Journal of New Media
DOI:10.32604/jnm.2021.018267
Article

Tech Science Press

Feature Selection Based on Distance Measurement

Mingming Yang[*] and Junchuan Yang

School of Computer & Software, Nanjing University of Information Science and Technology, Nanjing, 210044, China
[*]Corresponding Author: Mingming Yang. Email: yang1367353247@163.com

Received: 03 March 2021; Accepted: 05 March 2021

Abstract: Every day we receive a large amount of information through different social media and software, and this data and information can be realized with the advent of data mining methods. In the process of data mining, to solve some high-dimensional problems, feature selection is carried out in limited training samples, and effective features are selected. This paper focuses on two Relief feature selection algorithms: Relief and ReliefF algorithm. The differences between them and their respective applicable scopes are analyzed. Based on Relief algorithm, the high weight feature subset is obtained, and the correlation between features is calculated according to the mutual information distance measure, and the high redundant features are removed to obtain the feature subset with higher quality. Experimental results on six datasets show the effectiveness of our method.

Keywords: Feature selection; mutual information; distance measurement; relief

1 Introduction

With the rapid development of the Internet, we have entered the era of data explosion. Every day we receive a large amount of information through different social media and software, and this data and information can be realized with the advent of data mining methods. Data mining refers to extracting knowledge that people are interested in from a large database, which is implicit and realizes unknown and potentially useful information. Data mining is one of the most advanced research directions in the field of database and information decision-making. At present, the application of data mining and pattern recognition in the field of data set is moving toward large-scale and high dimension, such as text classification, image recognition, and biological gene expression array analysis applied in the processing of data set, in the process of data mining, to deal with and solve the problem of some high dimension, in the limited training samples, choose the useful feature for these problems, many experts have conducted extensive research.

There are two ways to implement dimensionality reduction for high-dimensional data sets: Feature extraction and feature selection. Feature extraction is the process of transforming the data set from high dimension to low dimension through mapping or transformation. Feature selection is to reduce the dimension of feature space by removing redundant and irrelevant features from the original feature space according to some evaluation function. The features obtained by feature extraction are linear or non-linear combinations of the original features, which causes the learning algorithm to measure all the original features in most cases, and will not reduce the computational workload of the learning algorithm, and the feature extraction from the combination of feature is not the actual physical significance, the characteristics of comprehensibility are poor, and feature extraction algorithm’s time complexity is higher than feature selection algorithm, so the feature selection has more advantages in the aspect of data dimension reduction, and is widely used in many industries.

Feature selection is an important way to improve the performance of learning algorithms. What’s more, it is a key component of data preprocessing in pattern recognition, data mining and machine learning. It cannot only reduce system complexity and processing time, but also improve system performance. Feature selection has three goals: to improve the classification performance of the classifier; to achieve a more effective and faster classifier, and to improve the intelligibility of the processing data generation mechanism. Feature selection reduces the dimension of the feature space by removing redundant and irrelevant features, and applies effective features to the algorithm, thereby improving the ability of deep data mining, improving the classification accuracy of the classifier, and further improving the comprehensibility of the classification results.

Feature selection research has been very active in the past ten years, and many feature selection methods have emerged. However, due to the diversity of feature selection methods and the complexity of the problems dealt with, there is no fixed selection mode and an effective method so far; the same Feature selection has not yet a unified and complete mathematical definition. Many scholars have defined feature selection according to different problems and target requirements. Finally, there is no uniform standard for the evaluation function of feature subsets, and many evaluation theories have not achieved the expected effects in the theory in practical applications. Therefore, feature selection is still in a process of continuous development and improvement, and more people need to study and explore.

According to the relationship with the learning algorithm, the feature selection methods are mainly divided into two categories: The Filter model and the Wrapper model. The Filter model mainly sorts the features by measurement, typically the distance between classes or the information entropy, to select the appropriate features; The wrapper model was first proposed by John to evaluate the selected feature set through a classifier to select the appropriate feature. The Filter model does not directly optimize the performance of any specified learning algorithm but evaluates the relevance and importance of features through specific metrics. Therefore, different feature selection algorithms will be generated based on different metrics, such as feature selection algorithms based on information entropy, Feature selection algorithm based on various distance metrics. In most cases, the Filter model is computationally more efficient, but its performance is worse than that of the Wrapper model.

