A self-adaptive selective method of remote sensing image classification algorithms

Xin Pan, Hongbin Sun*

School of Computer Project & Technology, Changchun Institute of Technology, Changchun City, Jilin Province, China, 130012

Received 28 January 2014, www.tsi.lv

Abstract

Remote sensing image classification algorithms, which can obtain information of land use\cover quickly and inexpensively have been widely used in the field of GIS. The quality of classification results is not only affected by the quality of remote sensing data, but also affected by the character of classification algorithm. At present, despite a lot of algorithms have been proposed, but users usually meet difficulties in algorithm selection due to single classification algorithm cat not applicable to all classification cases. This study proposes a self-adaptive selective method for remote sensing image classification algorithms based on data complexity evaluation, through data complexity evaluation, our method can distinguish remote sensing data's character even from same satellite sensor and give user recommendation of algorithm selection. Experiments indicate that the algorithms selected by this method can achieve higher classification accuracy, which provides the recommendation for the selection of appropriate classification models to users.

Keywords: Remote Sensing image, Classification, Algorithms Evaluate, Data Complexity

1 Introduction

Remote sensing image classification algorithms classify the entire images by using a few training samples, which can obtain information of land use quickly and inexpensively and have been widely used in the field of GIS. At present, many different algorithms (including Naïve Bayes, ID3, CART, KNN, Neural Networks, SVM etc.) have been applied to remote sensing image classification [1]. The quality of classification results is not only affected by the quality of remote sensing data, but also affected by the classification algorithm [2]. Just as like No Free Lunch theorem represent: "If algorithm A outperforms algorithm B on some cost functions, then loosely speaking there must exist exactly as many other functions, where B outperforms A" [3], data set's characteristics varying greatly and there are no single classification algorithm can applicable to all the cases [4], so select appropriate classification algorithm for remote sensing image classification is very important.

Many scholars have made research in classification algorithms selection field: Brodley proposed a knowledge based method to search an algorithm [5]; Gama proposed a linear regression method to predict algorithm's accuracy [6], Brazdil further presented a meta-learning method to select candidate classification algorithms [7]. Song gave an automatic recommendation framework for classification model selection [4]. However, the studies above are based on data characteristic of feature structure, type, range features; remote sensing image data will have same data structure information, which obtained from same satellite sensor, and have similar data structure and statistical information, which from different satellite sensor, this will lead a difficulty to distinguish the date using the above methods. Therefore, it is necessary to introduce new evaluation method to describe remote sensing data's character.

Data complexity is a method, which can characterize data measures on the training data instead of experimenting with train data [8]. A lot of data complexity measures, concerning statistical, geometrical and information theoretic descriptions have been proposed in past few years [9]. It can give a relation between classifier performance and training date character and we can further give recommendation for select appropriate classifiers by the help of data complexity evaluation [10-12]. Therefore, data complexity can be a more effective method to describe the characteristics of the data.

This study proposes a new method named Self-adaptive Selective Method of Remote Sensing Image Classification Algorithms based on Data Complexity Evaluation (SSMRSICADCE), which describes the characteristics of remote sensing images with data complexity, obtains the relationship between the characteristics of data and classification accuracy of algorithms through a large number of remote sensing datasets, and further proposes suggestions for the selection of classification algorithms on that basis. Experiments indicate that the algorithms selected by this method can achieve higher classification accuracy, which provides the knowledge for the selection of appropriate

**Corresponding author* e-mail: 101103991@qq.com

classification models in the study of land use.

The remainder of this paper is organized as follows: Section 2 provides remote sensing data characterization method based on data complexity evaluation, section 3 provides the algorithm of proposed method, section 4 gives experiments result and section 5 draws conclusions.

2 Remote sensing data characterization based on data complexity

There are many formulas in data complexity evaluation field; it is difficult for a single formula to describe remote sensing data thoroughly. Therefore, we use a data complexity evaluation vector, which has three indexes to describe the character of a remote data set.

1) Index of Fisher's discriminant ratio

Fisher's discriminant ratio can describe how separated classes according to features:

$$i1 = \frac{\sum_{i=1}^{c} n_i \times \delta(m, m_i)}{\sum_{i=1}^{C}\sum_{j=1}^{n_i} \delta(x_j^i, m_i)}, \quad (1)$$

where n_i denotes the number of samples in class i, δ is a metric, m is the overall mean, m_i is the mean of class i, and x^i_j represent the sample j belonging to class i [11].

