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 12 (23): 7409-7415, 2013
ISSN 1812-5638 / DOI: 10.3923/itj.2013.7409.7415

Research on Combination of Tri-training with Active Learning

[1,2]Yan Zhang, [1]Baoguo Wu and [2]Danjv Lv
[1]School of Information Science and Technology, Beijing Forestry University, Beijing, China
[2]School of Computer and Information, Southwest Forestry University, Kunming, China

Abstract: How to leverage the abundant unlabeled data with a few labeled training examples to construct a strong classification system is a focus issue. Both semi-supervised learning and active learning attempt to exploit the unlabeled data to improve the recognition rate of supervised learning algorithms and minimize the cost of data labeling. This paper proposed two approaches,Entropy Priority Sampling (EPS) and Simple Disagreement Sampling (SDS),to select samples in active learning , which are applied into the Tri-Training algorithm as Tri-EPS and Tri-SDS methods. Several experiments with these approaches on the UCI, remote sensing image and environmental audio datasets are carried out in order to illustrate the results of the proposed methods and compare their performance with that of Tri-Training algorithm. Experimental results show that the active learning combined with semi-supervised learning can effectively improve the performance.

Key words: Semi-supervised learning, active learning, Tri-training, Tri-EPS, Tri-SDS

INTRODUCTION

In many practical applications, such as text classification, bioinformatics, the remote sensing image classification and webpage classification, for example, in remote sensing image classification, the large number of spectral bands of modern sensors and the large number of land-cover classed of interest, require a number of training examples that are too expensive or tedious to acquire. Instead, collecting unlabeled data is often easy and inexpensive. In remote sensing fields, thanks to the high spatial resolution of modern sensors, a large number of unlabeled examples become available when new images are captures. It is easy to see that such application demand classification methods that achieve high accuracy using only a few labeled but many unlabeled examples (Zhu, 2006; Zhou and Wang, 2007; Zhou *et al*., 2007).

How to use a small labeled data and a lot of unlabeled data to improve the learning performance becomes the key problem, which the pattern recognition and machine learning researchers are focusing on. Semi-supervised learning (Seeger, 2001) is an effective way to utilize unlabeled data in assistance of supervised learning. The main idea is that an initial hypothesis is learned from the labeled set and then refined through information derived from the unlabeled set. Besides semi-supervised learning, active learning is another effective way to use the unlabeled examples. Different from semi-supervised learning choosing confident examples to label by itself, active learning actively chooses most confident examples from the unlabeled pool and asks a teacher for the labels. Semi-supervised learning and active learning are methods for exploiting unlabeled data in addition to labeled data automatically to improve learning performance when the labeled training data set is insufficient. They are the emerging fields of research in the machine learning and can obtain the better learning generalization performance and learning effect through the unlabeled examples information auxiliary to complete the learning model.

The rest of the paper is organized as follows. Section 2 briefly reviews semi-supervised learning and some related work with semi-supervised learning and active learning. Section 3 presents two approaches to combine semi-supervised and active learning. Section 4 reports the experiment results. Finally, section 5 concludes and issues some future work.

RELATED WORK

Semi-supervised learning and active learning tackle the same problem but they utilize unlabeled data in different way. Semi-supervised learning exploits what the underlying classifiers are most confident from the unlabeled data and active learning exploits the least confident ones (Hady and Schwenker, 2010). Therefore, their merits can be combined through some specifically designs.

Corresponding Author: Baoguo Wu, School of Information Science and Technology, Beijing Forestry University, Beijing, China

Two major techniques in active learning are uncertainty-base sampling (Lewis and Gale, 1994) and committee-base sampling (Seung *et al*., 1992). Query by Committee (QBC) (Freund *et al*., 1997) is a committee-based active learning algorithm, in which an ensemble of diverse classifiers is constructed. Then the ensemble members are applied to unlabeled examples. The most informative examples, the least confident ones on which the ensemble members are mostly disagree, are selected. Then an expert is asked to assign labels to these examples and then the committee is re-trained using the augmented training set.

McCallum and Nigam (1998) combined semi-supervised EM with committee-based sampling in text classification. The results have shown that combing QBC and semi-supervised EM outperform both of them. (Muslea *et al*., 2000, 2002) employed co-testing to choose unlabeled examples to query and used co-EM to boost the accuracy of the hypotheses. Zhou *et al.* (2004) combined co-training with co-testing in content-based image retrieval, which is used to exploit unlabeled images to improve the performance of content-based image retrieval that integrates the benefits of active learning and semi-supervised learning.

