A Review of Machine Learning Algorithms for Text Classification

Ruiguang Li[1,2(✉)], Ming Liu[2], Dawei Xu[1,3], Jiaqi Gao[3], Fudong Wu[3], and Liehuang Zhu[1]

[1] School of Cyberspace Science and Technology, Beijing Institute of Technology, Beijing, China
lrg@cert.org.cn
[2] National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing, China
[3] Changchun University, Changchun, China

Abstract. Text classification is a basic task in the field of natural language processing, and it is a basic technology for information retrieval, questioning and answering system, emotion analysis and other advanced tasks. It is one of the earliest application of machine learning algorithm, and has achieved good results. In this paper, we made a review of the traditional and state-of-the-art machine learning algorithms for text classification, such as Naive Bayes, Supporting Vector Machine, Decision Tree, K Nearest Neighbor, Random Forest and neural networks. Then, we discussed the advantages and disadvantages of all kinds of machine learning algorithms in depth. Finally, we made a summary that neural networks and deep learning will become the main research topic in the future.

Keywords: Natural language processing · Text classification · Machine learning · Neural network

Text classification is a basic task in the field of natural language processing. It is a process of extracting the key elements of sentences, judging the authors' views, intentions and emotions of the sentences. It is the basis of information retrieval, questioning and answering system, emotion analysis and other advanced tasks. It can be said that text classification technology is the basic technology of natural language processing and social network analyzing, which will be beneficial to information discovery, public opinion analysis in the web space.

Text classification is the earliest application of machine learning algorithm, and has achieved good results. In addition to the traditional methods based on statistical analysis such as Naive Bayes, Supporting Vector Machine and so on, neural networks and deep learning algorithms have also been widely used in the field of text analysis in recent years. Therefore, a comprehensive study and review of machine learning algorithms in the field of text classification is very valuable for academic research and engineering development.

W. Lu et al. (Eds.): CNCERT 2021, CCIS 1506, pp. 226–234, 2022.
https://doi.org/10.1007/978-981-16-9229-1_14

1 Introduction

Machine learning mainly studies how to learn unknown rules from given data, that is, to find some objective rules from observed data (samples), and use the learned rules (models) to analyze and predict irregular data or unknown events. The following figure illustrates the application principle of machine learning algorithm in text classification (Fig. 1).

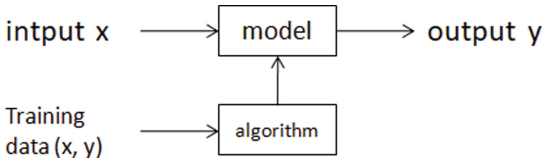

Fig. 1. Principle of text classification

As is shown in the above figure, the input is "x", and the output is "y". Through the model's processing, we got expected results from the irregular data. The "model" is previously trained by the "data (x, y)". That is the basic procedure of processing text content by machine learning algorithms.

At present, the main machine learning method is to conduct statistical analysis on the existing marked data, by which we can find laws and obtain models. Using these trained models, we can make prediction and analysis on unknown data to obtain classification results. Text classification algorithm based on machine learning is usually divided into four steps:

1) Features extracting and document modeling. Feature extracting is a very important task in machine learning. The traditional feature extraction method is usually manual extraction, the manual feature engineering is time-consuming, and the feature dimension is too large and the efficiency is low. In order to solve these problems, features extracted by traditional methods are usually dimensionality reduced, and a subset of features that can best represent text information and achieve the best classification effect are selected among all features.
2) Training the model. The machine learning model is trained by training data sets, and the labeling quality of training data sets has a great influence on the actual classification effect. Compared with the traditional machine learning algorithm based on statistics, deep learning has excellent feature learning ability, and the features acquired by learning have more essential characterization ability on data, which is conducive to visualization and classification.
3) Classifying the test text. Input the text data to be classified to the model, the classifier will judge the text data according to the learned rules, and give the classification result.
4) Evaluating the classification effect. After the completion of text classification, it is necessary to verify the quality of the classifier with some quantitative indicators, such as accuracy, recall rate and F measurement. The accuracy rate refers to the proportion

of samples with positive classification results that are truly positive; the recall rate refers to how many samples of all positive classes can be correctly classified by the classifier; the measurement value of F is the geometric average of the accuracy rate and recall rate.

