

Research on Machine Learning Feature Algorithm

Jing Liu¹, Ai Yang¹, Wenbo Jiang¹

¹ Haikou College of Economics, Haikou, Hainan, 571127

Keywords: machine learning; feature algorithm; computer technology

Abstract: The main idea of feature selection is to select feature subsets by removing features that contain little or no relevant information. Feature selection methods can be divided into three categories, filter, encapsulate and embed. Given the large number of feature selection algorithms currently available, in order to be able to properly decide which algorithm to use in a particular situation, it is necessary to propose standards that can be relied upon or determined. This paper is to review some basic feature selection algorithms, compare and classify feature selection methods and algorithms based on the existing theories and experimental results in the literature, and then propose a standard that can be dependent or determined.

1. Introduction

The essence of feature selection is the process of selecting the optimal subset from the original data set, and the quality of the optimal subset is measured by specific evaluation criteria. The theory of feature selection has experienced decades of development, and its research results have been widely applied to text classification, image extraction, and genome analysis. Combining with machine learning algorithms forms a complex algorithmic system. Due to the diversification and interdisciplinarity of algorithms, many research experts engaged in this field spend a lot of time to understand and study each algorithm. Based on this consideration, the Lists and summarizes existing feature selection algorithms, evaluates each algorithm objectively based on existing theories and experimental results, classifies them according to evaluation criteria, and finally proposes a method to guide personnel working in this field. There are technologies that can select appropriate algorithms to solve the practical problems of dependability or decision criteria.

2. Feature Selection

Features [1], variables and attributes are all part of the data. It is usually discrete data, continuous data, or boolean data. Text classification usually includes: (1) Related features: These features are very important for classification results. Large influences and cannot be replaced by other features; (2) Irrelevant features: These features have no effect on the classification result. For each instance of the class, its value is generated randomly; (3) Redundant features: No effect on the classification result or association with other features. The key to feature selection (or attribute selection) is how to remove redundant or unrelated features while selecting some useful inputs. The optimal feature subset should be that the feature space has the lowest dimension after removing redundant information, but it contains important key metric information. However, in actual problems, it contains a lot of noise data, unrelated and misleading features. In order to fully determine each feature, ideally all enumeration features should be tested, but in most cases it is not feasible because if there are n features there will be $2^n - 1$ feature subsets.

(1) Reduced feature space dimension, reduced demand space and accelerated algorithm speed; (2) Removed redundant, unrelated or “noisy” data; (3) Performed data analysis and shortened learning algorithm Running time; (4) The selected features are easier to understand; (5) increases the accuracy of the classification model; (6) the reduction of feature sets saves resources for the next round of data collection and utilization; (7) classification performance Improve the accuracy of predictions; (8) Data analysis helps to reveal the underlying information contained in the data

3. Feature Selection Algorithm

The purpose of the feature selection algorithm is to make the classification or regression model constructed by the selected optimal feature subset achieve similar or even better prediction accuracy before feature selection, and improve the generalization ability, comprehensibility, and computational efficiency of the model. Several factors of the feature selection algorithm need to be considered in the literature. Given these factors, the feature selection can be described as a search optimization problem in the hypothesis space. (1) Search strategy. The search strategy is related to the number of features in the feature space. A search algorithm uses a specific strategy to search for features. There are mainly three search types: exhaustive search, heuristic search, and random search. (2) The construction of feature subsets. The following five methods are commonly used to construct feature subsets: Forward Selection, Backward Selection, Bidirectional Selection, Weighting Method, and Random Selection. (3) Evaluation function. The candidate feature subset functions are evaluated. The evaluation functions mainly include: error rate, dispersion, dependence, distance measurement, accuracy, information measurement, and consistency. The basic framework of feature selection is given in [2].

4. Feature Selection Method

The feature selection algorithm can be divided into Filter [3] method, Wrapper [4] method and Embedded method according to different evaluation criteria. These feature selection methods [5] are based on mutually independent criteria for classification, and there are several methods that use simple correlation coefficients similar to Fisher's criteria. Others use mutual information or significance test (hypothesis test) (t-test, F-test, χ^2 test). The early filter-based evaluation method did not consider the correlation between features. In recent years, the MRMR (Medium Redundancy-Maximum Relevance) feature selection scheme has been proposed. This method uses the maximum correlation and minimum redundancy criteria to add features. The set of feature items optimizes feature subsets and improves their generalization capabilities.

The Wrapper method uses the classifier as a black box to store feature subsets based on the predictive power of feature items. Wrapper method based on support vector machine has been widely used in machine learning, and SVM-RFE (Support Vector Machine Recursive Feature Elimination) adopts a disadvantageous feature elimination method to recursively eliminate unnecessary feature items in feature subsets. This method has been applied to this method. Cancer research. In each recursion, the feature items are sorted according to the reduction of the objective function in the feature, and then the bottom feature items are eliminated. There are also some schemes that use backward elimination and variant methods of linear kernel functions.

In embedded feature selection, the feature selection algorithm is embedded as part of the learning algorithm. It is not necessary to divide the training text into a training set and a verification set. That is, there is no need to verify the intermediate results, and the feature selection and the training process are performed simultaneously. Directly use the classification model to decide whether to select features or reject features. The classical embedded algorithms are decision trees and artificial neural networks.