Hady and Schwenker (2010) proposed a new committee-based single-view Co-Training style algorithm for semi-supervised learning, named as CoBC, for application domains in which the available data is not described by multiple redundant and independent views. In addition, two new algorithms, QBC-then-CoBC and QBC-with-CoBC, combine the merits of committee-based active learning and committee-based semi-supervised learning.

COMBING ACTIVE LEARNING WITH TRI-TRAINING

Active learning methods: In active learning, the learning process repeatedly queries unlabeled samples to select the most informative examples and updates the training set on the basis of a supervisor who attributes the labels to the selected samples. The query function selects samples from the unlabeled pool, which have maximum ambiguity to belong to each class (Fig. 1).

The key point in active learning is sampling strategy. Given k classifiers, this paper applied committee-based sample selection techniques to select examples for training. Disagreement among the k committee members can be measured by the entropy of the classifications voted by each member:

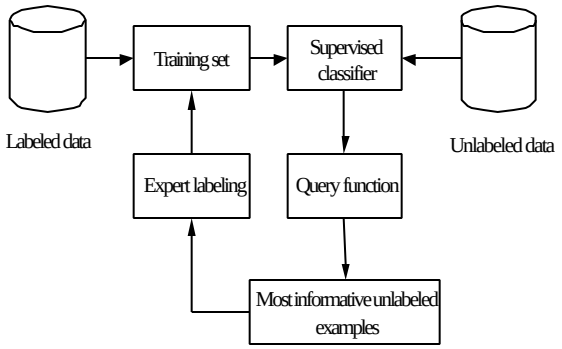

Fig. 1: Description of general active learning

Algorithm: EPS (Entropy Priority Sampling)
Input: L is original labeled data, U is unlabeled data,
$Learn_k$: ($k \geq 3$) learning algorithms
Output: H_{out} the final classifier
Begin
  Train k classifiers H_k on L with $Learn_k$
  Repeat N times
    L_a←Ø;
    For each $x_i \in U$
      H_k (x_i) (k = 1, 2, ...)
    Compute Entropy (x_i)
  End for
  L_a← {x| Entropy(x) is highest}
  U←U-L_a; L_a←Label (L_a);
  L←L∪L_a
  train k classifiers H_k on L with $Learn_k$ ($k \geq 3$)
End repeat
  H_{out}←Ensemble Method(H_k) (k=1,2,…)
End

Fig. 2: Description of Entropy Priority Sampling (EPS)

$$\text{Entropy}(x) = -\sum_{i=1}^{c} p_i \log p_i \quad (1)$$

$$p_i = \frac{V(i)}{k} \quad (2)$$

V(i) is the number of votes of class i, c is the total number of classes. Examples corresponding to higher entropy have priority of selection over others. We named the sampling process as Entropy Priority Sampling (EPS). The process is described in Fig. 2.

When the k learns with k = 2, contention points can be selected by disagreement among the two classifiers of the classification results. The approach is named as Simple Disagreement Sampling (SDS). Fig. 3 shows the description of SDS.

Active learning methods on Tri-training: In order to make better use of the unlabeled data to help the

Algorithm: SDS (Simple Disagreement Sampling)
Input: L is original labeled data, U is unlabeled data, $Learn_k$: (k = 2) learning algorithms
Output: H_{out} the final classifier
Begin
  For i = 1 to 2
    Si←Bootstrap (L)
Repeat N times //N iterations
  H_1←$learn_1$(S_1); H_2←$learn_2$ (S_2)
CPs←Ø
  For each x_i?U
    Cps←CPs ∪ {x_i|x_i?U and $H_1(x_i)$≠$H_2(x_i)$}
  U←U-{CPs}; // remove CPs from the U
  NewL←Label (CPs); //label the contention points
  S_1←S_1∪NewL; S_2←S_2∪NewL;
End Repeat
H_{out}←Ensemble Method (H_1, H_2)
End

Fig. 3: Description of Simple Disagreement Sampling (SDS)

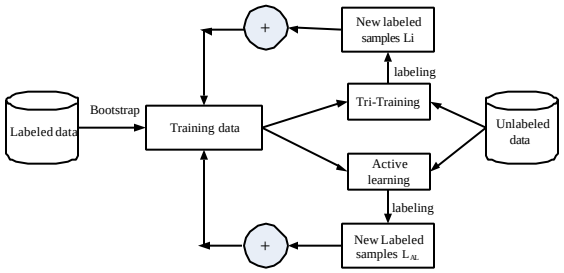

Fig. 4: Combining of selecting the labeled samples from Tri-Training and active learning

classification performance, combining active learning and semi-supervised learning is an effective way to exploit the merits of the unlabeled data.

