Research on Minimal Confusion Sets for Fine-Grained Visual Categorization

Junfeng Wu, Li Yao*, Bin Liu, Zheyuan Ding, Xiaosheng Feng

Science and Technology on Information System and Engineering Laboratory
National University of Defense Technology, Changsha, 410073, China

* Li Yao*: 371865833@qq.com

Abstract

Fine-Grained Visual Categorization (FGVC) is challenging due to the large intra-class and small inter-class variance, and classes which are prone to be confused with one or more similar classes are called as confusing classes. Hierarchical visual data structures are widely used for image classification with a large number of classes so that the sophisticated classification problem can be processed from coarse to fine. However, hierarchical structures are accompanied with error propagation. In order to reduce the error propagation caused by the hierarchical structure, we propose the concept of minimal confusion set and its metrics. Minimal confusion set is the minimal set of classes that are highly confused with other classes in the set and much less confused with classes outside the set. Based on minimal confusion sets, a flat two-layer hierarchical structure can be formed, which reduces the error propagation. An effective mining method of minimal confusion set is also proposed, which can also be used for hierarchical learning. The mining method of minimal confusion set is based on confusion matrix, but it is different from the traditional methods using the confusion matrix obtained from test data. This method is independent on test data, and experiments on multiple datasets demonstrate the effectiveness of the method.

Keywords: Fine-Grained Visual Categorization, confusing class, confusion matrix, minimal confusion set.

1. Introduction

There is a strong hierarchical relationship between classes in the real world, and as regards human learning, coarse classification is much simpler than fine-grained

classification⁽¹⁾. Given a hierarchy of classes, sophisticated classification can be performed in a coarse-to-fine fashion, which can not only improve accuracy, but also improve efficiency⁽²⁾. The hierarchical structure is also important and widely used in the image classification problem.

The methods of hierarchical learning for image classification can be classified into three categories: methods based on taxonomy^(3, 4), methods based on clustering with visual features⁽⁵⁻⁸⁾, and methods based on confusion matrix⁽⁹⁻¹¹⁾. However, hierarchical learning classification methods inevitably bring about the problem of error propagation; that is, instances that are misclassified in the intermediate layer tend to be misclassified in the end. The problem of error propagation is common to the above three categories of methods.

Therefore, error propagation should be considered when hierarchical models are used for image classification. Qu et al. transform the problem of class prediction into a task of finding the optimal path of a visual tree by maximizing a joint probability⁽²⁾. Rather than finding only one path, candidate paths corresponding to the top N joint probabilities are explored. However, the N path search in the method increases the computational complexity greatly. The closest related work reduces the error propagation by dividing confusing classes into some subsets, instead of generating a multi-layer structure for all classes⁽¹²⁾. However, this method still uses a multi-layer binary tree structure in each confusing subset, leading to large error propagation in subsets of confusing classes.

In addition to the common disadvantage of error propagation, these methods also have their own shortcomings. Taxonomy-related methods which use existing taxonomy like WordNet for classifying classes are widely used in image classification⁽³⁾, but they all ignore

visual features. Similar classes in the taxonomy may not be similar in terms of visual features. Therefore, the hierarchical structure based on taxonomy is not the best option to improve classification accuracy. The methods based on clustering with visual features are more inclined to find supersets of confusing classes, because similar in visual features is a necessary condition for different classes being confused, but not a sufficient one. That is to say, instances of confusing classes have similar visual features, but instances which have similar visual features may not be confused as long as some distinctive features can be distinguished.

In contrast, methods based on confusion matrix are more likely to obtain an ideal hierarchical structure, but the existing methods based on confusion matrix⁽⁹⁻¹¹⁾ use test data to compute the confusion matrix, which is susceptible to the distribution of training data and test data. Besides, it is always impossible to know the true labels of test data in practical applications in a short time. Therefore, such methods which use the confusion matrix based on test data have great limitations.

