Importance of Feature Extraction in Naïve Bayes Applied on Animal Classification

Tingliang Sun

College of Natural Science, Michigan State University, East Lansing, MI, 48825, USA

suntingliangha@gmail.com

Abstract. This study explores the use of the Naive Bayes classification algorithm for animal classification, particularly focusing on its computational efficiency and simplicity. The research evaluates the algorithm's performance using a dataset of 15,000 animal images across five species (cat, dog, elephant, horse, lion). The study investigates the impact of feature extraction techniques, such as Histogram of Oriented Gradients (HOG), on the classifier's accuracy. Results indicate that the Naive Bayes classifier, coupled with HOG feature extraction, performs well in terms of precision, recall, and F1-score. The classifier's simplicity and computational efficiency make it well-suited for large datasets, contributing to its scalability and usefulness in biological research. Moreover, the research emphasizes that combining Naive Bayes with other classifiers like Support Vector Machines (SVM) could mitigate some of its limitations, such as bias toward majority classes and feature independence assumptions. Overall, this study highlights the potential of Naive Bayes as a robust and efficient tool for animal classification and offers insights into its broader applicability in handling large biological datasets.

Keywords: Naïve Bayes, HOG, Animals Classification, Feature Extraction, Accuracy.

1. Introduction

This study is aim to delve into the innovative and efficient application of Naive Bayes classification algorithm in the field of animal classification, because the major way to do the animal classification is KNN or CNN and other algorithm, Naïve Bayes is hardly to seen in the animal classification field [1]. This study's primary goal is to evaluate the efficacy of the Naive Bayes algorithm and investigate the impact of feature extraction on what the algorithm would cause. Animal classification is an important task in biology and data science, because it provides a structured approach to understanding biodiversity and evolutionary relationships between different species. Efficient classification methods can help researchers organize and label large amounts of biological data more effectively, leading to significant advantages in these fields.

This study also highlights the computational efficiency of the Naive Bayes classifier, particularly its applicability to large datasets. As biological data increases exponentially with advances in data collection techniques and technologies, the ability to efficiently process and classify this data is critical, and the computational simplicity of the Naive Bayes classifier makes it ideal for processing such large datasets without excessive computational costs. This efficiency ensures that the classifier can be scaled to accommodate larger and larger datasets, thereby supporting ongoing research efforts in biology and related fields. In conclusion, this study tries to know more about the performance of Naïve Bayes in animal classification, showing its significant advantages in different fields [2].

2. Method

2.1. Dataset Description and Preprocessing

When the dataset is quite large, containing 15,000 figures from Google images. These data consist of unduplicated samples of animals from five species, cat, dog, elephant, horse, lion. This diversity of the dataset facilitates thorough analyses and has the potential to draw reliable and generalizable conclusions. The pre-processing of data is an important step to ensure the dataset has a great quality

and its usability for research purposes. Initially, the raw data files are subjected to a series of quality control checks to identify and remove any anomalous images.

In summary, the combination of extensive sample size and preprocessing steps ensured that the dataset used in this study was highly reliable and suitable for training simple animal classification models. The comprehensiveness of the dataset and the rigorous preprocessing protocols enhance the validity of the subsequent analyses and the conclusions that may be drawn from this study.

In animal classification, the size of the figures influences the accuracy and the efficiency of the computation. The resizing of the animal figures be 64*64 for all the case to reduce the time consuming of computation, and split the data set into trains data as 70% and test data as 30%.

2.2. Feature Extraction

Extracting relevant features of a dataset and feature extraction techniques used to optimise classification performance has some key steps. Firstly, pre-procession to ensure the data is ready further analysis. This typically involves dealing with missing values, categorical variables, and standardizing numerical features. These steps make sure the data is formatted consistently, that each feature contributes equally to the analysis and that no feature dominates due to size differences. Feature extraction consists of identifying the most relevant features from the dataset that effectively contribute to the classification task. This step is important because the quality of features directly affects the performance of the classifier. Feature extraction usually has techniques such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA) [3]. PCA will transform the dataset into an unrelated dataset feature to reduce the dimensionality, thus capturing the maximum variance in the data [4]. LDA tries to figure out the linear combination of features that can isolate the classes in the dataset.

This time Histogram of Oriented Gradients (HOG) to extract image features, which contain a lot of different feature vectors, like slope, color histogram and others, the key idea of HOG is a graph can be describe distribution of local intensity gradients [5]. After preprocessing, gamma correction and grayscale, then calculate the gradients of each pixel, then HOG divides the figure into several connected modules called cells, each cell has a corresponding descriptor. Several cells are composed into a block, and the descriptor of the block is obtained by combining the descriptors of all cells.

2.3. Naive Bayes Model

The Naïve Bayes Model is a classifier developed from Bayes' Theorem, which reports the likelihood of an event is dependent on the prior conditions of the event [2]. The construction of the model involves several important steps, each step comes from the statistical theory and probability.

