EVOLUTIONARY FEATURE AND INSTANCE SELECTION FOR VISUAL RECOGNITION

CHIH-FONG TSAI

Department of Information Management, National Central University, Taiwan
E-mail: cftsai@mgt.ncu.edu.tw

Abstract - The aim of image mining is to automatically extract semantically meaningful information (or knowledge) from image data or to identify other patterns that are not explicitly stored in the images. In particular, image mining must be approached by image preprocessing. The preprocessing step in image mining deals with some conventional image processing tasks, such as segmentation and low-level feature extraction, where the mining result is heavily dependent on the preprocessing result, e.g. the identification of spatial objects and their relationships. From the KDD point of view, the processed image data that are represented by some feature vectors are used for the data analysis or mining stage. However, data preprocessing of KDD is usually based on feature selection and/or instance selection, in which the former focuses on dimensionality reduction and the later for data sample reduction. In this paper, we examine the effects of performing feature and instance selection by the genetic algorithm (GA) on image classification as a very important problem of image mining. The results based on two image datasets show that image mining can take advantage of feature and instance selection. Specifically, GA largely reduces the training dataset size by over 50% dimensionality reduction and near 50% training data sample reduction. In addition, the classifier trained by the reduced dataset can provide comparable accuracy with the baseline classifier without data preprocessing.

Keywords - Feature Selection, Genetic Algorithms, Image Mining, Instance Selection, Visual Recognition.

I. INTRODUCTION

Knowledge discovery in databases (KDD), or data mining, generally contains a number of steps, such as dataset selection, data preprocessing, data analysis, and result interpretation and evaluation. In particular, data preprocessing is one of the most important steps in KDD, aimedat making the chosen dataset as 'clean' as possible for eventual analysis and evaluation. In other words, if there is no data quality, no quality mining results will be obtained [1, 2]. The two most commonly used ways to effectively pre-process the chosen dataset for data mining arefeature selection (or dimensionality reduction) and instance selection (or record reduction). This is because the number of features and data samples selected is usually very large in most real-world data mining problems. If too many instances are considered, it can result in large memory requirements, high disk access, slow processing speed, and a possible over-sensitivity to noise [3]. In addition, it is often the case that the data are not all equally informative and some data points will be further away from the sample mean than what is deemed reasonable. Data mining tasks such as classification or prediction performance, without the instance selection step will very likely produce poorer results[4].

On the other hand, if too many features are used for data analysis, it can cause a dimensionality problem [5]. Since not all of the pre-chosen features are informative, the objective of feature selection is to select more representative features which have more discriminative power over a given dataset. This is also called dimensionality reduction [6].

Genetic algorithms (GA), one of the most widely used techniques for feature and instance selection, can improve the performance of data mining algorithms [7, 8]. Cano et al. [9] have shown that a GA can obtain better results than many traditional and non-evolutionary instance selection methods in terms of better instance selection rates and higher classification accuracy. Moreover, it has been shown that GAs are suitable for large-scale feature selection problems [10]. Image mining (or knowledge discovery in image databases) has become an important part of data mining [11, 12].The main focus of data preprocessing in image mining is usually on how to extract image features for representation, such as segmentation, low-level visual feature extraction, etc. The data preprocessing is finished when an image is represented by one or several features used to describe its whole or partial content. However, no related study has examined the effects of conventional data preprocessing (including feature and instance selection) on image mining results. Therefore, the aim of this paper is to perform feature and instance selection by GA individually, to investigate their impact on the image mining results.

II. LITERATURE REVIEW

A. Feature Selection

It is usually the case that the number of features (or variables) collected in a dataset is relatively large (i.e., the curse of dimensionality), and not all of these

features are informative or can provide high discriminative power. [5].The aim of feature selection is thus to remove irrelevant and/or redundant features from the chosen dataset, thereby improving the performance of the classification and/or clustering algorithms. In addition, given a dataset, feature selection can help analysts understand which features are important as well as how the features are related.
Feature selection can be defined as the process of choosing a minimum subset of M features from the original dataset of N features (M<N), so that the feature space (i.e., the dimensionality) is optimally reduced.

