http://ansinet.com/itj

ITJ

ISSN 1812-5638

INFORMATION TECHNOLOGY JOURNAL

ANSInet
Asian Network for Scientific Information
308 Lasani Town, Sargodha Road, Faisalabad - Pakistan

Information Technology Journal 13 (1): 102-109, 2014
ISSN 1812-5638 / DOI: 10.3923/itj.2014.102.109

Image Annotation Based on Joint Feature Selection with Sparsity

Dongping Zhang, Yanjie Li, Huailiang Peng and Yafei Lu
College of Information Engineering, China Jiliang University, 310018, Hangzhou, China

Abstract: A large number of networked digital images have been explosively growing with the internet communication technology and the rapid development of digital imaging technology. Feature selection can represent the image information appropriately and obtain better accuracy, thus, it can improve the performance of annotation. In this study, the algorithm of feature selection is proposed to solve automatic image annotation. This algorithm is a reasonable and effective method of using a linear regression model for multi-label classification. By setting the appropriate threshold value, the final class label can be achieved with better performance to cope with continuous in the linear regression. Therefore, three methods of threshold are used to discretize prediction value. In addition, an iterative algorithm is exploited to optimize the objective function and two kinds of learning methods are utilized to train the algorithm. The results show that this algorithm is effective for feature selection and good performance of being applied to image annotation.

Key words: Feature selection, image annotation, sparsity, supervised learning, semi-supervised learning

INTRODUCTION

With digital cameras and other devices growing popular, the number of images is increasing rapidly. Therefore, how to manage and retrieve the network of multimedia information effectively becomes an urgent problem to be solved. In the past few decades, there's a lot of research on content-based image retrieval, restricted by semantic gap (Liu, 2013; Zhao and Wang, 2013), it cannot satisfy the needs of users completely. Users accustom to query keywords, but human-annotated is also a very laborious work which gave birth to the development of automatic image annotation. Automatic image annotation is to allow the computer to automatically add images to reflect the semantic content of words to no marked images. It is the key to achieve the image semantic retrieval research in the field of image retrieval.

Nie *et al.* (2010a) defined feature selection that selected the most effective feature from a subset in order to reduce the dimension of the feature space. Whether the samples contain unrelated or redundant information directly affect the performance of the classifier, so research on effective method of feature selecting is very important. Ye *et al.* (2013) proposed that data mining K-means algorithm was a similarity measure between samples based on indirect clustering method. From the perspective of improving the prediction accuracy, John and Kohavi (1994) defined feature selection was the process which could increase classification accuracy and reduce the characteristic dimension. The definition (Dash and Liu, 1997) have provided is to select a feature subset as small as possible.

According to available labeled training data, supervised feature selection is able to select discriminative features by using the correlation between labels and features. Duda *et al.* (2001) introduced the Fisher Score (FISHER) is a traditional algorithm used supervised feature selection. However, FISHER computes the weights of each feature and then selects features one by one. The algorithm would neglect the useful information of the association between different features. Wu *et al.* (2010) proposed another algorithm overcomes the shortage of selecting features individually. In unsupervised areas, there is no tag information can be directly used, so it is difficult to choose the discriminant features. Semi-supervised feature selection is increasing used in many applications both labeled and unlabeled data.

Sparsity-based feature selection can reduce the characteristic dimension was proposed by Zhao *et al.* (2010), Yang *et al.* (2011) and Yuan *et al.* (2012) also proposed a method to discriminating features with sparsity. Feng *et al.* (2013) proposed infrared and visible image fusion method based on sparsity which obtained a clear picture. Jiang and Li (2013) proposed semantic annotation similarity to improve and match services. Zhu *et al.* (2003) proposed using Gaussian fields to solve semi-supervised learning problems. Cohen *et al.* (2004) proposed semi-supervised learning method can reduce manual labor and learn the data structure utilized unlabeled data. The Structural Feature Selection with Sparsity (SFSS) exploiting the feature correlation which combines the strengths of semi-supervised learning and joint feature selection. The method is suitable for

Corresponding Author: Dongping Zhang, College of Information Engineering, China Jiliang University, 310018, Hangzhou, China

real-world multimedia understanding applications and considers the labeling cost, the characteristic of multimedia data and the computational efficiency.

