Visual Understanding by Learning From Web Data

Ziheng Zhang
ShanghaiTech University

Jia Zheng
ShanghaiTech University

Ruiyang Liu
ShanghaiTech University
{zhangzh, zhengjia, liury}@shanghaitech.edu.cn

Abstract

This report aims to study how to train a deep learning based classifier when only large scale noisy dataset is available. In order to overcome dataset noise, a series of training as well as testing methods are proposed, including bootstrapping method and ensemble method. Result shows that proposed methods achieve remarkable performance.

1. Introduction

The recent success of deep learning has shown that a deep architecture in conjunction with abundant quantities of labeled training data is the most promising approach for most vision tasks. However, annotating a large-scale dataset for training such deep neural networks is costly and time-consuming, even with the availability of scalable crowd-sourcing platforms like Amazons Mechanical Turk. As a result, there are relatively few public large-scale datasets (e.g., ImageNet and Places2) from which it is possible to learn generic visual representations from scratch.

Thus, it is unsurprising that there is continued interest in developing novel deep learning systems that train on low-cost data for image and video recognition. Among different solutions, crawling data from Internet and using the web as a source of supervision for learning deep representations has shown promising performance for a variety of important computer vision applications. However, the datasets and tasks differ in various ways, which makes it difficult to fairly evaluate different solutions, and identify the key issues when learning from web data.

We utilize a large scale web image dataset named WebVision for visual understanding by learning from web data. The datasets consists of 2.4 million of web images crawled from Interenet for 1,000 visual concepts. A lot of technology including bootstrapping, robust loss and dataset resampling are used to overcome label noise and dataset bias problem. Result shows that our approach competitive classification accuracy.

2. Related Work

There have been plenty of works about classification with noisy training dataset. Generally, approaches can be concluded as follows:

- **Dataset Cleaning.** A simple method to deal with label noise is to remove instances that appear to be mislabelled. Many such cleansing methods exist in the label noise literature. Similarly to outlier detection [3, 34] and anomaly detection [34], one can e.g. simply use methods based on ad hoc measures of anomaly and remove instances that are above a given threshold [74]. One can also remove instances that disproportionately increase the model complexity [23, 24].

  Model predictions may also be used to filter instances [24, 44] a simple heuristic is to remove training instances that are misclassified by a classifier [37], although this may remove too many instances [60, 31]. Iterative [38] and local model-based [8, 72] variants have been proposed, as well as voting filtering. With voting filtering [24, 44, 11, 78, 10], an instance is removed when all (or almost all) learners in an ensemble agree to remove it. Among other filtering methods, one may remove the instances that have an abnormally large influence on learning [54, 85], or which seem suspicious [33]. Many kNN-based methods have also been proposed (see e.g. [82, 81, 13] for surveys and comparisons), which are mainly based on heuristics [82, 13, 80, 27]. For example, the reduced nearest neighbours [27] removes instances whose removal does not cause other instances to be misclassified. Also, since AdaBoost tends to give large weights to mislabelled instances, several approaches use this unwelcome behaviour to detect label noise [78, 41].

  Hughes et al. [36] propose (i) to delete the label of the instances (and not the instances themselves) for which

Figure 1: Examples of Tench in the WebVision Dataset

experts are less reliable and (ii) to use semisupervised learning with both the labelled and the (newly) unlabelled instances. Surprisingly, this method has only been used in ECG segmentation; an open research question is whether it could be applied to other settings.

- **Robust Classification Model.** From a theoretical point of view, learning algorithms are seldom completely robust to label noise [56], except in some simple cases [71]. However, in practice, some of them are more robust than others [18, 45]. For example, bagging achieves better results than boosting [16] and several boosting methods are known to be more robust than AdaBoost [64, 66, 65]. For decision trees, the choice of the node splitting criterion can improve label noise-robustness [1]. In general, robust methods rely on overfitting avoidance to handle label noise [77].

- **Robust Training Methods.** In the probabilistic community, some authors claim that detecting label noise is impossible without making assumptions [22, 40, 75]. For example, [22] reports a probabilistic model taking label noise into account for which there is an infinite number of maximum likelihood solutions. In fact, for such identifiability issues [75], prior information is necessary to break ties. Bayesian priors on the mislabelling probabilities [40, 21] can be used, but they should be chosen carefully, for *the results obtained depend on the quality of the prior distribution* [48]. Beta priors [40, 21, 39, 30, 68, 12] and Dirichlet priors [70, 53] are common choices; Bayesian methods exist for logistic regression [12, 2, 61, 28], hidden Markov models [26] and graphical models [42]. Other approaches [69, 32, 68] are based on indicators which tell whether a given label has been flipped.

