

RESEARCH

Open Access

Research on semi-supervised multi-graph classification algorithm based on MR-MGSSL for sensor network



Yang Gang^{1*}, Zhang Na¹, Jin Tao¹, Wang Dawei¹, Kang Yin Zhu¹ and Gao Feng²

* Correspondence: caiyong@126.com

¹State Grid Shanxi Electric Power Research Institute, Taiyuan 030001, China

Full list of author information is available at the end of the article

Abstract

With the advent of the era of network information, the amount of data in network information is getting larger and larger, and the classification of data becomes particularly important. Current semi-supervised multi-map classification methods cannot quickly and accurately perform automatic classification and calculation of information. Therefore, this paper proposes an MR-MGSSL algorithm and applies it to the classification of semi-supervised multi-graph. By determining the basic idea and calculation framework of MR-MGSSL algorithm, the mining of optimal feature subsets in multi-graphs and the multi-graph vectorization performance time are taken as examples, and the proposed algorithm is compared with other semi-supervised multi-graph classification methods. The performance evaluation results show that compared with other classification calculation methods, MR-MGSSL algorithm has the advantages of low sensitivity to feature subgraph and short vectorization time. The method is used to extract and detect clouds in remote sensing images (GF-1 and GF-2).

Keywords: Sensors network, MR-MGSSL algorithm, Semi-supervised multi-graph, Feature subgraph

1 Introduction

With the rapid development of network information technology, the resulting network text and image information and other resources are increasing, although the number of such information resources is huge, that whose category attribute has been identified are only a small part, their dimensions are relatively high, and more training samples are needed to get an ideal classification calculation method [1]. Therefore, how to quickly and accurately classify the information resources is very important. Semi-supervised learning is used to obtain the learner with good performance, so as to realize the automatic classification of large-scale images. It not only can make up for the shortcomings of traditional information search, but also can classify information according to the similarity, so that information search becomes more simple and convenient [2].

At present, the methods based on semi-supervised multi-graph classification include mainly the decision tree method and Bayesian method [3]. These methods have high efficiency for the detection of text and image information resources but require the sample information to be detected and be marked according to its characteristics, and the unmarked information cannot be detected [4]. Some unsupervised multi-graph classification methods such as clustering can avoid the lack of decision tree method and Bayesian at information detection, but the detection rate is low and cannot be widely used in the classification of supervised multi-graphs. In addition, Wang Jing proposed a traffic classification method based on semi-supervised learning, which is based on the characteristic similarity of information resource to determine the mapping between clusters and traffic types through a small part of the labeled data in the clustering process, so as to realize the classification of information resources. This method not only reduces the requirement of the labeled data in the process of detection, but also ensures the accuracy of detection [5].

In this paper, a new MGSSL algorithm based on scoring function is proposed to solve the problem of semi-supervised small-scale multi-graph classification. In addition to solving the problem of semi-supervised large-scale multi-graph, on the basis of MGSSL algorithm and combined with MapReduce, we propose an MGSSL algorithm, which not only has high detection precision, but also can make up for the lack of other classification detection methods. It has far-reaching significance to the detection of largely existing text information and image resources. At the end of this article, the method was used to extract and detect clouds in remote sensing image (GF-1 and GF-2).

The specific contributions of this paper include:

- (1) An MR-MGSSL algorithm is proposed and applied to the classification of semi-supervised multi-graph.
- (2) The basic idea and calculation framework of MR-MGSSL algorithm are proposed.
- (3) By mining the optimal feature subsets in multiple graphs and taking the execution time of multi-graph vectorization as an example, the algorithm is compared with other semi-supervised multi-graph classification methods.
- (4) This method uses the semi-supervised calculation method to extract and detect remote sensing images.

The rest of this paper is organized as follows. Section 2 discusses related work, followed by the algorithm framework in Section 3. MR-MGSSL algorithm is discussed in Section 4. Section 5 shows the experiment, and Section 6 concludes the paper with a summary and future research directions.

2 Related work

2.1 Text classification algorithm based on semi-supervised learning

The text classification of semi-supervised learning refers to that the content of the text information is known, and it could automatically classify the text according to its characteristic similarity and the specified classification label. For the text that is not calibrated by the data label, it is mapped to the category which is calibrated by the data label according to its similarity, and mathematically, the process is text mapping. A

single text that has been determined can be associated with one or more texts according to its similarity [6]. Generally, the text classification of semi-supervised learning is realized by two parts, namely, training and classification process. The training process is based on the classifications that have been manually classified to construct the corresponding classifier by using a certain classification algorithm. The main task of the classification process is to classify the unclassified samples according to the classification calculation method and the classifier constructed after the training.

The text classification is widely used in the classification of semi-supervised multi-graphs [7]. It can not only reduce the detection range of the sample but also has a high detection accuracy, but there are still some deficiencies. The general semi-supervised multi-graph classification algorithm only focuses on the text information that is calibrated by the data label, and the attention to the unlabeled text is not high [8]. Although it can be correlated with its feature similarity to improve the accuracy of text classification calculation, often, to mark text is generally manually completed, and the cost is relatively high [9].

Image recognition is the use of computers to process, analyze, and comprehend images to identify various patterns of objects and objects [10]. It focuses on the study of the classification and description of various images. The purpose of image recognition is to allow the computer to automatically process the corresponding image information without the need for natural human intervention to accomplish the tasks of image recognition and classification [11]. The basic task of image recognition is to analyze and process the original input image so that one or more objects of interest in the image may be extracted [12].

In life, people will inadvertently complete the process of image recognition, but letting computers implement automatic image recognition has been a difficult problem for a long time in the past [13]. The main difficulties are as follows: First, the algorithm itself is not mature enough to complete the task of identifying complex images. In some classical image recognition frameworks, there are many steps including image pre-processing, target detection and segmentation, feature extraction, and classifier design [14]. Second, the limitations of the program operating environment mainly refer to some restrictions on computer hardware [15]. In recent years, digital image processing technology has continued to develop, and pattern recognition theory has been continuously introduced. The computer's CPU speed and memory capacity have also increased by several orders of magnitude, and the above two issues have gradually eased [16]. Image processing and recognition technology, with its extensive application research needs, will surely gain more attention from domestic and foreign scholars in the near future [17].