SFSS showed a good performance for multi-label classification in image annotation. This study aims to annotate the multi-label image exploiting adaptive threshold determination based on SFSS. The general analyzing process of the study in Fig. 1.

The study is geared towards better image annotation performance by exploiting feature selection. In this section, this study briefly reviews the research on feature selection and semi-supervised learning.

It is an effective tool in machine learning. Feature selection is one of the key problems of pattern recognition, it results a direct impact on classifier accuracy and generalization performance.

Feature selection is a process to select the most effective feature from a set of feature so that it can reduce the dimension of feature space. A good training is the key to learning sample classifier for pattern recognition system. Sample will affect the performance of classifier if the sample contains irrelevant or redundant information. Thus, it is heavy going of effective feature selection methods.

In the literature, there are a lot of feature extraction algorithms. Some typical methods such as FISHER discriminant which the basic idea is the projection, k groups and the p-dimensional data is projected to a direction, so that they are separated as far as possible in the projection between the groups. Although these algorithms typically have a good performance, but it also has some drawbacks. First, it calculates too much data and selects most discriminative feature from each feature individually. Second, calculation is costly when they evalueate features one by one.

In order to improve the method of classification, Ma *et al*. (2012a) proposed sparsity-based feature selection methods. In a lot about the sparse method, L2,1 norm regularization based algorithm has obtained more and more people interested in. Since, L2,1 norm regularization has some challenges because of non-smoothness for solving the optimization problems.

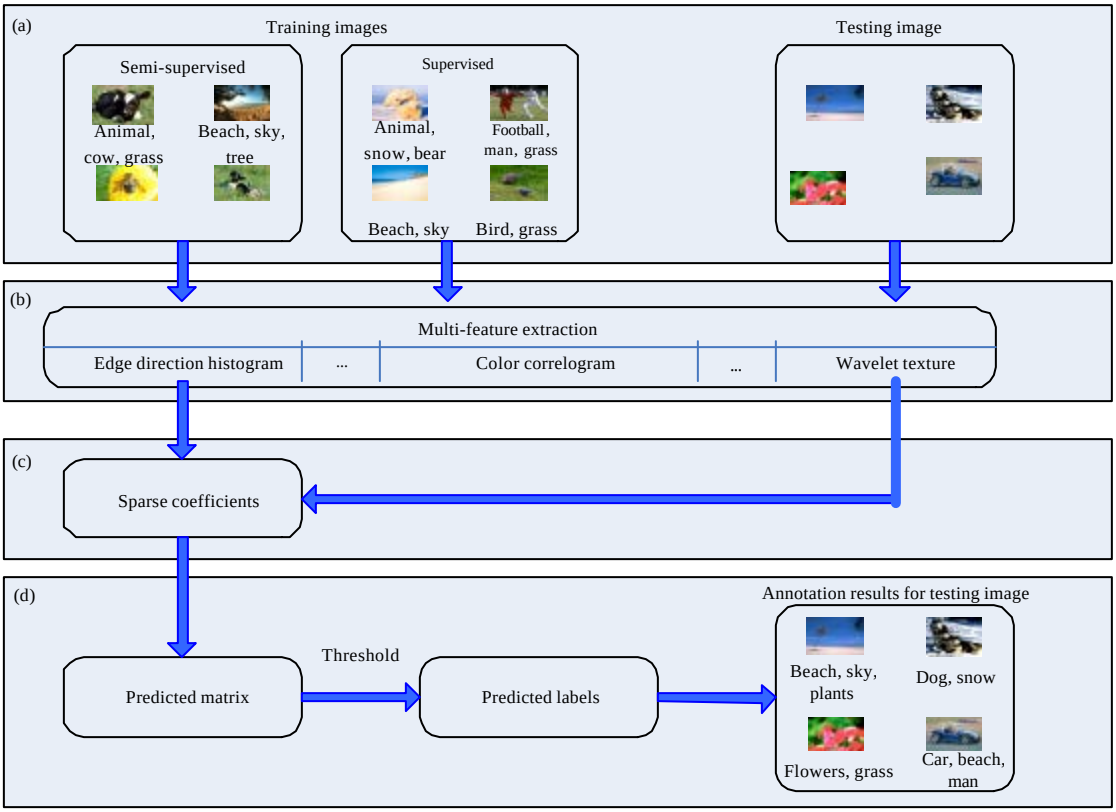

Fig. 1(a-d): General process of the study for image annotation (a) Images of datasets, (b) Muti-feature extraction, (c) Training algorithm and (d) Annotation for testing images

Nie *et al.* (2010b) proposed a method via L21-norm minimization to select feature with efficiently and robustly.