In summary, the hierarchical approaches based on confusion matrix in hierarchical learning have advantages, but the traditional methods rely on test data, which have great limitations in application. Besides, reducing the error propagation caused by the hierarchical structure is also very important.

This paper proposes the concept of minimal confusion set, based on which a flat two-layer hierarchical structure can be formed and the error propagation caused by this hierarchical structure is relatively small, due to the property of minimal confusion set that the classes in it are seldom confused with classes outside. The mining method of minimal confusion set utilize confusion matrix in a way that is independent of test data, and it is based on the assumption about the relation between class confusion and experiment data. The confusion matrices obtained by multiple experiments are used to mine the minimal confusion sets. Using confusion matrices obtained solely on the training data perfectly overcomes the deficiencies of existing confusion matrix-based methods. Research on minimal confusion set can also pave the way for ensemble learning. First, the method with the highest accuracy in the overall data can be used to perform rough classification and the minimal confusion sets can be obtained. Then another method that has the best performance on the minimal confusion sets can be used to further classify the classes in the minimal confusion sets.

The rest of the paper is organized as follows: Section 2 describes the concept of confusing class, minimal confusion set and their properties. Section 3 describes the mining method of minimal confusion set, while the experimental details are articulated in Section 4. Concluded remarks and the future directions of our work are described in Section 5.

2. Minimal Confusion Set

Before defining the minimal confusion set, we first define the elements in the minimal confusion set, confusing classes. The first thing to be clear is that confusion among classes is not only related to the property of the classes, but also related to the classification method used. Therefore, both confusing classes and the minimal confusion set are related to the classification method. For the convenience of description, it is no longer emphasized that the confusing classes and the minimal confusion set are obtained by using a certain classification method.

Confusion can occur amongst more than two classes. For example, if A is easily misclassified as B , B is easily misclassified as C , and C is easily misclassified as A , then A , B , and C are easily confused in the set they make up of. Whether a class is easily misclassified as other classes or other classes are easily misclassified as it, this class is regarded as a confusing class. That is, if a class has low *Recall* or low *Precision*, it is considered as a confusing class. A quantitative representation is expressed with the inequality,

$$P * R < T_1 \quad (1)$$

where P is *Precision*, R is *Recall*, and $P * R$ is the Confusion Coefficient (CC) we define, and T_1 is the threshold of CC .

Minimal confusion set is defined in the multi-class classification problem. It is a set of existing classes, in which each element is a confusing class. The classes in the minimal confusion set are highly confused with other classes in the set and much less confused with classes outside the set. Since the set cannot be reduced otherwise it no longer satisfies the property, it is named minimal confusion set. The minimal confusion set also cannot be extended or merged with other minimal confusion set; that is, adding classes of any number to a minimal confusion set makes it no longer a minimal confusion set.

The number of the minimal confusion set is dependent on the dataset and the classification method. It should be

noted that not every confusing class belongs to a minimal confused set. If there is no set (except the sets containing all confusing classes) which contains a certain confusing class satisfies the definition of the minimal confusion set, then the class won't belong to any minimal confusion set.

To illustrate the minimal confusion set, we take the confusion matrix obtained on CUB-200-2011 dataset⁽¹³⁾ with Bilinear CNN⁽¹⁴⁾ as an example.

class_number	59	60	62	64	65	66	TP	TP1	P
59	9	7	3	4	4	0	9	27	30
60	3	16	2	0	2	2	16	25	30
62	3	2	10	5	2	5	10	27	30
64	1	3	2	16	1	0	16	23	30
65	1	2	0	0	14	1	14	18	20
66	6	5	4	0	2	11	11	28	30
	...								
TP	9	16	10	16	14	11			
TP2	23	35	21	25	25	19			
P'	24	35	21	26	28	20			

Fig. 1. Minimal confusion set in the confusion matrix

The red part in Fig. 1 shows a part of the confusion matrix that was obtained on CUB-200-2011 dataset with the Bilinear CNN algorithm. The CUB-200-2011 dataset contains 200 species of birds. Most of these species have around 30 samples in the test set, except a few. For example, the 65th bird class has only 20 test samples. The confusion matrix is transformed so that the rows and columns which represent the bird classes that are easily confused are adjacent, as shown in Fig.1. TP represents the number of True Positive, which is on the diagonal line of the confusion matrix. P for each row represents the number of positive instances, and P' for each column represents the number of instances that are predicted to be the class represented by this column.