The model starts with Bayes' Theorem, which can be expressed mathematically as P(B|A) *P(A)/P(B), where the P(B|A) is the probability of A occurring when the event B is true, P(A) is the probability of event A happen, P(B) is the probability of event B happen.

The essential assumption of the Naïve Bayes Model is that features are independent of one another features, when the model gives the data a class label. This assumption usually violated in many real-world cases. This is always referred to as the 'Naïve' assumption. Naïve Bayes Classifiers tend to perform well in different complex tasks. This assumption allows the model to factorize the conditional probability into the product of individual probabilities.

The process of constructing the Naive Bayes model starts with computation of prior probabilities P(C) for each class C in the dataset. Those prior probabilities are estimated by calculating the ratio of training example to each class. The mathematic expression be like:

$P(C_i) = \frac{|X_i|}{|X|}$, $X_i$represent the example Xi, X represent the total example, || represent the number of examples. For continuous features, these probabilities are often modeled using a Gaussian distribution, whereas for categorical features, they simply involve counting occurrences and calculating frequencies.

After the prior probability, algorithm needs to build the mathematic expression for the posterior probability of specific feature, The core idea of classifying a new data point x with features (x1, x2, ..., xn) involves calculating the posterior probability for each class C given the features, which is

$$P(x|c_i) = P(x_1|c_i) * P(x_2|c_i) \dots P(x_p|c_i) = \prod_{j=1}^{p} P(x_{j=1}|c_i) \quad (1)$$

The Naive Bayes classifier then arranges x to the class with the highest posterior probability. Calculating P(C|x) involves taking the product of the prior probability of the class P(C) and probability of the features given the class, P(xi|C).

The Naive Bayes classifier then assigns x to the class with the highest posterior probability. Calculating P(C|x) needs to take the product of the prior probability of the class P(C) and the likelihood of the features given the class, P(xi|C). Cause the $P(x|c_i)$ need to do the successive multiplication, so if there are any 0 present in the $P(x_j|c_i)$, the final result of $P(x|c_i)$ also be 0. And also, if the probability of any feature is low, and need to do a successive multiplication, result shows a number close to 0, and after round to 3 digits the result shows 0, because the output is so small.

For the situation of $P(x_j|c_i) = 0$, can use the Laplacian smoothing to fix. Mathematic expression of Laplacian correction looks like:

$$P(C_i) = \frac{|X_i|+1}{|X|+N}, P(x_j|c_i) = \frac{|X_{ci,xj}|+1}{|c_i|+N} \quad (2)$$

In Laplace smoothing, the numerator added a tiny number (usually 1), and the total number of unique feature values is added to the denominator, which ensures that no probability is ever zero [6].

3. Experimental Results

3.1. Impact of feature extraction on classification accuracy

This section investigates the effect of feature extraction method on classification accuracy. Feature extraction is an important step for dataset preprocessing, which is aimed to reduce the number of variables and only keep the variables are most relevant to the task. Compare the classification accuracy between the Naïve Bayes classifier with HOG as feature extraction and the Naïve Bayes classifier with no feature extraction.

In this part, there are different parameters to quantize the accuracy of the Naïve Bayes. Precision is the ratio of correct prediction to the total predicted positives, Recall is the ratio of correct prediction to all observations in the actual class, F1-score is a weighted average of Precision and Recall, which is weighed by false positives and false negatives, Support is the number of actual times of happen of each class in the dataset.

The result in Table 1 is much higher than the result in Table 2, although I provide more dataset as the training dataset for the Naïve Bayes with no feature extraction. The use of HOG feature extraction dramatically improves classification performance, improve the accuracy from 38% to 76%. This demonstrates that feature engineering is crucial in image classification tasks, particularly for models like Naive Bayes, which assume statistical independence between features.

Table 1. Classification report for Naïve Bayes with HOG

Categorization	Precision	Recall	F1-Score	Support
Cat	0.92	0.92	0.92	635
Dog	0.79	0.82	0.8	590
Elephant	0.73	0.62	0.67	601
Lion	0.59	0.65	0.62	553
Accuracy			0.76	2379
Macro avg	0.76	0.75	0.75	2379
Weighted avg	0.76	0.76	0.76	2379

Table 2. Classification report for Naïve Bayes without feature extraction

Categorization	Precision	Recall	F1-Score	Support
Cat	0.4	0.37	0.38	833
Dog	0.54	0.3	0.38	785
Elephant	0.32	0.38	0.35	788
Lion	0.36	0.29	0.33	818
Accuracy			0.38	4043
Macro avg	0.4	0.38	0.37	4043
Weighted avg	0.39	0.38	0.37	4043

3.2. Combination with other classification algorithms

Naïve Bayes Classifier has some disadvantages. It performs bad when the dataset is biased, the result biased towards to the majority class, which makes the poor performance on classification of minority class. Cause the essential assumption of the Naïve Bayes is to ignore the relationship between each feature, which causes the oversimplified model, this leads to miss some important pattern in dataset.