A feature selection algorithm usually consists of four steps: subset generation, subset evaluation, stopping criterion, and result validation [13]. Subset generation is a search procedure which generates subsets of features for evaluation. Each subset generated is evaluated by a specific evaluation criterion and compared with the previous best one with respect to this criterion. If a new subset is found to be better, then the previous best subset is replaced by the new subset.

B. Instance Selection

As discussed in [3], one problem with using the original data points is that there may not be any data points located at the precise points that would make for the most accurate and concise concept descriptions. Therefore, the aim of instance selection, or record reduction, is to reduce the size of a dataset while still maintaining the integrity of the original dataset. In some cases, generalization accuracy can be increased when noisy instances are removed and when decision boundaries are smoothed to more closely match the true underlying function. The instances that are filtered out by the instance selection algorithm can also be regarded as outliers (or bad data). Specifically, outliers are data points which are highly unlikely to occur given a model of the data. One approach to performing this task is to derive the distances to neighboring data points by implementing a clustering algorithm [14].

Instance selection can be defined as follows. Let X_i be an instance where $X_i = (X_{i1}, X_{i2}, \ldots, X_{im}, X_{ic})$ meaning that X_i is represented by m-dimensional features and X_i belongs to class c given by X_{ic}. Now, assume that there is a training set TR which consists of M instances and a testing set TS composed of N instances. If $S \subseteq TR$ is the subset of selected samples that are produced by an instance selection algorithm, then we can classify a new pattern T from TS over the instances of S.

III. EXPERIMENTS

A. Experimental Setup

Three image datasets are used to determine the image mining (visual recognition) performance. They are Caltech 101 [1], German Traffic Sign Recognition Benchmark (GTSRB)[2], and NUS-WIDE-LITE [18]. The Caltech 101 dataset is used for object categorization, and contains pictures of objects belonging to 101 categories, with every category containing about 40 to 800 images. In total, there are 9146 images in Caltech 101. 30 images from each category are randomly selected for the training set, and the remaining images are used as the testing set. The random splits of training and testing sets were run ten times with the average of the ten results regarded as the final classification result.

On the other hand, GTSRB contains 39209 training and 12630 testing images in 43 classes, with image sizes varying between 15×15 to 250×250 pixels. NUS-WIDE-LITE contains 55615 Flickr images with each image belonging to one of 81 pre-defined categories, giving 27807 training and 27808 testing images. For image feature representation, the Bag-of-Words (BoW) feature [19] is extracted from the Caltech 101 images. Each image is first segmented into 16×16 grids. Then, the keypoints are detected by using a difference of Gaussian (DoG) detector and described by the Scale Invariant Feature Transform (SIFT) descriptor[3]. According to [20], each image is represented by a 200 dimensional BoW feature vector. For the GTSRB images, the histogram of Oriented Gradient (HOG) features [21] is used as the image descriptor. To extract the HOG feature from traffic signs, the images are scaled to a size of 40×40 pixels and converted to grayscale. As a result, each image is represented by a 1568-dimensional feature vector [22]. The color and texture features are extracted from each image in the NUS-WIDE-LITE dataset, to produce225-D block-wise color moments and 128-D wavelet textures. Consequently, each image is represented by a 353-dimensional feature vector. The k-NN classifier is usedwith different k values ranging from 1 to 21 examined in order to obtain the best k-NN. Furthermore, the SVM classifier using the RBF kernel function is constructed for comparison. Consequently, two types of classifiers are constructed:k-NN/SVM with GA and k-NN/SVM without data preprocessing.
Specifically, for each dataset, the feature and instance selection tasks are performed using the GA over the training dataset. Next, the reduced dataset is used to train the k-NN and the result is then evaluated by the testing set in terms of classification accuracy. In particular, since there are many variants of GAs, the MATLAB GA tool (Elitist Multi-objective) is used as the baseline GA.

