Society for Science and Education
United Kingdom

TMLAI Transactions on Machine Learning and Artificial Intelligence

Volume 5 No 4, 2017
ISSN 2054-7390

Data Editing for Semi-Supervised Co-Forest by the Local Cut Edge Weight Statistic Graph (CEWS-Co-Forest)

Nesma Settouti, Mohammed El Amine Bechar, Mostafa EL Habib Daho, Mohammed Amine Chikh
Biomedical Engineering Laboratory, Tlemcen University, Algeria
acmlis.conference@gmail.com

ABSTRACT

In order to address the large amount of unlabeled training data problem, many semi-supervised algorithms have been proposed. The training data in semi-supervised learning may contain much noise due to the insufficient number of labeled data in training set. Such noise may snowball themselves in the following learning process and thus hurt the generalization ability of the final hypothesis. If such noise could be identified and removed by some strategy, the performance of the semi-supervised algorithms should be improved. However, such useful techniques of identifying and removing noise have been seldom explored in existing semi-supervised algorithms. In this paper, we use the semi-supervised ensemble method "Co-forest" with data editing (we call it CEWS-Co-forest) to improve sparsely labeled medical dataset. The cut edges weight statistic data editing technique is used to actively identify possibly mislabeled examples in the newly-labeled data throughout the co-labeling iterations in Co-forest. The fusion of semi-supervised ensemble method with data editing makes CEWS-co-Forest more robust to the sparsity and the distribution bias of the training data. It further simplifies the design of semi-supervised learning which makes CEWS-co-forest more efficient. An experimental study on several medical data sets shows encouraging results compared with state-of-the-art methods.

Keywords-semi supervised learning, data editing, Co-forest, Ensemble methods, medical diagnosis.

1 Introduction

The goal of semi-supervised learning lies in understanding the combination of labeled and unlabeled data; this can alter learning behavior and design algorithms that benefit from such a combination. Semi-supervised learning is of great interest for automatic learning and data mining because it can use unlabeled data that are readily available on stage to improve supervised learning tasks when labeled data are scarce or expensive to obtain.

The study of semi-supervised learning is motivated by two factors: its practical value in building better computer algorithms, and its theoretical value in understanding learning in machines and humans. Semi-supervised learning has considerable practical value. In many tasks, there is a shortage of labeled data. Labels can be difficult to obtain because they require human experts, special devices and slow expensive experiments.

DOI: 10.14738/tmlai.54.3290
Publication Date: 15th August, 2017
URL: http://dx.doi.org/10.14738/tmlai.54.3290

Most semi-supervised learning strategies are based on the extension of supervised or unsupervised learning to include additional information typical of the other learning regime. More specifically, semi-supervised learning encompasses several different contexts, including:

- Semi-supervised classification: The purpose of the semi-supervised classification is to teach the hypothesis h from labeled and unlabeled data; so that it is preferable to the used hypothesis to be supervised and trained on the label data.
- Semi-Supervised Clustering: The objective here is to adapt existing clustering methods to support constraints or labeled data, in order to produce clusters for unlabeled data using the supervised information.
- The Active learning: refers to methods that select unlabeled examples that are the most important, and an oracle can be proposed for the labeling of these instances; the objective is to minimize the labeling data (Freund, Y. et al., 1997). Sometimes it is called selective sampling or sample selection [20].

Semi-supervised learning is the considered solution to the problem of manual annotation difficulty because it can use both labeled and unlabeled data to give us a more reliable estimate of the decision limit. Intuitively, the distribution of untagged data helps to identify regions with the same label, and the few labeled data provide then the actual labels. From a different perspective, semi-supervised learning can achieve the same level of performance as supervised learning, but with fewer labeled examples. This reduces the annotation effort, which leads to a reduced cost.