  Frequentist methods also exist to deal with label noise. A simple solution consists in using a mixture of a normal distribution and an anomalous distribution [55]. The latter is usually a uniform distribution on the instance domain, but other choices are possible. Lawrence et al. [49] have proposed a generative probabilistic model to deal with label noise. First, the true labels Y are drawn from a prior distribution p_Y . Then, the feature values are drawn from the conditional distribution $p_{X|Y}$ and the observed labels $\tilde{Y}$ from the conditional distribution $p_{\tilde{Y}|Y}$. The feature values and the observed labels are known, but the (hidden) true labels have to be inferred from the data. For example, Lawrence et al. [49] derive an EM algorithm to learn a Fisher discriminant while inferring the true labels. This has been extended to non-Gaussian conditional class distributions [51], multi-class problems [4], sequential data [19] and mutual information estimation [20]. Discriminative classifiers equipped with label noise probabilities have also been devised in [5, 6]. The model-based treatment of label noise is quite intuitive, however a theoretical analysis of the resulting algorithms is still in its infancy [7]. Instead, guarantees for risk minimisation under random label noise [62] lead to different procedures to modify a given loss function and obtain new noise-tolerant algorithms. Clustering can be used to detect mislabelled instances [67, 9], under the assumption that instances whose label is not consistent with the label of nearby clusters are likely to be mislabelled. An other solution consists in using belief functions [14, 15], since they allow modelling the confidence of the expert in its labels. When this information is not provided by the expert, several approaches have been proposed to infer beliefs directly from data [14, 15, 84]. Several other non-probabilistic models have been modified to become label noise-tolerant. For example, one can prevent instances to take too large weights in neural networks [47, 50, 43], support vector machines[25, 52] and ensembles obtained with boosting[17, 63, 29, 7]. Robust losses [59, 46, 83, 58, 73, 57] can also be used, and are theoretically shown to be less sensitive to outliers.

3. Our Approach

In order to train a classifier using noisy web data, we proposed a series of training and testing proceedure that are robust to dataset noise as shown in Figure 2. When training, the common training proceedure (in orange box) is firstly applied several iterations, followed by the bootstrapping proceedure (in greed box), which aims to clean the original noisy dataset. The two proceedures run alternately. When testing, several trained classifier are ensembled to make the final predition.

3.1. Re-sampling

We notice that WebVision training dataset suffers from category bias. The number of images per category of WebVision training dataset is shown in the Figure 3, which varies from several hundreds to more than 10,000.

In order to overcome dataset bias problem, we re-sample the dataset to guarantee that images of every class has the same probability to be sampled during training.

3.2. Data Augmentation

We use the following data augmentation method when training and testing our approach.

1. Scale and aspect ratio augmentation [76]
2. Color augmentation [35]
3. Multiple crop augmentation [76] (only when testing)

3.3. Base Model

We use ResNet-50, ResNet-152 and Inception-ResNet-v2 as our base classifier model.

3.4. Robust Loss

In order to enhance the discriminative power of deeply learned features, we investigate center loss proposed by [79], as formulated in Equation 1.

$$\mathcal{L}_C = \frac{1}{2}\sum_{i=1}^{N}\|f(x_i) - c_{y_i}\|_2^2 \tag{1}$$

where x_i denotes the i-th image in dataset, $f(x_i) \in \Re^d$ denotes deep feature of i-th image, $c_{y_i} \in \Re^d$ denotes the y_ith class center of deep embedded features.

We perform the update based on mini-batch. In each iteration, the centers are computed by averaging the features of the corresponding classes. To Avoid large perturbations caused by few mislabeled samples, we use a scale α to control the learning rate of the centers. The gradient of $\mathcal{L}_C$ with respect to x_i and update equation of centers c_{y_i} are computed as

$$\frac{\partial \mathcal{L}_C}{\partial x_i} = x_i - c_{y_i} \tag{2}$$

$$\Delta c_{y_i} = \frac{\sum_{i=1}^{m}\delta(y_i = j)\cdot(c_j - x_i)}{1 + \sum_{i=1}^{m}\delta(y_i = j)} \tag{3}$$

where $\delta(condition) = 1$ if the $condition$ is satisfied, and $\delta(condition) = 0$ if not.