Tri-training is a typical semi-supervised learning method. In the Tri-Training algorithm, three classifiers H1, H2 and H3, are in same learning algorithm but they differ from training labeled examples. For any classifier, an unlabeled example can be labeled for it as long as the other two classifiers agree on the labeling of this example. The condition of the labeling the unlabeled example is:

$$L_i = \{x \mid x \in U \text{ and } H_j(x) = H_k(x)\}$$

In another way, the contention points can be selected from the unlabeled examples, then labeled by an oracle and employed into the set of the L_i. This can enlarged labeling the unlabeled examples. According to the previous approaches of active learning, we put forward the two methods to enhance the performances of the classification.

Table 1: Information of datasets in the experiments

Data set	No. of instances	No. of attributes	No. of classes
Iris	150	4	3
Ionosphere	351	34	2
Vehicle	846	18	4
Heart_statlog	270	13	2
Tic_tac	958	9	2
wine	178	13	3
Can_tmr	8211	6	6
Envir_audio	6900	11	5

In the round of the each iteration of the Tri-training, the labeling of the unlabeled examples is involved in two ways. One is $L_i = \{x \mid x \in U \text{ and } H_j(x) = H_k(x)\}$, the other L_{ALi} is the selecting samples from approaches of the Entropy Priority Sampling or Simple Disagreement Sampling with active learning. That is, in the end of i-th round iteration, Hi is retrained with the training set as $L_i = L_i \cup L_{ALi}$ (for $i = 1, 2, 3$). In this way, the original training set expanded and new labeled examples produced by active learning sampling strategies in another way are good beneficial supplement to the training set. So the diversity is increasing in the training samples. The process of combining the Tri-Training and active learning to produce the new training set is shown in the Fig. 4.

EXPERIMENTS AND RESULT

In order to compare the performance of the above methods, some experiments are carried out to show the results of the classification. Six UCI datasets (Blake and Merz, 1998), a remote sensing image and environmental audio data are used in the experiments. The dataset can_tmr comes from the remote sensed data, TM image file as "can_tmr.img". The environmental data is acquired from network and field recording, with 8k sampling rate, 16 bits and mono-track. It includes five classes, such as the sound of different kinds of birds, frogs, wind, rain and thunder. The speech sound length amounts to almost 10 minutes. The removal of silence and noise are involved in the pre-processing. Information on these dataset is tabulated in Table 1.

For each dataset, it is 10 times sampling randomly with 25% as testing sample and 75% as training sample. In each pool, L and U are partitioned under different labeled rates including 10, 20, 40, 60 and 80%. Under each labeled rate, the original hypothesis is generated by the tri-training, which base classifier is J48 decision tree.

In order to compare the performance of the semi-supervised learning, active learning and combination

Table 2: Result of under 10% label rate

		Active learning		Tri-+AL	
Data set	Tri-J48	EPS	SDS	Tri-EPS	Tri-SDS
Iris	0.1103	0.1154	0.1333	0.0615	0.0590
Ionosphere	0.2011	0.1830	0.2193	0.2011	0.2034
Wine	0.2689	0.1822	0.1889	0.0756	0.0778
Vehicle	0.3887	0.3599	0.3802	0.3632	0.3325
Heart	0.2750	0.2662	0.2971	0.2176	0.2221
Tic_tac	0.2904	0.2963	0.2921	0.0992	0.0992
Can_tmr	0.0715	0.0576	0.0576	0.0538	0.0541
Envir_audio	0.2211	0.1709	0.1810	0.1791	0.1956

Table 3: Result of under 20% label rate

		Active learning		Tri-+AL	
Data set	Tri-J48	EPS	SDS	Tri-EPS	Tri-SDS
Iris	0.1103	0.0462	0.0564	0.0308	0.0410
Ionosphere	0.1477	0.1068	0.1386	0.1614	0.1636
Wine	0.1800	0.0800	0.1200	0.0511	0.0533
Vehicle	0.3637	0.2816	0.3217	0.3104	0.3042
Heart	0.2603	0.2353	0.2824	0.2206	0.2221
Tic_tac	0.1850	0.1604	0.1892	0.0475	0.0475
Can_tmr	0.0592	0.0496	0.0548	0.0507	0.0504
Envir_audio	0.1943	0.1594	0.1670	0.1656	0.1716

