Cross-lingual Opinion Analysis via Negative Transfer Detection

Lin Gui[1,2], Ruifeng Xu[1*], Qin Lu[2], Jun Xu[1], Jian Xu[2], Bin Liu[1], Xiaolong Wang[1]
[1]Key Laboratory of Network Oriented Intelligent Computation, Shenzhen Graduate School, Harbin Institute of Technology, Shenzhen 518055
[2]Department Of Computing, the Hong Kong Polytechnic University
guilin.nlp@gmail.com, xuruifeng@hitsz.edu.cn, csluqin@comp.polyu.edu.hk, xujun@hitsz.edu.cn, csjxu@comp.polyu.edu.hk,{bliu,wangxl}@insun.hit.edu.cn

Abstract

Transfer learning has been used in opinion analysis to make use of available language resources for other resource scarce languages. However, the cumulative class noise in transfer learning adversely affects performance when more training data is used. In this paper, we propose a novel method in transductive transfer learning to identify noises through the detection of negative transfers. Evaluation on NLP&CC 2013 cross-lingual opinion analysis dataset shows that our approach outperforms the state-of-the-art systems. More significantly, our system shows a monotonic increase trend in performance improvement when more training data are used.

1 Introduction

Mining opinions from text by identifying their positive and negative polarities is an important task and supervised learning methods have been quite successful. However, supervised methods require labeled samples for modeling and the lack of sufficient training data is the performance bottle-neck in opinion analysis especially for resource scarce languages. To solve this problem, the transfer leaning method (Arnold et al., 2007) have been used to make use of samples from a resource rich source language to a resource scarce target language, also known as cross language opinion analysis (CLOA).

In transductive transfer learning (TTL) where the source language has labeled data and the target language has only unlabeled data, an algorithm needs to select samples from the unlabeled target language as the training data and assign them with class labels using some estimated confidence. These labeled samples in the target language, referred to as the transferred samples, also have a probability of being misclassified. During training iterations, the misclassification introduces class noise which accumulates, resulting in a so called negative transfer that affects the classification performance.

In this paper, we propose a novel method aimed at reducing class noise for TTL in CLOA. The basic idea is to utilize transferred samples with high quality to identify those negative transfers and remove them as class noise to reduce noise accumulation in future training iterations. Evaluations on NLP&CC 2013 CLOA evaluation data set show that our algorithm achieves the best result, outperforming the current state-of-the-art systems. More significantly, our system shows a monotonic increasing trend in performance when more training data are used beating the performance degradation curse of most transfer learning methods when training data reaches certain size.

The rest of the paper is organized as follows. Section 2 introduces related works in transfer learning, cross lingual opinion analysis, and class noise detection technology. Section 3 presents our algorithm. Section 4 gives performance evaluation. Section 5 concludes this paper.

2 Related works

TTL has been widely used before the formal concept and definition of TTL was given in (Arnold, 2007). Wan introduced the co-training method into cross-lingual opinion analysis (Wan, 2009; Zhou et al., 2011), and Aue et al. introduced transfer learning into cross domain analysis (Aue, 2005) which solves similar problems. In this paper, we will use the terms source language and target language to refer to all cross lingual/domain analysis.

Traditionally, transfer learning methods focus on how to estimate the confidence score of transferred samples in the target language or domain (Blitzer et al, 2006, Huang et al., 2007; Sugiyama et al., 2008, Chen et al, 2011, Lu et al., 2011). In some tasks, researchers utilize NLP tools such as alignment to reduce the bias towards that of

*Corresponding author
Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Short Papers), pages 860–865, Baltimore, Maryland, USA, June 23-25 2014.

the source language in transfer learning (Meng et al., 2012). However, detecting misclassification in transferred samples (referred to as class noise) and reducing negative transfers are still an unresolved problem.

There are two basic methods for class noise detection in machine learning. The first is the classification based method (Brodley and Friedl, 1999; Zhu et al, 2003; Zhu 2004; Sluban et al., 2010) and the second is the graph based method (Zighed et al, 2002; Muhlenbach et al, 2004; Jiang and Zhou, 2004). Class noise detection can also be applied to semi-supervised learning because noise can accumulate in iterations too. Li employed Zighed's cut edge weight statistic method in self-training (Li and Zhou, 2005) and co-training (Li and Zhou, 2011). Chao used Li's method in tri-training (Chao et al, 2008). (Fukumoto et al, 2013) used the support vectors to detect class noise in semi-supervised learning.