3.5. Bootstrapping

We can use the prediction of base classifiers to find the potentially false label. The Bootstrapping algorithm is as follows

Input: training set
initialize an empty label mask table
foreach *image in training set* **do**
  predict label for the image
  if *predicted label equals to image label* **then**
    set mask label to image label
  else
    if *predicted probobality greater than a threshold* **then**
      set mask label to predicted label
    else
      set mask label to minus one
    end
  end
end
Output: label mask table

Algorithm 1: Bootstrapping Method

3.6. Model Ensemble

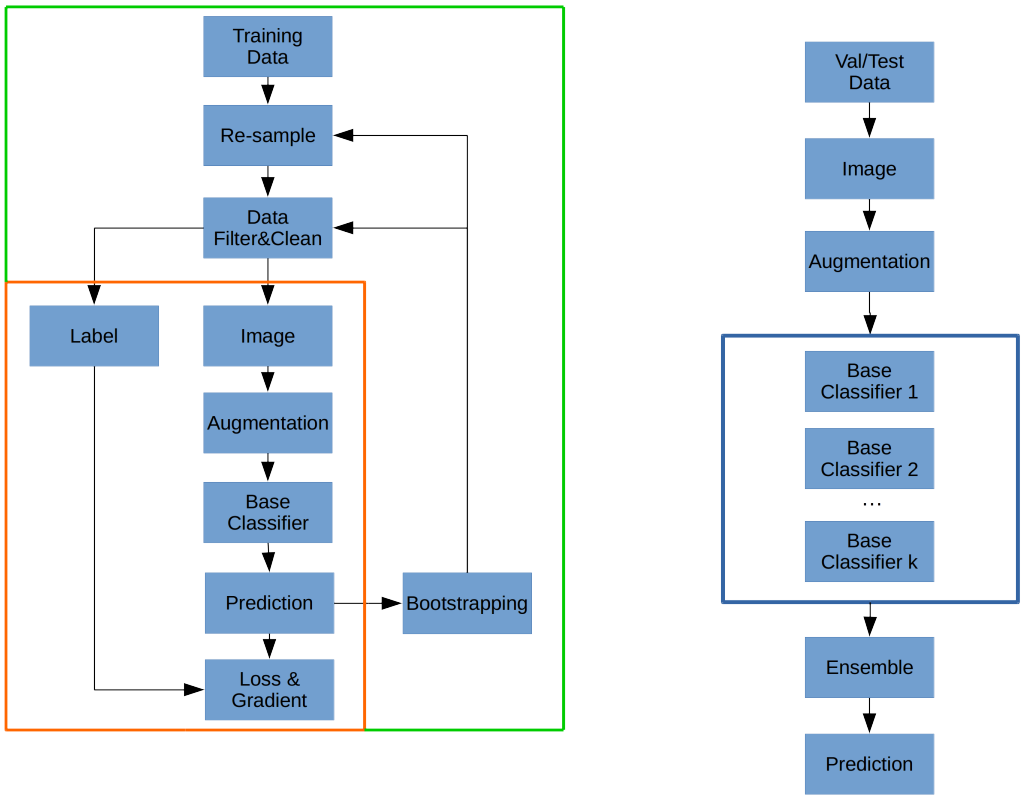
We noticed that if we train the same base classifier several times, their overall performance is similar, but their performance on each class slightly varies.

We can use multiple base classifiers to predict labels for each image when testing. This is called model ensemble. For now, we just simply average the output probabilities throughout all classifiers.

4. Experiments

We trained a ResNet152 classifier for one week, and sets learning rate to the initial learning rate decayed by 10 every 10 epochs. For now (28 epoch), The top-1 (and top-5) accuracies are $71.208\%(88.860\%)$ without ensemble, single crop, single bootstrapping. Baseline[1] is $58.98\%(79.30\%)$.

[1]Baseline: train the AlexNet models on this training set from scratch.