Table 4: Result of under 40% label rate

		Active learning		Tri-+AL	
Data set	Tri-J48	EPS	SDS	Tri-EPS	Tri-SDS
Iris	0.0513	0.0410	0.0462	0.0359	0.0359
Ionosphere	0.1193	0.0943	0.1239	0.1091	0.1114
Wine	0.1111	0.0644	0.0911	0.0267	0.0378
Vehicle	0.3278	0.2741	0.2840	0.2741	0.2693
Heart	0.2515	0.2279	0.2426	0.2059	0.2044
Tic_tac	0.0983	0.0746	0.0896	0.0067	0.0067
Can_tmr	0.0606	0.0451	0.0526	0.0468	0.0469
Envir_audio	0.1801	0.1495	0.1638	0.1416	0.1456

Table 5: Result of under 60% label rate

		Active learning		Tri-+AL	
Data set	Tri-J48	EPS	SDS	Tri-EPS	Tri-SDS
Iris	0.0538	0.0256	0.0385	0.0308	0.0333
Ionosphere	0.1080	0.1045	0.1102	0.1114	0.1136
Wine	0.1178	0.0289	0.0711	0.0200	0.0200
Vehicle	0.2925	0.2524	0.2575	0.2533	0.2608
Heart	0.2132	0.1985	0.2132	0.1971	0.2015
Tic_tac	0.0533	0.0196	0.0192	0.0004	0.0004
Can_tmr	0.0532	0.0460	0.0509	0.0476	0.0468
Envir_audio	0.1650	0.1474	0.1621	0.1409	0.1396

Table 6: Result of under 80% label rate

		Active learning		Tri-+AL	
Data set	Tri-J48	EPS	SDS	Tri-EPS	Tri-SDS
Iris	0.0436	0.0282	0.0308	0.0308	0.0333
Ionosphere	0.1023	0.0955	0.0977	0.1057	0.1045
Wine	0.0889	0.0178	0.0622	0.0200	0.0200
Vehicle	0.2896	0.2354	0.2368	0.2344	0.2425
Heart	0.2544	0.2162	0.2471	0.2338	0.2206
Tic_tac	0.0396	0.0046	0.0183	0.0000	0.0000
Can_tmr	0.0525	0.0439	0.0532	0.0436	0.0442
Envir_audio	0.1661	0.1417	0.1666	0.1357	0.1275

of them, each of them is employed in the experiments. Tri-J48 stands for the semi-supervised learning, the Entropy Priority Sampling (EPS) and the Disagreement Sampling (SDS) represent the active learning. The Tri-EPS and Tri-SDS are the combination the EPS and SDS with Tri-Training respectively. Among them, the EPS uses three different classifiers such as J48 decision tree, KNN and BP algorithm to select the samples. The SDS uses the two classifiers i.e., J48 decision tree and BP. The environment of the experiments is on the weak of the Java development (Witten *et al*., 2011).

The experimental results are the average of the ten times run. The averaged results are summarized in Table 2, 3, 4, 5 and 6, which present the classification error of the methods i.e. Tri-training (J48), EPS, SDS, Tri-EPS and Tri-SDS.

Compared the performance of Tri-J48 with that of active learning i.e., EPS and SDS, the active learning outperforms in most cases. It is critical that the committee be made up of consistent hypotheses that are very different from each other and an effecitve approach to selective sampling in which disagreement amongst an ensemble of hypotheses is used to select data for labeling. In the EPS and SDS methods, three or two different learning algorithms are uilitied to selelct sampling, so they can obtain the better performance. In most cases, EPS outperforms the SDS and some case they are similar to each other. Only under the 10 percent label rate, SDS obtains the better performance than EPS.

From the tables above, Combination Tri-training method with active learning can effectively improve the hypotheses under all the label rates, especially in Tri-EPS. The performance of Tri-SDS is second to the Tri-EPS and is better than SDS methods. With a small of training set, Tri-SDS and Tri-EPS is stable and all better than the Tri-training and active learning methods, they make full of the unlabeled examples to reduce the amount of training data needed to induce an accurate model.

