Report of the Coursework on Randomised Decision Forest

Ran Cheng
ran.cheng18@imperial.ac.uk

Yang Zhao
yang.zhao18@imperial.ac.uk

1. Introduction

Visual object categorisation has become a significant topic with the popularity in large-dimensional data processing. With proper labels, the information hidden in the image can be quantised effectively, which reduces the complexity and provides more possibilities for data processing. Essentially, the classification result depends on the decision boundaries in the high-dimensional space which can be learned by the training set. This report investigates two algorithms to obtain visual codebook: K-means clustering and Random forests (RF); then it examines the accuracy, complexity, and cost of RF classifier on Caltech101 dataset.

2. Implementation and Results

All results are averaged over 3 repetitions. We select equal amount of descriptors from each training image.

2.1. K-means Codebook

2.1.1 Vocabulary size

We investigated the influence of codebook size by varying the number of clusters of K-means algorithm $K = 32 - 560$. Each centroid corresponds a visual codeword of the training set. If too small, the classification can be vague for both training and testing data; if too large, the codebook can be more confident for the training data but the noise by the class-irrelevant information leads to overfitting and the performance on testing set is expected to reduce. Also, the complexity of clustering increases.

2.1.2 Bag-of-words Histograms of Example Images

Figure 1 shows distributions of the descriptors of the same class in the 256 clusters. It suggests that the histograms of images of the same class are similar, but for those of different classes the distribution can vary significantly. Based on the maximum-likelihood estimation, if the descriptors of a certain class falls in a cluster more often than the others, it is more likely that those descriptors in this cluster correspond to that class. However, the noise factors like colour distribution can influence clustering and reduce the accuracy. Moreover, for the testing data, some are biased to specific

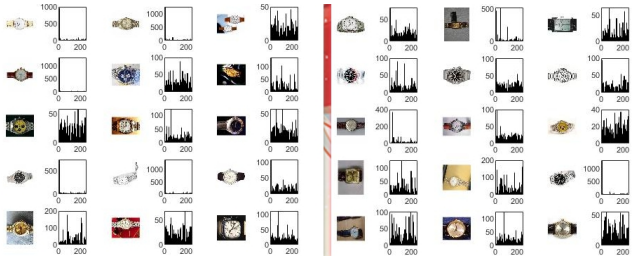

Figure 1. Bag-of-words 256D histograms of training and testing images of the class "watch"

codewords. It indicates that the degree of overlapping in between is not high. A possible approach to improve the system is to utilise more data for training. The major limitation is that the location information is not considered in bag-of-word method.

2.1.3 Vector Quantisation Process

K-means clustering classifying a large set of vectors (descriptors) into different groups and representing each group by its centroid index. First, the centroid are randomly selected in the image space. Then, descriptors are allocated to the corresponding cluster with the smallest Euclidean distance. Finally, the new centroid are calculated for each cluster. It keeps iterating until the result converges. The major problems are that the converging speed can be slow for large-scale data or even no converge. Also, it may stop at local optimal rather than global.

2.2. RF Classifier

2.2.1 Tree Number

The influence of tree number on training and testing set accuracy is shown in the first plot of Figure 2. It indicates that as the number of tree increases, the accuracy of the training set rises to a threshold then maintains at this level. For testing set, the optimum choice of tree numbers with the specified parameters is around 500, as a threshold between under-fitting and over-fitting. A good trade-off between performance and complexity is around 400.

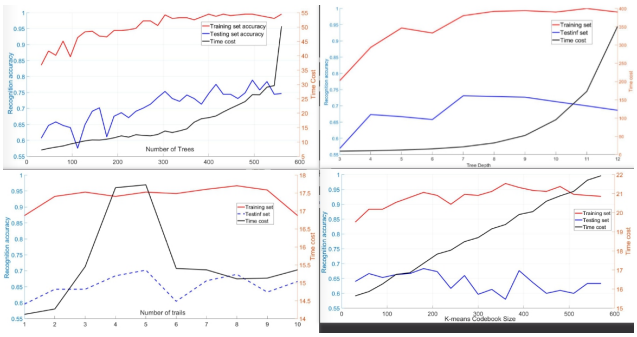

Figure 2. The influence of classification tree parameters on accuracy and time cost