Linear regression is a classical statistical analysis method which can find a relationship between the two variables and hope to get a mathematical model fitting system using the observed data. Therefore, a linear regression model is a reasonable and effective method for multi-label classification. This study establishes a mathematical model using linear regression method regression is:

$$Y = X^TW+1_nb^T \quad (1)$$

where, $X\epsilon R^{d\times n}$ are the training data, $Y\epsilon R^{n\times c}$ are the class labels accordingly. $W\epsilon R^{d\times c}$ is the projection matrix. $I_n\epsilon R^n$ denotes a column vector with all its n elements being 1. $b\epsilon R^{c\times 1}$ is the bias term. d is the dimension of the original feature, n is the number of the training data and c is the number of concepts. W is the projection matrix to correlate X with Y for feature selection.

The study aims to get the W when X and Y is known. So, the label of prediction will be obtained:

$$\hat{Y} = \dot{X}^TW+1_nb^T \quad (2)$$

where, $\dot{X}^T$ is the test matrix and $\dot{n}$ is the number of testing data.

The manifold regularization method is leveraged in many learning problem. Belkin *et al.* (2006) proposed the manifold regularization which exploits the geometry of the marginal distribution. Nie *et al.* (2010a) proposed the flexible manifold embedding which can make use of labeled data effectively. Which the Flexible Manifold Embedding (FME) named a function is $F = X^TW+1_nb^T$. The number of parameters in W does not depend on the number of samples. This function may be overstrict to fit the data sample from a nonlinear manifold.

Adaptive threshold is used to deal with the problem in many areas of the image. Such as the image is divided using the threshold value. This study, threshold-based segmentation, proposes the predicted label matrix is judged by binarization to classify for each image.

Because of the value of label predicted is continuous, the predictable results need to be discretized by setting a threshold value to obtain a final class label. Therefore, the threshold is a key issue in a multi-label classification methods which directly affects the final classification results.

This study proposes three methods of threshold. There are Otsu thresholding method, find the minimum number of the maximum and the method of averaging, respectively.

METHODS

Here, this study elaborates the approach about the process of the algorithm.

Objective function: This study aims to construct a projection matrix W to select feature which mostly related to the concepts of the training data.

Construction the objective function is:

$$\min_W \text{loss}(W) + \gamma\|W\|_{2,1} \quad (3)$$

where, loss (W) is the loss function and γ is a regularization parameter. The definition of $||W||_{2,1}$ is:

$$\|W\|_{2,1} = \sum_{i=1}^{d}\sqrt{\sum_{j=1}^{c} w_{ij}^2} \quad (4)$$

Least square regression: It is a new type of multivariate statistical data analysis method which is a major research on multiple dependent variables regression modeling of multiple independent variables. It can solve the problem that the number of samples is less than the number of variables and good for classification. Ma *et al.* (2012b) supposed that $X\epsilon R^{d\times n}$ indicate the training data, $Y = R^{n\times c}$ are the ground truth labels accordingly. It can solves the following optimization problem using traditional least square regression to obtain the projection matrix $W\epsilon R^{d\times c}$ and the bias $b\epsilon R^c$:

$$\min_{W,b}\sum\left\|x_i^TW+1_nb^T-y_i\right\|_2^2 \quad (5)$$

Laplacian eigenmap (LE): It is a nonlinear dimensionality reduction method and a kind of manifold learning proposed by Belkin and Niyogi (2003). In the observation space in close proximity of the two sample points mapped to the low-dimensional space is also in close proximity to the principle according to LE. It defines a nearest neighbor graph G to imitate and observe local topology space sample points.

