CLASS NOISE REMOVAL AND CORRECTION FOR IMAGE CLASSIFICATION USING ENSEMBLE MARGIN

Wei Feng, Samia Boukir

Bordeaux INP, G&E, EA 4592, F-33600, Pessac, France.
E-mail: name.surname@ensegid.fr

ABSTRACT

Mislabeled training data is a challenge to face in order to build a robust classifier whether it is an ensemble or not. This work handles the mislabeling problem by exploiting four different ensemble margins for identifying, then eliminating or correcting the mislabeled training data. Our approach is based on class noise ordering and relies on the margin values of misclassified data. The effectiveness of our ordering-based class noise removal and correction methods is demonstrated in performing image classification. A comparative analysis is conducted with respect to the majority vote filter, a reference ensemble-based class noise filter.

Index Terms— Class noise removal, class noise correction, ensemble margin, multiple classifier, mislabeled data identification.

1. INTRODUCTION

In supervised learning, the training set is an essential element of the learning process [1]. But the real data to be classified often include a certain amount of noise which can be mainly of two types: class noise (mislabeled data) and attribute noise [2]. Effective noise handling is one of the most difficult problems in inductive machine learning [3]. Cleaning the training data will result in a classifier with a higher predictive accuracy [4].

Mislabeled training data is a challenge to face in order to build a robust classifier whether it is an ensemble or not. Furthermore, learning from noisy data can create overfitting [5]. Labeling training instances is a costly and rather subjective task that usually induces some labeling errors in the training set [3, 6, 7]. Therefore, how to form an efficient training set is a main issue in supervised classification [7, 8].

Three different approaches exist for handling the mislabeling problem [9]: making algorithms that are more robust to noise [4, 10], filtering out the noise [6, 11], and correcting noisy instances [12]. It has been argued [3] that the first method is less effective than the other two methods.

The ensemble approach is a popular method to filter out mislabeled instances [5, 6, 11, 13, 14, 15]. It detects the mislabeled instances by considering the vote of each base classifier in the ensemble to each instance [9]. A typical approach is the majority vote filter. In this method [6], if more than half of all the base classifiers of the ensemble classify an instance incorrectly, then this instance is tagged as mislabeled. However, a majority vote filter not only eliminates mislabeled instances but also all the clean training instances that have been wrongly classified by the underlying ensemble classifier.

In the following, we use an ensemble margin-based class noise elimination method which can achieve a high mislabeled instance detection rate (*true positives*) while keeping the false detection rate (*false positives*) as low as possible [11]. This method, introduced in [11], is extended here by: 1) considering 4 different ensemble margins (instead of one), including a new one, hence significantly improving the previous noise removal performances; 2) tackling the class noise correction as well. A comparative analysis is also conducted with respect to the majority vote filter.

2. ENSEMBLE MARGIN

The ensemble margin [16] is a fundamental concept in ensemble learning. Several studies have shown that the generalization performance of an ensemble classifier is related to the distribution of its margins on the training examples [1, 16]. A good margin distribution means that most examples have large margins [17]. The decision by an ensemble for each instance is made by voting. The ensemble margin can be calculated as a difference between the votes [9] according to two different well-known definitions [18] in both supervised [16] and unsupervised [19, 20] ways.

1. A popular ensemble margin, which has been introduced by Shapire et al. [16], is defined by equation (1), where v_y is the number of votes for the true class y and v_c is the number of votes for any other class c. This ensemble margin is in the range [-1, +1] and the examples which are correctly classified have positive margin values.

$$margin(x) = \frac{v_y - \max_{c=1,\ldots,L \cap c \neq y}(v_c)}{\sum_{c=1}^{L}(v_c)} \quad (1)$$

   where L represents the number of classes.

2. The ensemble margin of a sample can also be obtained by the difference between the fraction of classifiers voting correctly and incorrectly, as in equation (2) [9, 18]. This second popular ensemble margin definition follows the same idea introduced by Schapire [16] but instead of using a max operation, it uses a sum operation [18].

$$margin(x) = \frac{v_y - \sum_{c=1,\ldots,L \cap c \neq y}(v_c)}{\sum_{c=1}^{L}(v_c)} \quad (2)$$

3. In [19, 20], the authors proposed an unsupervised version of Schapire's margin (equation (1)). This ensemble margin's range is from 0 to 1. It is defined by equation (3), where v_{c_1} is the votes number of the most voted class c_1 for sample x, and v_{c_2} is the votes number of the second most popular class c_2.