Table 1. Weak-learner type, accuracy and time cost

Weaker learner	Train accu.	Test accu.	Time cost
Axis-aligned	99.56%	68.67%	20.93s
2-pixel test	98.89%	64.44%	22.15s
Linear	98.89%	67.78%	25.49s

2.2.2 Tree Depth

As can be observed in the second plot of Figure 2, the testing set recognition accuracy firstly increases with the increase of tree depth. For tree depth larger than 7, the accuracy is decreasing with the increase of tree depth.

As the tree grows deeper, more splits nodes are generated and captures more features about the training data. Thus, the training data accuracy keeps increasing. However, the increase of tree depth will define some features of training set as the generalized features for all potential data, which is overfitting. Therefore, the testing data accuracy is increasing and then decrease.

2.2.3 Split Number

Split number plays a significant role in node dividing. As plot 3 of Figure 2 suggests, the optimum value is 7 after which the complexity grows exponentially while the accuracy reduces.

2.2.4 Weak-learners

In Axis-aligned splitting, a dimension of descriptors is randomly selected as a threshold for splitting data. For the 2-pixel test, only two dimensions of the K-dimensional descriptors are used. It approximates the gradient descent. The linear model can be more random and capture more genetical features of testing data with slightly higher complexity.

2.2.5 Confusion Matrix

As shown in Figure 3, the diagonal of the confusion matrix is brighter representing the RF classifier has a higher

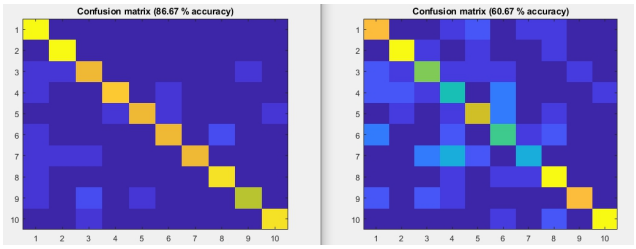

Figure 3. The confusion matrix of training and testing data of RF classification with K-means codebook

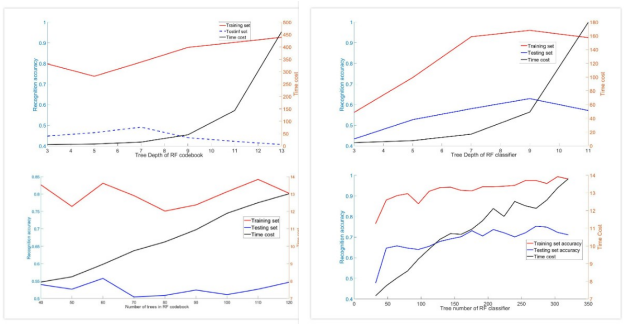

Figure 4. The influence of codebook tree parameters on accuracy and time cost

success rate than the rate of classifying the image into a different class.

The left plot corresponds the training data with accuracy 86.67% and the error is not biased into a particular class. The confusion matrix of testing data is on the right with the accuracy 60.67 %. The main error source is the images of class 7 which were wrongly categorised into class 4. A possible solution is to use more images of class 7 and less of class 4 to adjust the model.

2.2.6 Vocabulary Size and Accuracy

The vocabulary size should be large enough for characteristic distinction but not very large to reduce noise and avoid overfitting. The last plot of Figure 2 shows the decision accuracy and time cost for vocabulary size $K = 32 - 560$. It demonstrates that a small codebook is enough for the testing set. With the increase of codebook size, the accuracy does not improve which is not as expected. One possible reason is that the training data set is small. It result in the correlation within the cluster does not provide more useful information about the characteristic of classes.

3. RF Codebook

Compared with K-means codebook, the decision accuracy based on RF codebook is actually smaller. A possible explanation is that the leaves includes too much randomness which is irrelevant to the classification.

