Journal of Applied Sciences

ISSN 1812-5654

science alert

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 15 (8): 1133-1137, 2015
ISSN 1812-5654
© 2015 Asian Network for Scientific Information

RESEARCH ARTICLE OPEN ACCESS

DOI: 10.3923/jas.2015.1133.1137

Improvement of Outliers Detection in Image Classification

[1]Asma Lak, [2]Naser Parhizgar and [1]Mahbobeh Lak
[1]Department of Electrical Engineering, College of Engineering, Genaveh Branch, Islamic Azad University, Genaveh, Iran
[2]Department of Electrical Engineering, College of Engineering, Shiraz Branch, Islamic Azad University, shiraz, Iran

ARTICLE INFO
Article History:
Received: June 12, 2015
Accepted: August 29, 2015

Corresponding Author:
Asma Lak
Department of Electrical Engineering,
College of Engineering,
Genaveh Branch,
Islamic Azad University,
Genaveh, Iran

ABSTRACT
In this study a method has been proposed by using of the local information of image pixels in outlier identification to reduce the time of image classification. The applied algorithm is Expectation Maximization (EM algorithm) which is an iterative algorithm. In this algorithm, in each step, outliers are detected then removed in order to prevent error propagation in next steps. By decreasing the errors in each step the validation and accuracy of the classification is increased. Thus we proposed a method which use from the mean entropy of pixels which are in neighborhood of first, second and third edge pixels of mixture to the image. By using of this method the time of classification of a typical image (AVIRIS hyperspectral image) has been improved.

Key words: Outlier, EM algorithm, mixture, classification

INTRODUCTION

Labels are difficult to obtain and unlabeled data is abundant for image pixels, therefore semi-supervised learning is a good idea to reduce human labor and improve classification accuracy (Rosset *et al*., 2008). Von Luxburg *et al*. (2005) has been presented a method for hyperspectral images classification. Two data mining techniques to detect outliers are: The Bay's algorithm for distance-based outliers (Bay and Schwabacher, 2003) and a density-based local outlier algorithm (Breunig *et al*., 2000). There are some pixels in semi-supervised classification which are labeled wrongly. This problem affects on estimation, covariance matrix and means value and finally leads to decreasing the accuracy and validation of classification. These samples are called outlier. By using of outlier detection in multivariate analysis these pixels can be identified and before implementing of multivariate methods they can be used as a preprocessing. At the end of this prepossessing these pixels are removed and the destructive effects of them on classification are eliminated (Lak *et al*., 2013).

In this study edge pixels of mixtures (boundary samples) of any class of image are determined and mean entropy value of these boundary samples are used. This trend is done in any stage of EM algorithm and outliers in each stage are identified then removed and don't use in the next stages of the algorithm (Lak *et al*., 2013).

The aim of this study is to propose a method for more accuracy of outlier detection which reduces the repetition rate of the EM algorithm and reduces the time of classification finally.

MATERIALS AND METHODS

The image which is used in this study related to agriculture-forest area which is taken by AVIRIS in June 1992

in Indiana. This image has 145×145 pixels and radiometric accuracy is 8 bits. The main problem was that most plants of this area are corn and soya. They are at the beginning of the period of growth. In other hand, that area was an experimental site and some of the cultivated products of previous years remained on it. Based on cultivated plants and conditions of ground each area has 16 various classes (Rosset *et al*., 2008; Von Luxburg *et al*., 2005).

Outlier identification: When the image by using of number of classes is classified (these classes are determined formerly) the pixels are placed in the classes and any image is labeled. May be some of these estimation labels are wrong for some pixels. These samples which are labeled wrongly called "outlier". The EM algorithm has been used to classification (Acuna and Rodriguez, 2004). This algorithm is iterative and each stage uses the results of previous stage. Therefore, if an error is occurred, it will enter in next stage and repeat. As a result the errors propagate in next stages. Proposed method is comparison of value of "Entropy×weight" for any pixels with "Mean entropy×weight" for any sample of each class. If the image has i class and any class has m mixture, by estimation of mean vector μ_i and covariance matrix Σi for each class, i, outlier are detected as: one stage of EM is run and p(x|i) is calculated that is probability of belonging of x to i class.

