantly improved. For instance, in the CellTypist library, researchers have already established dozens of logistic regression models for various cell populations. When the research subject aligns with these models, scientists can directly retrieve and utilize them to make predictions for new studies, thereby streamlining the research process [4].

4. A Comparison of the k-Nearest Neighbors Algorithm and Logistic Regression in Specific Contexts

From the above introduction, it is evident that the KNNs algorithm directly utilizes data for predictive analysis without relying on a specific model, while logistic regression is a traditional method that constructs a model for prediction [12]. Below, this paper will compare and analyze these two methods from the perspective of specific research scenarios to provide better guidance for model selection.

4.1. Sample Size of Reference Data

Due to the larger number of parameters in logistic regression and the simpler structure of KNN, logistic regression often struggles to find suitable parameters with small sample sizes, leading to inferior performance compared to KNN. However, once the sample size reaches a certain threshold, logistic regression can identify appropriate parameters. Since logistic regression is computationally more efficient than KNN, it is preferred when other aspects of performance are comparable.

4.2. Number of Features

KNN relies on distance-based analysis, but in high-dimensional spaces, the concept of distance can suffer from the "curse of dimensionality," making it difficult to reflect the true similarity between cells. Although dimensionality reduction methods can enable KNN to make predictions, they often result in significant information loss. Other ways like Locality-Sensitive Hashing: through hash functions, similar data points are mapped to the same hash bucket, enabling the rapid identification of potential nearest neighbors. Although it can solve the dimension problem, it requires the selection of hash functions and parameter tuning, which diminishes the advantage of the k-nearest neighbors algorithm over logistic regression in terms of model complexity [13]. Logistic regression can use ways like Ridge and Lasso regularization to solve the dimension problem, but redundancy among highly correlated features will degrade its performance [14].

In low-dimensional spaces, since the KNN algorithm is sensitive to outliers, while logistic regression, with fewer parameters, is easier to implement and optimize, logistic regression holds a distinct advantage in such scenarios.

4.3. Linearity of Data Distribution

Logistic regression assumes that the distribution of samples follows a linear relationship. When the actual relationship between samples deviates significantly from linearity, logistic regression may fail to classify correctly. In contrast, KNN directly calculates distances between sample points, making it a non-parametric method that is not constrained by the underlying data relationships [15]. This allows KNN to handle various types of relationships effectively. The linearity of datasets can be assessed by constructing logistic regression models and comparing their performance with alternative methods, though this approach often demands significant computational resources. It’s stricter to use Linear Discriminant Analysis (LDA) to analyze the linearity of data distribution. LDA can utilize its inherent mathematical properties to help determine whether there is linear separability between different cell categories. But it assumes Gaussian-distributed data with equal class covariances, maximizing class separation, and may not work when data is not fit these assumptions [16]. If the LDA outcome shows the data distribution is far to linear, it’s better not to choose logistic regression unless in other parts, it is really unfit with KNN.

5. Conclusion

This paper examined the application of the k-nearest neighbors algorithm and logistic regression in cell type annotation, comparing their strengths and weaknesses in various contexts. The KNN algorithm excels in small datasets, low-dimensional spaces, and complex, nonlinear relationships due to its non-parametric nature. However, it struggles with high-dimensional data due to the "curse of dimensionality" and is sensitive to outliers. Techniques like dimensionality reduction and locality-sensitive hashing (LSH) can mitigate these issues but introduce additional complexity. In contrast, logistic regression is highly efficient for large-scale, high-dimensional datasets and benefits from regularization techniques like L1 and L2 to prevent overfitting. However, its assumption of linearity limits its effectiveness for nonlinear relationships, and it may underperform with small datasets due to parameter estimation challenges. In low-dimensional spaces, logistic regression's simplicity and robustness to outliers make it preferable, while KNN's adaptability suits complex data distributions. This study focused on two classical algorithms, excluding the comparison with advanced approaches such as deep neural networks or ensemble methods. Future research can expand the comparison to broader algorithm families and explore multimodal data integration or focus on improving logistic regression models for specific datasets, reducing overfitting in KNN for large datasets, and addressing scalability and data quality challenges. By understanding the strengths and limitations of each method, researchers can optimize cell type annotation and advance biological research, with further advancements in machine learning techniques enhancing their applicability in single-cell RNA sequencing or other fields.