Since the twenty-first century, image processing and recognition have been applied more and more in social networks, medical equipment, geographic information systems, information security, office automation systems, industrial automation, traffic control, postal systems, satellite photo transmission, and analysis [18]. In recent years, the computer technology, image processing technology, artificial intelligence, pattern recognition theory, etc. have become increasingly mature, and the image processing and recognition technology has been rapidly developed [19]. People are increasingly aware that image processing and recognition technologies have become inseparable from our daily lives [20].

2.2 Improved semi-supervised learning algorithm

There are some shortcomings in the existing classification methods for semi-supervised learning. The main improved methods of semi-supervised learning are as follows:

- (1) Dynamic clustering method: The dynamic clustering method is a process of marking a small part of the text as a training sample, and then taking them as a clustering center, through similar relevance to gather other text information. Through a small part of mixed text with and without the labels to build learning files, build text learning calculation classifier, further to finish text classification [21].
- (2) Multi-graph collaborative training method: The multi-graph collaborative training method is a kind of online video semi-supervised classification method based on multi-graph collaborative training. The specific process is to first select the representative two features of the text and the sight on the view, and then take the view feature vector as a network video classifier, thus building a classification calculation model. In order to obtain the classification prediction results, each view is propagated by linear domain propagation method. Use the co-training strategy to select unlabeled text between different views to update the classification calculator at any time, with a relatively high classification accuracy [22].
- (3) Traffic classification: traffic classification method is through the use of a small amount of tag data to support the clustering process to determine the mapping relationship between the cluster and traffic type, and ultimately to achieve the traffic classification of the application layer [23]. The traffic classification method can excavate the unknown area, and its coverage is extensive, which can make up for the shortcomings of other methods to the semi-supervised classification in the case of no label and improve the accuracy of the classification detection, and the requirement to the label of the data feature is relatively low [24].
- (4) Integrated direct push method [25]: The process of direct push integration method is first to form several stochastic subspaces within the internal of information resources, and then distinct semi-supervised space based on the subspace, and construct a neighborhood graph and train a classifier for each discriminant subspace, finally, to fuse these classifiers through vote [26]. Some experimental studies show that the integrated direct push method not only is more accurate, but also has more accurate selection of parameters and can better classify the information resources. In addition, it has an intuitive multiple map building strategy and could be coupled with other algorithms based on semi-supervised multi-graphs [27].

3 Algorithm framework

MR-MGSSL is a multi-graph classification algorithm for centralized tagged and unlabeled among the semi-supervised multi-graph classification algorithms [28]. The basic idea is to select some characteristic subgraphs from some of the multi-graph datasets. According to the characteristic of subgraphs, the multi-graphs are expressed through vectors and then classify the models by the existing semi-supervised learning methods. And it is summarized as two multi-map feature subgraph measurement model with label and without label [29].

3.1 Characteristic subgraph measurement

First, the establishment of the characteristic subgraph selection model is as follows:

According to the existing multi-map dataset $NF = \{NF_1, NF_2, \dots, NF_n\}$, map set $F_y = \{F | F \in NF_i, NF_i \in N_y\}$, subgraph collection of $F_y, YF = \{yf | yf \subseteq F, F \in F_y\}$, and Feature subgraph $R = \{r_1, \dots, r_n\} \subseteq YF$. The optimal feature subset is one of the most valuable feature subgraphs. The feature subgraph selection model is as follows:

$$RY = \arg \max Y(R), s.t. |R| = n \tag{1}$$

The value of feature subgraph R is evaluated through $Y(R)$. The larger the number of $Y(R)$, the higher the value of feature subgraph. In addition, feature subgraph should satisfy, respectively, must-link, cannot-link, and separation characteristics of the collection layer and map layer [30].

The value of feature subgraph R , $Y(R)$ is generally defined as [31]:

$$\begin{aligned} Y(R) = & \frac{1}{2B} \sum_{t=1}^n \sum_{xixj=-1} \left((y_i^{Et})^{Ny} - (y_j^{Et})^{Ny} \right)^2 \\ & - \frac{1}{2C} \sum_{t=1}^n \sum_{xixj=1} \left((y_i^{Et})^{Ny} - (y_j^{Et})^{Ny} \right)^2 \\ & \frac{1}{2|Ny^v|^2} \sum_{t=1, \forall NEj, NEi \in Ny^v}^n \sum \left((y_i^{Et})^{Ny} - (y_j^{Et})^{Ny} \right)^2 \\ & - \frac{1}{2|Ny^-|^2} \sum_{t=1, \forall Ej, Ei \in Ny^-}^n \sum \left((y_i^{Et})^{Ey} - (y_j^{Et})^{Ey} \right)^2 \\ & + \frac{1}{2|Ny^+|^2} \sum_{t=1, \forall Ej, Ei \in Ny^+}^n \sum \left((y_i^{Et})^{Ey} - (y_j^{Et})^{Ey} \right)^2 \end{aligned} \tag{2}$$

In which, $B = \sum_{xixj = -1} 1$, $C = \sum_{xixj = 1} 1$.

Through the value definition of feature subgraph, the problem of solving feature subgraph is transformed into solving the problem of n optimal feature subgraphs, and the auxiliary matrix is constructed as follows [32]:

$U_{Ny} = [u_{ij}^{Ny}]^{|Ny| \times |Ny|}$, $U_{Ey} = [u_{ij}^{Ey}]^{|Ey| \times |Ey|}$. u_{ij}^{Ny} and u_{ij}^{Ey} are defined as follows:

$$y_{ij}^{Ny} = \begin{cases} \frac{1}{B} & xixj = -1 \\ -\frac{1}{C} & xixj = 1 \\ \frac{1}{|Ny^v|^2} & Ei, Ej \in Ny^v \\ 0 & other \end{cases} \tag{3}$$

$$u_{ij}^{Ey} = \begin{cases} -\frac{1}{|Ny^-|^2} & EiEj \in Ny^- \\ \frac{1}{|Ny^+|^2} & EiEj \in Ny^+ \\ 0 & other \end{cases} \tag{4}$$

$$\begin{aligned}
 Y(R) &= Y(R)^{Ny} + Y(R)^{Ey} \\
 &= \frac{1}{2} \sum_{t=1}^n \sum_{xixj} \left((y_i^{Et})^{Ny} - (y_j^{Et})^{Ny} \right)^2 u_{ij}^{Ny} \\
 &\quad + \frac{1}{2} \sum_{t=1}^n \sum_{EiEj} \left((y_i^{Et})^{Ey} - (y_j^{Et})^{Ey} \right)^2 u_{ij}^{Ey}
 \end{aligned} \tag{5}$$

