APPLICATION OF RELIEFF ALGORITHM TO SELECTING FEATURE SETS FOR CLASSIFICATION OF HIGH RESOLUTION REMOTE SENSING IMAGE

Zhi Wang[1], *Yan Zhang*[1*], *Zhichao Chen*[1], *Huan Yang*[1], *Yaxin Sun*[2], *Jianmin Kang*[2], *Yong Yang*[2], *Xiaojun Liang*[2],

[1] Institute for Geo-Informatics and Digital Mine Research, College of Resources and Civil Engineering, Northeastern University, Shenyang 110819, China;
[2] East Anshan Iron Mine, Anshan Iron & Steel Group Mining Co., Ltd, Anshan 114041, China.
zhangyanneu@126.com(Y.Z.)

ABSTRACT

In classification, a large number of features often make it difficult to select appropriate classification features. In such situations, feature selection or dimensionality reduction methods play an important role in classification. ReliefF algorithm is one of the most successful filtering feature selection methods. In this paper, some shortcomings of the ReliefF algorithm are improved, on the problem of poor stability of neighbor samples selection, proposing the method of using the average value of multiple random selection to improve the anti-volatility of the algorithm. And redundant analysis is added to the ReliefF algorithm to eliminate the redundant features. The experimental results show that the improved ReliefF algorithm can effectively establish the classification feature sets, achieve the better classification accuracy.

Index Terms—*classification*, *feature selection*, *ReliefF algorithm*

1. INTRODUCTION

In recent years, with the continuous improvement of the resolution of the remote sensing image, the object-oriented classification making full use of the rich information of image has become the focus of the research. Theoretically speaking, all spectral, shape, texture and topological relation feature information of the object should be used for the classification of images. However, due to the "Dimension disaster", too many features will not only make the operation become more complex, the processing speed is greatly reduced; but also can reduce the classification accuracy in the case of limited samples [1]. Therefore, it is important to select the useful features of classification. But on the feature selection for classification, mainly according to the expert's experience knowledge, through a number of trials to select the more effective features for the classification [2]. This method is not only time consuming and laborious, but also has the interference of human factors. In the feature selection filed, one of the most successful individual feature filtering algorithms is the ReliefF algorithm. The Relief algorithm is proposed by Kira and Rendell in 1992 [3]. Kononenko, in 1994 extended the Relief algorithm, proposed the ReliefF algorithm to solve multi class problems [4]. This algorithm has been successfully used in many large subset feature selection tasks. Newton Spolaor proposed feature selection of hierarchical multi-label learning based on ReliefF algorithm [5]. L.Wang proposed a combination of ReliefF and mRMR algorithm to reduce remote sensing image features dimension [6]. Li Haixia combines the ReliefF algorithm and genetic algorithm for image feature selection, but its implementation process is more complex, the redundancy analysis is not integrated into the ReliefF algorithm. In this work, we mainly study improved ReliefF algorithm, evaluate its performance on the classification of UAV high resolution image data sets.

2. RELIEFF ALGORITHM

ReliefF algorithm has high efficiency and does not limit the characteristics of data types, the Relief enable it to deal with data sets with discrete or continuous. When dealing with multi class problems, the ReliefF algorithm selects the nearest neighbor samples from each of the samples in different categories. At first, we randomly select a sample x from the training sample, then to find out k nearest neighbor samples from the kind of sample x, and randomly find out k non similar nearest neighbor samples from neighbors of different classes. To adjust a feature weighting vector to give more weight to features by comparing within-class distance and between-class distance from neighbor samples. Repeat the above procedure on each feature dimension, finally get the weight value of each feature. The formula of ReliefF algorithm to update the weight value of feature is.

Supported by National Natural Science Foundation of China (41271013), the Fundamental Research Funds for the Central Universities(N150104002), Liaoning BaiQianWan Talents Program(2015-15).

978-1-5090-3332-4/16/$31.00 ©2016 IEEE

$$W_{\mathrm{f}}^{i+1}=W_{f}^{i}+\sum_{c\neq class(\mathrm{x})}\frac{\frac{p(\mathrm{x})}{1-p(\mathrm{class(x)})}\sum_{j=1}^{k}diff_{f}(\mathrm{x},\mathrm{M}_{j}(\mathrm{x}))}{m*k} \quad (1)$$

$$-\sum_{j=1}^{k}diff_{f}(\mathrm{x},\mathrm{H}_{j}(\mathrm{x}))/(\mathrm{m*k})$$

$diff_{f}()$ is the distance of two samples on the feature f

$\mathrm{H}_{j}(\mathrm{x})$ is the neighbor samples from the kind of sample x

$\mathrm{M}_{j}(\mathrm{x})$ is the neighbor samples from neighbors of different classes