In recent years, the research of feature selection has developed in a comprehensive and diversified direction. For example, Xing, etc., combined the advantages of Filter-type and Wrapper-type feature selection algorithms and proposed a hybrid feature selection algorithm to process high-dimensional data sets. Similarly, more and more selection techniques and metrics are beginning to be applied to feature selection, such as clustering techniques, fuzzy rough set theory, neural networks, and so on.

Feature selection algorithms can be divided into two categories: Single feature search and feature subset search. Single feature search In the feature evaluation process, only one feature is considered at a time, that is, a feature with the best performance is selected. The most representative of this type of algorithm is the Relief series of algorithms. With increasing the number of sample samples and the number of original features, the running time of the Relief algorithm increases linearly. Thus, it has high operating efficiency. This algorithm will give higher weight to the features, which have a high correlation with the category. But it ignores the redundancy between the features, and the original Relief algorithm cannot get satisfactory results.

The main research content of this article includes the following two aspects:

(1) Summarize and study two kinds of Relief algorithms: Relief and ReliefF algorithms. Analyzed the difference between them and the scope and type of their respective applications.

(2) Based on the Relief algorithm, the obtained high-weight feature subset is calculated according to the mutual information metric to calculate the correlation between the features, and the high-redundant features are removed, thereby obtaining a higher-quality feature subset.

2 Related Work

2.1 Distance Measurement

Distance measurement in feature space is also the core of pattern recognition. R. A. Fisher (1936) [1] attempted to use different measures to solve the classification problem. At the same time, LeCun et al. proposed a nonlinear feature extraction method based on distance measurement information [2]. Hatie et al. proposed an adaptive discriminant nearest neighbor classification algorithm [3]. After the 20th century, the concept of distance measurement learning was formally put forward. Distance measurement learning can be divided into supervised and unsupervised according to the effectiveness of training data. Distance measurement with supervision can be further divided into global measurement learning algorithms and local measurement learning algorithms. Global distance metric learning algorithms include: Xing et al. proposed a probability-based global distance metric learning algorithm (PGDM) by solving constrained convex programming [4]. Bar-Hillel et al. proposed the Correlation Component Analysis Algorithm (RCA) by selecting the structure of global data with equality constraints [5]. Local distance measurement learning algorithms include: Weinberger et al. proposed the Large Boundary Nearest Neighbour (LMNN) algorithm based on K-nearest neighbor classification [6]. Goldberger et al. estimated the conditional probability distribution of the sample and proposed the NCA algorithm [7]. Unsupervised distance measurement learning includes: Bar-Hillel et al. proposed principal component analysis algorithm (PCA) [8] by solving matrix decomposition that best preserves the original data structure. The ISOMAP algorithm proposed by Bar-Hillel et al. [8], etc.

2.2 Feature Selection

Since the 1960s, many researchers have done a lot of research on feature selection. Although it has been developed to this day, there is still no unified and perfect mathematical definition of feature selection. According to different problems and target requirements, scholars have different definitions of feature selection. According to whether the class label of the data set is available, feature selection methods can be divided into supervised feature selection, unsupervised feature selection, and semi-supervised feature selection. The class label of unsupervised feature selection is not available, so the target category cannot be used to evaluate the feature. It is necessary to design an evaluation standard to evaluate the quality of the feature. The data used in this paper are all class labels available, that is, the research is supervised feature selection.

Kira et al. [9] mainly defined feature selection from the aspect of target representation and recognition, and it is also the most ideal definition of feature selection: Find a feature subset that is sufficient to represent or recognize the target concept from the original space, and the feature set contains as few features as possible. Dash et al. [10] gave a more rigorous definition of feature selection from the perspective of classification accuracy and category probability distribution: Feature selection is to select a feature subset with the smallest number of features from the original feature space without significantly reducing the classification accuracy or changing the probability distribution of the target category. Starting from the purpose of feature selection, Lei et al. [11] defined feature selection: Feature selection is the process of removing category-independent features and redundant features from the original feature space, and retaining strong and weakly related features.

