ISSN: 2688-9323

Analysis of Particle Swarm Optimization based on Discrete Feature Selection

Xiaoxiao Liu

Chongqing University of Posts and Telecommunications, Chongqing 400000, China

Abstract

Feature extraction and feature selection technology has become an important approach to solve high-dimensional data, and has been widely used in information retrieval, text classification and other aspects. In this paper, the principle and method of feature selection are introduced, and the improved particle swarm optimization algorithm is proposed to solve the problem of feature selection based on discretization.

Keywords

feature selection; feature discretization; particle swarm optimization algorithm.

1. Introduction

The development of computer and network technology makes it more convenient for people to obtain a large number of data. With the explosive growth of data, how to use these huge data to mine out useful information has become a key problem. Faced with the problem of big data processing, we can use computer algorithms to help us analyze a large amount of data, thus the field of machine learning arises at the historic moment. Machine learning refers to the process in which a computer automatically improves or learns algorithm performance based on a large amount of data or previous experience. With these algorithms we can predict the data, analyze it, and help us make decisions. However, the complexity and diversity of problems in real work have many difficulties and challenges in machine learning and pattern recognition, also the complexity of data itself is a difficulty.

2. Feature Selection

2.1. Introduction to Feature Selection

Generally, in order to construct the real distribution model of data in a more comprehensive and accurate way, we hope to obtain as many features or attributes as possible. However, with the increase of data features and attributes, the complexity increases significantly. For example, in image processing, high latitude data will cause high complexity of processing time and space, increasing the difficulty of processing. In addition, there may be excessive redundant features or irrelevant features in a large amount of data, making these features become noises, which will have a certain impact on the subsequent analysis process and affect the credibility and authenticity of subsequent results. Therefore, how to find the most useful information from high-dimensional data and extract the most effective features from a large number of features for classification, clustering, recognition, regression and other tasks is a crucial issue. Therefore, dimensionality reduction data will be proposed to reduce the number of features in the data, and then be analyzed in fewer feature data sets.

Dimensionality reduction technology can reduce the complexity of data by extracting new features or discarding useless features. Dimensionality reduction technology can reduce the number of features or attributes for subsequent analysis and processing. At present, the common dimensionality reduction techniques include feature extraction and feature selection. Feature extraction is widely used in face recognition, image processing and other fields. And

ISSN: 2688-9323

feature selection is also called feature subset selection, or attribute selection. It refers to the process of selecting the most effective features from the original feature species to reduce the dimension of the data set. It is an important means to improve the performance of the learning algorithm and a key data preprocessing step in the recognition pattern.

Feature extraction algorithm is more used in image processing, biological information and other fields, is a relatively easy way to implement the dimensionality reduction algorithm. Feature selection is a process of selecting the best feature subset according to some search strategy and learning criteria. Feature selection is an inclusive relationship without changing the feature space of the original space.

2.2. Main Methods of Feature Selection

There are three main methods for feature selection. The following three methods are introduced in detail.

Filter Feature Select. The main idea of Filter method is to "score" the features of each dimension, that is, assign weights to the features of each bit, each weight represents the importance of the features of the dimension, and then rank them according to the weight. There are mainly Chi-square test (Chi-squared text), information gain (ID3) and correlation coefficient Sores.

Characteristics of encapsulation methods (Wrapper Feature Select) Wrapper method is the main idea is to put the subset selection as an optimization problem, generate different combinations, evaluate the combination again, comparing with other combination again, in this method the subset selection problem as an optimization problem, so you can use a lot of heuristic optimization algorithm to solve, such as genetic algorithm (GA) and particle swarm optimization algorithm (PSO) and differential evolution algorithm (DE), etc., Recursive feature elimination algorithm is mainly used.

The main idea of Embedded Feature Select is to learn the attributes that are best for improving the accuracy of the model in a given situation of the model, that is, to Select the attributes that are important for model training in the process of determining the model. There are regularization and so on.

2.3. Classification Method based on Search Strategy

Basic search strategies can be divided into global optimal search, random search and heuristic search according to the formation process of feature subsets. A specific search algorithm will adopt two or more basic search strategies, for example, genetic algorithm is a random search algorithm, but also a heuristic search algorithm.

2.3.1. Adopt the Feature Selection Method of Global Optimal Search Strategy

At present, the only search method to obtain the optimal result is the branch and bound method, which can ensure that the optimal subset relative to the designed criteria of separability can be found when the number of features in the optimal feature subset is determined in advance. Its search space is $O(2^N)$ (where N is the characteristic dimension). However, it is difficult to determine the number of optimal feature subsets; It is difficult to design the divisibility criterion which satisfies monotonicity, the time complexity of the algorithm is high when dealing with high-dimensional multi-class problems.

2.3.2. Adopt the Feature Selection Method of Random Search Strategy

In the process of computing the feature selection problem with simulated annealing algorithm, genetic algorithm, etc., or just a random resampling process, with probability reasoning and sampling process as the basis of algorithm, the validity of the estimated based on the classification, the algorithm in the operation of the characteristics of each given a certain weight; The importance of features evaluated according to user-defined or adaptive thresholds. When the weight of a feature exceeds this threshold, it is selected as an important feature to train the

ISSN: 2688-9323

classifier. The Relief algorithm series is a typical random search method that selects features according to weight. It can effectively remove irrelevant features, but cannot remove redundancy, and can only be used for two categories of classification. The random method can be subdivided into completely random method and probabilistic random method. Although the search space is still $O(2^N)$, the search space can be limited to less than $O(2^N)$ by setting the maximum number of iterations. For example, genetic algorithm USES heuristic search strategy, its search space is far less than $O(2^N)$, but it has high uncertainty, and only when the total number of cycles is large, can find better results. In the random search strategy, some parameters may need to be set, and the choice of parameters plays a great role in the final result.