$$margin(x) = \frac{v_{c_1} - v_{c_2}}{\sum_{c=1}^{L}(v_c)} \quad (3)$$

4. Finally, we propose a new unsupervised ensemble margin alternative defined as equation (4), where v_{c_1} is the votes number of the most voted class for sample x.

$$\begin{aligned} margin(x) &= \frac{v_{c_1} - \sum_{c=1,\ldots,L \cap c \neq c_1}(v_c)}{\sum_{c=1}^{L}(v_c)} \\ &= \frac{2v_{c_1} - T}{T} \end{aligned} \quad (4)$$

where T represents the number of base classifiers in the ensemble. This margin is an unsupervised version of the classic margin referred to as equation (2).

Naturally, for two-class problems these definitions are quite similar. However, a major concern needs to be solved in relation to multi-class problems. For example, by equation (2), the margins can represent a lower bound, since they can assume negative values even when the correct label gets the most of votes (when there is a plurality, but not a majority) [18]. In this study, we compare the performances of these different ensemble margins in label noise filter design.

3. ENSEMBLE MARGIN BASED CLASS NOISE REMOVAL

3.1. Class noise

A class noise (or mislabeling) is an instance whose label value conflicts with most of the other instance label values while having the same or similar attribution values. It implies that most base classifiers in the ensemble classified this instance as another class. In other words, this instance was classified wrongly with high confidence [9, 11].

Noisy distributions are application dependent and are generally unknown. We assume mislabeled instances to be uniformly distributed in associated classes like many other works investigating class noise [2, 5, 6, 11, 13].

3.2. Margin-based class noise ordering

Each training instance has a probability of being mislabeled. However, these probabilities are different depending on instance features and behavior in the training process. The objective of noise removal is to eliminate the most likely noisy instances. Ordering training instances according to their probability of being mislabeled is a simple and efficient method for noise removal [9]. In previous work [11], these probabilities rely on the margin values of training instances involving an unsupervised ensemble margin (equation (3)).

Let us consider an ensemble classifier C, and a set of n training data denoted as $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$, where x_i is a vector with feature values and y_i is the value of the class label. The mislabeled instance ordering approach, introduced in [11], simply relies on an ensemble margin's definition as a class noise evaluation function, slightly modified here, defined as (5). This method assesses only the training instances x_i whose attribution and label values are not consistent.

$$N(x_i) = |margin(x_i)| \quad \forall (x_i, y_i) \in S \mid C(x_i) \neq y_i \quad (5)$$

The higher $N(x_i)$, the higher the probability of x_i being mislabeled. Relying on the margin-based noise evaluation function, the ordering-based mislabeled instance elimination algorithm consists of the following steps [11]:

1. Constructing an ensemble classifier C with all the n training data $(x_i, y_i) \in S$.
2. Computing the margin of each training instance x_i.
3. Ordering all the training instances x_i, that have been misclassified, according to their noise evaluation values $N(x_i)$, in descending order.
4. Eliminating the first M most likely mislabeled instances x_i to form a new cleaner training set.
5. Evaluating the cleaned training set by classification performance, on a validation set.
6. Selecting the best filtered training set.

4. ENSEMBLE MARGIN BASED CLASS NOISE CORRECTION

Noise removal can discard some useful data, so we also attempt to automatically correct the training instances that have been identified as mislabeled (highest absolute margin misclassified instances). Noise correction has been shown to give better results than simply removing the noise from the data set in some cases [12]. In a data correction scheme, the noisy instances are identified, but instead of removing these instances out, they are repaired by replacing corrupted values with more appropriate ones [12]. The labels of the most likely mislabeled instances are changed to the predicted classes. Then, these corrected instances are reintroduced into the training set.

Data set	Train.	Valid.	Test	Variables	classes
Letter	5000	2500	5000	16	26
Optdigits	1000	500	1000	64	10
Pendigit	2000	1000	2000	16	10
Statlog	2000	1000	2000	36	6
Vehicle	200	100	200	18	4

Table 1. Data sets.

Our class noise correction method relies on an adaptive strategy that is similar to our class noise removal method. But, instead of removing an amount M of noise from training set at each step, it automatically corrects the detected noise using the predicted labels by the constructed bagging ensemble.