LE obtains the minimum loss function is to gain the required output data which set as Y. Minimize the loss function is defined as:

$$\varepsilon(Y) = \arg\ \min\sum_{i=1}^{n}\sum_{j=1}^{n}((y_i-y_j)^2K_{ij}) \quad (6)$$

where, K is the weight matrix between x_i and x_j as:

$$K_{ij} = \begin{cases} 1, & x_i \text{ and } x_j \text{ are k nearest neighbors} \\ 0, & \text{otherwise} \end{cases}$$

LE defines a Laplacian matrix through L = D-K. Where, D is a diagonal matrix of the equation:

$$D_{ii} = \sum_{j=1}^{n} K_{ij}$$

The minimization problem of loss function can be simplified as:

$$\begin{cases} \arg \min\limits_{Y} Y^T L Y \\ Y^T D Y = 1 \end{cases} \quad (7)$$

Because $Y = X^T W + 1_n b^T$ can be rewritten as:

$$\arg \min_{W} \mathrm{Tr}(W^T X L X^T W) \quad (8)$$

Consequently, by applying Manifold Regularization to the loss function in Eq. 3:

$$\arg \min_{W,b} \mathrm{Tr}(W^T X L X^T W) + \mu \left\| X_i^T W + 1_n b^T - Y_i \right\|_F^2 + \gamma \|W\|_{2,1} \quad (9)$$

where, Tr(.) denotes the trace operator. μ and γ are regularization parameters.

The algorithm aims to obtain the optimal W which affected by the ground truth Y unless all the training data on the label have contributed. Therefore, this study defines a predicted label matrix as $F = [f_1, f_2,..., f_n]^T \epsilon R^{n \times c}$ for all the training data in X. Consequently, the objective function becomes:

$$\begin{aligned} &\arg \min \mathrm{Tr}(F^T L F) + \mathrm{Tr}((F-Y)^T U (F-Y)) \\ &+ \gamma \|W\|_{2,1} + \mu \left\| X^T W + 1_n b^T - F \right\|_F^2 \end{aligned} \quad (10)$$

Otsu thresholding method: It was proposed on the basis of the least squares method by Ostu. Suppose M be the original gray level and n_i is the number of pixels in the gray level i, normalized gray value:

$$P_i = \frac{n_i}{M} \quad (11)$$

The gray-scale is divided into two categories and let t be split threshold. Probability of each column is:

$$w_0 = \sum_{i=0}^{t} P_i$$

and:

$$w_1 = \sum_{i=t+1}^{M-1} P_i$$

Average gray of each class are:

$$\mu_0 = \frac{\mu(t)}{w_0}$$

and:

$$\mu_1 = \frac{\mu_T(t) - \mu_t(t)}{1 - w_0}$$

Where:

$$\mu(t) = \sum_{i=0}^{t} i \times P_i \text{ and } \mu_T(t) = \sum_{i=t+1}^{M-1} i \times P_i$$

Then the between-class variance is defined as:

$$\sigma^2 = w_0(\mu_0 - \mu_T)^2 + w_1(\mu_0 - \mu_T)^2 \quad (12)$$

The between-class variance is the best segmentation threshold value when loop from 1-M increments.

Find the minimum number of the maximum: It is a kind of adaptive threshold and can be used to test the labeled sample. Y is known when X is trained sample in unsupervised. $Y = [y_1, y_2,..., y_n]^T \epsilon \{0, 1\}^{n \times c}$ is the lable matrix. Y_{ij} indicate the jth datum of y_i and $Y_{ij} = 1$ if x_i is in the jth class while $Y_{ij} = 0$ otherwise.

$M = [m_1, m_2,..., m_c] \epsilon R^{1 \times c}$ is the number of 1 in each column form Y and $\hat{Y} = [\hat{y}_1, \hat{y}_2,..., \hat{y}_n]^T$ is the predicted matrix. $\dot{n}$ is the number of test samples. Finding the number of m_j maximum numbers from $\hat{y}_j$, then, among $\hat{y}_j$ to find the minimum one is the threshold value of jth class.

The threshold value is $T_j = \inf E$, E denotes the m_j maximum numbers from $\hat{y}_j$.