It can be seen that the *Recall* and *Precision* of the above 6 classes are very low, so they are confusing classes, and these classes are easily confused within the set they make up of. If these classes are not differentiated within the set, and the elements in the red part are added by row and by column respectively, then we can obtain TP1 and TP2, and both TP1 and TP2 are already close to P and P' respectively.

In order to measure whether the classes in the set are highly confused with other classes in the set and much less confused with classes outside the set, the metric Internal and External confusion coefficient (*IECC*) is defined.

Internal confusion (*IC*) is the confusion that can be reduced by considering these classes as a whole, so the bigger it is the better. External confusion (*EC*) is the confusion that cannot be eliminated even after the minimal confusion set is formed, and this confusion will cause error

propagation, so the smaller it is the better. The number of classes in the minimal confusion set is denoted by *CI*, and the number of classes outside the minimal confusion set is denoted by *CO*. The degree of internal confusion is denoted by *DIC* and the degree of external confusion is denoted by *DEC*. The ratio of *DIC* to *DEC* is the Internal and External Confusion Coefficient (*IECC*). These measures can be computed as

$$DIC = IC / CI \quad (2)$$

$$DEC = EC / CO \quad (3)$$

$$IECC = DIC / DEC \quad (4)$$

In this case, $IC = 72$, $CI = 6$, $EC = 28$, $CO = 194$, $DIC = 72/6$, $DEC = 28/194$, and the *IECC* of this minimal confusion set is 83.14. In contrast, if a set of classes is randomly selected from a data set, its *IECC* should theoretically be close to 1. That is, the ratio of the number of internal confusions to the number of external confusions is approximately equal to the ratio of the number of internal classes to the number of external classes. Because the classes in the minimal confusion set are highly confused with other classes in the set and much less confused with classes outside the set, the minimal confusion set has a big *IECC*.

Taking a classification problem with 8 basic classes as an example, assume that there are 2 minimal confusion sets on the 8 classes. The hierarchical structure based on the minimal confusion sets is shown in Fig. 2 where v_0 represents all classes, v_{11} , v_{12} represent classes included in the minimal confusion set, and v_1 , v_2 , v_3 , v_4 , v_5 , v_6 , v_7 , and v_8 represent the 8 basic classes respectively. The advantage of hierarchical structure is that by grouping similar classes into one class, the classes in the same layer are distinguished easily, but at the same time the hierarchical structure will bring about errors propagation. Therefore, the advantages and disadvantages should be weighed in the process of constructing hierarchical structure. Assume that v_2 is also a confusing class, but it does not belong to any minimal confusion set. In other words, its confusion with other classes is more decentralized. If v_2 is combined with other classes to form a hierarchical structure, it will result in great error propagation. The minimal confusion set has a large *IECC*, thus classes in it will take more advantages than disadvantages in hierarchical classification models. Therefore, the hierarchical structure based on the minimal confusion set has advantages over existing hierarchical models.

Fig. 2. Hierarchy based on minimal confusion set

3. Mining method of Minimal Confusion Set

Ease of confusion is an inherent relationship between similar classes, but different image classification methods can also cause confusion due to their own defects.

We make the assumption that if a certain method tends to cause some classes to be confused, and when other conditions remain unchanged except that training data is randomly reduced, these classes are still confused. Based on this assumption, this paper proposes a method which only uses training data to mine minimal confusion set. The basic flow of the proposed method is as follows.