It could be obtained after resolving [33]:

$$\begin{aligned}
 Y(\text{Rea})^{Ny} &= \frac{1}{2} \sum_{t=1}^n \sum_{xixj} \left((y_i^{Et})^{Ny} - (y_j^{Et})^{Ny} \right)^2 U_{ij}^{Ny} \\
 &= \sum_{t=1}^n \sum_{xixj} \left(\left((y_i^{Et})^{Ny} \right)^2 U_{ij}^{Ny} - (y_i^{Et})^{Ny} (y_j^{Et})^{Ny} U_{ij}^{Ny} \right) \\
 &= \sum_{t=1}^n \left(\left(r_{Et}^{Ny} \right)^S C_{Ny} r_{Et}^{Ny} - \left(r_{Et}^{Ny} \right)^S U_{Ny} r_{Et}^{Ny} \right) \\
 &= \sum_{t=1}^n \left(r_{Et}^{Ny} \right)^S M_{Ny} r_{Et}^{Ny}
 \end{aligned} \tag{6}$$

In which, C_{Ny} is a diagonal matrix consisting of diagonal elements of $c_{ii}^{Ny} = \sum_{j=1}^{|Ny|} u_{ij}^{Ny}$. In $M_{Ny} = C_{Ny} - U_{Ny}$, $r_E^{Ny} = [r_E^{Ny1}, r_E^{Ny2}, \dots, r_E^{NE|Ny|}]^S$ indicates whether a multi-graph contains multiple sub-maps. When the weight is 1, the answer is yes. So [34]:

$$\begin{aligned}
 Y(\text{Rea})^{Ny} &= \frac{1}{2} \sum_{t=1}^n \sum_{xixj} \left((y_i^{Et})^{Ny} - (y_j^{Et})^{Ny} \right)^2 U_{ij}^{Ny} \\
 &= \sum_{t=1}^n \left(\left(r_{Et}^{Ny} \right)^S C_{Ny} r_{Et}^{Ny} - \left(r_{Et}^{Ny} \right)^S U_{Ny} r_{Et}^{Ny} \right) = \sum_{t=1}^n \left(r_{Et}^{Ny} \right)^S M_{Ny} r_{Et}^{Ny}
 \end{aligned} \tag{7}$$

It could be obtained by combining the above formula:

$$Y(\text{Rea}) = Y(R)^{Ny} + Y(R)^{Ey} = \sum_{t=1}^n \left(\left(r_{Et}^{Ny} \right)^S M_{Ny} r_{Et}^{Ny} + \left(r_{Et}^{Ey} \right)^S M_{Ey} r_{Et}^{Ey} \right) = \sum_{t=1}^n (r_{Et})^S M r_{Et} \tag{8}$$

Thus, the value of a single characteristic subgraph can be expressed as:

$$Y(E_t) = (r_{Et})^S M r_{Et} \tag{9}$$

So:

$$Y(R) = \sum_{t=1}^n Y(E_t) \tag{10}$$

$$R(Y) = \arg \max \sum_{t=1}^n Y(E_t) \tag{11}$$

3.2 Characteristic subgraph selection algorithm: MGSSL algorithm

MGSSC is a general method in the feature subgraph selection algorithm. The main calculation process is to select the feature subgraphs with weight 1 and weight 0 in the information resource, and then use this as the vector starting node, and conduct the

depth search in the information resource until the search is complete. In this paper, MGSSL algorithm is described in detail [35]:

MGSSL algorithm is to select the optimal feature subset R based on MGSSC algorithm, and then transform multi-map in the training dataset NY into a vector, and finally according to the traditional semi-supervised classification calculation method finish classification calculation [36].

The main calculation steps of MGSSL include input part and output part, as follows:

1. Input part: Train the multi-graph information set Ny , and then the obtained multi-graph dataset S is tested to finally obtain the support degree threshold r and the optimal characteristic subgraph number n of the multi-frequency occurrence subgraph [37].
2. Output part: The output part is mainly to test any one of the multi-map class tags contained in the multi-map dataset, mainly divided into the training phase and test phase.
 - (a) Training stage: First select the optimal feature subgraph $R = MGSSC(Ny, \gamma, n)$, and then represent multi-graph data in Ny with vector X , and finally get the classification model F .
 - (b) Test phase: The test phase is to transform the multi-map in S into the vector Xt , and then according to the model F to predict class label of Xt , and finally predict all the class labels.

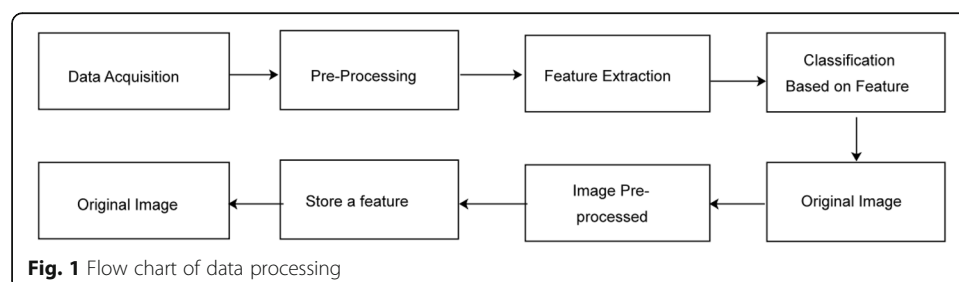
The data processing flow chart is shown in Fig. 1.