2) Index of volume of overlap region

The volume of the overlap region for two classes can be represented by the product of normalized lengths of overlapping ranges for all features [11]:

$$i2 = \prod_{K} \frac{\min\max k - \max\min k}{\max\max k - \min\min k}, \quad (2)$$

where k=1,2,3..n and $minmaxk = min\{max(f_k, c_1), max(f_k, c_2)\}$, $maxmink = max\{min(f_k, c_1), min(f_k, c_2)\}$, $maxmaxk = max\{max(f_k, c_1), max(f_k, c_2)\}$; $minmink = min\{min(f_k, c_1), min(f_k, c_2)\}$.

3) Index of pooled Mahalanobis distance

Pooled Mahalanobis distance can describe the distance between classes i and j:

$$D^2_{pool}(i, j) = (\mu i - \mu j)' \left(\frac{\sum_i + \sum_j}{2}\right)^{-1} (\mu i - \mu j), \quad (3)$$

where m is the mean vector of reflectance values, and Σ is the variance–covariance matrix. The index of pooled Mahalanobis distance can be describe as:

$$i3 = \sum_{i}^{C}\sum_{j}^{C} D^2_{pool}(i, j). \quad (4)$$

Through formulas above, we can describe remote sensing data's character by a vector from discriminant, overlap and classes distance level:

$$dcV=(i1,i2,\text{i3}). \quad (5)$$

Data sets which have similar data complexity evaluation vector would have similar requirement in classification, the distance of two data complexity evaluation vector in a data set d can have represented by the following formula:

$$dDCV(vi, vj, d) = \sum_{k}^{3} \frac{|v_{ik} - v_{jk}|}{2\delta(d1)}. \quad (6)$$

The following meta-information in a remote sensing image can distinguish between remote sensing images also plays an important role:
m1=number of bands;
m2=cellsize;
m3=Source Type;
m4=pixel Type;
m5=pixe Depth;
A meta-information vector can be represent as follows:

$$metaV=(m1,m2,m3,m4,m5); \quad (7)$$

the distance of two meta-information vector can have represented by the following formula:

$$dMI(mi1,mi2)= (\textit{number of difference items})/5; \quad (8)$$

A remote sensing data character can be represented by meta-information vector and data complexity evaluation vector:

$$RSC(\text{data})=\{\text{miv; dcv}\}=(\text{meta-information vector; data complexity evaluation vector}); \quad (9)$$

the *RSC* can describe a remote sensing date *d* and the differences between *R1* and *R2* can be calculate from following formulas:

$$\begin{aligned} dis\tan ce(R1, R2, d) &= \alpha 1 \times dM1(R1, R2) \\ &+ \alpha 2 \times dDCV(R1, R2, d). \end{aligned} \quad (10)$$

The process of remote sensing data characterization can be represented as the following algorithm:

Algorithm: Remote Sensing Image Data Characterization (*RSIDC*)

Input: Remote Sensing image *RS*, training samples' positions and catalogues *SPC*

Output: remote sensing data character *RSC*

1) *dataSet*=construct multi feature training data set from *RS* and *SPC;*
2) *mateInfo*= gather *number of bands, cellsize, Source Type, pixel Type, pixe Depth* information from RS;
3) *metaV*=construct vector by formula (7) from *mateInfo;*
4) *dcV*=construct vector by formula (5) from dataset;
5) *RSC*={ *metaV* ; *dcV* }
6) return RSC

End

From *RSIDC* Algorithm, we can characterize a remote Sensing Image Data and its training samples.

3 The self-adaptive selective method of remote sensing image classification algorithms

With the help of remote sensing data characterization from above section, we can realize the method of sensing image classification algorithms selection; the method can be described as follows:

Method: Self-adaptive selective method of remote sensing image classification algorithms (*SSMRSICADCE*)

Stage 1: *Training Stage*

Input: A large number of remote sensing image and their training samples

Ouput: Relationship data set *RDS*

RDS=Combine *remote sensing image, data character, classification algorithm, classification accuracy* together.

End

Stage 2:

Input: A remote sensing image *RS* and training samples *TS*

Output: recommend algorithms *RAS*

RAS= Self-adaptive recommend user to select classification algorithms with the help of *RDS*

End

Through *SSMRSICADCE* we can self-adaptive recommend user to select classification algorithms; it has two stages, as depicted in Fig 1, stage 1 aims at get the relationships from remote sensing image, data character, classification algorithm, classification accuracy.