Combining the Support Vector Machine (SVM) with Naïve Bayes can improve performance in some cases [7], this new hybrid model can be called Naïve Bayes SVM (NBSVM), we can have several potential benefits. First, NBSVM can lead to better performance, cause the Naïve Bayes can provide a strong baseline and SVM can provide strong decision boundaries. Secondly, Naïve Bayes is usually not the first choice for figure classification, because it assumes the features are irrelevant, but usually pixels in the figure are highly connected with the neighboring pixels, and SVM can handle high-dimensional data better than Naïve Bayes. Thirdly, SVM can be set with different kernels for different kinds of situations, in this case, we can use linear, polynomial, RBF or others to map the input dataset.

4. Result

4.1. Summary of key findings

Throughout the course of this research, various interesting findings have surfaced that not only demonstrate the power of the Naïve Bayes algorithm, but also show the extent to which different factors affect its performance. The Naïve Bayes algorithm is rooted in a statistic framework that operates on the principle that the features are irrelevant to each other. The fundamental assumption of feature independence simplifies the computation and makes the model particularly adept at handling large datasets. Despite its simplicity, the algorithm demonstrated a remarkable ability to achieve impressive classification accuracies. This success is attributed to the algorithm's ability to effectively exploit the probabilistic independence of features. For example, in animal classification, the shape and size of an animal is largely independent of its background and location, but both pieces of information are crucial for accurate classification. The Naive Bayes algorithm excels in this case, where the assumption of independent features is largely correct [8].

4.2. Implications for animal classification and related fields

An important implication of using Naive Bayes for animal classification is the method's ability to deal with incomplete or noisy data, which is common in biological research. Field studies often encounter data gaps due to difficulties in observing certain species or traits, usually the classification need to have the movement or behavior of animals [9], and Naive Bayes classifiers can effectively use any available data to predict the most likely animal classes, and increasing the reliability of classification in the presence of uncertainty. In addition, the use of Naive Bayes encourages the integration of machine learning techniques into traditional biological research and promotes interdisciplinary collaboration. Biologists, data scientists and computer scientists can work together to refine and optimize classification algorithms, driving innovation in the field. The collaboration between scientists from each field will help the improvement of accuracy of Naïve Bayes and also improve the way to distinguish the animals [10].

In conclusion, the implications of this research go beyond the immediate benefits of improved animal classification. By highlighting the potential of Naive Bayes as a viable approach to solving classification problems, this study emphasizes the transformative impact of integrating machine learning techniques into the biological sciences. The ability to handle large noisy datasets, scalability, and interdisciplinary collaboration are just a few of the strengths that Naive Bayes can offer, paving the way for furthering our understanding and conservation of biodiversity.

5. Conclusion

This study focuses on the Naïve Bayes classifiers' power on animal classification, especially the difference between Naïve Bayes Classifiers with HOG feature extraction and Naïve Bayes Classifiers without feature extraction, the HOG feature extraction helps the Naïve Bayes improve a lot in accuracy. Other than its simplicity and efficiency, the model faces limitations with biased datasets and feature independence assumptions.

References

[1] Kumar Y H S, Manohar N, Chethan H K. Animal classification system: a block-based approach. Procedia Computer Science, 2015, 45: 336-343.

[2] Chen S, Webb G I, Liu L, et al. A novel selective naïve Bayes algorithm. Knowledge-Based Systems, 2020, 192: 105361.

[3] Zhu F, Gao J, Yang J, et al. Neighborhood linear discriminant analysis. Pattern Recognition, 2022, 123: 108422.

[4] Greenacre M, Groenen P J F, Hastie T, et al. Principal component analysis. Nature Reviews Methods Primers, 2022, 2(1): 100.

[5] Mutlag W K, Ali S K, Aydam Z M, et al. Feature extraction methods: a review//Journal of Physics: Conference Series. IOP Publishing, 2020, 1591(1): 012028.

[6] Noto A P, Saputro D R S. Classification data mining with Laplacian Smoothing on Naïve Bayes method//AIP Conference Proceedings. AIP Publishing, 2022, 2566(1).

[7] Abdullah D M, Abdulazeez A M. Machine learning applications based on SVM classification a review. Qubahan Academic Journal, 2021, 1(2): 81-90.

[8] Wickramasinghe I, Kalutarage H. Naive Bayes: applications, variations and vulnerabilities: a review of literature with code snippets for implementation. Soft Computing, 2021, 25(3): 2277-2293.

[9] Khan M H, McDonagh J, Khan S, et al. Animalweb: A large-scale hierarchical dataset of annotated animal faces//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 6939-6948.

[10] Oliveira D A B, Pereira L G R, Bresolin T, et al. A review of deep learning algorithms for computer vision systems in livestock. Livestock Science, 2021, 253: 104700.