In TTL, however, training and testing samples cannot be assumed to have the same distributions. Thus, noise detection methods used in semi-supervised learning are not directly suited in TTL. Y. Cheng has tried to use semi-supervised method (Jiang and Zhou, 2004) in transfer learning (Cheng and Li, 2009). His experiment showed that their approach would work when the source domain and the target domain share similar distributions. How to reduce negative transfers is still a problem in transfer learning.

3 Our Approach

In order to reduce negative transfers, we propose to incorporate class noise detection into TTL. The basic idea is to first select high quality labeled samples after certain iterations as indicator to detect class noise in transferred samples. We then remove noisy samples that cause negative transfers from the current accumulated training set to retain an improved set of training data for the remainder of the training phase. This negative sample reduction process can be repeated several times during transfer learning. Two questions must be answered in this approach: (1) how to measure the quality of transferred samples, and (2) how to utilize high quality labeled samples to detect class noise in training data.

3.1 Estimating Testing Error

To determine the quality of the transferred samples that are added iteratively in the learning process, we cannot use training error to estimate true error because the training data and the testing data have different distributions. In this work, we employ the Probably Approximately Correct (PAC) learning theory to estimate the error boundary. According to the PAC learning theory, the least error boundary ε is determined by the size of the training set m and the class noise rate η, bound by the following relation:

$$\varepsilon \propto \sqrt{1/m(1-\eta)^2} \quad (1)$$

In TTL, m increases linearly, yet η is multiplied in each iteration. This means the significance of m to performance is higher at the beginning of transfer learning and gradually slows down in later iterations. On the contrary, the influence of class noise increases. That is why performance improves initially and gradually falls to negative transfer when noise accumulation outperforms the learned information as shown in Fig.1. In TTL, transferred samples in both the training data and test data have the same distribution. This implies that we can apply the PAC theory to analyze the error boundary of the machine learning model using transferred data.

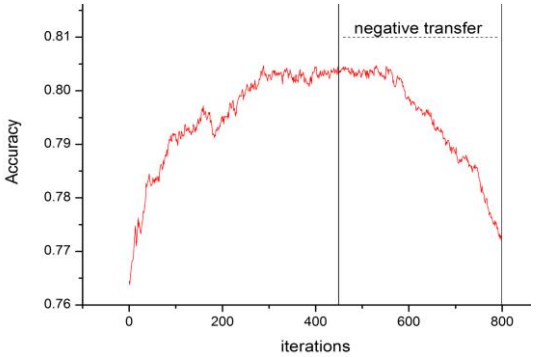

Figure 1 Negative transfer in the learning process

According to PAC theorem with an assumed fixed probability δ (Angluin and Laird, 1988), the least error boundary ε is given by:

$$\varepsilon = \sqrt{2\ln(2N/\delta)/(m(1-\eta)^2)} \quad (2)$$

where N is a constant decided by the hypothesis space. In any iteration during TTL, the hypothesis space is the same and the probability δ is fixed. Thus the least error boundary is determined by the size of the transferred sample m and the class noise of transferred samples η. According to (2), we apply a manifold assumption based method to estimate η. Let T be the number of iterations to serve as one period. We then estimate the least error boundary before and after each T to measure the quality of transferred samples during each T. If the least error boundary is reduced, it means that transferred samples used in this period are of high quality and can improve the performance. Otherwise, the transfer learning algorithm should stop.

3.2 Estimating Class Noise

For formula (2) to work, we need to know the class noise rate η to calculate the error boundary. Obviously, we cannot use conditional probabilities from the training data in the source language to estimate the noise rate η of the transferred samples because the distribution of source language is different from that of target language.

Consider a KNN graph on the transferred samples using any similarity metric, for example, cosine similarity, for any two connected vertex $(x_i, y_i)$and (x_j, y_j) in the graph from samples to classes, the edge weight is given by:

$$w_{ij} = sim(x_i, x_j) \quad (3)$$

Furthermore, a sign function for the two vertices $(x_i, y_i)$and (x_j, y_j), is defined as:

$$I_{ij} = \begin{cases} 0, if\ y_i = y_j \\ 1, if\ y_i \neq y_j \end{cases} \quad (4)$$

According to the manifold assumption, the conditional probability $P(y_i|x_i)$ can be approximated by the frequency of $P(y_i = y_j)$ which is equal to $P(I_{ij} = 0)$. In opinion annotations, the agreement of two annotators is often no larger than 0.8. This means that for the best cases $P(I_{ij} = 1)$ =0.2. Hence I_{ij} follows a Bernoulli distribution with p=0.2 for the best cases in manual annotations.