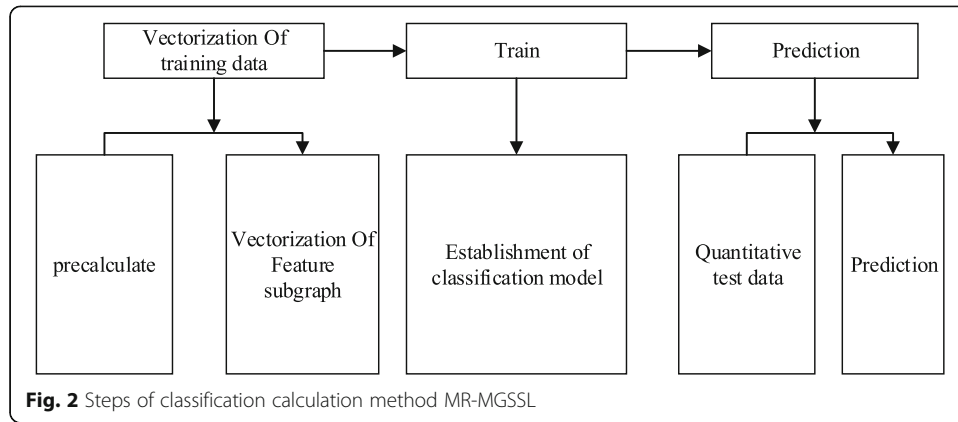
4 MR-MGSSL algorithm

MGSSL algorithm usually conducts classification by centralized processing, and it cannot directly calculate when dealing with semi-supervised large-scale multi-graph classification. In view of the emergence of such shortcomings, MR-MGSSL algorithm combining the MapReduce framework and MGSSL is proposed to conduct semi-supervised large-scale multi-map classification.

4.1 MR-MGSSL semi-supervised large-scale multi-graph classification algorithm

In the semi-supervised large-scale multi-map classification, $MR - MGSSL$ is generally divided into three steps shown below (Fig. 2).





4.2 Training data vectorization

The existing MGSSL algorithms cannot be directly applied to the semi-supervised multi-graph classification. We must first select the feature subgraphs, and transform the multi-graph data into eigenvectors, and then use the MGSSL algorithms to find the rules from the transformed eigenvectors. Construct subgraph model to conduct prediction of the calculation.

On the basis of the MR-MGSSL algorithm, an algorithm is proposed in the paper to select the optimal feature subset.

At present, the selection of feature subsets is determined by the single record of the scoring function. Therefore, in determining the semi-supervised multi-graph classification problem, we need to first determine the scores of the single frequent subgraph and then select N optimal characteristic subgraphs with the largest score.

In general, during the selection process of the feature subset, it first needs to select a subgraph appearing in multi-frequency and calculate its value, and the calculation of the score re_i needs to first understand the matrix M_{Ny} , M_{Ey} , r_{Ny} , and r_{Ey} . In a text message, M_{Ny} and M_{Ey} of the multi-frequency subgraph re_i is the same, so it is only necessary to compute the subgraphs included in the Ny and Et sum. And then calculate the value of each feature subgraph according to the formula $Y(re_i) = r_{Ny}^s L_{Ny} r_{Ny} + r_{Et}^s L_{Et} r_{Et}$. Finally, by calculating the partial optimal characteristic subgraphs, the value of the characteristic subgraph of all the text information is calculated and expressed by the vector.

Pre-calculate the matrix of M_{Ny} and M_{Ey} and the value of the multi-frequency characteristic subgraph.

4.2.1 Pre-calculation method

Calculate the matrix of M_{Ny} and M_{Ny} , id of text information and the list *Bag-list* and *Gra-list* of Et . The multi-graph is represented by the function record of the graph selection stage. In the multi-graph with labels, when the class label in the graph is positive, it is expressed as *input* $\langle 1 \cdot 1 \rangle$ and $\langle 4, |graph| \rangle$ (2~3line): if the output is negative, it is expressed as *output* $\langle 2 \cdot 1 \rangle$ and $\langle 5, |graph| \rangle$ (4~5line). The unlabeled multi-graph is expressed as *output* $\langle 3 \cdot 1 \rangle$ and $\langle 6, |graph| \rangle$ (6line). The role of keys 1 to 8 is to produce a synergistic effect on the calculation of $|Ny^+|$, $|Ny^-|$, $|Ny^v|$, $|Et^+|$, $|Et^-|$, $|Et^v|$, *Bag-list*, and *Gra-list*. And then, according to the above calculated key

value to calculate $|Ny^+|$, $|Ny^-|$, $|Ny^v|$, $|Et^+|$, $|Et^-|$, $|Et^v|$, *Bag-list*, and *Gra-list* in line 12 to 14. Finally, in the calculation of these key values, M_{Ny} and M_{Ey} is calculated.

Use MR-MGSSL algorithm to pre-calculate.

MR-MGSSL algorithm:

Input: train the text information data set Ny .

Output: M_{Ny} , M_{Ey} , *Bag-list* and *Gra-list*

1. //map stage
2. *Map(Offset, MG)*
3. If MG is positive
4. *emit* $\langle 1, 1 \rangle$ and $\langle 4, |graph| \rangle$
5. If MG is negative
6. *emit* $\langle 2, 1 \rangle$ and $\langle 5, |graph| \rangle$
7. *emit* $\langle 2, 1 \rangle$ and $\langle 5, |graph| \rangle$
8. If *emit* $\langle 3, 1 \rangle$ and $\langle 6, |graph| \rangle$ // no label
9. *emit* $\langle 7, (bagld, bag-class-label) \rangle$
10. *emit* $\langle 8, (gld_1, gld_2, \dots, gld_{|graph|}; bag-class-label) \rangle$;
11. //reduce stage
12. *Reduce(key, value)*
13. Calculate $|Ny^+|$, $|Ny^-|$, $|Ny^v|$, $|Et^+|$, $|Et^-|$, $|Et^v|$
14. Calculate *List* $(bagld, bag-class-label), emit(List(bagld))$
15. Calculate *List* $(bagld, bag-class-label), emit(List(gld))$
16. *cleanup*
17. $B = |Ny^+| \times |Ny^-|$, $C = |Ny^+|^2 \times |Ny^-|^2$
18. Calculate U_{Ny} , U_{Et} , U_{Ny} , U_{Et} , C_{Ny} , C_{Et} , F_{Ny} , F_{Et}
19. *emit* F_{Ny} and F_{Et}

In the prediction method, it is necessary to obtain the multi-graph and the super multi-map first, and then determine whether the frequency of the multi-frequency sub-graph has been calculated. If it is calculated, it is output directly according to the calculation step; otherwise, it needs to be judged again until the output is calculated. Finally, the calculated frequency is compared with its threshold, and the multi-graph and super multi-graph of multi-frequency subgraph are output.