- 1) Divide the training data randomly into k folds in average;
- 2) Classification algorithm trains models on with $k-1$ folds and tests with 1 fold to obtain the confusion matrix CM ;
- 3) Transform the confusion matrix CM to obtain the Confusion Relation Matrix (CRM), so that each row contains the confusion relation between one class with other classes;
- 4) Take each row which represents a class in the CRM as an instance and cluster them into *sub-clusters*;
- 5) Calculate the $IECC$ of each *sub-cluster*, and pick out the *sub-clusters* whose $IECC$ exceed the threshold T_2 ;
- 6) Take 1 of the k folds of data obtained in step 1 in turn as test data, and repeat steps 2-5 a total of k times to obtain a set of *sub-clusters*, and the set is named CS ;
- 7) Take each *sub-cluster* in CS as a transaction⁽¹⁵⁾ and each class as an item⁽¹⁵⁾, and find the maximal frequent k itemsets⁽¹⁵⁾ in CS .

The maximal frequent k itemsets obtained through the above process are the minimal confusing sets to acquire. In step 4, classes which are easily confused in the set they

make up of will be clustered into the same sub-cluster. Since step 2-5 have been repeated k times, each class in CS will appear k times at most, and based on the assumption we make, classes that are easily confused should appear together in different *sub-clusters* for k times. Hence, frequent pattern mining is used in step 7 to find the frequent k itemsets, among which the maximal frequent k itemsets are the minimal confusing sets. The representation of confusion between classes in step 3 is very important, which is the basis for clustering in step 4. The detail of step 3 is described in Algorithm 1.

Algorithm 1: CM_Trans(CM)

Input:

- CM , the Confusion Matrix
- T_1 , the threshold of CC

output:

- CRM , Confusion Relation Matrix
- 1 $Confusingclass = \{ \}$ // The set of confusing classes is set empty initially
 - 2 **for each** $class$ in CM **do**
 - 3 COCalculate ($class$) // Calculate CC for each $class$
 - 4 **if** COCalculate($class$) $< T_1$
 - 5 $Confusingclass = Confusingclass + class$ // add the new class whose CC is below T_1 into the set of confusing classes
 - 6 **end if**
 - 7 **end for**
 - 8 CM is normalized by row to get $CM1$
 - 9 Extract the elements in $CM1$ to form a new matrix $CM2$ according to the row and column of $Confusingclass$.
 - 10 Add $CM2$ to its transposed matrix to get CRM
-

Algorithm 1 is simple, but there is a trick in it. The elements in confusion matrix represent a directed relation between two classes; that is, ‘ A is misclassified as B ’ is different from ‘ B is misclassified as A ’. However, confusion is an undirected relation, therefore $CM2$ is added to its transposed matrix in the last step. The advantage of doing so is to better express the confusion between different classes, so that classes which are highly confused can be clustered into the same sub-cluster easily.

4. Experiments

Experiments are carried out on three widely used FGVC datasets. Since the mining method of the minimal

confusion set is based on the assumption we make in section 3. The main purpose of the experiment is to verify whether the proposed method can acquire the minimal confusion set on multiple datasets and whether the set acquired through the mining method satisfy the property of minimal confusion set we propose.

4.1 Datasets

CUB-200-2011⁽¹³⁾, STANFORD CARS⁽¹⁶⁾, And FGVC-AIRCRAFT⁽¹⁷⁾ are used for experiments. Their main descriptions are given in Table 1.

The training data are used to mine the minimal confusion set, and the test data are mainly used to test the *IECC* of the minimal confusion set acquired. *IECC* is used for measuring the quality of the minimal confusion set.

Table 1. Datasets

Name	# Category	# Training	# Test
CUB-200-2011	200	5994	5794
STANFORD CARS	196	8144	8041
FGVC-AIRCRAFT	100	6667	3333

4.2 Experimental Setup

Deep neural networks have achieved the best experimental results in FGVC. This paper uses an advanced deep learning framework PFNet, which has better performance on the STANFORD CARS and FGVC-AIRCRAFT data sets than the Macnn⁽¹⁸⁾, the state-of-the-art. K-means⁽¹⁹⁾ is used for clustering, and Aprior⁽¹⁵⁾ is used for frequent itemsets mining.