$p(\mathrm{x})$ is the probability of class

In this study, the Euclidean distance is used to calculate the within-class distance and between-class distance of the samples. Euclidean distance is widely used in the field of image processing. The Euclidean distance reflects the degree of similarity between two pixels, the smaller the value, the difference between two pixels is smaller. Euclidean distance formula is.

$$D(\mathrm{x},\mathrm{y})=\left[\sum_{i=1}^{d}(\mathrm{x}_{i}-y_{i})^{2}\right]^{2} \quad (2)$$

3. METHOD

3.1 Data

This study uses the UAV high resolution image of East Anshan open-pit mines and the surrounding urban areas to validate the ReliefF feature selection algorithm. the UAV high resolution image can reach the resolution of 0.3m, not only has the spectral information, but also contains a wealth of shape, texture information. The image contains both the open-pit mine area and the surrounding city, its complex surface features types can be used to validate the effectiveness of the ReliefF feature selection algorithm better.

3.2 Feature Extraction and Selection

Feature extraction and selection are important steps in classification, an optimum feature set should have effective and discriminating features, which can reduce the redundancy of features to avoid ‘‘curse of dimensionality’’ problem.

In this study, we improved the ReliefF algorithm to measure the weight of the classification feature. In view of the deficiency of ReliefF algorithm, the redundant analysis is added to the algorithm, and the method for calculating the average value of a number of times random selection is adopted to increase the reliability of the nearest neighbor sample selection problem. The improved ReliefF algorithm mainly includes the following steps.

1) Original data of sample normalization. normalization value = (sample value - sample mean)/ sample variance. At the same time, calculate the number of samples in each class.
   [stand_data, sort] = standardization(data);
2) The weight vector for feature. The formula of feature weighting is formula (1). within-class and between-class distance is Euclidean distance. The improvement of using the average value of multiple random selection in the selection of neighbor samples, improving effectively the anti-volatility of the algorithm.
   [weight] = weighted(sort, stand_data);
3) Painting the distribution map and statistical histogram of features' weight.
   [max_weight] = dot_weight (weight);
4) Feature selection. according to the set of weight threshold, selecting a set of features with the maximum classification weight.
   [important_data, important_order, important_w, stand_important] = choose (weight, g_numbertotal, data, stand_data);
5) Redundancy analysis. Analysis the relevance of the larger weight features, retaining the feature with higher weight and eliminating the feature with lower weight in the large correlation of features, the purpose is to remove the interference of redundant features, form the final feature sets.
   [feature, feature_order] = redundance (data, important_order, important_w, stand_important, g_numbertotal).

3.3 Classification

This study uses the fuzzy classification method of software eCogntion to classify the target image based on the feature set be selected by the ReliefF feature selection algorithm. The classification method uses the membership function or probability to represent the possibility of each object may belong to a category. The values of membership function between 0-1, in the form of probability to divide categories. The optimal segmentation scale is different for different objects, so we use the multi-scale hierarchical classification to complete the classification of complex objects in open pit mine.

3.4 Accuracy Evaluation

At the present stage, the method of accuracy evaluation is varied, in this paper, we choose the method of confusion matrix to evaluate the accuracy of the classification results. The classification accuracy of each category is reflected by the producer accuracy and user accuracy. Kappa coefficient uses the information of the whole error matrix to evaluate the accuracy of the classification.

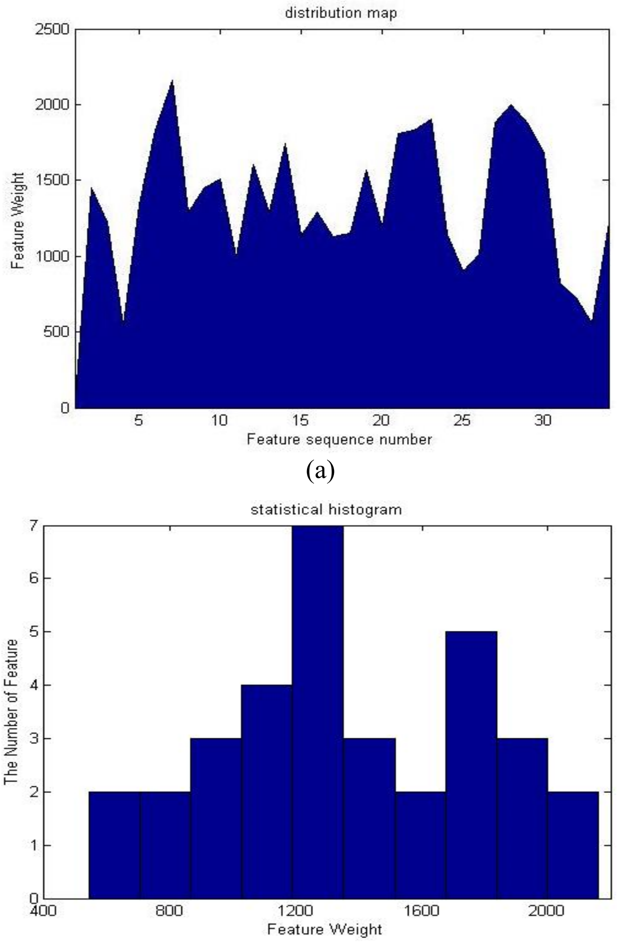
4. EXPERIMENTAL RESULTS AND ANALYSIS

This part mainly lists the results of feature selection and the classification results based on feature sets.