B. Results for Feature Selection

Table 1 shows the feature selection results given by GA over the two datasets. These results demonstrate

[1] http://www.vision.caltech.edu/Image_Datasets/Caltech101/
[2] http://benchmark.ini.rub.de/?section=gtsrb&subsection=dataset
[3] http://www.cs.ubc.ca/~lowe/keypoints/

that feature selection by GA performs very well in image mining. In other words, it can not only reduce the dimensionality of the dataset with a data reduction of over 50%, but also allows the classifier to provide similar classification accuracy to the one without feature selection capabilities. Table 2 shows the required training and testing times (in seconds) during classifier training and testing with and without feature selection[4]. We can observe that the classifier training and testing times for both k-NN and SVM are significantly reduced when feature selection is performed. Although performing feature selection is likely to lead to a slight degradation in the classification accuracy for the three image datasets, it relieves the need for storage space and reduces the time cost of training and testing a classifier.

	With feature selection			Without feature selection		
	features	Accuracy		features	Accuracy	
		k-NN	SVM		k-NN	SVM
Caltech 101	95	14.64% (k=7)	31.19%	200	16.22 % (k=9)	32.54%
GTSRB	385	75.31% (k=19)	75.7 %	1568	75.87 % (k=17)	78.23%
NUS-WIDE	353	3.27% (k=1)	46.63%	163	3.1% (k=3)	47.16%

Table 1 Performance of GA for feature selection

	With feature selection		Without feature selection	
	k-NN	SVM	k-NN	SVM
Caltech 101	0.34/0.014	30.2/0.67	0.39/0.005	107/2.03
GTSRB	16.69/34.62	517/251	46.31/30.29	4288/1734
NUS-WIDE	22.96/4.69	418/329	28.01/42.46	803/602

Table 2 Classifier training and testing times (training time/testing time)

To illustrate the discriminatory power of the selected features through the GA, multidimensional scaling (MDS) [23] is used. MDS is usually used for exploring similarities or dissimilarities in data. In other words, it provides a visual representation of a given set of objects in terms of their similarities or distances.

We randomly choose three classes from Caltech 101 and GTSRB as examples by using MDS for data visualization [5]. Note that more classes are not considered here, because using too many classes will lead to more data points in the visualization results, which could make it very difficult to distinguish between different classes. Figure 1 shows the MDS results for the original features and the selected features given by GA, respectively. For Caltech 101 without feature selection (Figure 2(a)), more data samples of the '+' class are closer to the 'Δ' and 'O' classes in comparisonto Caltech 101 with feature selection (Figure 2(b)). In other words, performing feature selection shows a number of data samples of the '+' class to be far away from the other two classes. Moreover, despite several data samples that were very close in the original feature space being changed after feature selection, there was not much degradation in the classification accuracy. On the other hand, for the GTSRB dataset, many of the data samples belonging to the 'Δ' and 'O' classes overlap. However, looking at the dark line used to distinguish between the two classes, performing feature selection does make the area of overlap smaller.

[4] The computer environment is based on an Intel Core i7-2600 with 16 GB RAM.

[5] The categories of Caltech 101 are gerenuk, strawberry, and trilobite. The class IDs of GTSRB are 30, 31, and 32.

C. Results for Instance Selection

Table 3 shows the GA instance selection results over the two datasets. We can observe that the GA method can lead to about a 50% data reduction for different dataset sizes. However, after instance selection, there is less similarity in theperformance of the classifiers than without instance selection. On average, there is a difference in performance between the classifiers with and without instance selection of about 5%, which is larger than with feature selection, i.e., around 2%.