The experimental parameters include: T_1 , the threshold of *CC*; T_2 , the threshold of *IECC*; K , the number of experiment times; C , the number of clusters in K-means; and min_sup , the minimal support in Aprior algorithm.

According to experience, T_1 is set to 0.8 and T_2 is set to 15, which ensures that only a set with big *IECC* will be selected as a minimal confusion set. The number of clusters C is set to $\sqrt{N}$, where N is the number of confusing classes in the dataset, and is determined by the dataset, classification method and T_1 . The number of experiment times K is set to 3, which means that the data in training set are divided into 3 parts in average. Since the amount of the training data is relatively limited and only one fold of data are used for test, if K is too big, the test data will be too few. K cannot be too small either, otherwise the training data are too few to get a good classification model. Thus, the selection of K in this experiment is 3. The steps 2-5 in

the experiment are repeated for 3 times, and $3*\sqrt{N}$ sub-clusters are obtained as items. Different from traditional frequent itemset mining, each class will appear at most only 3 times instead of $3*\sqrt{N}$ times in the set of items, while the classes that are easily confused should appear together in different itemsets for 3 times. Therefore, the min_sup is set to $1/\sqrt{N}$.

4.3 Experimental Results

The minimal confusion sets obtained on each dataset are shown in Table 2-4. It can be seen that the *IECC* of all the minimal confusion sets are very big, which means that the classes in the minimal confusion sets are highly confused with other classes in the set and much less confused with classes outside the set.

Table 2. The minimal confusion sets on CUB-200-2011

The minimal confusion sets	<i>IECC</i>
{59, 60, 62, 64, 65, 66}	89.24
{143, 144, 145, 146, 147}	47.52

Table 3. The minimal confusion sets on STANFORD CARS

The minimal confusion sets	<i>IECC</i>
{16, 17, 18}	193
{14, 19, 25}	279

Table 4. The minimal confusion sets on FGVC-AIRCRAFT

The minimal confusion sets	<i>IECC</i>
{11, 12, 13, 14}	162
{24, 25, 26, 27}	45.33
{28, 29, 30, 31}	62
{38, 39}	490
{44, 54}	114.3

The red part in Fig. 3 shows the minimal confusing set with the biggest *IECC*. It can be seen that the 38th class and the 39th class of FGVC-AIRCRAFT are highly confused, while they are seldom confused with other classes outside the minimal confusing set that they make up of. Therefore, the *IECC* of this minimal confusing set is extremely big.

class	38	39	TP	TP1	P
38	29	4	29	33	33
39	6	27	27	33	34
TP	29	27			
TP2	35	31			
P*	35	31			

Fig. 3. The minimal confusing set with the biggest *IECC*

5. Conclusions

The advantage of hierarchical structure is to make the classes in the same layer be separated from other classes as easy as possible by grouping similar classes into a set. However, at the same time, the hierarchical structure will bring about error propagation. In this paper, we propose the concept of minimal confusion set and its important metric *IECC* so that a flat two-layer hierarchical structure can be formed, in which classes in minimal confusion sets will take more advantages than disadvantages in hierarchical classification models. A general mining method of the minimal confusion set is proposed based on the assumption about the relation between class confusion and experiment data. Our method only needs data of the training set, thus overcoming the limitation of the traditional confusion matrix-based methods. The assumption is verified by experiments on multiple FGVC datasets.

The minimal confusion set can also be used for ensemble learning. By combining the overall optimal method with the optimal method on the minimal confusion sets, an ensemble learning which focus on the confusing classes can be achieved. However, the part of ensemble learning has not been realized yet, which is the next step of our research work.

Acknowledgment

The presented work is framed within the National Natural Science Foundation of China (No. 71371184). We also thank Bin Liu and Zheyuan Ding for giving us revision advice.