Figure 1 shows the distribution map and statistical histogram of features' weight obtained by the improved ReliefF feature selection algorithm. From the figure we can directly see the weight distribution.

Table 1 shows the final classification feature set through removing redundant features by redundancy analysis.

Figure 2 shows the object-oriented classification result of the UAV high resolution image based on the feature sets obtained by the improved ReliefF algorithm.



(a)

(b)

Figure 1 Statistical Histogram

Table 2 shows the confusion matrix, producer accuracy, user accuracy and Kappa coefficient of the classification result. The overall accuracy reached 81.6%, kappa coefficient reached 0.791.

It can be seen from the final classification results, the improved feature selection algorithm based on ReliefF algorithm plays the role of automatic selection of feature sets in object-oriented classification, solves the problem of the interference of redundant features and random sample selection. Whether it is the object of the mining area or the surrounding cities, it is able to achieve better classification results. The accuracy evaluation results showed that the overall accuracy reached 81.6%, kappa coefficient reached 0.791, achieve better classification quality.

Table 1 Feature Sets

Feature	Weight Value	property
Density	2161.32314	Shape Feature
Asymmetry	2000．74564	Shape Feature
GLCM_StdDe	1881．03951	Texture Feature
Mean_Layer1	1831．28629	Shape Feature
Length/width	1681．08873	Shape Feature
RI	1901．90045	Custom Feature

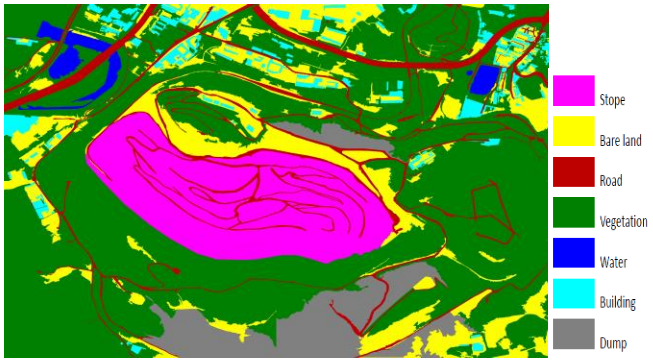


Figure 2 Classification Results

Table 2 Accuracy Evaluation Results

category	Vegetation	Road	Bare land	Stope	Building	Dump	Sum	User Accuracy
Confusion Matrix								/
Vegetation	55	3	2	1	0	1	62	0.902
Road	0	32	1	1	0	0	33	0.970
Bare land	6	8	34	4	10	1	63	0.619
Stope	0	1	0	19	0	0	20	0.950
Building	0	0	0	0	35	0	35	1
Dump	0	0	9	0	0	16	25	0.640
Sum	61	44	46	23	45	18	237	/
Production Accuracy	0.902	0.727	0.739	0.826	0.778	0.889	/	/
Overall Accuracy	81.6%							
Kappa	0.791							

5. REFERENCES

[1] T. Blaschke, "Object based image analysis for remote sensing," ISPRS Journal of Photogrammetry and Remote Sensing, vol. 65, pp. 2-16, 2010.

[2] L. Bruzzone and C. Persello, "A novel approach to the selection of spatially invariant features for the classification of hyperspectral images with improved generalization capability," IEEE Transactions on Geoscience and Remote Sensing, vol. 47, pp. 3180-3191, 2009.

[3] K. Kira and L. A. Rendell, "The feature selection problem: traditional methods and a new algorithm," in AAAI-92. Tenth National Conference on Artificial Intelligence , Menlo Park, CA, USA, pp. 129-34, 1992.

[4] I. Kononenko, "Estimating attributes: analysis and extensions of RELIEF," in Machine Learning: ECML-94. European Conference on Machine Learning, Berlin, Germany, pp. 171-82, 1994.

[5] N. Spolar and M. C. Monard, "Evaluating ReliefF-based multi-label feature selection algorithm," Lecture Notes in Computer Science, vol. 8864, pp. 194-205, 2014.

[6] L. Wang, "Multiple features remote sensing image classification based on combining ReliefF and mRMR," Chinese Journal of Stereology and Image Analysis, vol. 19, pp.250-256,2014.