Many semi-supervised ensemble learning (SSL) algorithms have been proposed, among which the ”Co-forest” algorithms are widely used. In this work, we present an improvement of Co-forest algorithm. It uses a filtering method to identify and correct the examples possibly mislabeled throughout the co-labeling iterations. Despite all the improvement of Co-forest by data editing techniques, our method CEWS Co-forest proposes a filtering method that permits to remove mislabeled examples through the control of neighborhood data, where the local cut edge weight statistic is used to help estimate whether a newly labeled example is reliable or not, and only the reliable examples are used to enlarge the labeled training set. We have conducted a comparative study that has indicated an overall significant improvement of our method compared to the existing data editing semi-supervised approach.

This paper isorganized as follow: a review of some ensemble methods in the semi-supervised field is performed in section2. We then describe in section 3, the Co-forest algorithms and its existing data-editing version DE-Co-forest. Section 4 is about the general process of our proposed approach by local cut edge weight statistic editing training data. We validate our algorithm and the choices we have made in an experimental phase in section 5. Finally, we endup with a conclusion that summarizes the contributions made and the tracks defining possible opportunities for future work.

2 Ensemble Methods in Semi-Supervised Learning

The first to be born in this category is the co-Training algorithm proposed by Blum and Mitchell [4] for semi-supervised classification web pages.

The idea of co-learning is that the feature space can be divided into two sub-spaces providing each a good learning environment. Thus, initially two classifiers are trained with the labeled data on two different subspaces. Then, each obtained classifier for each subspace is used to determine the probable class of the unlabeled data that will be used to re-train the other classifier.

To use the co-training method, you must have two different views of data to be classified, and these two different views have to be compatible and independent, and each view is used to extract the characteristics, so each view gives rise to a characterization Different from the forms to be recognized.

Compatibility makes it possible to have the same label for a given example according to each view considered independently. As far as independence is concerned, one wants for a given example, that there is no correlation between the characteristics resulting from the two different views. However, in practice, it is not always possible to obtain two independent subsets of attributes relative to the label, which makes it difficult to generalize co-Training.

To overcome this difficulty, Li and Zhou introduced in [7] a new algorithm that extends the paradigm of co-Training using Random Forest [6]. This algorithm named Co-forest uses N classifiers. N−1 classifiers are used to determine examples of trust, called concomitant Set Hi = HN-1. The confidence measure of the newly labeled example can be simply estimated by the degree of agreement on labeling, i.e. the number of classifiers that agree on the label assigned by Hi.

The approaches proposed by Blum and Mitchell [4], Zhou and Li [9] show the advantage of using multiple classifiers. Which means, learning these classifiers involves predicting the unlabeled examples before using them. Therefore, the algorithm of Li and Zhou, offers the best compromise in the semi supervised approach. Jiang and Zhou in their paper [11], provide an improvement Co-forest with very interesting results by integrating an filtering method ”DATA Editing”, the algorithm is calledDE-Co-forest.

DE-Co-forest uses the RemoveOnly editing approach [11] to identify and eliminate the ”suspect” noisy mislabeled examples in the subset of newly certified learning ones.

The data editing (filtering) approaches have the advantages that they are very efficient and robust against the over fitting. However, they tend to select examples with rather than redundant information, and do not take into account the interactions between the elements.

Therefore, in this work we are also interested in improving Co-forest, but unlike the RemoveOnly editing approach DE-Co-forest, which is based on calculating distance to the removal of noisy elements, we propose a filtering method that permits to remove mislabeled examples through the control of neighborhood data, with the local cut edge weight statistic graph strategy.

3 Methods

Many semi-supervised learning (SSL) algorithms have been proposed, among which the ”Co-forest” algorithms are widely used.

3.1 Co-forest algorithm

Co-forest was proposed by Li and Zhou [7]. This proposal is an extended version of the co-Training paradigm [4] by the ensemble method Random Forest [6]. In Co-forest, a set of N classifiers is used instead of two in co-Training. In this way, we can effectively improve the confidence estimate by each classifier. If we want to consider the labeled instance, the most confident by a classifier hi (i = 1, 2 . . . N) of the set H*, the all other classifiers are used except hi, called concomitant ensemble of hi and denoted by Hi. Therefore, the confidence level is calculated as the degree of agreement on the label, i.e. the number of classifiers agree on the label assigned by Hi. The general idea of Co-forest is to learn a set of classifiers.