In 1997, Dash et al. [10] gave the basic framework of feature selection. First, the original data set obtains a feature subset according to a certain search strategy, and then uses the evaluation function to evaluate it. According to the score of the feature subset or other constraint conditions, it is judged whether the feature subset meets the termination condition, and if so, the loop is ended, and the feature subset is verified. Otherwise, according to the search strategy again, search for the next subset until the termination condition is met.

3 Method

3.1 Relief Algorithm

The Relief algorithm first proposed by Kira is mainly for two classification problems. This method is a feature weighting algorithm. According to the relationship of each feature and category, it can assign different weights to features. The meaning of weight is to subtract the feature difference of the same category and add the feature difference of different categories. You can specify a threshold r, you only need to select the feature with a greater weight than r, or you can specify the number of features you want to select k, and then select the k features with the largest weight.

The correlation between features and categories in the Relief algorithm is determined by the power of distinguishing short-distance samples. A sample Z is firstly selected by the Relief algorithm from the training set T. Then, the nearest neighbor sample T from samples of the same type as Z is found, called Near Hit. Meanwhile, the nearest sample F from samples of different types from Z is found, called Near Miss. According to the following rule updates the weight of each feature: If the distance between Z and Near Hit on the feature is less than the distance between Z and Near Miss, it means that the feature is useful to distinguish the nearest neighbors of the same type from different types, so increase feature Conversely, if the distance between Z and Near Hit on the feature is greater than the distance between Z and Near Miss, it means that the feature hurts distinguishing the nearest neighbors of the same type from different types, so reduce the importance of the feature. Because when a feature is beneficial to classification, the similar samples are closer to the feature, while the heterogeneous samples are farther away from the feature. The above process is repeated m times, that is, the number of sampling is n times, and we can obtain the average weight of each feature in the end. The weight of a feature indicates the classification ability of the feature.

The pseudo-code of the Relief algorithm is as follows:

Suppose training data set T, feature number S, sample sampling times n, feature weight threshold ∂, and the feature weights of each feature are W:

```
1. Initialize the feature weight of all features $W(A)$ to 0, and $L$ is an empty set;
2. For $i=1$ to $n$ do:
2.1) Randomly select a sample $Z$;
2.2) Find the nearest neighbor sample $T$ of $Z$ from the sample set of the same kind, and find the
     nearest neighbor sample $F$ from the sample set of different types;
2.3) For $O=1$ to $S$ do:
```

$$W(O)=W(O)-diff(O,Z,T)/n+diff(O,Z,F)/n \tag{1}$$

$$diff(O,T_1,T_2)=\frac{|T_1[O]-T_2[O]|}{\max(O)-\min(O)} \tag{2}$$

```
3. For $O=1$ to $S$ do:
3.1) If $W(O)\geq\partial$:
3.2) Add the feature $O$ to $L$
4. end;
```

Relief is a well-known filtering feature selection method. In 1994, Kononeill extended the Relief algorithm to obtain the ReliefF algorithm, which can handle multi-category problems. This algorithm is used to deal with the problem of continuous value regression for the target attribute.

When the Relief algorithm deals with multi-class problems, a sample Z is randomly selected from the training sample set each time, and c nearest neighbor samples of Z are found from the sample set of the same class of Z, and c nearest neighbor samples are found from the sample set of different classes of Z. Then, we update the weight of each feature, which is defined as the following formula:

$$W(O)=W(O)-\sum_{j=1}^{c}\frac{diff(O,Z,T_j)}{nc}+\sum_{C\notin class(Z)}\left[\frac{p(C)}{1-p(class(Z))}\sum_{j=1}^{k}diff(O,Z,F_j(C))\right]/(nc) \tag{3}$$

where $p(C)$ is the proportion of class C and $p(class(Z))$ refers to the proportion of the categories of a randomly selected sample Z.