(a) Training proceedure (b) Testing proceedure

Figure 2: Proposed training and testing proceedure

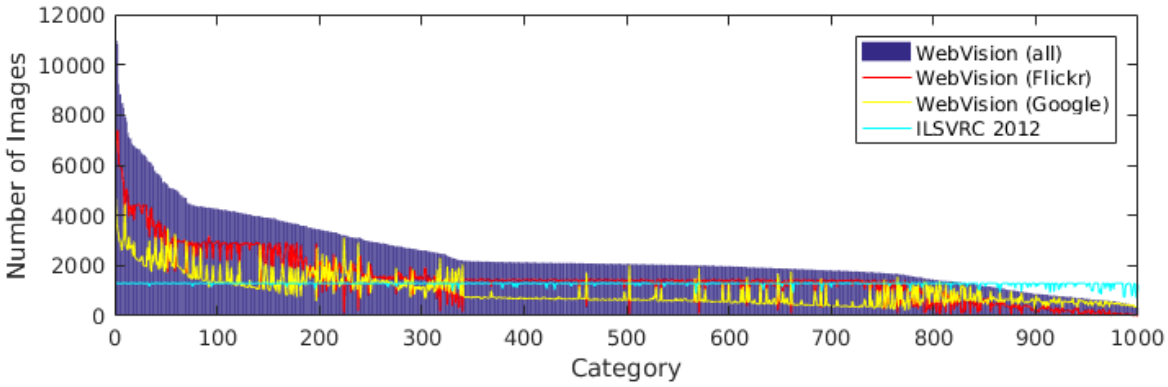


Figure 3: The number of images per category in the WebVision training dataset.

References

[1] J. Abellán and S. Moral. Building classification trees using the total uncertainty criterion. *International Journal of Intelligent Systems*, 18(12):1215–1225, 2003.

[2] J. A. Achcar, E. Z. Martinez, and F. Louzada-Neto. Binary data in the presence of misclassifications. In *Proc. 16th Symp. Int. Assoc. Statist. Comput*, pages 581–587, 2004.

[3] R. J. Beckman and R. D. Cook. Outlier. s. *Technometrics*, 25(2):119–149, 1983.

[4] J. Bootkrajang and A. Kabán. Multi-class classification in the presence of labelling errors. In *ESANN*, pages 345–350. Citeseer, 2011.

[5] J. Bootkrajang and A. Kabán. Label-noise robust logistic regression and its applications. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 143–158. Springer, 2012.

[6] J. Bootkrajang and A. Kabán. Classification of mislabelled microarrays using robust sparse logistic regression. *Bioinformatics*, page btt078, 2013.

[7] J. Bootkrajang and A. Kabán. Learning a label-noise robust logistic regression: Analysis and experiments. In *International Conference on Intelligent Data Engineering and Automated Learning*, pages 569–576. Springer, 2013.

[8] L. Bottou and V. Vapnik. Local learning algorithms. *Neural computation*, 4(6):888–900, 1992.

[9] C. Bouveyron and S. Girard. Robust supervised classification with mixture models: Learning from data with uncertain labels. *Pattern Recognition*, 42(11):2649–2658, 2009.

[10] C. E. Brodley and M. A. Friedl. Identifying mislabeled training data. *Journal of artificial intelligence research*, 11:131–167, 1999.

[11] C. E. Brodley, M. A. Friedl, et al. Identifying and eliminating mislabeled training instances. In *AAAI/IAAI, Vol. 1*, pages 799–805, 1996.

[12] C. Daniel Paulino, P. Soares, and J. Neuhaus. Binomial regression with misclassification. *Biometrics*, 59(3):670–675, 2003.

[13] S. J. Delany, N. Segata, and B. Mac Namee. Profiling instances in noise reduction. *Knowledge-Based Systems*, 31:28–40, 2012.

[14] T. Denoeux. A k-nearest neighbor classification rule based on dempster-shafer theory. *IEEE transactions on systems, man, and cybernetics*, 25(5):804–813, 1995.

[15] T. Denoeux. A neural network classifier based on dempster-shafer theory. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 30(2):131–150, 2000.

[16] T. G. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Machine learning*, 40(2):139–157, 2000.

[17] C. Domingo and O. Watanabe. Madaboost: A modification of adaboost. In *COLT*, pages 180–189, 2000.

[18] A. Folleco, T. M. Khoshgoftaar, J. Van Hulse, and L. Bullard. Identifying learners robust to low quality data. In *Information Reuse and Integration, 2008. IRI 2008. IEEE International Conference on*, pages 190–195. IEEE, 2008.