More specifically, Co-forest is an iterative process, the concomitant ensemble hi will test each unlabeled example. Thereafter, if the number of classifiers that agree on a particular label exceeds a predefined threshold θ, this new label is assigned to the example and then it will be copied in the new labeled set. In the next iteration, the new labeled set is used for refining hi. Hereafter, we note that the unlabeled examples are not deleted, so they can be selected by other Hj (j≠i) in the following iterations. Consult [7, 20] for more details on the understanding of co-Forest.

3.2 DE-Co-forest algorithm

In semi-supervised learning, there is a problem that may affect Co-forest as well as other algorithms such as co-Training, which is the unlabeled examples may be mislabeled and introduced into the learning process. This is due to the limited number of examples initially labeled that usually generates low classifiers, which lacks precision and diversity. Based on this observation, a new algorithm that combines Co-forest with a data editing technique called DE-Co-forest is used. DE-Co-forest uses a data editing technique to identify and possibly eliminate mislabeled examples through iterations of co-labeling. In DE-Co-forest the RemoveOnly data editing technical [11] is used to identify mislabeled data.

Its principle is that the label of each unlabeled instance is not only determined by multiple classifiers, but also by the nearest neighbor rule. If the label is compatible with those selected by a minimum of k' nearest neighbor data, the unlabeled instance data with the greatest confidence are added to the training set. Otherwise, they are rejected and removed from the set of re-learning.

This method drove improvements to enrich the learning set, which is based on the k-nearest neighbor (k-NN). In this context, Cover and Hart [12] studied the asymptotic optimality of the nearest neighbor (NN) rule [13] and they proved that the NN rule is asymptotically optimal when different classes do not overlap in the input space. Otherwise, it may seem as one of the sub-optimality of the NN rule and it can overcome a bad classification. To decrease this error of the optimality, we propose the implication of CEWS [14] filtering method that permits to remove mislabeled examples through the control of neighborhood data, where the local cut edge weight statistic is used to help estimate whether a newly labeled example is reliable or not. Thereby, only the reliable examples are used to enlarge the labeled training set. We conduct a comparativestudy that indicate an overall significant improvement of our method compared to the existing data editing semi-supervised approach.

4 The Proposed Method CEWS Co-Forest Algorithm

The Cut edges weight statistic (CEWS) [14], this filtering method permits to remove mislabeled examples through the control of neighborhood data. At the beginning, it is necessary to build a geometrical connected graph like Toussaint's Relative Neighborhood Graph [15] on all examples of the training set. By definition a neighborhood graph G = (V, E) [16] is represented by vertex V and there exists an edge E between two vertices's xi and xj if the distance between xi and xj satisfies Eq. (1).

$$R_i = J_i / I_i \tag{2}$$

Where, Ii is the sum of weights relative to edges for sample xi Eq(3), Ji is the sum of weights relative to cut edges for sample xi Eq(4) and wij is the weighting distance of each edge Eq(5).

$$I_i = \sum_{j \in Neighborhood(x_i)} w_{ij} \tag{3}$$

$$J_i = \sum_{j \in Neighborhood(x_i), y_j \neq y_i} w_{ij} \tag{4}$$

$$w_{ij} = 1/\left(1 + dist\left(x_i, x_j\right)\right) \tag{5}$$

5 Experiments and Results

We have selected a set of seven databases from ASU repository [18] and UCI [19] their characteristics are summarized in Table I. To study the effectiveness of CEWSCo-forest incomparison to the performance of Co-forest and DE-Co-forest.