Criteria	Training set accuracy	Testing set accuracy	Time cost
SIFT	91.87%	42.73%	14.03s
DSIFT	92.00%	62.27%	171.83s
COVDET	94.13%	52.13%	24.36s
PHOW	94.67%	70.27%	14.44s

Criteria	Train accu.	Test accu.	Time cost
IG	93.93%	62.73%	12.20s
IGR	91.80%	63.13%	11.89s
Gini	94.93%	65.07%	11.77s

3.1. Tree Number

Figure 4 indicates the impact of tree number of codebook forest on the classification accuracy and time cost. As the tree number increases, there are more independent leaves acting as decision machine but the accurate first increase then decrease. The result suggests that 60 trees is the balanced point.

The number of trees in classifier forest still have a positive correlation to the accuracy. As the number of trees increases, the correct rate on training and testing data grows rapidly to a saturation region with about 50 trees and then increasing slowly. It corresponds to the discussion above.

3.2. Tree Depth

The optimum depth of codebook forest is 7 suggested by Figure 4. The complexity grows exponentially as we increase the depth but the accuracy reaches the optimum in the middle. It indicates that proper randomness of individual decision machine can benefit the classification.

When combined with codebook forest, the classifier forest prefer a deeper structure that favours 9 as the depth. The accuracy reaches 60% before overfit.

3.3. Criteria for Descriptors

We tried different methods to obtain descriptors. SIFT and DSIFT are grid-search method while COVDET and PHOW detects upright scale and translation covariant features. All approaches work well on the training set but the performance varies at testing data. PHOW outperforms the rest as expected, and DSIFT as a dense traversing method can approximate details but the time cost is ten times higher than PHOW. The accuracy of the rest are relatively low for the testing data.

3.4. Criteria in Node Splitting

Three standards to evaluate the purity of data are applied respectively: Information Gain (IG), Information Gain Ratio (IGR) and Gini Index (Gini). The result is shown in Table 3.4.

The higher value of IG means the split data has higher purity. IG criteria prefer for the splitting with more sub-nodes. The preference may have a negative impact on classification. To avoid this, IG is divided by the intrinsic value of the selected dimension in IGR method. For Gini, a different strategy is performed. It randomly selected two descriptors from the sub-node and recorded the probability that the descriptors are not the same. The complexity is lower.

3.5. Evaluation

The descriptors selected by SIFT are not affected by scale, location and rotation. Thus, the trained model's generalisation ability is improved. However, the spatial information of descriptors is missing, such as the relative position between descriptors.

Moreover, the models have very high accuracy of classifying training data does not have a good performance in classifying testing data in most trails. Because these models define some features of the training set as the generalised features for all potential testing data. It results in overfitting and hence high generalisation error in classifying testing set.

To reduce the effect of overfitting, bagging and pruning are performed in RF algorithm. Sampling with the replacement of bagging decreases the correlation between trees to a low level. However, overfitting may still exist after bagging. To reduce overfitting further, pre-pruning and post-pruning are used in pruning method. Pre-pruning happens during node splitting; the node will be set as a leaf node if the node split is unable to improve the generalisation ability of the corresponding tree. Post-pruning is applied after tree growth, its pruning from the bottom of the tree to the top. A sub-node of a node will be set to a leaf if this increase the generalisation ability.

We also developed some other pruning strategies like confidence and sample ratio that can be used to adjust the decision model but due to the limitation of space they are not listed. Relevant code are provided in the appendix.

3.6. Conclusion

Two main types of image classifier are implemented and evaluated in this report: RF classifier with K-means codebook and RF classifier with RF codebook. Increasing tree size enable RF to fit more complex images with the increase of overfit probability. Increasing the number of trees is not causing overfit. However, the increase of recognition accuracy is neglectable when tree number exceeds 230. Different criteria for feature detection and splitting node are implemented, "IG" and "PHOW" criteria achieve best recognition accuracy, respectively. The performance of the two algorithm is limited by SIFT, due to the lack of spatial information. In further development, the relative position between descriptors can be added to classify images.

4. References