Value of entropy for all pixels of image is:

$$\text{Entropy}(x) = -\sum_{k=1}^{i} p(x \mid i) \log p(x \mid i)$$

By taking in account of training samples the mixture of each class is determined and by using an edge detector the boundary samples are identified. The boundary samples of mixture of each class means that, pixels which are placed at the boundary of a mixture in the training labeled samples. For boundary samples of each mixture, the neighbor pixels are determined by:

Fi = [x y-1, x y+1, x-1 y, x-1 y-1,
x-1 y+1, x+1 y, x+1 y-1, x+1 y+1]

Se = [x y-2, x y+2, x+1 y-2, x+1 y+2, x-1 y-2, x-1 y+2,
x+2 y-2, x+2 y+2, x+2 y-1, x+2 y+1, x+2 y, x-2 y-2,
x-2 y+2, x-2 y-1, x-2 y+1, x-2 y]

Th = [x+3 y-3, x+3 y-2, x+3y-1, x+3 y, x+3 y+1,
x+3 y+2, x+3 y+3, x+2y-3, x+1 y-3, x y-3,
x-1 y-3, x-2 y-3, x-3 y-3, x-3y-2, x-3 y-1,
x-3 y, x-3 y+1, x-3 y+2, x-3 y+3, x-2y+3,
x-1 y+3, x y+3, x+1 y+3, x+2 y+3]

where, Fi, Se and Th are first, second and third neighborhood, respectively. Since, some of samples of neighbor of pixel "x" that are not labeled may be belonging to another class, the entropy of each neighbor is multiplied in pixels weight. The mean is calculated for these three obtained values. The final mean is calculated for mixtures of each class. To outlier identification in next step, the entropy of each x pixel is compared with final mean of each i class. If the entropy value is smaller than mean value, x pixel is identified as an outlier i for class.

The detected outliers of each class are removed (Becker and Gather, 1999, 2001) in the next stage are classified again. The identification of outliers in each stage of EM algorithm is done. The number of repetition is equal to the number of EM repetition and continues to stop condition of algorithm. In this study 1346 training sample and 2105 experimental samples has been used.

EM algorithm: Mixture model has long been used for semi-supervised learning, e.g. Gaussian Mixture Model (GMM). More information can be found in Ratsaby and Venkatesh (1995), Nigam and Ghani (2000) and Castelli and Cover (1996). In this model training is typically done with the EM algorithm. It has several advantages: The model is inductive and handles unseen points naturally and it is a parametric model with a small number of parameters.

In typical mixture model for classification, the generative process is the following 6: one first select a class *y*, then chooses a mixture component $m \in \{1,..... M\}$ with p(m|y) and finally a point x according to p(m|y) is generated. Thus:

$$p(x \mid y) = \sum_{m=1}^{M} p(y)\, p(m \mid y)\, p(x \mid m)$$

RESULTS

Figure 1a-b shown the classified image in the first and final stages of the algorithm. Actual classification of image is according to Fig. 2. By the comparison between these three images it can be seen the effect of the improved algorithm on classification improvement. X and Y axis show the coordinates of each pixel of the hyperspectral image.

Figure 3 shows the variations of classification validity for some class and total accuracy for classification.

Figure 4 shows the comparison between proposed method and previous ones in related to spent time for each repetition of algorithm. As it can be seen by improving algorithm, the time of each repetition and total time has been reduced by reduction of number of repetition.