[19] B. Frénay, G. de Lannoy, and M. Verleysen. Label noise-tolerant hidden markov models for segmentation: application to ecgs. *Machine learning and knowledge discovery in databases*, pages 455–470, 2011.

[20] B. Frénay, G. Doquire, and M. Verleysen. Estimating mutual information for feature selection in the presence of label noise. *Computational Statistics & Data Analysis*, 71:832–848, 2014.

[21] A. Gaba. Inferences with an unknown noise level in a bernoulli process. *Management science*, 39(10):1227–1237, 1993.

[22] A. Gaba and R. L. Winkler. Implications of errors in survey data: a bayesian model. *Management Science*, 38(7):913–925, 1992.

[23] D. Gamberger and N. Lavrač. Conditions for occam's razor applicability and noise elimination. *Machine Learning: ECML-97*, pages 108–123, 1997.

[24] D. Gamberger, N. Lavrac, and C. Groselj. Experiments with noise filtering in a medical domain. In *ICML*, pages 143–151, 1999.

[25] A. Ganapathiraju, J. Picone, et al. Support vector machines for automatic data cleanup. In *INTERSPEECH*, pages 210–213, 2000.

[26] M. J. García-Zattera, T. Mutsvari, A. Jara, D. Declerck, and E. Lesaffre. Correcting for misclassification for a monotone disease process with an application in dental research. *Statistics in medicine*, 29(30):3103–3117, 2010.

[27] G. W. Gates. observations). direct calculation of pn*(co) was possible for small values of n. for larger values of n it became necessary to upper-bound pn*(co). a chernoff bound on p,*(m) is. 1972.

[28] R. Gerlach and J. Stamey. Bayesian model selection for logistic regression with misclassified outcomes. *Statistical Modelling*, 7(3):255–273, 2007.

[29] V. Gómez-Verdejo, M. Ortega-Moral, J. Arenas-García, and A. R. Figueiras-Vidal. Boosting by weighting critical and erroneous samples. *Neurocomputing*, 69(7):679–685, 2006.

[30] P. Gustafson, N. D. Le, and R. Saskin. Case–control analysis with partial knowledge of exposure misclassification probabilities. *Biometrics*, 57(2):598–609, 2001.

[31] I. Guyon, N. Matic, V. Vapnik, et al. Discovering informative patterns and data cleaning., 1996.

[32] D. Hernández-Lobato, J. M. Hernández-Lobato, and P. Dupont. Robust multi-class gaussian process classification. In *Advances in neural information processing systems*, pages 280–288, 2011.

[33] T. Heskes. The use of being stubborn and introspective. In *Prerational Intelligence: Adaptive Behavior and Intelligent Systems Without Symbols and Logic, Volume 1, Volume 2 Prerational Intelligence: Interdisciplinary Perspectives on the Behavior of Natural and Artificial Systems, Volume 3*, pages 1184–1200. Springer, 2000.

[34] V. Hodge and J. Austin. A survey of outlier detection methodologies. *Artificial intelligence review*, 22(2):85–126, 2004.

[35] A. G. Howard. Some improvements on deep convolutional neural network based image classification. *arXiv preprint arXiv:1312.5402*, 2013.

[36] N. P. Hughes, S. J. Roberts, and L. Tarassenko. Semi-supervised learning of probabilistic models for ecg segmentation. In *Engineering in Medicine and Biology Society, 2004. IEMBS'04. 26th Annual International Conference of the IEEE*, volume 1, pages 434–437. IEEE, 2004.

[37] P. Jeatrakul, K. W. Wong, and C. C. Fung. Data cleaning for classification using misclassification analysis. *Journal of Advanced Computational Intelligence and Intelligent Informatics*, 14(3):297–302, 2010.

[38] G. H. John. Robust decision trees: Removing outliers from databases. In *KDD*, pages 174–179, 1995.

[39] L. Joseph and T. W. Gyorkos. Inferences for likelihood ratios in the absence of a" gold standard". *Medical decision making*, 16(4):412–417, 1996.

[40] L. Joseph, T. W. Gyorkos, and L. Coupal. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *American Journal of Epidemiology*, 141(3):263–272, 1995.