The ReliefF pseudo algorithm is as follows:

Suppose training data set D, feature number N, sample sampling times m, the number of the nearest neighbor samples k, and the feature weights of each feature are T:

Suppose training data set T, feature number S, sample sampling times n, feature weight threshold ∂, and the feature weights of each feature are W:

1) Initialize the feature weight of all features $W(O)$ to 0, and L is an empty set;

2) For $i=1$ to n do:

2.1) Randomly select a sample Z;

2.2) Find the c nearest neighbors of Z from the same sample set of Z, $T_j\left(j=1,2,\cdots,c\right)$, and find the c nearest neighbors from each sample set of different types, $F_j\left(j=1,2,\cdots,c\right)$;

2.3) For $O=1$ to S do:

$$W(O)=W(O)-\sum_{j=1}^{c}\frac{diff\left(O,Z,T_j\right)}{nc}+\sum_{C\notin class(Z)}\left[\frac{p(C)}{1-p\left(class(Z)\right)}\sum_{j=1}^{k}diff\left(O,Z,F_j(C)\right)\right]/(nc)$$

2.4) Sort O by $W(O)$

3) End

3.2 ReliefF Algorithm Based on Mutual Information

Relief algorithm will give high weight to all features with high relationship with categories, but the disadvantage is that it ignores the correlation between features, which results in high redundancy of the obtained feature subset. ReliefF is improved from the perspective of removing redundant features and combining ReliefF with mutual information.

3.2.1 Maximum Correlation Minimum Redundancy Algorithm (mRMR)

mRMR is a filtering feature selection algorithm proposed by Peng et al.Uses: image recognition, machine learning, etc. A commonly used feature selection method is to maximize the correlation between features and categorical variables, which is to select the top k variables that have the highest correlation with categorical variables. However, the combination of the single good feature may not increase the accuracy of the classifier, because the features may be highly correlated, which leads to feature redundancy. Thus came MRMR, which maximizes the correlation between features and categorical variables, and minimizes the correlation between features. This is the core idea of MRMR.

3.2.2 Mutual Information

Mutual information is used to describe the correlation between things in information theory. Mutual information can be thought of as the amount of information contained in a random variable about another random variable, or the uncertainty of a random variable since another random variable is known. The relationship between two random variables, X and Y, can be defined by the amount of mutual information. The degree of correlation between random variables is described by mutual information. It measures the amount of common information between two variables.

Suppose two random variables X and Y, we define the mutual information $I(X;Y)$ between them as:

$$I(X;Y)=\iint p(X,Y)\log\frac{p(X,Y)}{p(X)p(Y)}dXdY \tag{4}$$

where $p(X)$ and $p(Y)$ are the edge probability distributions of random variables X and Y, respectively, and $p(X,Y)$ is the joint probability distribution of random variables X and Y. According to the definition, when the random variables X and Y are independent or completely unrelated, their mutual information is 0, that is, there is no common information between the two variables. On the contrary, the greater the mutual information $I(X;Y)$ is, the higher the correlation degree of random variables X and Y is, and the more common information they have.

Mutual information can be equivalent to:

$$I(X;Y)=H(X)-H(X\mid Y)=H(X,Y)-H(X\mid Y)-H(Y\mid X) \tag{5}$$

where $H(X)$ is the edge of entropy, $H(X\mid Y)$, $H(Y\mid X)$ is conditional entropy, and $H(X,Y)$ is the entropy.

3.2.3 Improved Relief *Algorithm*

Considering that the Relief algorithm ignores the correlation between features and has high redundancy, it is better to combine the Relief algorithm and MRMR algorithm to get the Relief algorithm based on mutual information. Firstly, a threshold value δ is set, and the feature subset obtained by the Relief algorithm is sorted from high to low according to the weight size. Then, the mutual information value between the two features is calculated successively. When the mutual information value is higher than the threshold value, the low-weighted feature subset is removed and the high-weighted feature subset is retained, and then the optimal feature subset is obtained.