The Fig. 5 showed the performance of corresponding methods of dataset. Generally, the Tri-EPS, Tri-SDS are better than Tri-Training and the curve trends are smooth and stable. The combination of semi-supervised and active learning classification can get the ideal and stable state while the ratio of the amount of labeled data to the total sample data between 10 and 35%. With the increasing of the number of the labeled examples, the performance of the Tri-EPS and Tri-SDS reduce in some cases, especially the dataset such as heart. Because the unlabeled examples may often be wrongly labeled during the learning process, the enlarged labeled data set for the learner to learn at each iteration could contain much noise and the mislabeled examples will keep on affecting the

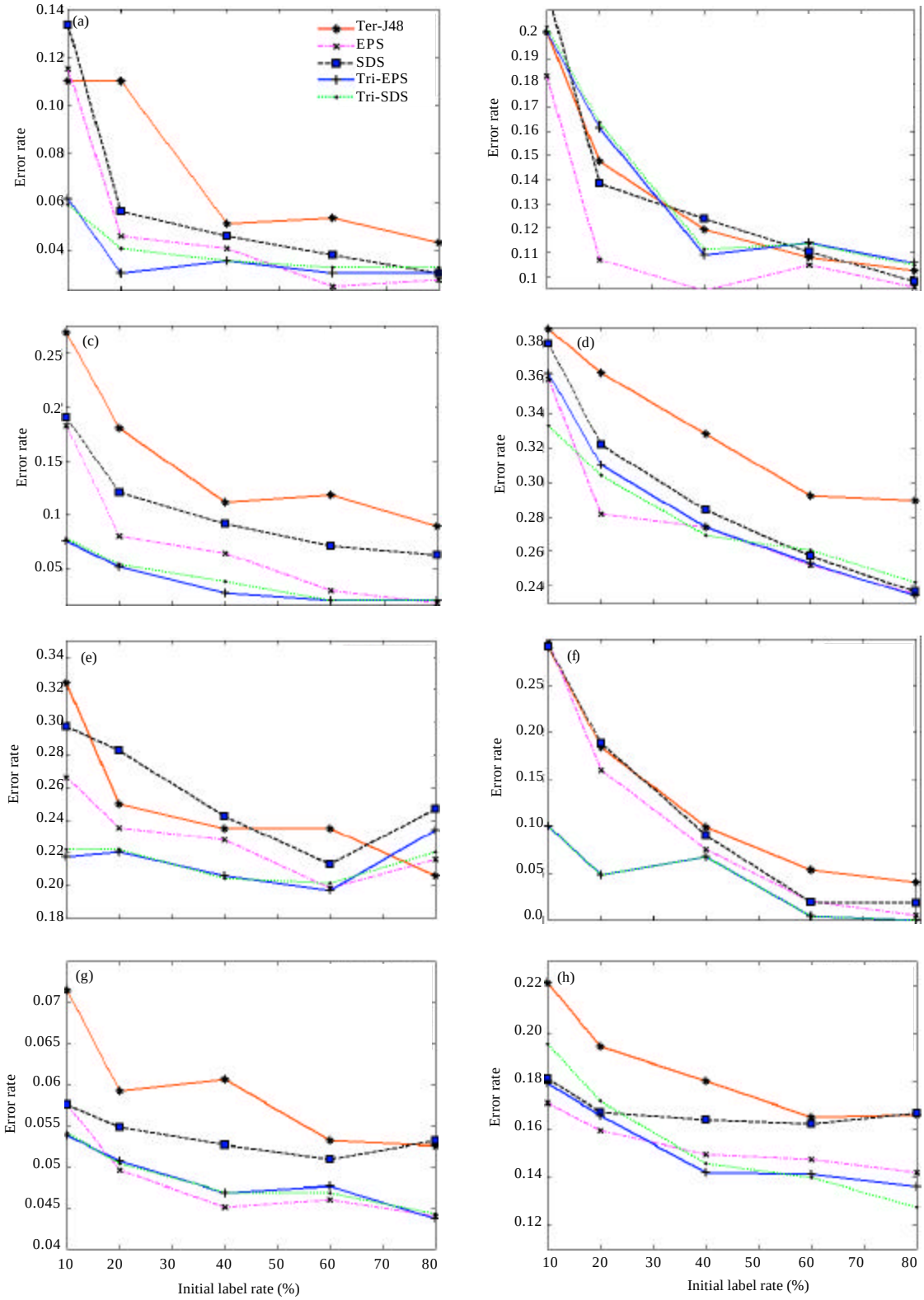

Fig. 5(a-h): Performance of corresponding methods of dataset, (a) Lris, (b) Inosophere, © Wine, (d) Vehicle, (e) Heart, (f) Tic_tac, (g) Can_tmr and (h) Envir_Audio

learner in the subsequent iterations. The data editing approach (Li and Zhou, 2005; Deng and Guo, 2007) can solve the problem. Due to the other two methods are based on the tri-training algorithm, therefore, they have the similar trends in fluctuation of curves.