From Figure 1, the detail of Stage 1 can be represented as follows:

1) Construct a training database from Remote sensing images and training samples. In the database, each group of remote sensing image and corresponding samples can construct a train training sample set, and each set can obtain remote sensing data character through *RSIDC* algorithm;
2) Lots of classical classification algorithms were gathered and construct a Classification algorithms database.
3) Classification algorithms database provide algorithm and classification each training sample set and obtain their classification accuracy.
4) Each training sample set and each classification algorithms together with their data character and corresponding classification accuracy were combine into relationship, we can obtain the relation describe as follows:

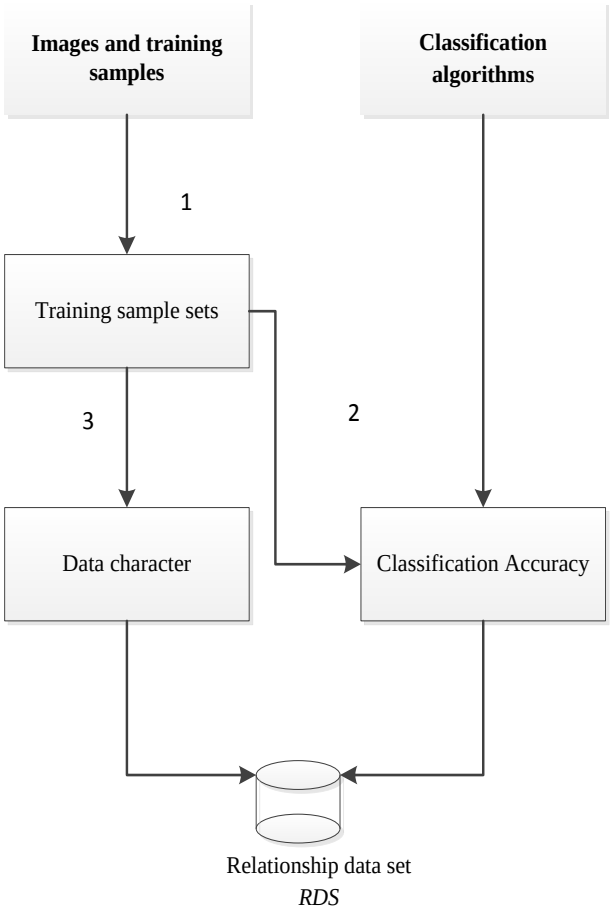

FIGURE 1 Get the relationships from remote sensing image, data character, classification algorithm, classification accuracy

Further save all the data mentioned above into Relationship data set *RDS*. Through stage 1, we can obtain a lot of relationships which can be regarded as knowledge to recommend classification algorithm selection. In stage 2 this knowledge are firstly used to select similar data character (see Figure 2).

Algorithm: Find similar data character (*FSDC*)

Input: Relationship data set *RDS*, A remote sensing image RS and training samples *TS*

Output: founded data characters *FDC*

1) *T_RSC*= get data character from RS and TS by algorithm RSIDC;
2) *FDC*= select *RSC* from *RDS* where *RDS.DC*. miv=*T_RSC;*
3) if *FDC*<>NULL Then return *FDC*; return data characters with exactly same data structure end if;

4) *FDC_distance* = select all the *DC* from *RDS*, and calculate formula (10) with $\alpha 1$=1 and $\alpha 2$=0;
5) *FDC*= select the items from *RDS* with FDC_distance < a threshold;
6) return *FDC*; return data character with similar data structure.

End

Data characters: DC

DC_ID: primary key
MIV: meta-information vector
DCV: meta-information vector

Classification algorithms: CA

CA_ID: primary key
Name: algorithm's name

Relationship: R

R_ID: primary key
DC_ID: corresponding data character
CA_ID: corresponding algorithm
ACCURACY: accuracy with specific data character and algorithm

FIGURE 2 The table structure of relationship

The algorithm *FSDC* we can find similar data characters, in the next step stage 2 need select most similar data character and select algorithms with may be archive higher classification accuracy (see Figure 3).