References

[1] *Jovic, Dragomirka, Xue Liang, Hua Zeng, Lin Lin, Fengping Xu, and Yonglun Luo. "Single-Cell Rna Sequencing Technologies and Applications: A Brief Overview." Clinical and Translational Medicine 12, no. 3 (2022): e694.*

[2] *Liu, Menglin. "Research On Cell Type Annotation Methods for Single-Cell Rna-Sequencing Data.", Hunan University, 2022.*

[3] *Pasquini, Giovanni, Jesus Eduardo Rojo Arias, Patrick Schaefer, and Volker Busskamp. "Automated Methods for Cell Type Annotation On Scrna-Seq Data." Computational and Structural Biotechnology Journal 19 (2021): 961-69.*

[4] *Domínguez Conde, C., C. Xu, L. B. Jarvis, D. B. Rainbow, S. B. Wells, T. Gomes, S. K. Howlett, O. Suchanek, K. Polanski, H. W. King, L. Mamanova, N. Huang, P. A. Szabo, L. Richardson, L. Bolt, E. S. Fasouli, K. T. Mahbubani, M. Prete, L. Tuck, N. Richoz, Z. K. Tuong, L. Campos, H. S. Mousa, E. J. Needham, S. Pritchard, T. Li, R. Elmentaite, J. Park, E. Rahmani, D. Chen, D. K. Menon, O. A. Bayraktar, L. K. James, K. B. Meyer, N. Yosef, M. R. Clatworthy, P. A. Sims, D. L. Farber, K. Saeb-Parsy, J. L. Jones, and S. A. Teichmann. "Cross-Tissue Immune Cell Analysis Reveals Tissue-Specific Features in Humans." SCIENCE 376, no. 6594 (2022): eabl5197.*

[5] *Dasarathy, Belur V. "Nearest Neighbor (Nn) Norms: Nn Pattern Classification Techniques." IEEE Computer Society Tutorial (1991).*

[6] *Prakisya, Nurcahya Pradana Taufik, Febri Liantoni, Puspanda Hatta, Yusfia Hafid Aristyagama, and Andika Setiawan. "Utilization of K-Nearest Neighbor Algorithm for Classification of White Blood Cells in Aml M4, M5, and M7." Open Engineering 11, no. 1 (2021): 662-68.*

[7] *Bai, Xiuxiu, Xiaoshe Dong, and Yuanqi Su. "Edge Propagation Kd-Trees: Computing Approximate Nearest Neighbor Fields." IEEE SIGNAL PROCESSING LETTERS 22, no. 12 (2015): 2209-13.*

[8] *Wu, Yingquan, Krassimir Ianakiev, and Venu Govindaraju. "Improved K-Nearest Neighbor Classification." PATTERN RECOGNITION 35, no. 10 (2002): 2311-18.*

[9] *Li, Jia, Yu Shyr, and Qi Liu. "Aknno: Single-Cell and Spatial Transcriptomics Clustering with an Optimized Adaptive K-Nearest Neighbor Graph." GENOME BIOLOGY 25, no. 1 (2024).*

[10] *Kleinbaum, David G., K. Dietz, M. Gail, Mitchel Klein, and Mitchell Klein. Logistic Regression: Springer, 2002.*

[11] *Dreiseitl, S., and L. Ohno-Machado. "Logistic Regression and Artificial Neural Network Classification Models: A Methodology Review." JOURNAL OF BIOMEDICAL INFORMATICS 35, no. 5-6 (2002): 352-59.*

[12] *Peterson, Leif E. "K-Nearest Neighbor." Scholarpedia 4, no. 2 (2009): 1883.*

[13] Lee, K. M., and K. M. Lee. "A Locality Sensitive Hashing Technique for Categorical Data." In INDUSTRIAL INSTRUMENTATION AND CONTROL SYSTEMS, PTS 1-4, edited by P. Yarlagadda and Y. H. Kim, 3159-64. International Conference on Measurement, Instrumentation and Automation (ICMIA 2012), 2013.

[14] Boonyakunakorn, P., C. Nunti, W. Yamaka, and ACM. "Forecasting of Thailand's Rice Exports Price: Based On Ridge and Lasso Regression." In PROCEEDINGS OF 2019 2ND INTERNATIONAL CONFERENCE ON BIG DATA TECHNOLOGIES (ICBDT 2019), 354-57. 2nd International Conference on Big Data Technologies (ICBDT) / 3rd International Conference on Business Information Systems Workshop (ICBIS), 2019.

[15] Abdelaal, Tamim, Lieke Michielsen, Davy Cats, Dylan Hoogduin, Hailiang Mei, Marcel J. T. Reinders, and Ahmed Mahfouz. "A Comparison of Automatic Cell Identification Methods for Single-Cell Rna Sequencing Data." GENOME BIOLOGY 20, no. 1 (2019).

[16] Luecken, M. D., and F. J. Theis. "Current Best Practices in Single-Cell Rna-Seq Analysis: A Tutorial." Molecular Systems Biology 15, no. 6 (2019).