Let $C_{ij} = \{(x_j, y_j)\}$ be the vertices that are connected to the i^{th} vertex, the statistical magnitude of the i^{th} vertex can be defined as:

$$J_i = \sum_j w_{ij} \cdot I_{ij} \quad (5)$$

where j refers to the j^{th} vertex that is connected to the i^{th} vertex.

From the theory of cut edge statics, we know that the expectation of J_i is:

$$\mu_i = P(I_{ij} = 1) * \sum_j w_{ij} \quad (6)$$

And the variance of J_i is:

$$\sigma_i{}^2 = P(I_{ij} = 0)P(I_{ij} = 1) * \sum_j w_{ij}{}^2 \quad (7)$$

By the Center Limit Theorem (CLT), J_i follows the normal distribution:

$$\frac{(J_i - \mu_i)}{\sigma_i} \sim N(0,1) \quad (8)$$

To detect the noise rate of a sample (x_i, y_i), we can use (8) as the null hypothesis to test the significant level. Let p_i denotes probability of the correct classification for a transferred sample. p_i should follow a normal distribution,

$$p_i = \frac{1}{\sqrt{2\pi}\sigma_i} \int_{J_i}^{+\infty} e^{-\frac{(t-\mu_i)^2}{2\sigma_i^2}} dt. \quad (9)$$

Note that experiments (Li and Zhou, 2011; Cheng and Li, 2009; Brodley and Friedl, 1999) have shown that p_i is related to the error rate of the example (x_i, y_i), but it does not reflect the ground-truth probability in statistics. Hence we assume the class noise rate of example (x_i, y_i) is:

$$\eta_i = 1 - p_i \quad (10)$$

We take the general significant level of 0.05 to reject the null hypothesis. It means that if η_i of (x_i, y_i) is larger than 0.95, the sample will be considered as a class noisy sample. Furthermore, η_i can be used to estimate the average class noise rate of a transferred samples in (2).

In our proposed approach, we establish the quality estimate period T to conduct class noise detection to estimate the class noise rate of transferred samples. Based on the average class noise we can get the least error boundary so as to tell if an added sample is of high quality. If the newly added samples are of high quality, they can be used to detect class noise in transferred training data. Otherwise, transfer learning should stop. The flow chart for negative transfer is in Fig.2.

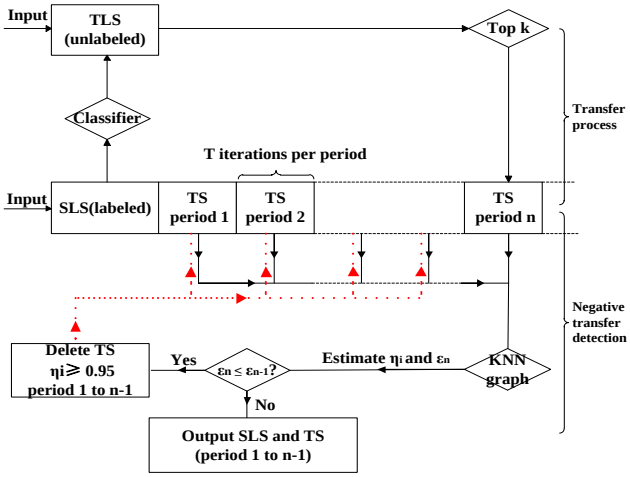

Figure 2 Flow charts of negative transfer detection

In the above flow chart, SLS and TLS refer to the source and target language samples, respectively. TS refers to the transferred samples. Let T denote quality estimate period T in terms of iteration numbers. The transfer process select k samples in each iteration. When one period of transfer process finishes, the negative transfer detection will estimate the quality by comparing and either select the new transferred samples or remove class noise accumulated up to this iteration.