A comparative analysis is conducted between our margin-based mislabeled data identification method and the majority filter, also an ensemble-based mislabeled training data identification approach [6]. Both class noise removal and correction schemes are involved in the comparison. Each of the four different ensemble margins, defined in section 2, are involved in the validation of our algorithms.

5. EXPERIMENTAL RESULTS

In all our experiments, we used *bagging* [21] to create an ensemble involving Classification and Regression Trees (CART) [22] as base classifiers. Two well-known noise sensitive classifiers were used to assess the quality of both class noise removal and class noise correction algorithms: K-Nearest Neighbor (*K-NN*) [23] and *Adaboosting.M1* [24]. *K* was set to 1 in *K-NN* classifier. *Boosting* and *bagging* ensembles were implemented with 200 and 100 decision trees respectively.

5.1. Data sets

We applied the class noise removal and correction methods on 5 image data sets from UCI Machine Learning repository [25] (table 1). Each data set has been divided into three parts: training set, validation set and test set, as shown on table 1. We randomly chose a subset of 20% from the whole sets of training set and validation set respectively. The class label values of these selected examples were randomly labeled to another label. For a fairer comparison, we included the validation in the training data when the validation set was not necessary (fixed and majority filters).

5.2. Class noise removal performance

Tables 2 and 3 show respectively the accuracy of *Adaboosting.M1* and *1-NN* without noise filtering, and by noise filtering for both majority vote and margin-based methods. The margin-based approach involves the two popular definitions of ensemble margin: equations (1) and (2) (section 2) and their unsupervised versions. Two noise removal strategies are experimented. The first one is adaptive and involves the elimination of an amount of ordered potential mislabeled instances equal to the one that led to maximum accuracy on validation set. The second one just eliminates a fixed amount equal to the noise rate (20%). These tables show that the margin-based mislabeled data removal scheme significantly outperforms the majority vote filter. The accuracies achieved by adaptive filtering and by a fixed amount of filtering are slightly in favor of the fixed strategy (at least for the best performances, indicated in bold). However, the adaptive strategy does not require the knowledge of the noise rate (which is generally unknown) and leads to a more automated noise filtering procedure.

Supervised margins are more effective for class noise identification than unsupervised margins. Among the two possible definitions of ensemble margin, the second one, based on a sum operation, is the most successful for mislabeled data removal. This result is rather expected as a sum operation is more robust to noise than a max operation, at the core of the 1^{st} definition of ensemble margin. The new introduced margin, although not as efficient as its supervised counterpart (equation (2)), is also effective to identify mislabeled data and outperforms the majority vote filter. It has an appealing advantage over the supervised margin though: *it can be involved in a semi-supervised ensemble learning scheme*. Furthermore, it is more effective than the max-based unsupervised margin we proposed in previous work [11, 20] for noise removal as shown on tables 2 and 3.

5.3. Class noise correction performance

In tables 4 and 5, organised as tables 2 and 3, we attempt to correct the training data identified by margin-based or majority vote methods as mislabeled. A comparison of the results of *Adaboosting.M1* and *1-NN* on the original training data (no noise correction) and on the corrected training data reveals that margin-based algorithms reach higher accuracy while the ability of the majority vote correction to retain the baseline accuracy decreases. Unlike in noise removal, the adaptive scheme turns out more effective than the fixed one for noise correction. While the sum-based definition of ensemble margin (equation (2)) remains beyond all doubt the most appropriate for *1-NN* classifier, it is not the case for *Adaboosting.M1* classifier for which similar performances are obtained for the two different (supervised) margin definitions. Our new unsupervised margin outperforms the max-based unsupervised margin in *1-NN* classifier noise correction but is less effective with *Adaboosting.M1* classifier. Unsurprisingly, the class noise removal scheme outperforms its correction counterpart, for both majority vote and margin-based methods, the noise correction being a more challenging task. Indeed, noise correction algorithms are at high risk of inducing additional noise, and retaining bad data hinders performance more than throwing out good data.