4 Experiment

Based on Matlab platform to realize four various algorithms in this article, Relief, ReliefF, and Relief algorithm and ReliefF algorithm based on mutual information, will get feature subsets and the original data sets using KNN classifier calculates the precision of the algorithm, respectively compared with relief based on mutual information algorithm, as well as ReliefF algorithm and ReliefF algorithm based on mutual information. To make the experimental results more representative and accurate, the selected data sets are diverse, including multi-instance, multi-feature, and multi-category.

4.1 Datasets

Since the Relief algorithm can only solve binary classification problems, and ReliefF can solve multi-category problems, the data sets WBCD, SPECT, and German will be used for Relief and improved Relief algorithm experiments, while the data sets Iris, Abalone, and Newthyr will be used for ReliefF and improved ReliefF algorithm experiments. For mutual information, since it only has a lower limit of 0 and no upper limit, the threshold should be adjusted for different data sets according to the actual situation without using fixed threshold evaluation criteria.

Table 1: Datasets

Dataset	Instance	Feature	Categories
Wbcd	569	30	2
Spect	267	44	2
German	1000	25	2
Iris	150	4	3
Abalone	4177	8	3
Newthyr	215	5	3

4.2 Results

4.2.1 Relief and Relief Algorithm Based on Mutual Information Experiments

Since the number of features in each data set is different to ensure an appropriate number of selected feature subsets and reduce errors, I set the number of features in the selected feature subset to 80% of the total number of features in the original feature set at the beginning, that is, the features are sorted in descending order according to weight, and then the first 80% are selected. Tab. 2 shows the experimental results.

Table 2: The number of features selected by the Relief algorithm

	Instance	Features	Relief Features
Wbcd	569	30	24
Spect	267	44	35
German	1000	25	19

Table 3: The selection feature number and accuracy of Relief and Relief based on mutual information on Wbcd dataset

Threshold	Relief		Relief (mutual information)	
	Accuracy	Features	Accuracy	Features
0.7	0.9227	24	0.8682	8
0.8	0.9279	24	0.8524	13
0.9	0.9278	24	0.8840	15
1.0	0.9315	24	0.9244	24

Table 4: The selection feature number and accuracy of Relief and Relief based on mutual information on Spect dataset

Threshold	Relief		Relief (mutual information)	
	Accuracy	Features	Accuracy	Features
0.6	0.7154	35	0.7193	16
0.7	0.7191	35	0.7453	20
0.8	0.6890	35	0.7119	30
0.9	0.7040	35	0.7157	35

Table 5: The selection feature number and accuracy of Relief and Relief based on mutual information on German dataset

Threshold	Relief		Relief (mutual information)	
	Accuracy	Features	Accuracy	Features
0.6	0.6870	19	0.6890	17
0.7	0.6850	19	0．7080	17
0.8	0.6810	19	0.6860	18
0.9	0.6830	19	0.6830	19

In the Relief algorithm experiment, the experimental results on the WBCD data set are not ideal. After constantly changing the threshold and reducing the number of features, the accuracy of the algorithm decreases significantly, and the algorithm performance is not optimized. Experiments on SPECT data sets, in most cases, not only reduce the number of features, but also slightly improve the accuracy of the algorithm, and achieve true optimization. On the German data set, the number of features is reduced under the condition of ensuring the same or close accuracy, which is also a successful experiment of algorithm optimization.

4.2.2 ReliefF and ReliefF Algorithm Experiments Based on Mutual Information

Table 6: Number of features selected by ReliefF algorithm

	Instance	Features	ReliefF Features
Iris	150	4	3
Abalone	4177	8	3
Newthyr	215	5	3

Table 7: The selection feature number and accuracy of ReliefF and ReliefF based on mutual information on Iris dataset

Threshold	ReliefF		ReliefF (mutual information)	
	Accuracy	Features	Accuracy	Features
0.5	0.9397	3	0.6462	1
0.6	0.9265	3	0.7799	2
0.7	0.9399	3	0.6800	2
0.8	0.9667	3	0.7533	2
0.9	0.9668	3	0.9268	3

