Analysis of article type category based on KNN

Qiushi Wang

Department of computer science, University of Manchester, Manchester, United Kindom

qiushi.wang-4@student.manchester.ac.uk

Abstract. Natural Language Processing (NLP) boasts a rich historical background, evolving since the 1950s at the crossroads of artificial intelligence and linguistics. Over time, it has seamlessly converged with Information Retrieval (IR), adopting diverse approaches encompassing symbolic, statistical, and connectionist methods. This study focuses on the utilization of the k-nearest neighbours (KNN) algorithm for the categorization of articles. It delves into the feasibility of accurately classifying articles when provided with ample datasets and clearly defined category labels. Through the development of a model and the integration of the KNN algorithm, this experiment successfully conducted content-based article classification. By selecting an appropriate value for k, employing a confusion matrix for performance assessment, and predicting article categories, the experiment achieved an accuracy rate of 0.96. Nonetheless, limitations arise when dealing with small datasets or imbalanced article distributions. This paper delves into the intricacies of article type classification, emphasizing the pivotal roles of data quality and feature engineering in this process. Furthermore, it underscores the potential for in-depth exploration in various contexts using alternative methodologies, such as Bayesian, Support Vector Machines, and deep learning, providing valuable references for future research endeavours.

Keywords: Article category, KNN, confusion Matrix.

1. Introduction

NLP originated in the 1950s as the intersection of artificial intelligence and linguistics [1,.and Don Walker, Jane Robinson and Karen Spark Jones discuss when to conduct research on NLP at an ACL conference in 1987 [2]. NLP was initially distinct from text information retrieval (IR), which uses highly scalable statistical-based techniques to efficiently index and search large amounts of text. However, over time, NLP and IR have converged to some extent. Through the extension of natural language processing, the Prolog language was invented in 1970 [3]. Its syntax is particularly suitable for writing grammars, although in the simplest implementation mode (top-down parsing), the wording of the rules must be different from that of the yacc-style parser (i.e., right recursion). Top-down parsers are easier to implement than bottom-up parsers (they do not require generators), but they are much slower [1].

In a natural language processing system, the actual occurrence can be most accurately explained through a language-level approach. This is also known as the synchronic model of language, which assumes that human language processing levels follow each other in a strict sequential fashion. Psycholinguistic research suggests that language processing is dynamic because these levels can interact in various sequences. Another part is about methods of natural language processing, which is roughly divided into four categories: symbolic, statistical, connected, and mixed. Symbolic and statistical methods existed before the field even began, with connectionism emerging in the 1960s, and in the 1980s, statistical methods re-emerged in popularity due to the availability of key computational resources and the need to handle a wide range of real-world environments [4].

First of all, KNN (K-Nearest Neighbours) is a fundamental algorithm in machine learning for classification tasks. It is a learning algorithm based on neural network, which means that it stores the entire dataset in its memory for later use. The algorithm works by comparing a new, unknown object to the information stored in the dataset to determine its class. KNN’s work have three steps. First is Determine the k-nearest neighbours. Second is Aggregating class labels. Third is Making a prediction.

There are many applications based on KNN methods, such as new facial expression recognition, article type classification, and more. In facial expression recognition, Facial expression recognition includes three key parts: detection and localization, extraction and representation, recognition, and classification. The classifier plays a very important role in this application. Traditional classifiers include KNN, SVM [5, 6] and FSVM. The FSVM algorithm can effectively solve the problem of unclassifiable regions in multiple classification, improving the accuracy of classification [7, 8]. However, the disadvantage is that it introduces a fuzzy membership function, which increases the calculation amount. The classification algorithm KNN uses the Euclidean distance formula to calculate the distance to represent the similarity between two objects, but this is not accurate and the classification accuracy is not high [9]. As a matter of fact, facial images are loaded and face locations are located. If there are no faces detected in the images, an empty result is returned. Subsequently, the KNN classifier determines facial features in the test images and utilises the KNN model to identify the optimal matches for the target faces. Finally, the KNN classifier makes predictions for classification, removing those classifications outside of the threshold range [10].

The application in another direction is article type recognition, which selects an appropriate calculation method by comparing Euclidean distance and cosine distance to implement the KNN algorithm, separates article data, compares performance under different K values, and then creates a confusion matrix to examine the performance of the model. Subsequently, the behaviour of the KNN classifier on novel classes is observed.