***Algorithm**:* get recommend algorithms (*GRA*)
***Input**:* founded data characters *FDC*, relationship data set *RDS*, data character *T_RSC*, number threshold *NT*;
***Output**:* recommend algorithms *RAS*

1) *DC_distance*= *T_RSC* get all the distance from *FDC*,calculate formula (10) with $\alpha 1$=0.5 and $\alpha 2$=0.5;
2) *FDC*= select top *NT* data characters from *FDC* with DC_distance ascend order;
3) Relations=select all the relationships RDS.R where *RDS.R .DC_ID* in (*FDC*);
4) *OrderedGroup*= Relations split into groups by *RDS.R. CA_ID* and each group's corresponding accuracy grade=average(accuracy in this group)+ (current group member number)/(*FDC* member), and all the group arrange in descending order by accuracy grade;
5) *RAS*= select classification algorithms from *RDS.CA* where *RDS.CA.CA_ID* in (OrderedGroup);
6) return *RAS*.

End

The flowchart of classification algorithm recommendation and obtain a classification result can be seen as follows:

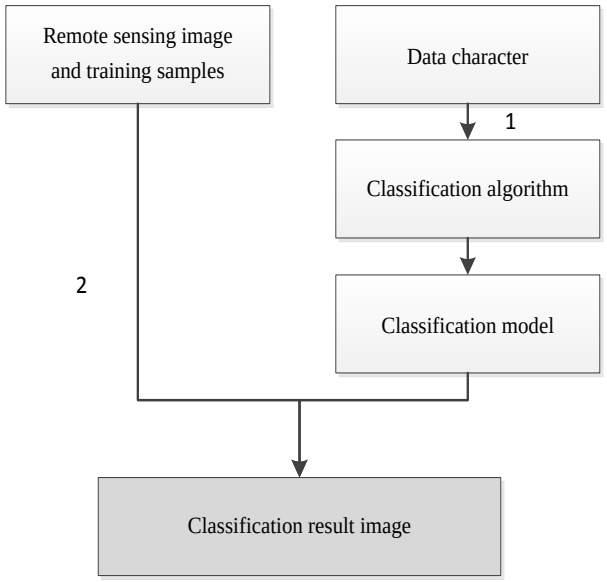

FIGURE 3 Classification algorithm recommendation and classification result

As can be seen from Figure 3:

1) remote sensing image and it's training samples are characterized by RSIDC algorithm and then find similar characters from Relationship data set by FSDC algorithm, through GRA algorithm get Recommend algorithms with prediction accuracy descending order;
2) select corresponding algorithms and train it by raining samples to obtain a Classification model, this classification model can be used to classify the whole remote sensing image.

4 Result of experiments

Self-adaptive method of remote sensing image classification algorithms selection has two stages. In stage 1, the method need to construct relationship data set *RDS*; in this study *RDS* will be generated through the following data set: There are two different types of the remote sensing images in the data set list as follows:

TABLE 1 The details of data which to construct *RDS*

Sensor type	Content	Number of sub-images
Landset TM	forest, grass and water	12
	building and farmland	20
SPOT5	wetland, grass and farmland	31
	building and road	15

Landset TM and SPOT5 are different remote sensing sensor which have different resolution, and images are further cut into sub-images with 200×200 pixel size from whole scene image, and each sub-image will be designated the training samples in manual interpretation way. Each group of sub-image and training samples are characterized by *RSIDC* Algorithm. The Classification algorithms database contains following 5 classical algorithms: Naïve Bayes, ID3 Tree, CART Tree, SVM and ANN. The entire classification algorithm classify all the sum-images (features maybe discretized for some algorithms) and obtain the corresponding classification accuracy, all the data collected and stored into database to construct *RDS*.

In order to verify the correctness of method proposed, this study has introduced a remote sensing images and cut into 10 sub-images as test data set, use 5 classification algorithm to classify and obtain the accuracy as Table 2:

TABLE 2 Test data sets classification accuracy

Data set	Classification Algorithms (%)				
	Naïve Bayes	ID3 Tree	CART Tree	SVM	ANN
1	70	87	84	86	86
2	83	80	80	86	83
3	82	85	87	89	87
4	75	77	76	85	83
5	90	93	93	93	93
6	99	98	98	99	99
7	85	87	88	90	89
8	78	89	81	81	82
9	60	85	92	90	90
10	89	90	90	94	90

The one marked with underlined in the table is the highest classification accuracy. Further, the elf-adaptive method of remote sensing image classification algorithms selection is utilized to select algorithm, which may have highest classification accuracy.