	With feature selection			Without feature selection		
	storage	Accuracy		storage	Accuracy	
		k-NN	SVM		k-NN	SVM
Caltech 101	3720	13.48% (k=5)	27.7 %	7809	16.22% (k=9)	32.54%
GTSRB	18738	73.07% (k=13)	73.63%	39209	75.87% (k=17)	78.23%
NUS-WIDE	3594	1.68 % (k=1)	41.81%	27807	3.1 % (k=3)	47.16%

Table 3 Performance of GA with instance selection

Table 4 shows the computational times for training and testing the k-NN and SVM classifiers with and without instance selection. As we can see, the classifiers trained by the reduced training set after instance selection require less training time than the baseline classifiers trained by the original training set. Although there is a significant reduction in the classifier training time after instance selection, the final classification accuracy is certainly different from that of the baseline classifiers, compared with the classification results after feature selection.

	With feature selection		Without feature selection	
	k-NN	SVM	k-NN	SVM
Caltech 101	0.15/0.24	20/8	0.39/0.005	107/2.03
GTSRB	11.97/9.48	287/286	46.31/30.29	4288/1734
NUS-WIDE	4.84/1.67	15.88/91.03	28.01/42.46	803/602

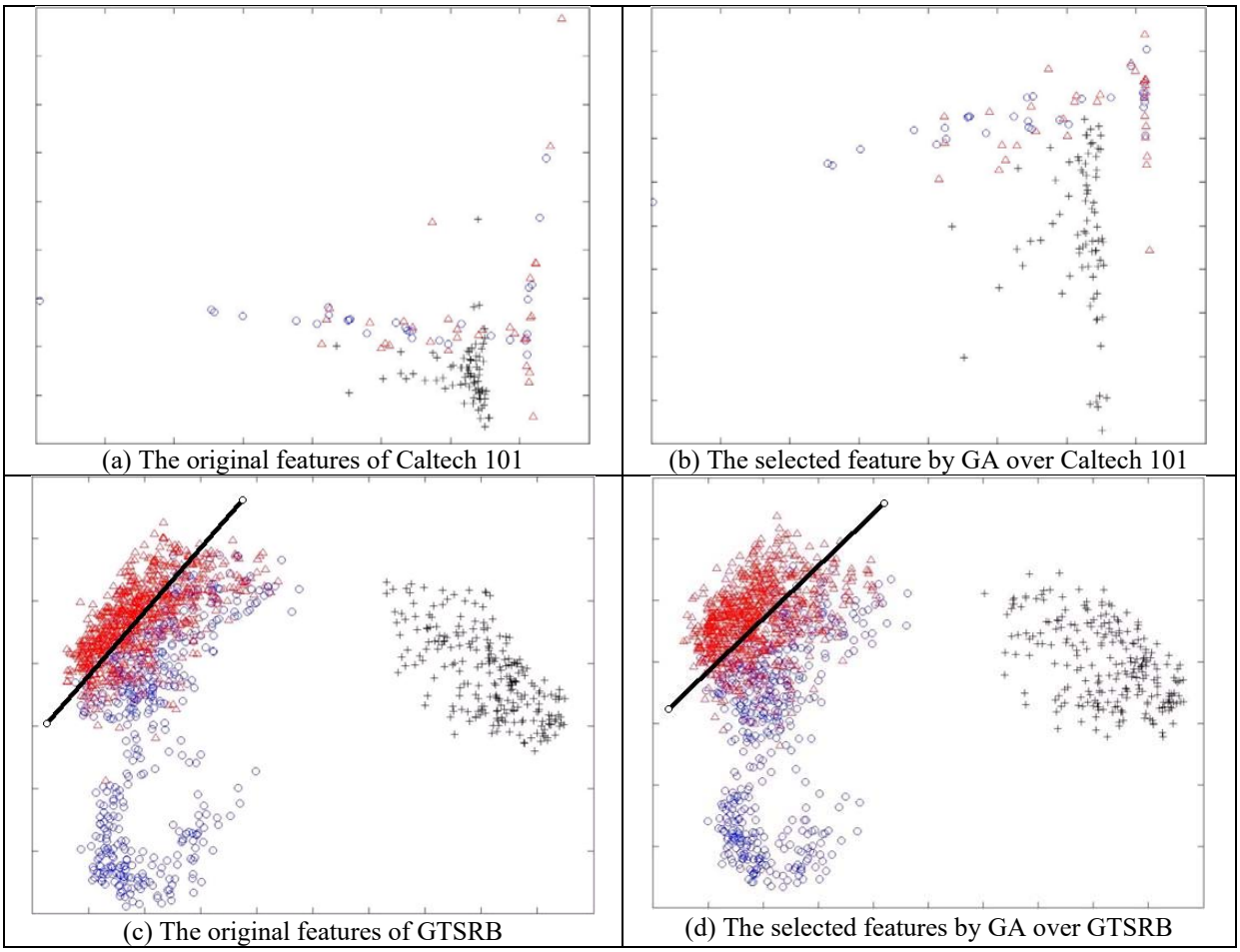
Table 4 Classifier training and testing times (training time/testing time)