[41] A. Karmaker and S. Kwek. A boosting approach to remove class label noise1. *International Journal of Hybrid Intelligent Systems*, 3(3):169–177, 2006.

[42] F. O. Kaster, B. H. Menze, M.-A. Weber, and F. A. Hamprecht. Comparative validation of graphical models for learning tumor segmentations from noisy manual annotations. In *International MICCAI Workshop on Medical Computer Vision*, pages 74–85. Springer, 2010.

[43] R. Khardon and G. Wachman. Noise tolerant variants of the perceptron algorithm. *Journal of Machine Learning Research*, 8(Feb):227–248, 2007.

[44] T. M. Khoshgoftaar and P. Rebours. Generating multiple noise elimination filters with the ensemble-partitioning filter. In *Information Reuse and Integration, 2004. IRI 2004. Proceedings of the 2004 IEEE International Conference on*, pages 369–375. IEEE, 2004.

[45] T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano. Supervised neural network modeling: an empirical investigation into learning from imbalanced data with labeling errors. *IEEE Transactions on Neural Networks*, 21(5):813–830, 2010.

[46] N. Krause and Y. Singer. Leveraging the margin more carefully. In *Proceedings of the twenty-first international conference on Machine learning*, page 63. ACM, 2004.

[47] W. Krauth and M. Mézard. Learning algorithms with optimal stability in neural networks. *Journal of Physics A: Mathematical and General*, 20(11):L745, 1987.

[48] M. Ladouceur, E. Rahme, C. A. Pineau, and L. Joseph. Robustness of prevalence estimates derived from misclassified data from administrative databases. *Biometrics*, 63(1):272–279, 2007.

[49] N. D. Lawrence and B. Schölkopf. Estimating a kernel fisher discriminant in the presence of label noise. In *ICML*, volume 1, pages 306–313. Citeseer, 2001.

[50] Y. Li and P. M. Long. The relaxed online maximum margin algorithm. *Machine Learning*, 46(1-3):361–387, 2002.

[51] Y. Li, L. F. Wessels, D. de Ridder, and M. J. Reinders. Classification in the presence of class noise using a probabilistic kernel fisher method. *Pattern Recognition*, 40(12):3349–3357, 2007.

[52] C.-f. Lin et al. Training algorithms for fuzzy support vector machines with noisy data. *Pattern recognition letters*, 25(14):1647–1656, 2004.

[53] J. Liu, P. Gustafson, N. Cherry, and I. Burstyn. Bayesian analysis of a matched case–control study with expert prior information on both the misclassification of exposure and the exposure–disease association. *Statistics in medicine*, 28(27):3411–3423, 2009.

[54] A. Malossini, E. Blanzieri, and R. T. Ng. Detecting potential labeling errors in microarrays by data perturbation. *Bioinformatics*, 22(17):2114–2121, 2006.

[55] Y. Mansour and M. Parnas. Learning conjunctions with noise under product distributions. *Information Processing Letters*, 68(4):189–196, 1998.

[56] N. Manwani and P. Sastry. Noise tolerance under risk minimization. *IEEE transactions on cybernetics*, 43(3):1146–1151, 2013.

[57] H. Masnadi-Shirazi, V. Mahadevan, and N. Vasconcelos. On the design of robust classifiers for computer vision. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 779–786. IEEE, 2010.

[58] H. Masnadi-Shirazi and N. Vasconcelos. On the design of loss functions for classification: theory, robustness to outliers, and savageboost. In *Advances in neural information processing systems*, pages 1049–1056, 2009.

[59] L. Mason, J. Baxter, P. L. Bartlett, M. Frean, et al. Functional gradient techniques for combining hypotheses. *Advances in Neural Information Processing Systems*, pages 221–246, 1999.

[60] N. Matic, I. Guyon, L. Bottou, J. Denker, and V. Vapnik. Computer aided cleaning of large databases for character recognition. In *Pattern Recognition, 1992. Vol. II. Conference B: Pattern Recognition Methodology and Systems, Proceedings., 11th IAPR International Conference on*, pages 330–333. IEEE, 1992.

[61] P. McInturff, W. O. Johnson, D. Cowling, and I. A. Gardner. Modelling risk when binary outcomes are subject to error. *Statistics in medicine*, 23(7):1095–1109, 2004.