The selection of the optimal feature subgraph and the value calculation: the characteristic subgraph refers to the multi-frequency subgraph that occurred with the highest frequency in the text information, and the selection of multi-frequency feature sub-map first needs to calculate the frequency of the subgraph that occurred in the text information and then according to the frequency, determine the multi-map and super

multi-map of multi-frequency subgraph. In general, the text information is divided into pieces, and then its frequency in the multi-frequency subgraph has been determined; when determined, output, if not sure, needs to re-calculate the frequency subgraph, until it is determined and then output. Finally, the frequency of all the text information is obtained according to the known output frequency of each block, and then the optimal feature subset existing in the whole text information is determined according to the comparison with the maximum and minimum thresholds.

In general, the selection of the optimal feature subgraph mainly uses the MR-MGSSL algorithm.

MR -MGSSL algorithm

```

Input: ordinal candidate multi-frequency subgraphs, multi-graphs and super-multi-graph data sets
Output: multi-frequency subgraphs and its  $NE_u^1$  and  $NYE_u^1$ 
1.Read SCS from the cache, read the super multi map from DFS://map stage
2. listgraphPartition ;
3. Map(Offset, MG >)
4.Add all the maps in MG to graphPartition
5. cleanup
6. for  $u \in SCS$  do
7.If the frequency in graphPartition has been calculated
8. emit ( $\langle u, f_u \rangle$ )
9.Otherwise decode  $u$ 
10.Calculate the frequency  $f_u$  of graphPartition
11. emit ( $\langle u, (f_u, NE_u, NRE_u) \rangle$ )
12.//reduce stage
13. Reduce ( $\langle u, list(f_u), (f_u, NE_u, NRE_u) \rangle$ )
14.  $NE_u^1 = \Phi, NYE_u^1 = \Phi$ ;
15. for  $value \in list(f_u), (f_u, NE_u, NRE_u)$  do
16. if  $value$  is  $f_u$ 
17.Find list( $NE_u^2$ ) and list( $NYE_u^2$ ) from SG
18.  $NE_u^1 = NE_u^1 \cup list(NE_u^2), NYE_u^1 = NYE_u^1 \cup list(NYE_u^2)$ 
19.Otherwise  $NE_u^1 = NE_u^1 \cup list(NE_u), NYE_u^1 = NYE_u^1 \cup list(NYE_u)$ 
20.Calculate the sum of  $f_u$ 
21. if  $\sum f_u \geq \delta$ 
22. emit( $\langle u, NE_u^1, NRE_u^1 \rangle$ )

```

The score N of the multi frequency subgraph is calculated by the following steps:

```

Input:  $M_{N_y}, M_{N_y}, Bag-list$  and  $Gra-list$ , multi frequency subgraph and its  $NE_u^1$  and  $NYE_u^1$ 
Output: Scores of all the multi frequency subgraphs  $N(u)$ 
1.Read the matrix of  $M_{N_y}, M_{N_y}, Bag-list$  and  $Gra-list$ 
2. Map( $\langle u, (NE_u^1, NRE_u^1) \rangle$ )
3.Calculate  $N(u) = F_{E_s}^s F_{E_s^1(u)} + F_{N_y}^s F_{N_y^1(u)}$ ;
4. emit( $\langle u, (N(u), NRE_u^1) \rangle$ )

```

The method of solving the optimal feature subgraphs is usually with a small see big. The basic idea is to output the multi-frequency subgraphs of each part first, and then

obtain the characteristic subgraphs of the partial frequency subgraphs, and finally obtain the optimal characteristic subgraph of the whole text information. The specific calculation method is as follows.

Input: information of optimal characteristic subgraph
 $H = list(u, N(u), NRE_u^1), N_y = \{NE_1, \dots, NE_{NY}\}$ and $E_y = \{E_1, \dots, E_{NY}\}$
 Output: Optimal characteristic subgraph H and NE , Feature vector set U based on H .

1. $U = \varphi$
2. When $NE_1 \in NG_y$, continue
3. Zero dimensional vector of H is represented with θ
4. $u_h \in H^1$, continue
5. When $NE_1 \in YNE_{u_h}^1$, continue
6. Set 1 as the weight of θ
7. $U = U \cup \{\theta\}$;

Map vectorization generally through the following steps to test.

Input: test multi map $N_y = \{NE_1, \dots, NE\}$.
 Output: Test the corresponding matrix of multi map,

1. $U_s = \varphi$;
2. When $NE_i \in NE_s$, continue
3. Set the corresponding vector of NE_i as u_i
4. $u_i = EU(HE, NE)$
5. $U_s = U_s \cup \{u_i\}$.

Map vectorization is realized by the vector of each block multi-frequency subgraph, namely in the first end part of the above input and output for each feature sub-block multi-frequency subgraph; then, at the reduced end, get Bag – list and Gra – list, finally obtain all the sub-images of text information, and conduct vectorization of the trained multi-map.

5 Experiment

Evaluate the performance of the MR-MGSSL algorithm by comparing it with the algorithm baseline and the MGSSL+M algorithm, which is mainly based on the two indicators of the mining time and the quantization time.