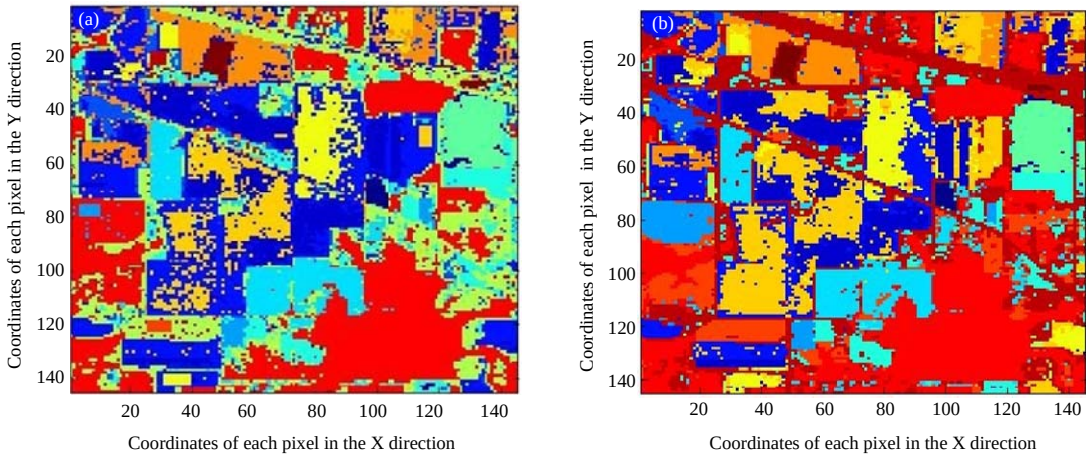

Fig. 1(a-b): (a) First step and (b) Final step of the proposed algorithm

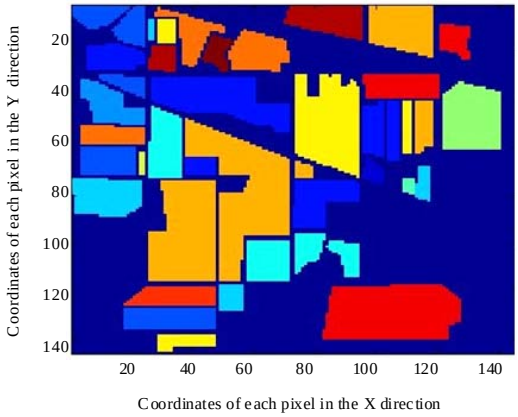

Fig. 2: Thematic image

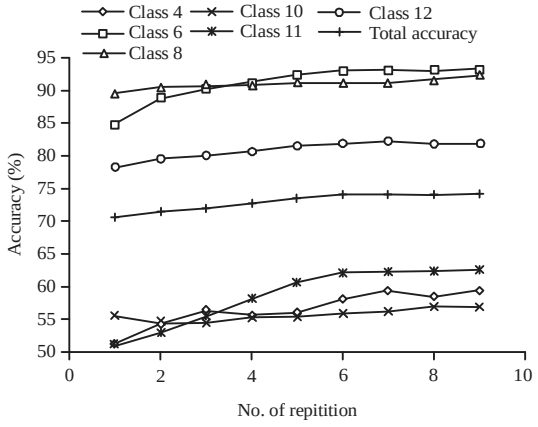

Fig. 3: Accuracy for some classes and total accuracy

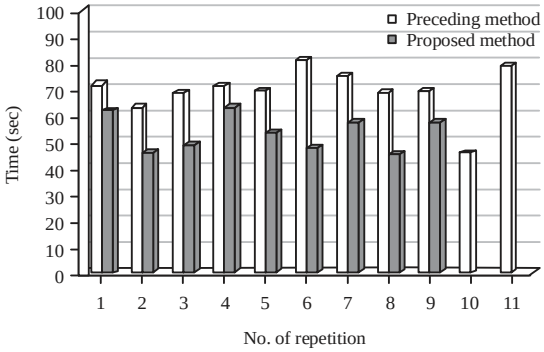

Fig. 4: Comparison of the time of each step of EM algorithm with previous method

DISCUSSION

Previous study in this area includes modifying the clustering objective function so that it includes a term for satisfying specified constraints (Demiriz *et al*., 1999) and enforcing constraints to be satisfied during the cluster assignment in the clustering process (Wagstaff *et al*., 2001).