Adaptive threshold of averaging: This is a very high accuracy rate algorithm in this study. There are $\dot{n}$ numbers in the jth datum of $\hat{Y} = [\hat{y}_1, \hat{y}_2,..., \hat{y}_n]^T$ and to calculate the average number $\dot{n}$ of column j. The jth average is the threshold value of jth class. The threshold value T_j is:

$$T_j = \frac{\sum_{i=1}^{n} \hat{Y}_{ij}}{\dot{n}} \quad (13)$$

The final new prediction label is:

$$Z = \begin{cases} 1, & \hat{Y}_{\cdot j} \geq T_j \\ 0, & \hat{Y}_{\cdot j} < T_j \end{cases} \quad (14)$$

Where:

$$Z = [z_1, z_2, ..., z_n]^T \epsilon \{0, 1\}^{n \times c}$$

RESULTS

Here, this study tests the performance of SFSS. There are three experiments about the selected feature of image annotation performance.

Experimental datasets: There are two datasets MSRA-MM 2.0 and NUS-WIDE are used in this study. The following is a brief description of the two datasets:

- **NUS-WIDE:** The dataset is a web image dataset created by Lab for Media Search in National University of Singapore. The dataset includes 269,648 images and the associated tags from Flickr, with a total number of 5,018 unique tags. six types of low-level features extracted from these images include (1) 64-D color histogram, (2) 144-D color correlogram, (3) 73-D edge direction histogram, (4) 128-D wavelet texture and (5) 225-D block-wise color moments and 500-D bag of words based on SIFT descriptions. Ground-truth for 81 concepts that can be used for evaluation. The images are represented by portfolio consists of Edge Direction Histogram, Color Correlogram and Wavelet Texture
- **MSRA-MM:** The dataset consists of two parts of the image and video. A total of 65,443 images include 68 concepts, each about 1,000 images. All images are Microsoft's online collection, each image has a related standard, namely: Very association, association, not associated with the corresponding three associated values are 2, 1, 0. For the image data, the effective features include (1) 225D black-wise color moment, (2) 64D HSV color histogram, (3) 256D RGB color histogram, (4) 144D color correlogram (5) 75D edge histogram, (6) 128D textures and (7) 7D facial features and so for each sample is 899-dimensional. Three feature types used in this study, namely Edge Direction Histogram, Color Correlogram and Wavelet Texture

Compared algorithms: The study compares the performance of FISHER algorithm. It is a classical method and selects the most discriminative features by evaluating the importance of each feature individually.

Experimental parameters: In the experiments, researchers must adjust two types of parameters. K is a parameter which defines the K nearest neighbor is used in the calculation LE and fixed it at 15. The other one is the regularization parameters which are represented as μ and γ in Eq. 9. The study tunes them from $\{10^{-3}, 10^{-2}, 10^{-1}, 1, 10, 10^2, 10^3\}$ and report the best results.

This study use Mean Average Precision (MAP) as the evaluation metric to evaluate the classification performance.

Experimental datasets: This study selects six different percentages which are labeled sample proportion of the total sample. Specific details of the settings in Table 1.

For semi-supervised training: This study selects some samples with labeled and unlabeled to experiment in the case of semi-supervised. The results can be seen about the SFSS performance changes.

The performance comparison SFSS and FISHER shows in Fig. 2. The following observations from the experimental results (1) Classification performance is improved with the labeled increase in the proportion of the training data, (2) SFSS classification performance is much higher than FISHER and (3) Performance will increase rapidly when the training sample accounted for 50%.

Select features performance: This experiment was to test the performance and the calculated validity about SFSS selected features in the case of supervised for the two image datasets.

Figure 3 shows the performance variation w.r.t the number of selected features in terms of MAP. The observations can be found from the experimental result (1) The MAP is the lowest when the number of selected features is too small, (2) MAP increases as the number of selected features increases, (3) MAP is growing quickly along with the growing number of selected features and (4) MAP largest position in the number of features that all feature.