Table 8: The selection feature number and accuracy of ReliefF and ReliefF based on mutual information on Abalone dataset

Threshold	ReliefF		ReliefF (mutual information)	
	Accuracy	Features	Accuracy	Features
0.6	0.5205	6	0.4671	1
0.8	0.5183	6	0.4800	1
0.9	0.5095	6	0.4759	1
1.0	0.5162	6	0.5224	6

Table 9: The selection feature number and accuracy of ReliefF and ReliefF based on mutual information on Newthyr dataset

Threshold	ReliefF		ReliefF (mutual information)	
	Accuracy	Features	Accuracy	Features
0.4	0.9071	4	0.7860	1
0.5	0.9117	4	0.8977	2
0.6	0.9071	4	0.8931	2
0.7	0.8930	4	0.8837	3
0.8	0.9255	4	0.9115	4

The same ReliefF algorithm selects 80% of the feature number, and then makes the selection based on mutual information. The results of the three experiments did not reach the expected effect.

4 Conclusion

Experimental results of the improved Relief algorithm show that the improved algorithm is effective when the number of features is large, the accuracy of the original data set is not high, and the number of instances is large. However, when the accuracy is already very high or the number of features is relatively small, the algorithm accuracy will be greatly reduced if the number of features is reduced by improving the algorithm. The improved experiment of the ReliefF algorithm is not satisfactory. In a word, the main reason for this experimental result is that the number of features in the selected data set is too small. Among them, the experimental results of the Abalone data set are of certain significance. The features of this data set have high redundancy, and the accuracy difference between one feature and six features is not very big, which shows the effectiveness of the algorithm improvement from a certain aspect.

Funding Statement: The author(s) received no specific funding for this study.

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] W. L. G. Koontz and K. Fukunaga, "A nonlinear feature extraction algorithm using distance information," *IEEE Transactions on Computers*, vol. 21, no. 1, pp. 56–63, 1972.

[2] Y. LeCun, P. Y. Simard and J. Decker, "Efficient pattern recognition using a new transformation distance," in *Proc. NIPS,* Denver, Colorado, USA, pp. 50–58, 1992.

[3] T. Hatie and R. Tibshirani, "Discriminant adaptive nearest neighbor classification," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, no. 6, pp. 607–616, 1996.

[4] E. Xing, A. Ng, M. Jordan and S. Russell, "Distance metric learning, with application to clustering with side-information," in *Proc. NIPS,* Vancouver, British Columbia, Canada, pp. 505–512, 2002.

[5] A. Bar-Hillel, T. Hertz, N. Shental and D. Weinshall, "Learning distance functions using equivalence relations," in *Proc. ICML,* Washington DC, USA, pp. 11–18, 2003.

[6] K. Q. Weinberger and L. K. Saul, "Distance metric learning for large margin nearest neighbor classification," *Journal of Machine Learning Research*, vol. 10, no. 1, pp. 207–244, 2009.

[7] J. Goldberger, S. F. Wong, B. Steng, J. Kittler and R. Cipolla, "Incremental linear discriminant analysis using sufficient spanning set approximations," in *Proc. CVPR,* Minneapolis, Minnesota, USA, 2007.

[8] A. Bar-Hillel, T. Hertz, N. Shental and D. Weinshall, "Learning distance functions using equivalence relations," in *Proc. ICML,* Washington, DC, USA, pp. 11–18, 2003.

[9] K. Kira and L. A. Rendell, "The feature selection problem: traditional methods and a new algorithm," in *Proc. AAAI,* San Jose, CA, USA, pp. 50–58, 1992.

[10] M. Dash and H. Liu, "Consistency-based search in feature selection," *Artificial Intelligence*, vol. 151, no. 1–2, pp. 155–176, 2003.

[11] Y. Lei and H. Liu, "Efficient feature selection via analysis of relevance and redundancy," *The Journal of Machine Learning Research*, vol. 5, no. 1–2, pp. 1205–1224, 2004.