The same three-class example of each dataset is used for the MDS analysis. Figure 2 shows the MDS results for the original data and the instances selected by GA. Similar to Figure 1, the data distributions of the processed GTSRB dataset for GA based instance selection are similar to the original GTSRB dataset. Specifically, in some areas, the numbers of 'overlapping' data in the cases where GA is used are less than the original data over the GTSRB dataset (see the circled areas). However, the above results indicate that removing some of the overlapping data samples does not make the classifiers better able to distinguish between different classes since their classification accuracy is certainly lower than that of the baseline classifiers.

IV. CONCLUSION

The experimental results show that GA is very good at feature selection in terms of the dimensional reduction rate and classification accuracy. In particular, after performing feature selection over a given (training) dataset, the classifiers trained by the reduced dataset perform very similar to the ones trained by the original dataset. On the other hand, performing instance selection is likely to make the classifiers perform more poorly than is the case without instance selection.

In other words, data preprocessing, especially feature selection, is advantageous for image mining and visual recognition. That is, using a well-designed algorithm, such as GA, can largely reduce the dimensionality of a given (training) dataset, which can not only lower the storage



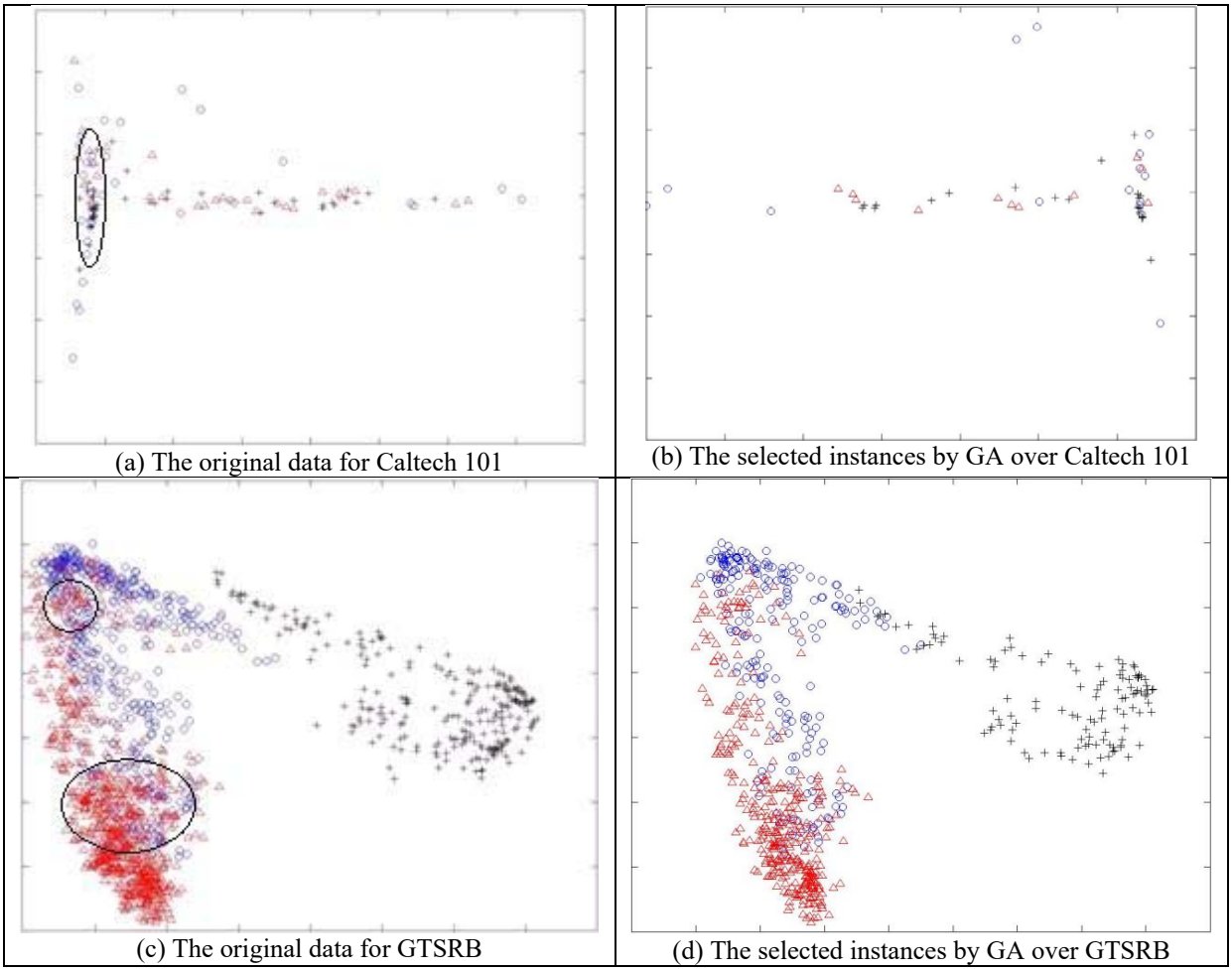
(a) The original features of Caltech 101

(b) The selected feature by GA over Caltech 101

(c) The original features of GTSRB

(d) The selected features by GA over GTSRB