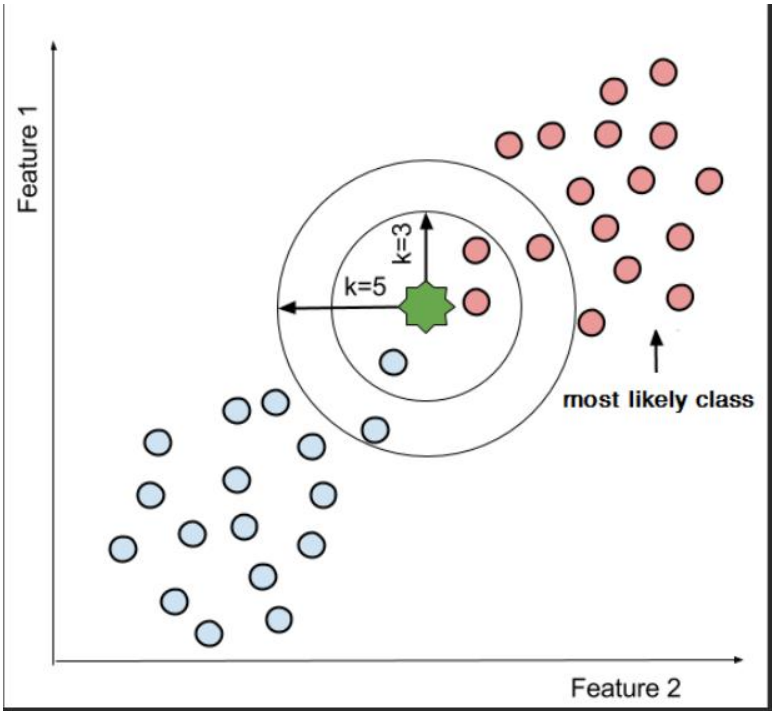

Figure 1. A sketch of KNN [7].

2. Data and method

The data source is a total of 800 articles from Reuters newswire, which belong to 4 categories: "earn" (0), "crude" (1), "trade" (2), and "interest" (3); there are 200 articles in each category. Each article is characterized by the occurrence of words. This data will be used for research on article type recognition. The K nearest neighbour (KNN) classifier is an extension of the simple nearest neighbor (NN) classifier system and in this research, used KNN algorithm is an important method, Euclidean distance d1 and cosine distances are good way to calculate the distance:

$$\mathrm{d}_1(\mathrm{x}, \mathrm{y}) = (\sum_{i=1}^{N} |x_i - y_i|)^{1/2} \tag{1}$$

$$d_{cos}(x, y) \ = 1 - \frac{\vec{x}.\vec{y}}{|x|.|y|} \tag{2}$$

The performance of the Cosine distance measurement is better when compared to the other calculate method. Cosine distance is a measure of the angle between two vectors in an n-dimensional space. It measures the similarity between two vectors in terms of the cosine of the angle between them. Cosine distances are especially useful when comparing text data. Euclidean distance, on the other hand, measures the linear distance between two points in n-dimensional space. It measures the magnitude of the difference between two vectors, considering the difference in each dimension. Euclidean distance is particularly useful when working with continuous data such as images or audio signals, so it would be better to use cosine distance for this experiment. When selecting a computational method, the next step is to divide the database into training and testing sets, with the range of k values set between 1-50 to observe the performance of different k values. Then, the confusion matrix is used to assess the performance of the model. Additionally, database separation occurs, followed by the division into training and testing sets, with a specified k value. Finally, new articles are inserted into the model to predict their types. Subsequently, a new class "sport" is added, with the new data randomly split into a training set containing a certain number of articles on "earn", "crude", "trade", and "interest", with only 3 articles from "sport". The remaining articles are retained for testing. Test the performance of a new 3-NN classifier. A sketch of KNN is shown in Fig. 1.

3. Results and discussion

After the model selection is completed, it is essential to compare different values of k by partitioning the dataset to determine the appropriate k value within the range of 1 to 50. Observations from experimental data indicate that, under the same k value, the accuracy of the testing data is slightly higher than that of the training data. This discrepancy can be attributed to the increasing standard deviation with the increment of k, signifying a simultaneous increase in the variance of error rates with k. As k grows larger, error rates exhibit a general upward trend. This suggests a decrease in classification accuracy with the increasing number of neighbors. Both the training and testing samples exhibit an increase in error rates as k increases, thus indicating that a value around 10 for k is a reasonable choice (seen from Fig. 2 and Fig. 3).

From the experiment, the result are the first Articles has been classified as crude, the second has been classified as crude, the third has been classified as interest, the forth has been classified as trade and the fifth has been classified as earn. Through random data splitting, retaining the remaining articles for testing, and populating a confusion matrix that records the classification model's performance, an accuracy of 0.96 was achieved, indicating a favourable ratio of successfully classified testing data and demonstrating positive experimental results. When three sports articles were introduced, this process was repeated six times. Through experimentation, it was observed that without a sufficient amount of training data, especially in the 'sports' category, the article accuracy significantly decreased. This phenomenon stemmed from the classifier's inadequate understanding of the 'sports' category. The lack of adequate training data raised the risk of misclassifying articles into other categories, potentially leading to errors in the confusion matrix and a reduced overall accuracy. Given the absence of samples for sports training, this scenario could be considered a zero-shot learning task. Thus, it can be concluded that small-sample learning in this context requires the addition of limited labelled samples for effectively recognizing the sports category.

From the experimental results, it can be inferred that when there is a sufficient dataset to support it, predictions can effectively yield the classification outcomes for different article categories. However, in scenarios where an ample dataset is not available and new article types are introduced, obtaining highly accurate results becomes challenging. This situation can be described as a case of small-sample learning.