Table 1. The average accuracy of the compared algorithms under different labeled rateμ

Labeled rate	*Methods*	*data_C*	*Leukemia*	*Lung*	*Musk*	*Ovarian*	*Pancreatic*	*Prostate*
***μ*=80%**	**Co-Forest**	0,5800	0,8615	0,8732	0,8138	0,8111	0,5452	0,8457
	DE-Co-Forest	0,5500	0,8538	0,8648	0,8175	0,8444	0,5194	0,8800
	CEWS-Co-Forest	0,5900	0,9077	0,8761	0,8200	0,8889	0,5097	0,9029
***μ*=60%**	**Co-Forest**	0,6100	0,8615	0,8507	0,8213	0,8000	0,4871	0,8800
	DE-Co-Forest	0,6100	0,8308	0,8507	0,8088	0,8222	0,4968	0,8400
	CEWS-Co-Forest	0,6700	0,8846	0,8648	0,8250	0,8333	0,4903	0,8571
***μ*=40%**	**Co-Forest**	0,5800	0,8154	0,8225	0,7763	0,8444	0,5032	0,7600
	DE-Co-Forest	0,5800	0,8154	0,8338	0,7788	0,8444	0,5032	0,7200
	CEWS-Co-Forest	0,5900	0,8538	0,8394	0,7875	0,8667	0,5194	0,7714
***μ*=20%**	**Co-Forest**	0,6400	0,7615	0,7352	0,7138	0,5333	0,4903	0,6800
	DE-Co-Forest	0,6500	0,7538	0,7493	0,6988	0,5556	0,4903	0,6686
	CEWS-Co-Forest	0,6800	0,8154	0,7718	0,7188	0,5778	0,5194	0,6857

Table 2. Description of Experimental High Dimensional Datasets

Datasets	*#instances*	*#features*	*#class*
Data C	60	7130	2
Leukemia	73	7129	2
Lung	203	12600	5
Musk	476	166	2
Ovarian	54	1536	2
Pancreatic	119	6771	2
Prostate	102	12533	2

For each dataset, a 10 cross validation is carried out for evaluation. The training data are randomly divided into two sets: L labeled and unlabeled U determined by a rate (μ), which is calculated by the size of L on the size of L ∪ U. To simulate different amounts of unlabeled data, four different unlabeled rates μ = 20 %, 40 %, 60 % and 80 %, are studied.

The distributions of class in L and U are maintained similar to the original set. In these experiments, the value of N is 6 trees. Confidence level θ is set at 0.75, i.e., a newly labeled example is considered trusted if more than three quarters of the trees are agreements on its assigned label. For the RemoveOnlydata editing method, we fixed the number of neighbors k at 3, and the minimum number of neighbors equals to 2.

In term to estimate the accuracy on each dataset, we have predetermined a set of labeled examples. For each set, the algorithm is evaluated on its ability to correctly predict the labels of unlabeled examples. The labeled samples were randomly selected, with the only constraint being the presence of at least one example of each class for each set.

To compare the performances of CEWS Co-forest to Co-forest and DE-Co-forest, for each dataset with a specific labeled rate μ, a cross-validation is repeated ten times, and the results are averaged and recorded. Table 2 shows the average accuracy results and the ranking of each testing algorithm obtained. Specifically, it shows the overall results of the analyzed algorithms over the seven used datasets with 20, 40, 60 and 80 % rate of unlabeled data.

Our proposition outperforms the other algorithms; except CEWS-Co-Forest's performance was degraded on Pancreatic dataset with 80% and 60% of unlabeled rate (Table II). This can be explained by poor learning of the initial hypothesis and by the addition of misclassified data especially in training set. However, the CEWS-Co-Forest benefits much from the unlabeled data since the performances are evidently improved over all the seven datasets compared to other algorithms.

The immersion of the Cut edges weight statistic (CEWS) into the co-forest algorithm process allows enhanced confidence labeling to improve the classification accuracy, so we can deduce that CEWS Co-forest algorithm gives a good result in comparison with other algorithms.

6 Conclusion

The presented algorithm is an improvement of the Co- forest method [7] for semi-supervised classification. The aim of data editing in CEWS Co-forest is to identify and remove the noise contained in labeling step and thus to improve the overall performance. Our basic consideration is to implement a filtering method that permits to remove mislabeled examples through the control of neighborhood data, while, we are fully utilizing the advantage of ensemble learning in order to incur less computation complexity when improving the accuracy.

Experiments on high biomedical data sets show that data editing is a very useful technique for improving the performance of sparsely labeled data classification, and it makes the algorithm more efficient. For future work, we will further explore new techniques to cope with the training data sparsity and trainingdata bias for sparsely labeled data classification, e.g. semi-supervised clustering aided techniques.