[62] N. Natarajan, I. S. Dhillon, P. K. Ravikumar, and A. Tewari. Learning with noisy labels. In *Advances in neural information processing systems*, pages 1196–1204, 2013.

[63] N. C. Oza. Aveboost2: Boosting for noisy data. In *International Workshop on Multiple Classifier Systems*, pages 31–40. Springer, 2004.

[64] G. Rätsch, T. Onoda, and K. R. Müller. An improvement of adaboost to avoid overfitting. In *Proc. of the Int. Conf. on Neural Information Processing*. Citeseer, 1998.

[65] G. Rätsch, T. Onoda, and K.-R. Müller. Soft margins for adaboost. *Machine learning*, 42(3):287–320, 2001.

[66] G. Rätsch, B. Schölkopf, A. J. Smola, S. Mika, T. Onoda, and K.-R. Müller. Robust ensemble learning for data mining. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 341–344. Springer, 2000.

[67] U. Rebbapragada and C. E. Brodley. Class noise mitigation through instance weighting. In *European Conference on Machine Learning*, pages 708–715. Springer, 2007.

[68] R. Rekaya, K. Weigel, and D. Gianola. Threshold model for misclassified binary responses with applications to animal breeding. *Biometrics*, 57(4):1123–1129, 2001.

[69] K. Robbins, S. Joseph, W. Zhang, R. Rekaya, and J. Bertrand. Classification of incipient alzheimer patients using gene expression data: Dealing with potential misdiagnosis. *Online Journal of Bioinformatics*, 7(1):22–31, 2006.

[70] M. Ruiz, F. Girón, C. Pérez, J. Martín, and C. Rojano. A bayesian model for multinomial sampling with misclassified data. *Journal of Applied Statistics*, 35(4):369–382, 2008.

[71] P. Sastry, G. Nagendra, and N. Manwani. A team of continuous-action learning automata for noise-tolerant learning of half-spaces. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 40(1):19–28, 2010.

[72] N. Segata, E. Blanzieri, S. J. Delany, and P. Cunningham. Noise reduction for instance-based learning with a local maximal margin approach. *Journal of Intelligent Information Systems*, 35(2):301–331, 2010.

[73] G. Stempfel and L. Ralaivola. Learning svms from sloppily labeled data. *Artificial Neural Networks–ICANN 2009*, pages 884–893, 2009.

[74] J.-w. Sun, F.-y. Zhao, C.-j. Wang, and S.-f. Chen. Identifying and correcting mislabeled training instances. In *Future generation communication and networking (FGCN 2007)*, volume 1, pages 244–250. IEEE, 2007.

[75] T. B. Swartz, Y. Haitovsky, A. Vexler, and T. Y. Yang. Bayesian identifiability and misclassification in multinomial data. *Canadian Journal of Statistics*, 32(3):285–302, 2004.

[76] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–9, 2015.

[77] C.-M. Teng. A comparison of noise handling techniques. In *FLAIRS Conference*, pages 269–273, 2001.

[78] S. Verbaeten and A. Van Assche. Ensemble methods for noise elimination in classification problems. In *International Workshop on Multiple Classifier Systems*, pages 317–325. Springer, 2003.

[79] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. In *European Conference on Computer Vision*, pages 499–515. Springer, 2016.

[80] D. L. Wilson. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Transactions on Systems, Man, and Cybernetics*, 2(3):408–421, 1972.

[81] D. R. Wilson and T. R. Martinez. Instance pruning techniques. In *ICML*, volume 97, pages 403–411, 1997.

[82] D. R. Wilson and T. R. Martinez. Reduction techniques for instance-based learning algorithms. *Machine learning*, 38(3):257–286, 2000.

[83] L. Xu, K. Crammer, and D. Schuurmans. Robust support vector machine training via convex outlier ablation. In *AAAI*, volume 6, pages 536–542, 2006.

[84] Z. Younes, T. Denœux, et al. Evidential multi-label classification approach to learning from data with imprecise labels. In *International Conference on Information Processing and Management of Uncertainty in Knowledge-Based Systems*, pages 119–128. Springer, 2010.

[85] C. Zhang, C. Wu, E. Blanzieri, Y. Zhou, Y. Wang, W. Du, and Y. Liang. Methods for labeling error detection in microarrays based on the effect of data perturbation on the regression model. *Bioinformatics*, 25(20):2708–2714, 2009.