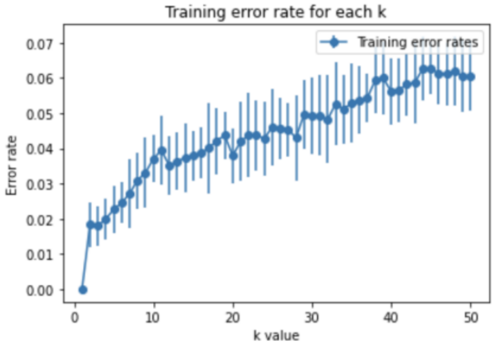

Figure 2. The training error rate in different k value (Photo/Picture credit: Original).

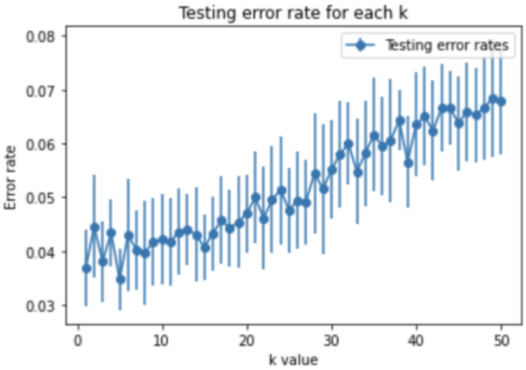

Figure 3. The testing error rate in different k value (Photo/Picture credit: Original).

4. Limitations and prospects

In this experiment, 800 articles containing a total of 6,428 words were utilized for article classification across four distinct categories. Although the experiment yielded relatively favourable results, several limitations remain evident. Firstly, there is an issue of data imbalance, as the distribution of both the quantity and types of articles in the dataset is uneven. Specifically, certain categories contain a significantly larger number of articles than others, causing the model to perform better on the more abundant categories and less effectively on the less represented ones. For instance, introducing a sports category with only three articles introduces a scenario akin to small-sample learning. Secondly, overfitting is a concern, as the dataset employed in this experiment should not be excessively small to prevent the model from fitting noise and exhibiting poor generalization. Thirdly, the practical application of the model in specific contexts may introduce additional complexity and noise, such as the presence of numerous abbreviations or emoticons in real-life articles, further complicating the classification task.

This experiment was conducted using the K-Nearest Neighbours (KNN) classification method. In the future, alternative methods can be explored to enhance the classification task. For example, the Naive Bayes classifier, a probability-based classification method that assumes independence among features, can be employed for handling text datasets. Support Vector Machines (SVM), a binary and multi-class classifier that separates different data points with an optimal hyperplane, represent another viable option.

Furthermore, deep learning approaches such as Convolutional Neural Networks (CNN) or Recurrent Neural Networks (RNN) can be considered, especially in the context of large and complex text classification tasks, as these models excel in learning intricate feature representations. By experimenting with these diverse methods, a more comprehensive and in-depth analysis of article category classification can be achieved. A comparative study can also be undertaken to assess the distinctions in results obtained from the KNN method in contrast to these alternatives.

5. Conclusion

To sum up, this paper investigates an experiment on article category classification based on the K-Nearest Neighbours (KNN) method. The experimental results demonstrate that, under the presence of a sufficient dataset and well-defined type labels, accurate categorization of articles can be achieved. The experiment involves model establishment, KNN algorithm integration, selection of the appropriate value for k, performance evaluation using confusion matrices, and article classification prediction. The experiment exhibits a high level of completeness and yields favourable results. However, it is important to acknowledge certain limitations. The dataset for this experiment consists of 800 articles, and issues such as small dataset size or imbalanced distribution of articles among categories may lead to overfitting and varying model performance. Future research can explore alternative methods, including Naive Bayes, Support Vector Machines, and deep learning, to further enhance the applicability of article classification. This study is significant in its in-depth exploration of article type classification issues and emphasizes the critical roles of data quality and feature engineering in article classification tasks. It provides valuable insights for future research in this domain.

References

[1] Nadkarni P M, Ohno-Machado L and Chapman, W W 2011 Journal of the American Medical Informatics Association vol 18(5) pp 544-551
[2] Jones K S 1994 Current issues in computational linguistics: in honour of Don Walker pp 3-16,
[3] Clocksin W F and Mellish C S 2003 Programming in Prolog: Using the ISO Standard 5th edn New York: Springer.
[4] Liddy E D 2001 Natural language processing. New York: Springer.
[5] Fradkin D 2006 Theoretical Computer Science vol 70 pp 13-20.
[6] Chang C C and Lin C J 2011 ACM transactions on intelligent systems and technology (TIST) vol 2(3) pp 1-27.
[7] Theodoridis S and Koutroumba, K 2006 Pattern Recognition Elsevier vol 12 p 20.
[8] Bishop C M and Nasrabadi N M 2006 Pattern Recognition vol 4(4) p 738.
[9] Wang X H, Liu A and Zhang S Q 2015 Optik vol 126(21) pp 3132-3134.
[10] Guo X 2021 International Conference on Communications, Information System and Computer Engineering (CISCE) pp 292-297.