2.3.3. Adopt the Feature Selection Method of Heuristic Search Strategy

This kind of feature selection method mainly has the single optimal feature combination, sequence forward selection method (SFS), generalized sequence forward selection method (GSFS), sequence backward selection method (SBS), generalized sequence backward selection method (GSBS). Floating search method. The method is easy to implement and fast, its search space is O (N2), USES the floating generalized backward selection method (FGSBS) is more conducive to the practical application of a kind of feature selection search strategies, it is considering the statistical correlation between features, and with floating method to ensure that the algorithm runs fast stability, but although heuristic search strategy with high efficiency, it is at the expense of the global optimal.

Each search strategy has its own advantages and disadvantages. In the practical application, we can find an optimal balance point according to the specific environment and criterion function. For example, if the number of features is small, the global optimal search strategy can be adopted. If the global optimum is not required, but the computing speed is fast, then the heuristic strategy can be adopted. If you need a subset of high performance and don't mind computation time, you can use a random search strategy.

3. Discrete Feature Selection - Particle Swarm Optimization Algorithm

3.1. Feature Discretization

Many discretization methods with different strategies are proposed in order to determine the segmentation points of the eigenvalues into discrete values. Within the range of eigenvalues, the segmentation points are true values that are used to segment the range to several intervals. Existing discretization methods can be classified using different criteria. In the direct method, the interval is generated based on predefined parameters. Incremental methods, on the other hand, recursively separate (or merge) intervals until some criteria are met, resulting in a split (or merge) method. They are also known as top-down or bottom-up methods. The discretization method is supervised or unsupervised according to whether class labels are used in the discretization process. If the entire instance space is used in each discrete step, or if each discrete step USES only a subset of instances, then it will be global. A method is also univariate or multivariate, depending on whether the feature is discrete or discretization of multiple features, taking into account the interaction between the features. By discretization, some small fluctuations and possible noises in the data can be ignored to improve the effectiveness and efficiency of the algorithm. There are many discretization methods, but the most commonly used discretization method is the univariate method. When there is no feature interaction, such methods are very effective for a certain feature at a certain time. However, univariate discretization may destroy the information of feature interaction, so in practical application, it is a feasible method to combine univariate discretization and feature selection as a separate process.

ISSN: 2688-9323

3.2. Particle Swarm Optimization Algorithm

Particle swarm optimization algorithm is a kind of heuristic algorithm, the stochastic optimization algorithm based on population, belongs to a kind of evolutionary algorithms, particle swarm optimization algorithm for feature selection is the main idea will choose as a subset of search optimization problem (Wrapper method), the new method of a kind of improved particle swarm optimization algorithm (evolve particle swarm optimization (EPSO)), the use of a known as the "bare bare-bones" PSO (BBPSO) PSO derivation method to discrete and feature selection at the same time, In PSO, PSO is usually an n-dimensional vector corresponding to N features, and the range of each value is [0,1]. If it is greater than a predetermined threshold, the feature is chosen to believe, and vice versa, regardless of how large or small it is compared to the threshold. Therefore, two different evolutionary vectors may produce the same subset of features. On the other hand, in discretization, a slightly different point of evolution may lead to a different discrete characteristic. Therefore, finding a good pointcut requires a fine-tuned search mechanism, which can be found in BBPSO. In this derived VERSION of PSO, a Gaussian random generator is used to sample the new location, centered on the standard deviation of the individual best location (PBest) and its neighbor's best location (GBest) and the distance between them.

EPSO uses BBPSO to achieve discretization and feature selection. Every characteristic has a turning point. Since a pointcut can be any value in the feature range, the number of possible solutions for discretization is much larger than the feature selection. Therefore, the entropy-based cut points derived from it are used as initial or potential cut points to narrow the search space. The method has achieved good results. However, because of this representation, the search space is still too large for BBPSO to achieve better performance. In order to narrow the search space, a potential particle swarm optimization algorithm using BBPSO is proposed. A new fitness function and scaling mechanism is proposed to improve the performance of this method.

4. Conclusion

In this paper, a method combining discretization and feature selection is proposed for high dimensional data. The next work is to verify the feasibility analysis of PSO in feature selection based on discretization, and to study how to conduct multivariate discretization and feature selection in a single process to improve the recognition ability of feature set.

References

[1] Liao Weilin, CHENG Shan, SHANG Dongdong, WEI Zhaobin. Particle Swarm Optimization Algorithm with Multi-Strategy Fusion [J/OL]. Computer Engineering and Application :1-10[2020-06-29].

[2] Huang Xin, MO Haimao, ZHAO Zhigang, ZENG Min. Study on Discrete Enhanced Fireworks Algorithm and kNN in Feature Selection [J/OL]. Computer Engineering and Application :1-8[2020-06-29.

[3] Laniel, sun chaoli, he xiaojuan, tan ying. Improved particle swarm optimization for solving large-scale multi-objective problems [J]. Journal of taiyuan university of science and technology, 2020, 41 (04):249-256.

[4] Gu xiaolin, huang Ming, liang xu, jiao xuan. A chaotic particle swarm optimization algorithm with improved inertia weight [J]. Journal of dalian jiaotong university,2020,41(03):102-106+113.