Zhu *et al*. (2003) has been introduced an approach to semi-supervised learning based on a Gaussian random field model defined with respect to a weighted graph representing labeled and unlabeled data. Promising experimental results have been presented for text and digit classification, demonstrating that the framework has the potential to effectively exploit the structure of unlabeled data to improve classification accuracy.

Zho (2005) has been introduced an EM algorithm for semi-supervised classification of a hyperspectral image. This method is iterative. In each step of algorithm, a label has been estimated for pixels without label and the classification has been finished after reaching the stop condition.

Lak *et al*. (2013) has been improved the previous algorithm. In cases that the training samples are not good enough (are few) for a class, a method has been presented for improving the covariance matrix estimation. The outliers have been detected by using of the local information of first and

Table 1: Accuracy for each class and total accuracy

	No. of Iterations								
Class No.	1	2	3	4	5	6	7	8	9
1	87.0370	87.0370	87.0370	87.0370	87.0370	87.0370	87.0370	87.0370	87.0370
2	65.0628	65.899	66.3180	66.4575	66.8061	66.7364	67.0851	66.8759	66.7364
3	73.6211	74.1007	73.6211	73.0216	72.9017	73.2614	72.7818	71.5827	71.9424
4	51.2821	54.2735	56.4103	55.5556	55.9829	58.1197	59.4017	58.5470	59.4017
5	81.8913	82.0926	81.4889	81.4889	81.2877	81.2877	81.4889	81.2877	81.2877
6	84.7390	88.8889	90.2276	91.1647	92.3695	93.0388	93.0388	93.1727	93.0388
7	100.0000	100.0000	100.0000	100.000	100.0000	100.0000	100.0000	100.0000	100.0000
8	89.5706	90.5930	90.7975	90.7975	91.2065	91.2065	91.2065	91.6155	92.4335
9	100.0000	100.0000	100.0000	100.000	100.0000	100.0000	100.0000	100.0000	100.0000
10	55.5785	54.5455	54.4421	55.3719	55.3719	55.9917	56.3017	56.9215	57.0248
11	51.1345	52.9579	55.4700	58.2253	60.6969	62.1151	62.2366	62.3987	62.6013
12	78.3388	79.6417	80.1303	80.7818	81.5961	81.9218	82.2476	81.7590	81.9218
13	99.5283	98.5849	97.1698	96.2264	96.2264	96.2264	96.2264	96.2264	96.2264
14	97.2179	97.4498	97.7589	97.7589	97.8362	98.0680	98.0680	98.2226	98.2226
15	67.1053	61.0526	56.5789	55.2632	51.5789	50.5263	50.2632	48.4211	47.6316
16	91.5789	92.6316	92.6316	92.6316	92.6316	92.6316	92.6316	92.6316	92.6316
Total accuracy	70.7119	71.5030	72.1107	72.8439	73.4999	74.0208	74.1366	74.0305	74.1270

second neighbor pixels. Also, a method has been proposed to determine the number of mixtures of each class.

In this work, the results are improved for outlier samples estimation. In this method the local information of the pixels of neighbors is used more than previous method. In the proposed method, the third neighbor is used too. and the outlier samples are detected by it. The results show the smaller number of iteration of EM algorithm and the time of running is reduced to half of previous model.

CONCLUSION

In this study a method has been proposed to detect outlier samples in each step of algorithm and remove them to estimate parameters in next stage. The proposed method by using more information of neighbor pixels of training samples leads to reduce the number of repetition of EM algorithm and in fact by this method the time of reach to the desirable accuracy and validation has been reduced in image classification. Figure 1 and 2 is shown the comparison between proposed algorithm and other methods in outlier sample detection for class 4. In this figures, the red color pixels shows the detected samples.