2 Principles of Machine Learning Algorithms

We made a review of the traditional and state-of-the-art machine learning algorithms for text classification including Naive Bayes, Supporting Vector Machine, Decision Tree, K Nearest Neighbor, Random Forest and neural networks, and so on. We introduce the principles of various machine learning algorithms, and discuss the advantages and disadvantages in depth.

2.1 Naive Bayes

Naive Bayes Classifier is a common statistical classification algorithm, which classifies input samples based on Bayes' theorem and the independence of the eigenvector elements.

Its working principle is as follows [1]: First, text vector is generated from the text to be classified, that is, text vector of document x is constructed based on a given dictionary, where n is the number of elements in dictionary set W, and represents the number of occurrences in document d.

$$W = \{w_1, w_2, \cdots, w_n\}$$

$$X = \{x_1, x_2, \cdots, x_n\}$$

Given the training data, the classification vector indicates that the text can be divided into m classes.

$$D = \{d_1, d_2, \cdots, d_m\}$$

$$Y = \{y_1, y_2, \cdots, y_m\}$$

D is used to train the naive bayesian algorithm. On this basis, the classification with the maximum probability is to solve argmaxP:

$$\text{argmax}P(y_j|X)$$

According to bayes' theorem,

$$P\left(y_j|X\right) = \frac{P(y_j)P\left(X|y_j\right)}{P(X)} = \frac{P(y_j)\prod_{i=1}^{n} P\left(x_i|y_j\right)}{\prod_{i=1}^{n} P(x_i)}$$

From the given training data $D = \{d_1, d_2, \cdots, d_m\}$, we can calculate $P(y_j)$, and each $P(x_i)$. So, for the new input data X, we can calculate all $P\left(y_j|X\right)$, in which the highest probability value is the classification result of document X.

The naive bayes algorithm assumes that the attributes of the data set are independent from each other, so the logic of the algorithm is very simple. And the algorithm is relatively stable. When the data show different characteristics, the classification performance of the naive bayes will not have much difference. In other words, the simplicity bayes algorithm is relatively robust and will not show too much difference for different types of data sets.

2.2 Supporting Vector Machine (SVM)

Supporting Vector Machine is a new statistical classification algorithm proposed by Vapnik and his team at Bell Labs in 1995, which is mainly used to solve binary classification problems. The basic idea is to find a hyper-plane in the sample feature space, which can divide all the sample data well and make the distance between the sample point and the hyper-plane maximum.

Its working principle is as follows [2]: Text vector is also generated from the text to be classified, that is, the text vector of document x is constructed based on the given dictionary, where n is the number of elements in the dictionary set W, and x_i represents the number of x_i occurrences in document d.

$$W = \{w_1, w_2, \cdots, w_n\}$$

$$X = \{x_1, x_2, \cdots, x_n\}$$

The classification vector $Y = \{y_1, y_2, \cdots, y_m\}$, where $y_i \in \{-1, 1\}$ respectively represents the negative class and the positive class in the dichotomy.

$$Y = \{y_1, y_2, \cdots, y_m\}\, y_i \in \{-1, 1\}$$

Assuming that there is a decision boundary in the characteristic space of the input sample data, all sample points can be separated by a hyper-plane according to positive class and negative class, and the distance between any sample point and the plane is greater than 1, then the classification problem is said to be linearly separable.

The Decision boundary:

$$w^T X + b = 0$$

Point to plane short:

$$y_i\left(w^T X_i + b\right) \geq 1$$

In the following figure, the solid line in the figure is the hyper-plane $w^T X + b = 0$, and w and b are the normal vector and intercept of the hyper-plane respectively. The solid red dot constructs the upper interval boundary $\left(w^T X_i + b\right) \geq 1$, the hollow red dot constructs the lower interval boundary $\left(w^T X_i + b\right) \leq -1$, the distance between the two interval boundaries $d = \frac{2}{\|w\|}$. The positive class and the negative class samples on the interval boundary are the support vectors (Fig. 2).