Prediction accuracy of the label: Using three different threshold methods for prediction accuracy of labels for different samples judgment unsupervised. The observations are found in the Fig. 4. (1) Ostu threshold method has the highest accuracy rate in the NUS-WIDE,

Table 1: Settings of the training sets which generated consisting of n samples from NUS-WIDE and MSRA-MM

Parameter	Size (n)[a]	Labeled (m)[b] (%)
NUS-WIDE	3000	1, 5, 10, 25, 50, 100
MSRA-MM	3000	1, 5, 10, 25, 50, 100

a: Note that the size is the No. of samples, b: Note that labeled percentage is the percentage of labeled images in the samples

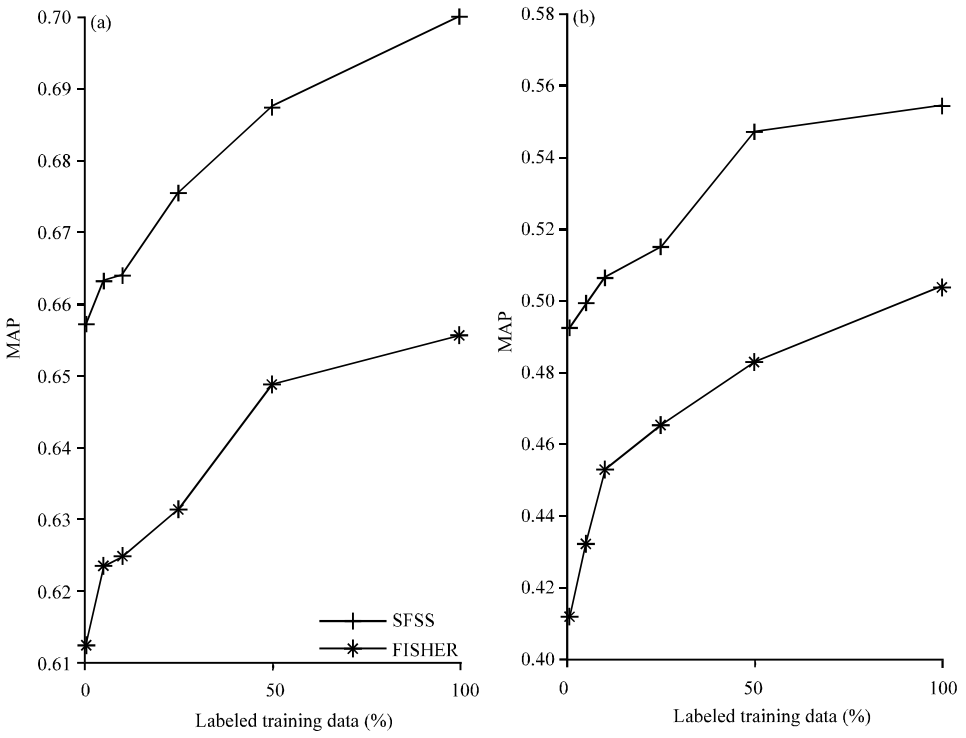

Fig. 2(a-b): Performance comparison of image annotation (a) NUS-WIDE and (b) MSRA-MM