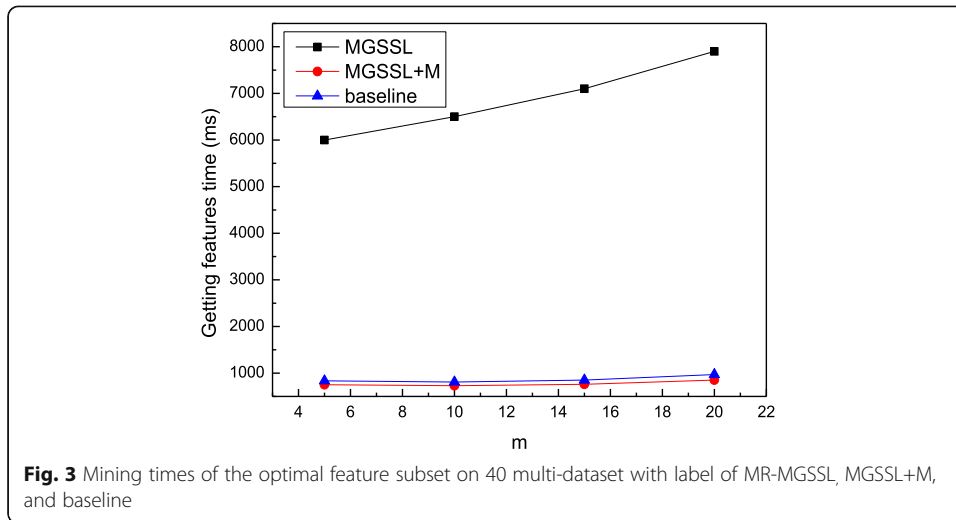
5.1 Evaluation of mining time

The following figure shows the mining times of the optimal feature subset on 40 multi-datasets with label of MR-MGSSL, MGSSL+M, and baseline (Fig. 3).

By 40 multi-datasets with label DBLP, we can see that when the number of multi-feature subset and the threshold are the same, the MR-MGSSL algorithm needs more time than the algorithm baseline and MGSSL+M in the same conditions, the baseline algorithm only needs to dig out the feature map, and MR-MGSSL algorithm not only needs to dig out feature subgraph algorithm but also still need to dig out the characteristic sub-map of E_t . And the mining time increases with the increase of text information.

5.2 Vectorization time performance evaluation

The following figure shows the vectorization times of the optimal feature subset on 40 multi-map dataset with label of MR-MGSSL, MGSSL+M, and baseline.



It is clearly evident from Fig. 4 that vectorization time of the *MR-MGSSL* algorithm is shorter than the other methods, t in the process it only needs the vectorization of characteristic subgraph, so as to realize vectorization of the entire information text. The other two methods also need to test the similarity of all the data in the text information. In addition, when the characteristic subgraph mining out from text information is more, the other two methods need a longer time to multi-map vectorization. In general, the sensitivity of the other two methods of sub-images is higher than that of the *MR-MGSSL* algorithm.

5.3 Algorithm application

The regional growth method and support vector machine method are selected as references, and GF-1 and GF-2 remote sensing images are selected to perform cloud detection experiments in the image. The experimental data are shown in Table 1, and there are two aspects of visual effects and detection accuracy. The region growth method and support vector machine method are compared with the method in this

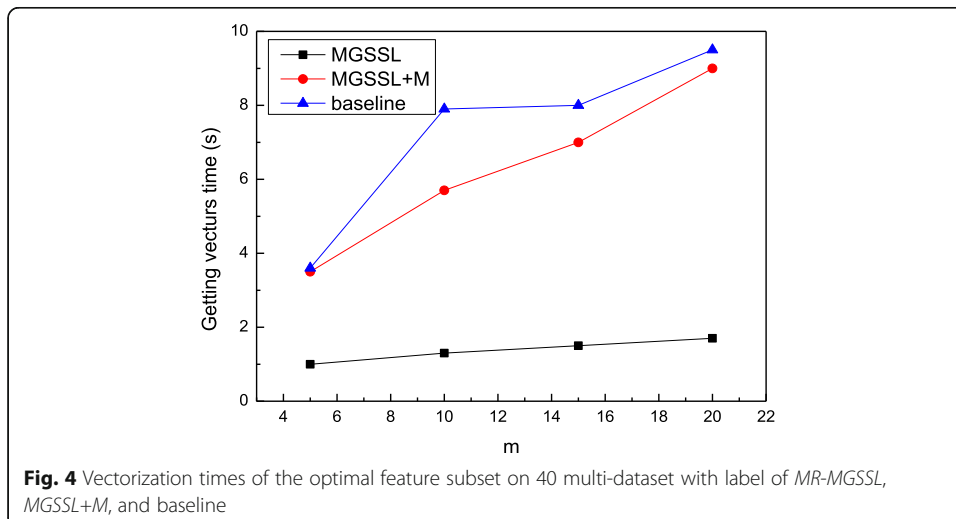


Table 1 Experimental data

Number of remote sensing image	Satellite	Image size	Surface type	Cloud type
1601	GF-1	2000 × 2000	Mountains and towns	Sparse cloud
1602	GF-2	3000 × 3000	Farmland	Sparse and dense clouds

paper. The experimental results are shown in Fig. 4. The red part in the figure is the detected cloud area.

Figure 4 compares the experimental results of the region growing method, support vector machine method, and the method in this paper. In the figure, the orange circle is the missed cloud area, and the blue circle is the missed cloud area. It can be seen that the visual effect of the method in this paper is the best. In the first picture (1601), there is a small amount of thin cloud that missed detection in the support vector machine method (Fig. 5). In the second (1602) image, there are a large number of thin clouds that missed detections in the area growing algorithm. This proves that the method proposed in this paper effectively improves the accuracy of cloud detection.

In the experiment, the actual cloud area was manually drawn. The accuracy of cloud detection was evaluated using three indicators: accuracy, recall, and error. The calculation formula is

$$PR = \frac{TC}{FA} \quad (12)$$

$$RR = \frac{TC}{TA} \quad (13)$$

$$ER = \frac{TF + FT}{NA} \quad (14)$$

in which, PR is the precision rate, TC is the number of true cloud pixels that can be accurately identified, FA is the total number of cloud pixels identified, RR is the recall rate. TA is the number of true cloud pixels. ER is the error rate, TF is the number of pixels that have been misjudged as non-cloud by true cloud, FT is the number of pixels that have been misjudged by cloud as non-cloud, and NA is the total number of pixels. The final results are shown in Table 2.