4 Experiment

4.1 Experiment Setting

The proposed approach is evaluated on the NLP&CC 2013 cross-lingual opinion analysis (in

short, NLP&CC) dataset[1]. In the training set, there are 12,000 labeled English Amazon.com products reviews, denoted by Train_ENG, and 120 labeled Chinese product reviews, denoted as Train_CHN, from three categories, DVD, BOOK, MUSIC. 94,651 unlabeled Chinese products reviews from corresponding categories are used as the development set, denoted as Dev_CHN. In the testing set, there are 12,000 Chinese product reviews (shown in Table.1). This dataset is designed to evaluate the CLOA algorithm which uses Train_CHN, Train_ENG and Dev_CHN to train a classifier for Test_CHN. The performance is evaluated by the correct classification accuracy for each category in Test_CHN[2]:

$$Accuracy_c = \frac{\#correctly\ classified\ samples\ in\ c}{4000}$$

where *c* is either *DVD, BOOK* or *MUSIC.*

Team	DVD	Book	Music
Train_CHN	40	40	40
Train_ENG	4000	4000	4000
Dev_CHN	17814	47071	29677
Test_CHN	4000	4000	4000

Table.1 The NLP&CC 2013 CLOA dataset

In the experiment, the basic transfer learning algorithm is co-training. The Chinese word segmentation tool is ICTCLAS (Zhang et al, 2003) and Google Translator[3] is the MT for the source language. The monolingual opinion classifier is SVM[light4], word unigram/bigram features are employed.

4.2 CLOA Experiment Results

Firstly, we evaluate the baseline systems which use the same monolingual opinion classifier with three training dataset including Train_CHN, translated Train_ENG and their union, respectively.

	DVD	Book	Music	Accuracy
Train_CHN	0.552	0.513	0.500	0.522
Train_ENG	0.729	0.733	0.722	0.728
Train_CHN +Train_ENG	0.737	0.722	0.742	0.734

Table.2 Baseline performances

It can be seen that using the same method, the classifier trained by Train_CHN are on avergage 20% worse than the English counter parts.The combined use of Train_CHN and translated Train_ENG, however, obtained similar performance to the English counter parts. This means the predominant training comes from the English training data.

In the second set of experiment, we compare our proposed approach to the official results in NLP&CC 2013 CLOA evaluation and the result is given in Table 3. Note that in Table 3, the top performer of NLP&CC 2013 CLOA evaluation is the HLT-HITSZ system(underscored in the table), which used the co-training method in transfer learning (Gui et al, 2013), proving that co-training is quite effective for cross-lingual analysis. With the additional negative transfer detection, our proposed approach achieves the best performance on this dataset outperformed the top system (by HLT-HITSZ) by a 2.97% which translate to 13.1% error reduction improvement to this state-of-the-art system as shown in the last row of Table 3.

Team	DVD	Book	Music	Accuracy
BUAA	0.481	0.498	0.503	0.494
BISTU	0.647	0.598	0.661	0.635
HLT-HITSZ	0.777	0.785	0.751	0.771
THUIR	0.739	0.742	0.733	0.738
SJTU	0.772	0.724	0.745	0.747
WHU	0.783	0.770	0.760	0.771
Our approach	**0.816**	**0.801**	**0.786**	**0.801**
Error Reduction	**0.152**	**0.072**	**0.110**	**0.131**

Table.3 Performance compares with NLP&CC 2013 CLOA evaluation results

To further investigate the effectiveness of our method, the third set of experiments evaluate the negative transfer detection (NTD) compared to co-training (CO) without negative transfer detection as shown in Table.4 and Fig.3 Here, we use the union of Train_CHN and Train_ENG as labeled data and Dev_CHN as unlabeled data to be transferred in the learning algorithms.

		DVD	Book	Music	Mean
NTD	Best case	0.816	0.801	0.786	0.801
	Best period	0.809	0.798	0.782	0.796
	Mean	0.805	0.795	0.781	0.794
CO	Best case	0.804	0.796	0.783	0.794
	Best period	0.803	0.794	0.781	0.792
	Mean	0.797	0.790	0.775	0.787

Table.4 CLOA performances

Taking all categories of data, our proposed method improves the overall average precision (the best cases) from 79.4% to 80.1% when compared to the state of the art system which translates to error reduction of 3.40% (p-value≤0.01 in Wilcoxon signed rank test). Although the improvement does not seem large, our

[1]http://tcci.ccf.org.cn/conference/2013/dldoc/evdata03.zip
[2]http://tcci.ccf.org.cn/conference/2013/dldoc/evres03.pdf
[3]https://translate.google.com
[4]http://svmlight.joachims.org/