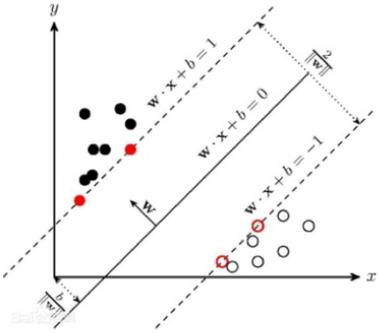

Fig. 2. Principle of supporting vector machine

The Supporting Vector Machine classification algorithm has good generalization ability and learning ability [3]. It aims at minimizing structural risk, and the solution obtained is the global optimal solution. This algorithm overcomes the problem of "dimension disaster". It is widely used in text automatic classification, face recognition, gene expression, handwriting recognition and other fields.

2.3 Decision Tree

Decision tree is a method to classify texts by constructing a tree-like decision system on the basis of known probabilities. Because the branch of the decision system is drawn like the branches of a tree, it is called Decision Tree. The processing of decision tree algorithm can be divided into three steps: Feature selection, Decision tree generation and Pruning:

1) Feature selection: Feature selection refers to the selection of some features from text vectors as the decision basis. These features will affect the branch shape of the decision tree and have an important impact on the classification results.
2) Decision tree generation: This step is the main decision process. The root node is a specific word sequence, that is, there is only one word, which has the best classification error rate among all words and the highest probability for a certain category. The subsequent child nodes are divided into left and right sub-trees according to the above decision. If the coefficient is not zero or the word sequence has no child sequence, the decision will stop; if it is not zero and not unique, the decision will continue in the possible category.
3) Pruning: The decision tree will form a very large tree-like system according to all the training samples, which has a high accuracy in the training samples and a poor accuracy in the test samples, forming a phenomenon of over fitting. The solution of the over-fitting phenomenon requires manual observation and debugging, observation and control of the size of the decision tree at each layer, setting the number of samples of the smallest leaf node, adjusting the minimum weight of the leaf node, etc.

2.4 KNN (K-Nearest Neighbor)

KNN is one of the simplest classification algorithms, but it is also one of the most commonly used classification algorithms [4]. KNN algorithm is a supervised classification algorithm, which is similar to another machine learning algorithm "K-means". KNN (K-nearest neighbor) algorithm is to find K texts in the training set that are most similar to the target text according to the known data category and the text to be tested, and then score the candidate category according to K samples.

The advantage of KNN algorithm is that it is simple and easy to use. Compared with other algorithms, KNN is a relatively simple and clear algorithm. Even without a high mathematical foundation, we can find out its principle. The model has fast training time and good prediction effect. The disadvantage of KNN algorithm is that it requires high memory, because it stores all training data, the prediction stage may be very slow, and it is sensitive to irrelevant functions and data scale.

2.5 Random Forest

Random forest is a classifier that contains multiple decision trees, and the classification results of its output are determined by the number of votes of all the output results of the tree. By means of integration, the random forest integrates multiple trees together. Its basic unit is the decision tree, and each decision tree is a classifier. For an input sample, N trees will have N classification results. The random forest integrates all the classification voting results and specifies the category with the most votes as the final output. This is the simplest Bagging idea. The principle of random forest is:

1) N is used to represent the number of training cases (samples), and M is used to represent the number of features.
2) The number of input features m is used to determine the decision result of a node in the decision tree; Where m should be much less than m.
3) A training set (i.e., bootstrap sampling) was formed by sampling N times from N training cases (samples), and the prediction was made with the unselected use cases (samples) to evaluate the error.
4) For each node, m features are randomly selected, and the decision of each node in the decision tree is determined based on these features. According to these m characteristics, the optimal splitting mode is calculated.
5) Each tree will grow intact without pruning, which may be adopted after the construction of a normal tree classifier.

2.6 Neural Network

Neural network is an adaptive nonlinear dynamic network system composed of a large number of neurons connected with each other. The classifier establishes a neural network for each category document, inputs feature word vector, and finally outputs text category after repeated learning.

At present, CNN (Convolutional Neural Network) and RNN (Recursive Neural Network) are two important components of deep learning [5]. CNN based on convolution

is good at identifying the structure of the target task, while RNN has advantages in sequence recognition modeling due to its memory function. For natural language processing tasks, CNN and RNN have their own advantages. In 2014, Kim proposed the Text CNN algorithm for text classification based on CNN, applied the convolutional neural network to the text classification task, used the convolution kernel with different channel number and different size to extract the key information in the sentence (similar to n-gram with multi-window size), and could better capture the local correlation. However, Text CNN can only extract sentence features in a limited window, and cannot consider the long distance dependence on information and word order information, so it will lose some semantic information. RNN is characterized by automatic learning and memory of text sequence features, and has achieved great success in text processing in natural language processing.