Figure 1 Caltech 101 and GTSRB with and without feature selection



(a) The original data for Caltech 101

(b) The selected instances by GA over Caltech 101

(c) The original data for GTSRB

(d) The selected instances by GA over GTSRB

Figure 2Caltech 101 and GTSRB with and without instance selection

requirements, but also allow the construction of the classifiers with lower computational cost, i.e.,more efficient classifier training. Moreover, the classification accuracy of the trained classifier based on the reduced (training) dataset is comparable to that of the one trained using the original (training) dataset.

Several issues can be considered in future. First, since there are many feature and instance selection algorithms proposed in the literature, it would be valuable to compare GA with other related algorithms for image classification. Second, it has been found that by combining feature selection and instance selection in a two-stage data preprocessing approach, it may be possible to produce a more representative dataset [24]. Therefore, the effect of performing both feature and instance selection by GA on image classification can also be examined. Third, some classifiers, such as decision trees, can handle feature selection internally and can also be compared for further examination.

REFERENCES

[1] J. Han and M. Kamber, Data mining: concepts and techniques. Morgan Kaufmann, 2000.

[2] S.F. Crone, S. Lessmann, and R. Stahlbock, “The impact of preprocessing on data mining: an evaluation of classifier sensitivity in direct marketing,” European Journal of Operational Research, vol. 173, no. 3, pp. 781-800, 2006.

[3] D.R. Wilson and T.R. Martinez, “Reduction techniques for instance-based learning algorithms,” Machine Learning, vol. 38, no. 3, pp. 257-286, 2000.