As can be seen from Table 3, there are 6 times in 10 selection process find the best algorithm, and 3 times select the 2nd algorithm, only 1 time select the 3rd algorithm. This proved that the proposed method has good algorithm selection ability.

TABLE 3 Selected algorithm and its rank

Data set	Selected algorithm	Highest accuracy in Table 2	Rank
1	ID3 Tree	ID3 Tree	1
2	SVM	SVM	1
3	SVM	SVM	1
4	ANN	SVM	2
5	ID3 Tree	ID3 Tree	1
6	Naïve Bayes	Naïve Bayes	1
7	ANN	SVM	2
8	CART Tree	ID3 Tree	3
9	SVM	CART Tree	2
10	SVM	SVM	1

5 Conclusions

To select appropriate classification algorithm for a remote sensing image classification is very important, but the similar data structure of remote sensing image data and statistical information hinders the traditional algorithm selection method, this research introduced data complexity evaluation into remote sensing image classification algorithm selection field, and proposed a self-adaptive selective method of remote sensing image classification algorithms (SSMRSICADCE). The method has two stages: in stage 1, users can input a large number of remote sensing image and their training samples, method can combine remote sensing image, data character, classification algorithm, classification accuracy together and save these information into Relationship data set RDS, the RDS is the knowledge of algorithm selection; in stage 2, A self-adaptive selective algorithm mechanism was proposed with the help of RDS. Experiments indicate that the algorithms selected by this method can achieve higher classification accuracy, which provides the recommendation for the selection of appropriate classification models to users.

Acknowledgments

This research was supported by the National Natural Science Foundation Youth Fund of China (41101384); Natural Science Foundation of Jilin Provincial Science & Technology Department (No. 20140101178JC).

References

[1] Shao Y, Lunetta R S 2012 *ISPRS Journal of Photogrammetry and Remote Sensing* **70** 78-87
[2] Pan X, Zhang S, Zhang. H, Na X, Li X 2010 *Computers & Geoscience* **36**(12) 1466-73
[3] Wolpert D H, Macready W G 1997 *IEEE Transactions on Evolutionary Computation* **1** 67-82
[4] Song Q, Wang G, Wang C 2012 *Pattern Recognition* **45** 2672–89
[5] Brodley C E 1993 Addressing the selective superiority problem: automatic algorithm/model class selection *Proceedings of the 10th International Conference on Machine Learning* **1** 17-24
[6] Gama J, Brazdil P B 1995 *Regress in Artificial Intelligence* **1** 189–200
[7] Brazdil P B, Soares C, Da Costa J P 2003 *Machine Learning* **50** 251–77
[8] Ho T K, Baird H S 1998 *Computer Vision and Image Understanding* **70** 101-10
[9] Ho T K, Basu M 2002 *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24**(3) 289–300
[10] Mollineda R A, Sánchez J S, Sotoca J M 2005 *Lecture Notes in Computer Science* **3523** 27–34
[11] Baumgartner R, Somorjai R L 2006 Data complexity assessment in undersampled classification *Pattern Recognition Letters* **27** 1383–89
[12] Luengo J, Herrer F 2010 *Fuzzy Sets and Systems* **16** 13–9

Authors

Xin Pan, born on January, 1978, Changchun City, Jilin Province, P.R. China

Current position, grades: an associate professor in the Changchun Institute of Technology, DSc
University studies: BSc degree in computer science from Changchun University of Technology (2000),
MSc degree from Changchun University of Technology (2005),
DSc degree in remote sensing and geographic information sciences from Chinese Academy of Sciences (2009)
Scientific interest: His research interest fields include Data mining, GIS, Remote sensing image analysis
Publications: more than 10 papers published in various journals
Experience: He has teaching experience of 8 years, has completed two scientific research projects

Hongbin Sun, born on May, 1969 , Changchun City, Jilin Province, P.R. China

University studies: BSc and MSc in Electrical Engineering from Huabei Electric University, China (1991, 1997),
Ph.D. in Electrical Engineering from Donghua University Shanghai, China (2007)
Scientific interest: His research interest fields include include complex network systems, Data mining
Publications: more than 15 papers published in various journals.
Experience: He has teaching experience of 10 years, has completed five scientific research projects.