REFERENCES

[1] O. Chapelle, B. Sch¨ and A. Zien, Semi-Supervised Learning,MIT Press, Cambridge, MA, 2006.

[2] Antoine Cornu´ and Laurent Miclet, Apprentissage artificiel : Concepts et algorithmes, Eyrolles, June 2010.

[3] Xiaojin Zhu, “Semi-Supervised learning literature survey,” Tech. Rep.,Computer Sciences, University of Wisconsin-Madison, 2005.

[4] Avrim Blum and Tom Mitchell, “Combining labeled and unlabeled datawith co-training,” in Proceedings of the eleventh annual conference onComputational learning theory, New York, NY, USA, 1998, COLT’ 98,pp. 92–100.

[5] L. G. Valiant, “A theory of the learnable,” Commun. ACM, vol. 27, no.11, pp. 1134–1142, Nov. 1984.

[6] L. Breiman, “Random forests,” Machine Learning, vol. 45, pp. 5–32,2001.

[7] Ming Li and Zhi-Hua Zhou, “Improve computer-aided diagnosis withmachine learning techniques using undiagnosed samples,” Trans. Sys.Man Cyber. Part A, vol. 37, no. 6, pp. 1088–1098, Nov. 2007.

[8] Yan Zhou and Sally Goldman, “Democratic co-learning,” in Proceedingsof the 16th IEEE International Conference on Tools with ArtificialIntelligence, Washington, DC, USA, 2004, ICTAI ’04, pp. 594–202,IEEE Computer Society.

[9] Zhi-Hua Zhou and Ming Li, “Tri-training: Exploiting unlabeled datausing three classifiers,” IEEE Trans. on Knowl. and Data Eng., vol. 17,no. 11, pp. 1529–1541, Nov. 2005.

[10] Boaz Leskes and Leen Torenvliet, “The value of agreement a newboosting algorithm,” J. Comput. Syst. Sci., vol. 74, no. 4, pp. 557–586,June 2008.

[11] Yuan Jiang and Zhi hua Zhou, “Editing training data for knn classifierswith neural network ensemble,” in Lecture Notes in Computer Science,Vol.3173. 2004, pp. 356–361, Springer.

[12] T. Cover and P. Hart, “Nearest neighbor pattern classification,” IEEETrans. Inf. Theor., vol. 13, no. 1, pp. 21–27, Sept. 2006.

[13] Fix and J. L. Hodges, “Discriminatory analysis. Nonparametricdiscrimination: consistency properties,” Jr. U.S. Air Force Sch. AviationMedicine, Randolf Field, vol. Rep.4, pp. Project 21–49–004, ContractAF 41 (128)–31, 1951.

[14] Fabrice Muhlenbach, St[U+FFFD]ane Lallich, and Djamel A. Zighed,“Identifying and handling mislabelled instances.,” J. Intell. Inf. Syst.,vol. 22, no. 1, pp. 89–109, 2004.

[15] Godfried T. Toussaint, “The relative neighbourhood graph of a finiteplanar set,” Pattern Recognition, vol. 12, pp. 261–268, 1980.

[16] L. Devroye, L. Gy and G. Lugosi, A Probabilistic Theory of PatternRecognition, Springer, 1996.

[17] Yu Wang, Xiaoyan Xu, Haifeng Zhao, and Zhongsheng Hua, “Semi-supervised learning based on nearest neighbor rule and cut edges,”Know.-Based Syst., vol. 23, no. 6, pp. 547–554, Aug. 2010.

[18] Reza Zafarani and Huan Liu, “Asu repository of social computingdatabases,” 1998.

[19] D.J. Newman, S. Hettich, C.L. Blake, and C.J. Merz, “Uci repositoryof machine learning databases,” 1998.

[20] Lytras, Miltiadis D., and Paraskevi Papadopoulou. "Applying Big Data Analytics in Bioinformatics and Medicine." IGI Global, 2018. 1-402. Web. 6 Apr. 2017. doi:10.4018/978-1-5225-2607-0.