Both short-term and long-term memory network (LSTM) is a kind of special used for time series data of network [6], the traditional RNN neural networks of neurons is after applying input function to calculate the output of the unit, and LSTM will neurons into the memory unit, each memory unit door to door, forgotten by the input and output, the LSTM this one design solved the RNN gradient disappeared and gradient explosion problems of the network. Bidirectional LSTM (BiLSTM) USES two LSTM networks to process the sequence from front to back, which can extract the context information of the sequence more comprehensively, thus making the classification effect better.

3 Comparative Study of the Machine Learning Algorithms

The Naive Bayesian algorithm assumes that attributes are independent from each other, which is often not true in practical applications, which has some influence on the correct classification of NBC model. When the relation between attributes of data set is relatively independent, the Naive Bayesian classification algorithm will have better effect. However, when the independence of attributes of data sets is difficult to satisfy in many cases, because the attributes of data sets are often correlated with each other, the classification effect will be greatly reduced.

Support vector machine (SVM) in tackling small sample, nonlinear and high dimension data is a great advantage, but it also has disadvantages: limited to small cluster sample, can only handle the second class classification problem, for many classification problem to solve the effect is not good, and in dealing with a specific classification problem to select the correct effective kernel function, in turn, affects the efficiency of classifier.

The Decision Tree classifier has the following disadvantages: it is difficult to predict the continuous fields; For the data with time sequence, it needs a lot of preprocessing. Errors may increase more quickly when there are too many categories; General algorithm classification, just according to a field to classify, accuracy is not enough.

The main shortcomings of KNN algorithm are as follows: first, when the samples are unbalanced, for example, the samples of one class are large, while the samples of other classes are small. As the samples of K neighbors of the new sample are large in size, it may lead to classification errors. Second, the algorithm is computationally intensive, because for each text to be classified, the distance from it to all known samples must be calculated to obtain its K nearest neighbor points.

The advantage of random forest algorithm is high accuracy. Because the integrated algorithm is adopted, its accuracy is better than most single algorithms, so it has high accuracy, especially in the test set. Due to the introduction of two random elements, the random forest is not easy to fall into over fitting. The disadvantage of random forest algorithm is that when there are many decision trees in random forest, the space and time required for training will be large. In addition, there are many unexplainable places in the random forest, which is a bit of a black box model. On some noisy sample sets, the random forest model is easy to fall into over fitting.

The neural network classifier has a high prediction accuracy and certain parallel processing and learning capabilities. It also has strong robustness and fault tolerance against noise. But the neural network requires a large number of parameters, long learning time, large runtime resource consumption, and low interpretability of output results.

4 Conclusion

This paper discusses various machine learning algorithms, including Naive Bayes, Support Vector Machines, KNN algorithm, Decision Tree algorithm, Random Forest, Neural Network model and Deep Learning model. In general, these methods have been widely studied and applied, and have achieved good practical results. In practice, the suitable algorithm can be selected flexibly according to the advantages and disadvantages of various algorithms and the actual situation of the text data to be classified.

In general, the machine learning algorithm based on neural network is the main research direction in the future due to the traditional machine learning algorithm based on statistical theory in the aspects of feature automatic learning, classification accuracy, operation robustness and fault tolerance.

Acknowledgement. This work was supported by the National Natural Science Foundation of China under Grant U2003206 and 62106060.

References

1. Liu, Y.: The application of naive bayes in text classification preprocessing. Comput. Inf. Technol. (2010)
2. Zhang, X.: The summary of text classification based on support vector machines. Sci. Technol. Inf. (2008)
3. Zhao, D.: Research on the vector space model based text automatic classification system. Int. J. Digit. Content Technol. Appl. **7**(3), 381–388 (2013)
4. Zhou, Y., Li, Y., Xia, S.: An improved KNN text classification algorithm based on clustering. J. Comput. **4**(3) (2009)
5. Zhou, Y., Chen, S., Wang, Y., et al.: Review of research on lightweight convolutional neural networks. In: 2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC). IEEE (2020)
6. Niu, S., Chai, X., Deqi, L.I., et al.: A text classification algorithm based on neural network and LDA. Comput. Eng. (2019)