References

- (1) Niki Martinel : "Biologically-Inspired Hierarchical Learning Architectures for Image Classification," Thesis, the University of Udine, 2017
- (2) Yanyun Qu, Li Lin, Fumin Shen, Chang Lu, Yang Wu, Yuan Xie and Dacheng Tao : "Joint Hierarchical Category Structure Learning and Large-Scale Image Classification," *IEEE Transactions on Image Processing*, vol. 26, pp. 4331-4346, 2017
- (3) Min Sun, Wan Huang and Silvio Savarese : "Find the best path: An efficient and accurate classifier for image hierarchies," *IEEE International Conference on Computer Vision (ICCV)*, pp. 265-272, 2013
- (4) Yanming Guo, Yu Liu, Erwin M Bakker, Yuanhao Guo and Michael S Lew : "CNN-RNN: a large-scale hierarchical image classification framework," *Journal of Multimedia Tools and Applications*, pp. 1-21, 2017
- (5) Peixiang Dong, Kuizhi Mei, Nanning Zheng, Hao Lei and Jianping Fan : "Training inter-related classifiers for automatic image classification and annotation," *Journal of Pattern Recognition*, vol. 46, pp. 1382-1395, 2013
- (6) Hao Lei, Kuizhi Mei, Nanning Zheng, Peixiang Dong, Ning Zhou and Jianping Fan : "Learning group-based dictionaries for discriminative image representation," *Journal of Pattern Recognition*, vol. 47, pp. 899-913, 2014
- (7) Ning Zhou and Jianping Fan : "Jointly Learning Visually Correlated Dictionaries for Large-Scale Visual Recognition Applications," *IEEE transactions on pattern analysis and machine intelligence*, vol. 36, pp. 715-730, 2013
- (8) Calvin Murdock, Zhen Li, Howard Zhou and Tom Duerig : "Blockout: Dynamic model selection for hierarchical deep networks," *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2583-2591, 2016
- (9) Samy Bengio, Jason Weston and David Grangier : "Label embedding trees for large multi-class tasks," *Advances in Neural Information Processing Systems*, pp. 163-171, 2010
- (10) Gregory Griffin and Pietro Perona: "Learning and using taxonomies for fast visual categorization," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1-8, 2008
- (11) Baoyuan Liu, Fereshteh Sadeghi, Marshall Tappen, Ohad Shamir and Ce Liu: "Probabilistic label trees for efficient large scale image classification," *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 843-850, 2013
- (12) ZongYuan Ge, Chris McCool, Conrad Sanderson, Alex Bewley, Zetao Chen and Peter Corke: "Fine-grained bird species recognition via hierarchical subset learning," *IEEE International Conference on Image Processing*, pp. 561-565, 2015
- (13) Catherine Wah, Steve Branson, Peter Welinder, Pietro Perona and Serge Belongie: "The caltech-ucsd birds-200-2011 dataset," 2011
- (14) Tsung-Yu Lin, Aruni RoyChowdhury and Subhransu Maji: "Bilinear cnn models for fine-grained visual recognition," *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1449-1457, 2015

- (15) D. Burdick, M. Calimlim and J. Gehrke: "MAFIA: a maximal frequent itemset algorithm for transactional databases," International Conference on Data Engineering, p. 443, 2001
- (16) Jonathan Krause, Jia Deng, Michael Stark and Li Feifei: "Collecting a Large-scale Dataset of Fine-grained Cars," 2013
- (17) Subhransu Maji, Esa Rahtu, Juho Kannala, Matthew Blaschko and Andrea Vedaldi: "Fine-grained visual classification of aircraft," arXiv preprint arXiv:1306.5151, 2013
- (18) Heliang Zheng, Jianlong Fu, Tao Mei and Jiebo Luo: "Learning Multi-attention Convolutional Neural Network for Fine-Grained Image Recognition," IEEE International Conference on Computer Vision, pp. 5219-5227, 2017
- (19) J. A. Hartigan and M. A. Wong: "Algorithm AS 136: A K-Means Clustering Algorithm," Journal of the Royal Statistical Society, vol. 28, pp. 100-108, 1979