[4] T. Reinartz, “A unifying view on instance selection,” Data Mining and Knowledge Discovery, vol. 6, no. 2, pp. 191-210, 2002.

[5] W.B. Powell, Approximate dynamic programming: solving the curses of dimensionality. Wiley-Interscience, 2007.

[6] I. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” Journal of Machine Learning Research, vol. 3, pp. 1157-1182, 2003.

[7] J. Derrac, S.García, and F. Herrera, “A survey on evolutionary instance selection and generation,” International Journal of Applied Metaheuristic Computing, vol. 1, no. 1, pp. 60-92, 2010.

[8] M.L. Raymer, W.F. Punch, E.D. Goodman, L.A. Kuhn, and A.K. Jain, “Dimensionality reduction using genetic algorithms,” IEEE Transactions on Evolutionary Computation, vol. 4, no. 2, pp. 164-171, 2000.

[9] J.R. Cano, F. Herrera, and M. Lozano, “Using evolutionary algorithms as instance selection for data reduction: an experimental study,” IEEE Transactions on Evolutionary Computation, vol. 7, no. 6, pp. 561-575, 2003.

[10] M. Kudo and J. Sklansky, “Comparison of algorithms that select features for pattern classifiers,” Pattern Recognition, vol. 33, pp. 25-41, 2000.

[11] W. Hsu, M. Lee, and J. Zhang, “Image mining: trends and developments,” Journal of Intelligent Systems, vol. 19, no. 1, pp. 7-23, 2002.

[12] C.A. Bhatt and M.S. Kankanhalli, “Multimedia data mining: state of the art and challenges,” Multimedia Tools and Applications, vol. 51, no. 1, pp. 35-76, 2011.

[13] M. Dash and H. Liu, "Feature selection methods for classifications," Intelligent Data Analysis, vol. 1, no. 3, pp. 131-156, 1997.
[14] Ghosting, A., Parthasarathy, S., and Otey, M.E. (2008) Fast mining of distance-based outliers in high-dimensional datasets. Data Mining and Knowledge Discovery, vol. 16, pp. 349-364.
[15] D.E. Goldberg, Genetic algorithms in search optimization and machine learning. Addition Wesley, 1989.
[16] P.G. Espejo, S. Ventura, and F. Herrera, "A survey on the application of genetic programming to classification," IEEE Transactions on Systems, Many, and Cybernetics – Part C: Applications and Reviews, vol. 40, no. 2, pp. 121-144, 2010.
[17] R. Sikora and S. Piramuthu, "Framework for efficient feature selection in genetic algorithm based data mining," European Journal of Operational Research, vol. 180, no. 2, pp. 723-737, 2007.
[18] T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y. Zheng, "NUS-WIDE: a real-word web image database from National University of Singapore," ACM International Conference on Image and Video Retrieval, article no.: 48, 2009.
[19] J. Sivic and A. Zisserman, "Video Google: a text retrieval approach to object matching in videos," IEEE International Conference on Computer Vision, pp. 1470-1477, 2003.
[20] S. Lazebnik, C.Schmid, and J. Ponce, "Beyond bags of features: spatial pyramid matching for recognizing natural scene categories," IEEE International Conference on Computer Vision and Pattern Recognition, pp. 2169-2178, 2006.
[21] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," IEEE Conference on Computer Vision and Pattern Recognition, pp. 886-893, 2005.
[22] J. Stallkamp, M.Schlipsing, J.Salmen, and C. Igel, "The German traffic sign recognition benchmark: a multi-class classification competition," International Joint Conference on Neural Networks, pp. 1453-1460, 2011.
[23] T.F. Cox and M.A.A. Cox, Multidimensional scaling. Chapman and Hall, 2001.
[24] C.-F. Tsai, W.Eberle, andC.-Y. Chu, "Genetic Algorithms in Feature and Instance Selection," Knowledge-Based Systems, vol. 39, pp 240-247, 2013.

★★★