By implementing of this method, accuracy and validation of classification are 73.5 and 74.2, respectively. The number of repetition of algorithm has been reduced from 11-9 and the time of run reduced from 1370-560 sec. The variation of accuracy of image classes is shown in Table 1.

REFERENCES

Acuna, E. and C.A. Rodriguez, 2004. A meta analysis study of outlier detection methods in classification. Technical Paper, Department of Mathematics, University of Puerto Rico at Mayaguez.

Bay, S.D. and M. Schwabacher, 2003. Mining distance-based outliers in near linear time with randomization and a simple pruning rule. Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, August 24-27, 2003, Washington, DC., USA., pp: 29-38.

Becker, C. and U. Gather, 1999. The masking breakdown point of multivariate outlier identification rules. J. Am. Stat. Assoc., 94: 947-955.

Becker, C. and U. Gather, 2001. The largest nonidentifiable outlier: A comparison of multivariate simultaneous outlier identification rules. Comput. Stat. Data Anal., 36: 119-127.

Breunig, M.M., H.P. Kriegel, R.T. Ng and J. Sander, 2000. LOF: Identifying density-based local outliers. Proceedings of the International Conference on Management of Data, May 15-18, 2000, Dallas, TX., USA., pp: 93-104.

Castelli, V. and T. Cover, 1996. The relative value of labeled and unlabeled samples in pattern recognition with an unknown mixing parameter. IEEE Trans. Inf. Theory, 42: 2102-2117.

Demiriz, A., K.P. Bennett and M.J. Embrechts, 1999. Semi-supervised clustering using genetic algorithms. Proceedings of the Artificial Neural Networks in Engineering Conference, November 7-10, 1999, ASME Press, pp: 809-824.

Lak, M., A. Keshavarz and H. Pourghassem, 2013. Graph-Based Hyperspectral Image Classification Using Outliers Detection Based on Spatial Information and Estimating of the Number of GMM Mixtures. (IEEE) International Conference on Communication Systems and Network Technologies, April 6-8, 2013, Washington, DC., USA., pp: 196-200.

Nigam, K. and R. Ghani, 2000. Analyzing the Effectiveness and Applicability of Co-Training. Ninth International Conference on Information and Knowledge Management, November 6-11, 2000, McLean, VA., USA., pp: 86-93.

Ratsaby, J. and S. Venkatesh, 1995. Learning from a mixture of labeled and unlabeled examples with parametric side information. Proceedings of the 8th Annual Conference on Computational Learning Theory, July 5-8, 1995, Santa Cruz, CA., USA., pp: 412-417.

Rosset, S., J. Zhu, H. Zou and T. Hastie, 2008. A Method for Inferring Label Sampling Mechanisms in Semi-Supervised Learning. In: Advances in Neural Information Processing Systems, Saul, L.K., Y. Weiss and L. Bottou (Eds.)., Chapter 17, MIT Press, Cambridge, MA., pp: 1161-1168.

Von Luxburg, U., O. Bousquet and M. Belkin, 2005. Limits of Spectral Clustering. In: Advances in Neural Information Processing Systems 17, Saul, L.K., Y. Weiss and L. Bottou (Eds.). MIT Press, Cambridge, MA., pp: 857-864.

Wagstaff, K., C. Cardie, S. Rogers and S. Schrodl, 2001. Constrained K-means clustering with background knowledge. Proceedings of the 18th International Conference on Machine Learning, June 28-July 1, 2001, San Francisco, CA., USA., pp: 577-584.

Zho, X., 2005. Semi-supervised learning with graphs. Ph.D. Thesis, Language Technologies Institute , School of Computer Science, Camegie Mellon University, Pennsylvania.

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.