Data	No filter	Majority filter	Margin-based noise removal							
			Classic margin (1)		Unsupervised margin (1)		Classic margin (2)		Unsupervised margin (2)	
			Fixed	Adapt.	Fixed	Adapt.	Fixed	Adapt.	Fixed	Adapt.
Letter	46.72	47.83	49.91	52.04	49.92	49.6	52.36	**56.88**	48.54	50.54
Optdigits	89.32	90.84	94.53	93.43	93.07	93.15	**94.74**	94.14	93.92	93.23
Pendigit	90.34	92.95	93.24	95.27	92.67	94.24	**95.85**	95.4	92.77	93.87
Statlog	83.38	85.68	86.27	**88.75**	85.56	86.68	88.33	88.66	85.56	86.98
Vehicle	72.2	73.7	73.5	72.3	72.6	72.05	**74.05**	72.05	72.8	73

Table 2. Accuracy of $Adaboosting.M1$ classifier with no filter, with majority vote filtered and with four margin-based filtered training sets, using both an adaptive and a fixed amount of filtering.

Data	No filter	Majority filter	Margin-based noise removal							
			Classic margin (1)		Unsupervised margin (1)		Classic margin (2)		Unsupervised margin (2)	
			Fixed	Adapt.	Fixed	Adapt.	Fixed	Adapt.	Fixed	Adapt.
Letter	74.62	59.08	81.1	79.94	78.14	78.18	**87.12**	85.12	79.16	77.6
Optdigits	77.9	93.3	92.5	93	91.1	92.9	**93.6**	93.1	91	93
Pendigit	79.75	95.05	94.9	**96.3**	94.1	95.9	96.15	95.85	93.5	94.2
Statlog	73.35	84.95	86.95	87.1	86.3	86.4	86.9	**87.15**	86.25	86.85
Vehicle	59	66.5	68.5	63	68.5	64	**70**	**70**	69	66

Table 3. Accuracy of $1-NN$ classifier with no filter, with majority vote filtered and with four margin-based filtered training sets, using both an adaptive and a fixed amount of filtering.

Data	No filter	Majority filter	Margin-based noise correction							
			Classic margin (1)		Unsupervised margin (1)		Classic margin (2)		Unsupervised margin (2)	
			Fixed	Adapt.	Fixed	Adapt.	Fixed	Adapt.	Fixed	Adapt.
Letter	46.72	45.59	43.74	**50.44**	43.59	**50.44**	43.57	48.4	43.62	50.32
Optdigits	89.32	87.98	91.7	**93.23**	89.22	92.3	92.38	92.21	87.83	91.33
Pendigit	90.34	89.56	91.87	93.57	90.53	91.84	**94**	93.64	89.34	91.57
Statlog	83.38	83.66	85.79	87.8	84.96	86.39	87.89	**88.15**	84.77	86.38
Vehicle	72.2	**74.3**	72.15	73.85	72.25	72	72.6	73.55	72.05	72.8

Table 4. Accuracy of $Adaboosting.M1$ classifier with no filter, with majority vote corrected and with four margin-based corrected training sets, using both an adaptive and a fixed amount of noise correction.

Data	No filter	Majority filter	Margin-based noise correction							
			Classic margin (1)		Unsupervised margin (1)		Classic margin (2)		Unsupervised margin (2)	
			Fixed	Adapt.	Fixed	Adapt.	Fixed	Adapt.	Fixed	Adapt.
Letter	74.62	63.2	73.38	76.44	71.32	75.72	78.5	**80.44**	72.68	75.98
Optdigits	77.9	88.7	90.4	89.1	89.2	89.2	**91.1**	90.4	89.7	88
Pendigit	79.75	89.8	93.45	92.6	92.1	91.75	**94.25**	93.65	91.35	92.05
Statlog	73.35	81.4	85.3	85.05	84.6	84.45	85.55	**85.8**	84.45	84.7
Vehicle	59	66.5	67.5	64	67	63.5	**68**	64.5	65.5	61.5

Table 5. Accuracy of $1-NN$ classifier with no filter, with majority vote corrected and with four margin-based corrected training sets, using both an adaptive and a fixed amount of noise correction.

6. CONCLUSION

We have presented a mislabeled training data identification algorithm that handles both the removal and the correction of noisy labels, based on ensemble margin. Two popular ensemble margin definitions, as well as their unsupervised alternatives that we have proposed, are assessed in our margin-based handling of the mislabeling problem. Our approach has been demonstrated to be effective for the classification of image data and significantly more accurate than the majority vote method. Future work will investigate more realistic ways of introducing artificial class noise in the data sets than random switching of class labels.

7. ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China (61162013) and China Scholarship Council.

8. REFERENCES

[1] S. Boukir, L. Guo, and N. Chehata, "Classification of remote sensing data using margin-based ensemble methods," in *ICIP'2013, IEEE International Conference on Image Processing*, Sept 2013, pp. 2602–2606.