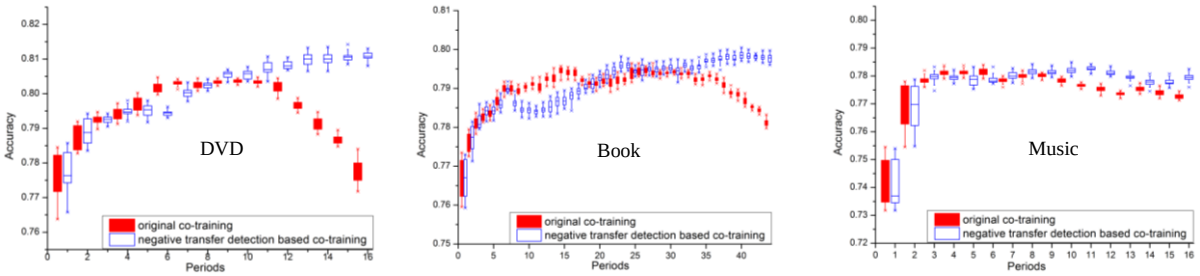

Figure 3 Performance of negative transfer detection vs. co-training

algorithm shows a different behavior in that it can continue to make use of available training data to improve the system performance. In other words, we do not need to identify the tipping point where the performance degradation can occur when more training samples are used. Our approach has also shown the advantage of stable improvement.

In the most practical tasks, co-training based approach has the difficulty to determine when to stop the training process because of the negative transfer. And thus, there is no sure way to obtain the above best average precision. On the contrary, the performance of our proposed approach keeps stable improvement with more iterations, i.e. our approach has a much better chance to ensure the best performance. Another experiment is conducted to compare the performance of our proposed transfer learning based approach with supervised learning. Here, the achieved performance of 3-folder cross validation are given in Table 5.

	DVD	Book	Music	Average
Supervised	0.833	0.800	0.801	0.811
Our approach	0.816	0.801	0.786	0.801

Table.5 Comparison with supervised learning

The accuracy of our approach is only 1.0% lower than the supervised learning using 2/3 of Test_CHN. In the BOOK subset, our approach achieves match result. Note that the performance gap in different subsets shows positive correlation to the size of Dev_CHN. The more samples are given in Dev_CHN, a higher precision is achieved even though these samples are unlabeled. According to the theorem of PAC, we know that the accuracy of a classifier training from a large enough training set with confined class noise rate will approximate the accuracy of classifier training from a non-class noise training set. This experiment shows that our proposed negative transfer detection controls the class noise rate in a very limited boundary. Theoretically speaking, it can catch up with the performance of supervised learning if enough unlabeled samples are available. In fact, such an advantage is the essence of our proposed approach.

5 Conclusion

In this paper, we propose a negative transfer detection approach for transfer learning method in order to handle cumulative class noise and reduce negative transfer in the process of transfer learning. The basic idea is to utilize high quality samples after transfer learning to detect class noise in transferred samples. We take cross lingual opinion analysis as the data set to evaluate our method. Experiments show that our proposed approach obtains a more stable performance improvement by reducing negative transfers. Our approach reduced 13.1% errors than the top system on the NLP&CC 2013 CLOA evaluation dataset. In BOOK category it even achieves better result than the supervised learning. Experimental results also show that our approach can obtain better performance when the transferred samples are added incrementally, which in previous works would decrease the system performance. In future work, we plan to extend this method into other language/domain resources to identify more transferred samples.

Acknowledgement

This research is supported by NSFC 61203378, 61300112, 61370165, Natural Science Foundation of GuangDong S2013010014475, MOE Specialized Research Fund for the Doctoral Program of Higher Education 20122302120070, Open Projects Program of National Laboratory of Pattern Recognition，Shenzhen Foundational Research Funding JCYJ20120613152557576, JC201005260118A, Shenzhen International Cooperation Research Funding GJHZ20120613110641217 and Hong Kong Polytechnic University Project code Z0EP.

Reference

Angluin, D., Laird, P. 1988. Learning from Noisy Examples. Machine Learning, 2(4): 343-370.

Arnold, A., Nallapati, R., Cohen, W. W. 2007. A Comparative Study of Methods for Transductive Transfer Learning. In Proc. 7th IEEE ICDM Workshops, pages 77-82.

Aue, A., Gamon, M. 2005. Customizing Sentiment Classifiers to New Domains: a Case Study, In Proc. of t RANLP.