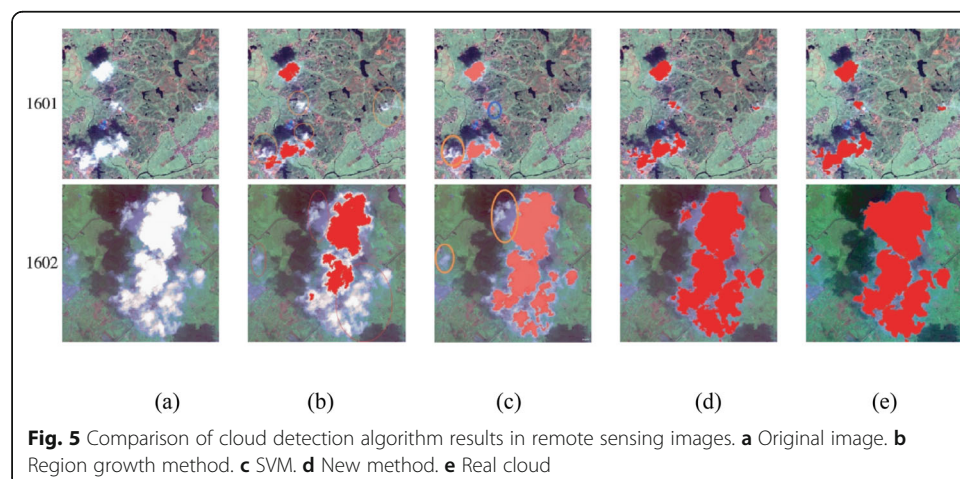


Table 2 Comparison of accuracy indicators of different cloud detection algorithms (%)

Number of remote sensing image	Accuracy index	Region growth method	SVM	MR-MGSSL
1601 (GF-1)	PR	99.22	89.74	99.09
	RR	49.92	72.71	91.13
	ER	2.54	2.17	1.34
1602 (GF-2)	PR	98.43	98.73	99.21
	RR	53.37	80.62	89.54
	ER	10.55	6.27	2.91

Quantitative analysis of cloud detection results is in the figure with Table 2. The area growth algorithm is affected by the selection of seeds and similar region determination criteria, and it is easy to miss thin clouds at the edges, which leads to fewer accurately identified true cloud pixels TC and fewer total cloud pixels FA , and the true cloud accuracy rate is both above 90%, but the recall rate is low. The results of the support vector machine method are affected by the selection and training of the samples. Although the recall rate is improved compared to the area growth algorithm, the overall error rate is higher. In the first picture, the accuracy rate of the area growth method is as high as 99.22%, but the recall rate is only 49.92, because there are large areas of cloud edge misses and thin cloud misses; the support vector machine method has misjudged the house as cloud situation. The algorithm in this paper has obvious superiority in recall rate and error rate. The recall rate is around 90%, the highest error rate is 6.03%, and the lowest error rate is only 0.89%.

6 Conclusion

Based on the analysis of the existing problems of semi-supervised multi-map classification the $MR-MGSSL$ algorithm is proposed, the calculation steps of each factor in the semi-supervised classification algorithm are determined and the evaluation system is established. Based on the comparison of the proposed algorithm and other classification methods on mining time and vectorization time, the proposed algorithm has a longer mining time of the optimal feature subgraph and the time increases with the increase of text information; on the other hand, the proposed algorithm has a shorter time of the subgraph vectorization and has positive correlation relationship with the number of the optimal feature subgraph and lower sensitivity to the number of sub-images. It affirmed the feasibility of $MR-MGSSL$ algorithm in semi-supervised multi-map classification, so as to reduce the cost of communication and improve the efficiency of the algorithm.

Funding

Supported by the science and technology project of the State Grid Corporation of China, research on intelligent infrared image diagnosis of substation equipment (520530190003).

Availability of data and materials

The datasets used and/or analyzed during the current study are available from the corresponding author on reasonable request.

Competing interests

We have no competing interests.

Author details

¹State Grid Shanxi Electric Power Research Institute, Taiyuan 030001, China. ²Modest Moistens & Harmonious Technology Co. Ltd, Beijing 100193, China.