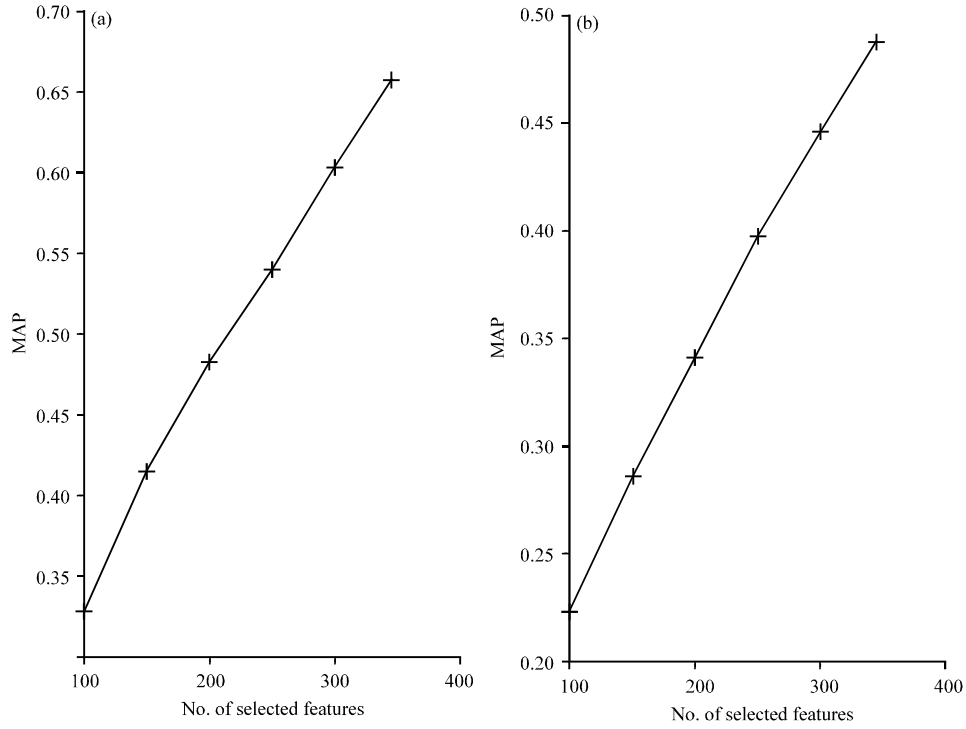

Fig. 3(a-b): Performance variation w.r.t the No. of selected features in terms of MAP (a) NUS-WIDE and (b) MSRA-MM

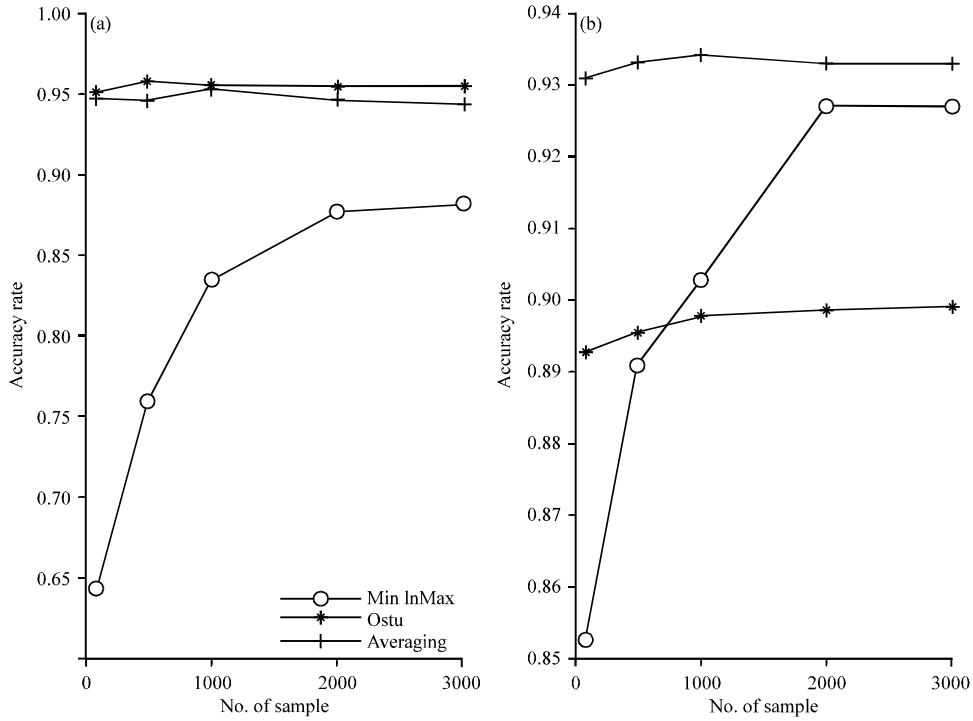

Fig. 4(a-b): Accuracy of three kinds of threshold (a) NUS-WIDE and (b) MSRA-MM

(2) Averaging threshold method is the highest accuracy rate in the MSRA-MM, (3) Both Ostu and averaging of the accuracy are very stable and little variation and (4) Overall, the average threshold method is the best because of high accuracy and high stability.

CONCLUSION

In this study, researchers propose a feature selection method and appliy it into image annotation. The method takes advantage of manifold regularization, joint feature selection with sparsity and transductive classification. Optimization algorithms have been proposed to solve the non-smoothness of the objective function and be applied to image annotation with good performance. The results of experiment have demonstrated that the method of this study outperforms the FISHER algorithms and the average threshold method is the best for high accuracy and stability.

ACKNOWLEDGMENTS

This study was supported by Zhejiang Provincial NSF (Grant No. Y1110506) Zhejiang Provincial Science and Technology Research Program of Zhejiang Province (Grant No. 2013C33046) and Zhejiang Provincial Natural Science Foundation of China (LY13H180011).