5. Basic Description of Feature Selection Algorithm

The CHI statistical method is a statistical test method for measuring the degree of correlation between terms and document categories. The most basic idea is to determine the correctness of the theory by observing the deviation between the actual value and the theoretical value. In the statistics, the χ^2 test is used to test the degree of deviation of two mutually independent events A, B. Event A, B is defined as if $P(AB) = P(A)P(B)$, in feature selection, Events A and B represent the frequency of occurrence of terms and classes, respectively. If χ^2 is small enough, it is considered that the error is caused by the measurement means being not precise enough or happening accidentally. The two are indeed independent and accept the original hypothesis; if χ^2 is large enough, such an error is

unlikely to be accidentally generated. Or because of inaccurate measurements, one can assume that the two are actually related, that is, reject the null hypothesis and accept the alternative hypothesis.

The information gain measures the information gain (entropy) of a term for a certain class by examining the frequency of the document in which the term appears and does not appear in the document category. Given a feature t , the amount of information each time it is considered and not considered, and the difference between the two is the amount of information that this feature brings to the system, ie gain. The more information a feature brings, the more important this feature is. Intuitively, this makes sense, because strings with maximum entropy cannot be compressed, and a highly ordered string can use a short string to describe the information it carries. In text categorization, if an instance is randomly assigned to multiple classes, the entropy of the information carried by this instance is high, but if all the instances are distributed in a single class, the information entropy will be low.

Genetic Algorithm is an efficient and widely used global optimization algorithm. It simulates natural evolution by establishing a dynamic population solution model. Members of the population call chromosomes and encode them as features to encode selected features. The characteristics of each chromosome establish a model. Through training data sets, model errors are quantified and used as fitness functions. This algorithm allows the most adaptable chromosomes to survive and multiply, thus greatly reducing the error functions of subsequent generations. The success of a genetic algorithm depends on several factors: one is the crossover and mutation operator; the second is the number of chromosomes that survive; and the third is the number of initial populations, and Shannon's entropy algorithm combined with graphical analysis can solve this problem.

The distributed estimation algorithm is a brand-new evolutionary model, without the crossover and mutation operations of traditional genetic algorithms. Instead, it is the learning and sampling of probability models. The distributed estimation algorithm solves a wide range of optimization problems by learning the new candidate solutions that have been proposed so far to show the probabilistic model scheme and the model established by sampling. It integrates advanced machine learning algorithms with genetic and evolutionary algorithms that can solve many challenging problems and is clearly superior to traditional genetic algorithms and other machine learning techniques. In the past decade, remarkable results have been produced in the design, theoretical research and application of EDA.

Decision tree learning is a tree structure. Each non-leaf node on the tree represents a test feature. Each branch represents the result of the test. Each leaf represents a class label. The popularity of decision tree classifiers is because building a decision tree does not require knowledge in any field, and the knowledge expressed in a tree format is easy to understand; in addition, the classification step of decision tree induction is simple and rapid. The most famous and most representative algorithm is ID3, which is described in detail as follows: When selecting attributes at each level node of the decision tree, information gain (IG) is used as the selection criterion of the attribute, so that when testing each non-leaf node, Get the largest category information about the record being tested. The specific method is: Detect all the features, select the feature with the greatest information gain to generate a decision tree node, establish a branch from the different values of the feature, and then recursively call the method on the subset of branches to establish a branch of the decision tree node until All subsets contain only data of the same category. Although each feature is selected in a greedy algorithm, the characteristics of the test are irrelevant to the original features.

Support vector machine is based on statistical learning theory VC theory and structural risk minimization principle. The basic idea is to establish an optimal hyperplane as the decision surface, so that the classification interval between positive and negative samples is maximized. Selecting the appropriate kernel function is the key. Using different kernel functions will result in different SVM algorithms. The main kernel functions are: polynomial kernel, Gaussian radial basis kernel, Sigmoid kernel and so on. The advantage of SVM is that it is sparse, two-dimensional space learning ability, with a small number of training samples to minimize the scope of the empirical error and the complexity of the classifier, to avoid the local minimum point, and can effectively solve the learning problem, Has good promotion and better classification accuracy. Contrary to the

neural network, the objective function is a unimodal function, which can reach a global optimal solution to a large extent. The truly valuable application of the SVM approach is to solve the problems of small sample, nonlinear and high-dimensional pattern recognition.

6. Conclusion

The feature selection algorithm mainly provides a solution to the problem of feature selection based on related definitions. The FSA proposed in the literature is based on different criteria (such as the evaluation method used to explore the search space, etc.). The main work in this paper is to propose a solution to evaluate FSA based on the overall performance of FSA on related, irrelevant, redundant, and artificial sample data sets. To achieve this goal, some experiments using artificially generated datasets were designed and implemented, then compared with the subsets generated by the FSA using an optimal subset scheme, and finally, based on an evaluation criterion. Comparing the similarity between the results obtained by the FSA and the actual results, the final result of the experiment can be used as a useful information to obtain which algorithm to use in a particular situation. And we can see the effect of different algorithms on experimental results and the risk of relying on a single algorithm. Therefore, researching new hybrid methods or combining more reliable feature-related assessment methods is the next research direction.

References

- [1] Karegowda A G, Manjunath A S, Jayaram M A. Comparativestudy of attribute selection using gain ratio and correlation based feature selection [J]. International journal of information technology and knowledge management, 2010, 2 (2): 271-274.
- [2] Chena Y, Abrahama A, Yanga B. Feature selection and classification using flexible neural tree[J]. Neurocomputing,2006,70(1): 306 -308.
- [3] Koller D, Sahami M. Toward optimal feature selection[C] / /Proceedings of international conference on machine learning. [s. l.]: [s. n.],1996: 162-187.
- [4] Forman G. An extensive empirical study of feature selectionmetrics for text classification[J]. Machine research learning,2003 (3): 1289-1305.
- [5] Sheikhi N, Rahmani A, Mohsenzadeh M. An unsupervised feature selection method based on genetic algorithm [J]. International journal of computer science and information security, 2011, 24 (3): 117-120.