Received: 23 August 2019 Accepted: 3 June 2020

Published online: 22 June 2020

References

1. W.J. Zheng, L.I. Lei, S.O. Science, Research on combined semi-supervised SVM cluster kernel algorithm based on graph. *Computer Technology & Development* (2014)
2. L. Jia, Semi-supervised multi-class classification algorithm based on local learning. *J Comput Appl* **32**(12), 3308–3310 (2012)
3. J. Lv, Semi-supervised multi-class classification algorithm based on local learning// information engineering and applications. Springer London (2012)
4. X.Q. Wang, Research on multi-view semi-supervised learning algorithm based on co-learning// international conference on machine learning and cybernetics. *IEEE* **20**(6), 1276–1280 (2016)
5. Y. Zhao, G. J. Wang, A multi-classification algorithm of semi-supervised support vector data description based on pairwise constraints// proceedings of 2013 Chinese intelligent automation conference. Springer Berlin Heidelberg **20**(5), 531–538 (2013).
6. D.Q. Xue, The research on semi-supervised support vector data description multi-classification algorithm. *Adv. Mater. Res.* **26**(5), 1115–1120 (2011)
7. S. Ding, H. Jia, L. Zhang, Research of semi-supervised spectral clustering algorithm based on pairwise constraints. *Neural Comput. Applic.* **24**(1), 211–219 (2014)
8. K. Mardia, J. Kent, J. Bibby, *Multivariate analysis*. Academic Press, San Diego, CA, 300–325 (1980)
9. M. Grbovic, C. Dance, S. Vucetic, Sparse principal component analysis with constraints //Proc. of 26th AAAI , 935–941(2012).
10. W. Yue, K.C. Ho, Unified near-field and far-field localization for AOA and hybrid AOA-TDOA positionings. *IEEE Trans. Wirel. Commun.* **17**(11), 1242–1254 (2018)
11. Z. Yi, Y. Wu, J. Yan, H. Wang, 3D inversion of full gravity gradient tensor data in spherical coordinate system using local north-oriented frame. *Earth Planets Space* **70**(12), 58–58 (2018)
12. J. Wang, X.J. Cheng, J.Q. Liu, Y.J. Wen, A enhanced algorithm based on RSSI and quasi Newton method for the node localization in wireless sensor networks. *Comput. Knowl. Technol.* **12**(8), 222–225 (2016)
13. G.Q. Zhou, L.J. YANG, Z. Liu, Analysis on the influence of base station layout on the fuzzy region distribution and positioning accuracy based on TDOA positioning. *J. Nav. Univ. Eng.* **29**(11), 96–101 (2017)
14. Y. Tuo, S. Wang, Wang, reliability-based robust online constructive fuzzy positioning control of a turret-moored floating production storage and offloading vessel. *IEEE Access.* **6**(8), 36019–36030 (2018)
15. Y. Tuo, Y. Wang, S. Wang, Reliability-based robust online constructive fuzzy positioning control of a turret-moored floating production storage and offloading vessel. *IEEE Access.* **6**(10), 36019–36030 (2018)
16. S. Song, W. Zhang, P. Han, D. Zou, Sliding window method for vehicles moving on a long track. *Veh. Syst. Dyn.* **56**(1), 113–127 (2018)
17. A.N.Z. Rashed, A. Mohammed, H.A. Sharshar, A.M. El-Eraki, Fast routing algorithm in optical multistage interconnection networks using fast window method. *Int J Advanced Res Electron Commun Eng* **6**(1), 37–43 (2017)
18. J. Kasza, R. Hemming, R. Hooper, J. Matthews, A. Forbes, Impact of non-uniform correlation structure on sample size and power in multiple-period cluster randomised trials. *Stat. Methods Med. Res.* **28**(3), 703–716 (2019)
19. I. Hanasaki, C. Hosokawa, Non-uniform stochastic dynamics of nanoparticle clusters at a solid-liquid interface induced by laser trapping. *Japanese Journal of Applied Physics* **58**(SD), 07 (2019)
20. S. M. M. Gilani, T. Hong, W. Jin, G. Zhao, H. M. Heang, C. Xu, Mobility management in IEEE 802.11 WLAN using SDN/NFV technologies. *EURASIP J. Wirel. Commun. Netw* **67**(12), 56–62 (2017).
21. K. Nahida, C. Yin, Y. Hu, Z.A. Arain, C. Pan, I. Khan, Y. Zhang, G.M.S. Rahman, Handover based on AP load in software defined Wi-fi systems. *J. Commun. Netw.* **19**(6), 596–604 (2017)
22. T. Zahid, X. Hei, W. Cheng, A. Ahmad, P. Maruf, On the tradeoff between performance and programmability for software defined WiFi networks. *Wirel. Commun. Mob. Comput* **35**–41 (2018).
23. L. Li, G. Oikonomou, M. Beach, R. Nejabati, D. Simeonidou, in Paper presented at IEEE International Conference on Communications. An SDN agent-enabled rate adaptation framework for WLAN (Shanghai, 2019).
24. K. Kostal, R. Bencel, M. Ries, P. Truchly, I. Kotuliak, High performance SDN WLAN architecture. *Sensors* **19**(8), 18–25(2019).
25. E. Coronado, S.N. Khan, R. Riggio, 5G-EmPOWER: A software-defined networking platform for 5G radio access networks. *IEEE Trans. Netw. Serv. Manag.* **16**(2), 715–728 (2019)
26. E. Coronado, E.T. Garriga, J. Villalon, A. Garrido, L. Goratti, R. Riggio, SDN@play: Software-defined multicasting in enterprise WLANs. *IEEE Commun* **57**(7), 85–91 (2019)
27. A. Sen, K. M. Sivalingam, Testbed evaluation of a seamless handover mechanism for an SDN-based enterprise WLAN. *Sadhana Acad* **44**(12), 243 (2019).
28. B. Dezfouli, V. Esmaealzadeh, J. Sheth, M. Radi, A review of software-defined WLANs: Architectures and central control mechanisms. *IEEE Commun* **21**(1), 431–463 (2019)
29. S. Zhu, Z. Sun, Y. Lu, L. Zhang, Y. Wei, G. Min, Centralized QoS routing using network calculus for SDN-based streaming media networks. *IEEE Access* **7**(12), 146566–146576 (2019)
30. X. Zhong, L. Zhang, Y. Wei, Dynamic load-balancing vertical control for large-scale software-defined internet of things. *IEEE Access* **7**(12), 140769–140780 (2019)
31. P. Dong, K. Gao, J. Xie, W. Tang, N. Xiong, A. Vasilakos, Receiver-side TCP countermeasure in cellular networks. *Sensors* **19**(12), 27–32 (2019)
32. Z. Kuang, G. Liu, G. Li, X. Deng, Energy efficient resource allocation algorithm in energy harvesting-based D2D heterogeneous networks. *IEEE Internet Things J.* **6**(1), 557–567 (2019)
33. Z.H. Huang, X. Xu, H.H. Zhu, M.C. Zhou, An efficient group recommendation model with multiattention-based neural networks. *IEEE Transactions on Neural Networks and Learning Systems* (2020)
34. R. Jiang, M. Y. Shi, W. Zhou, A privacy security risk analysis method for medical big data in urban computing. *IEEE Access* **7**(12), 143841–143854(2019).

35. Y. Sun, C. Xu, G.F. Li, W.F. Xu, J.Y. Kong, D. Jiang, B. Tao, D.S. Chen, Intelligent Human Computer Interaction Based on Non Redundant EMG Signal Alexandria Engineering Journal (2020)
36. W. Wei, H. Song, W. Li, P. Shen, A. Vasilakos, Gradient-driven parking navigation using a continuous information potential field based on wireless sensor network. *Information Sciences* 408(2), 100-114(2017).
37. Z. Wan, N. Xiong, N. Ghani, A. V. Vasilakos, L. Zhou, Adaptive unequal protection for wireless video transmission over IEEE 802.11 e networks. *Multimedia Tools and Applications* 72(1), 541-571(2014).

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Submit your manuscript to a SpringerOpen[®] journal and benefit from:

- ▶ Convenient online submission
- ▶ Rigorous peer review
- ▶ Open access: articles freely available online
- ▶ High visibility within the field
- ▶ Retaining the copyright to your article

Submit your next manuscript at ▶ [springeropen.com](https://www.springeropen.com)