REFERENCES

Belkin, M. and P. Niyogi, 2003. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput., 15: 1373-1396.

Belkin, M., P. Niyogi and V. Sindhwani, 2006. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. J. Mach. Learn. Res., 12: 2399-2434.

Cohen, I., F. Cozman, N. Sebe, M.C. Cirelo and T.S. Huang, 2004. Semisupervised learning of classifiers: Theory, algorithms and their application to human-computer interaction. IEEE Trans. Pattern Anal. Mach. Intell., 26: 1553-1567.

Dash, M. and H. Liu, 1997. Feature selection for classification. Intell. Data Anal., 1: 131-156.

Duda, R.O., P.E. Hart and D.G. Stork, 2001. Pattern Classification. 2nd Edn., John Wiley and Sons, New York.

Feng, X., X. Wang, J. Dang and Y. Shen, 2013. Fusion method for visible light and infrared images based on compressive sensing of non-subsampled contourlet transformation sparsity. Inform. Technol. J., 12: 672-679.

Jiang, X. and Y. Li, 2013. Web service matching based on natural semantic annotation. Inform. Technol. J., 12: 857-861.

John, G. and R. Kohavi, 1994. Irrelevant features and the subset selection problem. Proceedings of the 11th International Conference on Machine Learning, July 10-13, 1994, Morgan Kaufmann Publishers, pp: 121-129.

Liu, X., 2013. Semantics oriented inference of keyword search intention over XML documents. Inform. Technol. J., 12: 51-60.

Ma, Z., F. Nie, Y. Yang, J.R.R. Uijlings and N. Sebe, 2012a. Web image annotation via subspace-sparsity collaborated feature selection. IEEE Trans. Multimedia, 14: 1021-1030.

Ma, Z., F. Nie, Y. Yi, J.R.R. Uijlings, N. Sebe and A.G. Hauptmann, 2012b. Discriminating joint feature analysis for multimedia data understanding. IEEE Trans. Multimedia, 14: 1662-1672.

Nie, F., D. Xu, I.W. Tsang and C. Zhang, 2010a. Flexible manifold embedding: A framework for semi-supervised and unsupervised dimension reduction. IEEE Trans. Image Process., 19: 1921-1932.

Nie, F., H. Huang, C. Xiao and C. Ding, 2010b. Efficient and Robust Feature Selection via Joint L21-Norms Minimization. In: Advances in Neural Information Processing Systems 23, Lafferty, J. (Ed.). Curran Associates Inc., USA., pp: 1813-1821.

Wu, F., Y. Yuan and Y. Zhuang, 2010. Heterogeneous feature selection by group lasso with logistic regression. Proceedings of the International Conference on Multimedia, October 25-29, 2010, Firenze, Italy, pp: 983-986.

Yang, Y., H.T. Shen, Z. Ma, Z. Huang and X. Zhou, 2011. L2, 1-norm regularized discriminative feature selection for unsupervised learning. Proceedings of the 22nd International Joint Conference on Artificial Intelligence, July 16-22, 2011, Barcelona, Catalonia, Spain, pp: 1589-1594.

Ye, L., C. Qiuru, X, Haixu, L. Yijun and Z. Guangping, 2013. Customer segmentation for telecom with the k-means clustering method. Inform. Technol. J., 12: 409-413.

Yuan, X., X. Liu and S. Yan, 2012. Visual classification with multitask joint sparse representation. IEEE Trans. Image Process., 21: 4349-4360.

Zhao, P. and X. Wang, 2013. Semantic event mining based on hierarchical structure for soccer video. Inform. Technol. J., 12: 113-119.

Zhao, Z., L. Wang and H. Liu, 2010. Efficient spectral feature selection with minimum redundancy. Proceedings of the 24th AAAI Conference on Artificial Intelligence, July 11-15, 2010, Atlanta, GA., USA., pp: 673-678.

Zhu, X., Z. Ghahramani and J. Lafferty, 2003. Semi-supervised learning using gaussian fields and harmonic functions. Proceedings of the 20th International Conference on Machine Learning, August 21-24, 2003, Washington, DC., USA., pp: 912-219.