[2] X. Zhu and X. Wu, "Class noise vs. attribute noise: A quantitative study," *Artificial Intelligence Review*, vol. 22, no. 3, pp. 177–210, 2004.

[3] D. Gamberger, N. Lavrac, and S. Dzeroski, "Noise detection and elimination in preprocessing: Experiments in medical domains," *Applied Artificial Intelligence*, vol. 14, no. 2, pp. 205–223, 2000.

[4] J.R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986.

[5] S. Verbaeten and A. Van Assche, "Ensemble methods for noise elimination in classification problems," in *4th international workshop on Multiple classifier systems, MCS'03*, 2003, pp. 317–325.

[6] C.E. Brodley and M.A. Friedl, "Identifying mislabeled training data," *Journal of Artificial Intelligence Research*, vol. 11, pp. 131–167, 1999.

[7] A. Mellor, S. Boukir, A. Haywood, and S. Jones, "Exploring issues of training data imbalance and mislabelling on random forest performance for large area land cover classification using the ensemble margin," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 105, pp. 155 – 168, 2015.

[8] A. Mellor, S. Boukir, A. Haywood, and S. Jones, "Using ensemble margin to explore issues of training data imbalance and mislabeling on large area land cover classification," in *ICIP'2014, IEEE International Conference on Image Processing*, 2014, pp. 26–29.

[9] L. Guo, *Margin framework for ensemble classifiers. Application to remote sensing data*, PhD thesis, University of Bordeaux 3, France, 2011.

[10] G.H. John, "Robust decision trees: Removing outliers from databases," in *First International Conference on Knowledge Discovery and Data Mining*, 1995, pp. 174–179.

[11] L. Guo and S. Boukir, "Ensemble margin framework for image classification," in *ICIP'2014, IEEE International Conference on Image Processing*, 2014, pp. 4231–4235.

[12] C.M. Teng, "Correcting noisy data," in *Proceedings of the Sixteenth International Conference on Machine Learning*, 1999, pp. 239–248.

[13] B. Sluban, D. Gamberger, and N. Lavrac, "Ensemble-based noise detection: noise ranking and visual performance evaluation," *Data Mining and Knowledge Discovery*, pp. 1–39, 2013.

[14] X. Zhu, X. Wu, and Q. Chen, "Eliminating class noise in large datasets," in *In Proceeding of International Conference on Machine Learning (ICML 2003)*, 2003, pp. 920–927.

[15] T.M. Khoshgoftaar, S. Zhong, and V. Joshi, "Enhancing software quality estimation using ensemble-classifier based noise filtering," *Intelligent Data Analysis*, vol. 9, no. 1, pp. 3–27, 2005.

[16] R.E. Schapire, Y. Freund, P. Bartlett, and W.S. Lee, "Boosting the margin: a new explanation for the effectiveness of voting methods," *The Annals of Statistics*, vol. 26, no. 5, pp. 1651–1686, 1998.

[17] Q. Hu, L. Li, X. Wu, G. Schaefer, and D. Yu, "Exploiting diversity for optimizing margin distribution in ensemble learning," *Knowledge-Based Systems*, vol. 67, no. 0, pp. 90 – 104, 2014.

[18] M.N. Kapp, R. Sabourin, and P. Maupin, "An empirical study on diversity measures and margin theory for ensembles of classifiers," in *Information Fusion, 2007 10th International Conference on*, July 2007, pp. 1–8.

[19] L. Guo, S. Boukir, and N. Chehata, "Support vectors selection for supervised learning using an ensemble approach," in *ICPR'2010, 20th ICPR International Conference on Pattern Recognition*, 2010, pp. 37–40.

[20] L. Guo and S. Boukir, "Margin-based ordered aggregation for ensemble pruning," *Pattern Recognition Letters*, vol. 34, no. 6, pp. 603–609, 2013.

[21] L. Breiman, "Bagging predictors," *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[22] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone, *Classification and Regression Trees*, Publisher: Wadsworth, 1984.

[23] R.O. Duda, P.E. Hart, and D.G. Stork, *Pattern Classification*, John Wiley & Sons, $2nd$ edition, 2001.

[24] Y. Freund and R.E. Schapire, "Experiments with a new boosting algorithm," in *The 13th International Conference on Machine Learning, ICML'96*, 1996, pp. 148–156.

[25] A. Asuncion and D.J. Newman, "UCI machine learning repository," 2007.