CONCLUSION

Both Semi-supervised learning and active learning are helpful for improving learning generalization under the small training examples. In this paper we proposed two approaches to select samples in active learning and the combination of those two approaches with the semi-supervised learning Tri-Training algorithm. In each iteration, the training data are enlarged by two parts. One comes from the original Tri-Training and the other is from the EPS or SDS with active learning. The experimental results show that in general, the Tri-EPS and Tri-SDS obtain better performance than that Tri-Training method does. Combining active learning techniques with semi-supervised leaning, it makes full use of unlabeled examples from different directions to improve the learning generalization and get a better prediction effect. Although, EPS and SDS are simple, they lead to more effective sample selection.

Increasing diversity is the key to selective sampling in active learning and a more extensive study needs to be done to employ the diverse ensemble in active learning. Further research work about the effectively exploiting the information from the unlabeled data with the combining active learning and semi-supervised learning to build the better learning model are underway.

ACKNOWLEDGMENT

This study was supported by the Research Fund Projects, National "863" Plan Project, the Grant No. 2012AA102003.

REFERENCES

Blake, C. and C. Merz, 1998. UCI Repository of Machine Learning Databases. University of California, Irvine.

Deng, C. and M.Z. Guo, 2007. ADE-Tri-training: Tri-training with adaptive data editing. Chinese J. Comput., 30: 1213-1226.

Freund, Y., H.S. Seung, E. Shamir and N. Tishby, 1997. Selective sampling using the query by committee algorithm. Machine Learn., 28: 133-168.

Hady, M.F.A. and F. Schwenker, 2010. Combining committee-based semi-supervised learning and active learning. J. Comput. Sci. Technol., 25: 681-698.

Lewis, D.D. and W.A. Gale, 1994. A sequential algorithm for training text classifiers. Proceedings of the 17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, July 3-6, 1994, Dublin, Ireland, pp: 3-12.

Li, M. and Z.H. Zhou, 2005. SETRED Self-training with editing. Proceedings of the 9th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining, May 18-20, 2005, Hanoi, Vietnam, pp: 611-621.

McCallum, A. and K. Nigam, 1998. Employing EM and pool-based active learning for text classification. Proceedings of the 15th International Conference Machine Learning, July 24-27, 1998, Madison, USA., pp: 350-358.

Muslea, I., S. Minton and C.A. Knoblock, 2000. Selective sampling with redundant views. Proceeding of the 17th National Conference on Artificial Intelligence and 12th Conference on Innovative Applications of Artificial Intelligence, July 30-August 3, 2000, Austin, Texas, pp: 621-626.

Muslea, I., S. Minton and C.A. Knoblock, 2002. Active + semi-supervised learning = Robust multi-view learning. Proceeding of the 19th International Conference on Machine Learning, July 8-12, 2002, Sydney, Australia, pp: 435-442.

Seeger, M., 2001. Learning with labeled and unlabeled data. Technical Report, University of Edinburgh, Edinburgh, UK.

Seung, H.S., M. Opper and H. Sompolinsky, 1992. Query by committee. Proceedings of the 5th Annual Workshop on Computational Learning Theory, July 27-29, 1992, Pittsburgh, PA., USA., pp: 287-294.

Witten, I.H., E. Frank and A.H. Mark, 2011. Data Mining: Practical Machine Learning Tools and Techniques. 3rd Edn., Morgan Kaufmann, San Francisco, CA., USA.

Zhou, Z.H. and Y. Wang, 2007. Machine Learning and Application. Tsinghua University Press, Beijing, China.

Zhou, Z.H., D.C. Zhan and Q. Yang, 2007. Semi-supervised learning with very few labeled training examples. Proceedings of the 22nd National Conference on Artificial Intelligence, July 22-26, 2007, Vancouver, Canada, pp: 675-680.

Zhou, Z.H., K.J. Chen and Y. Jiang, 2004. Exploiting unlabeled data in content-based image retrieval. Proceedings of the 15th European Conference on Machine Learning, September 20-24, 2004, Pisa, Italy, pp: 525-536.

Zhu, X., 2006. Semi-supervised learning literature survey. Technical report No. 1530, Department of Computer Sciences, University of Wisconsin at Madison, Madison, WI., USA.