Blitzer, J., McDonald, R., Pereira, F. 2006. Domain Adaptation with Structural Correspondence Learning. In Proc. EMNLP, 120-128.

Brodley, C. E., Friedl, M. A. 1999. Identifying and Eliminating Mislabeled Training Instances. Journal of Artificial Intelligence Research, 11:131-167.

Chao, D., Guo, M. Z., Liu, Y., Li, H. F. 2008. Particpatory Learning based Semi-supervised Classification. In Proc. of 4th ICNC, pages 207-216.

Cheng, Y., Li, Q. Y. 2009. Transfer Learning with Data Edit. LNAI, pages 427–434.

Chen, M., Weinberger, K. Q., Blitzer, J. C. 2011. Co-Training for Domain Adaptation. In Proc. of 23th NIPS.

Fukumoto, F., Suzuki, Y., Matsuyoshi, S. 2013. Text Classification from Positive and Unlabeled Data using Misclassified Data Correction. In Proc. of 51st ACL, pages 474-478.

Gui, L., Xu, R., Xu, J., et al. 2013. A Mixed Model for Cross Lingual Opinion Analysis. In CCIS, 400, pages 93-104.

Huang, J., Smola, A., Gretton, A., Borgwardt, K.M., Scholkopf, B. 2007. Correcting Sample Selection Bias by Unlabeled Data. In Proc. of 19th NIPS, pages 601-608.

Jiang, Y., Zhou, Z. H. 2004. Editing Training Data for kNN Classifiers with Neural Network Ensemble. In LNCS, 3173, pages 356-361.

Li, M., Zhou, Z. H. 2005. SETRED: Self-Training with Editing. In Proc. of PAKDD, pages 611-621.

Li, M., Zhou, Z. H. 2011. COTRADE: Confident Co-Training With Data Editing. IEEE Transactions on Systems, Man, and Cybernetics—Part B: Cybernetics, 41(6):1612-1627.

Lu, B., Tang, C. H., Cardie, C., Tsou, B. K. 2011. Joint Bilingual Sentiment Classification with Unlabeled Parallel Corpora. In Proc. of 49th ACL, pages 320-330.

Meng, X. F., Wei, F. R., Liu, X. H., et al. 2012. Cross-Lingual Mixture Model for Sentiment Classification. In Proc. of 50th ACL, pages 572-581.

Muhlenbach, F., Lallich, S., Zighed, D. A. 2004. Identifying and Handling Mislabeled Instances. Journal of Intelligent Information System, 22(1): 89-109.

Pan, S. J., Yang, Q. 2010. A Survey on Transfer Learning, IEEE Transactions on Knowledge and Data Engineering, 22(10):1345-1360.

Sindhwani, V., Rosenberg, D. S. 2008. An RKHS for Multi-view Learning and Manifold Co-Regularization. In Proc. of 25th ICML, pages 976–983.

Sluban, B., Gamberger, D., Lavra, N. 2010. Advances in Class Noise Detection. In Proc.19th ECAI, pages 1105-1106.

Sugiyama, M., Nakajima, S., Kashima, H., Buenau, P.V., Kawanabe, M. 2008. Direct Importance Estimation with Model Selection and its Application to Covariate Shift Adaptation. In Proc. 20th NIPS.

Wan, X. 2009. Co-Training for Cross-Lingual Sentiment Classification, In Proc. of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, 235–243.

Zhang, H. P., Yu, H. K., Xiong, D. Y., and Liu., Q. 2003. HHMM-based Chinese Lexical Analyzer ICTCLAS. In 2nd SIGHAN workshop affiliated with 41th ACL, pages 184-187.

Zhou, X., Wan X., Xiao, J. 2011. Cross-Language Opinion Target Extraction in Review Texts. In Proc. of IEEE 12th ICDM, pages 1200-1205.

Zhu, X. Q., Wu, X. D., Chen, Q. J. 2003. Eliminating Class Noise in Large Datasets. In Proc. of 12th ICML, pages 920-927.

Zhu, X. Q. 2004. Cost-guided Class Noise Handling for Effective Cost-sensitive Learning In Proc. of 4th IEEE ICDM, pages 297-304.

Zighed, D. A., Lallich, S., Muhlenbach, F. 2002. Separability Index in Supervised Learning. In Proc. of PKDD, pages 475-487.