1 Some literature on accounting for misclassified labels in machine learning - a reading guide

Arnout van Delden (April 2025)

1.1 Introduction

The labels that are used for training machine learning models might not always be error-free. In some situations it may already be quite difficult to actually determine by humans what is the correct label. So, we are interested to learn which possibilities for fitting machine learning models there are, when there are errors in the labels of the training set; we also refer to this as label noise.

1.2 Kind of label noise

In many papers, such as Eskin (2002), Sigurdsson et al. (2002) and Rantalainen & Holmes (2011) one assumes a completely random label noise. The noise is completely random in two ways a) the units that are misclassified are selected randomly from the population and b) the wrong category that they have is selected randomly from the set of categories of the classification variable. Kolcz & Cormack (2009) and Sarma & Palmer (2004) are examples of studies in which the probability that a unit is misclassified is no longer completely random.

1.3 Approaches for dealing with misclassifications

Frénay & Verleysen (2014) gives an interesting overview of literature on the application of supervised learning in the presence of noisy labels. The authors clusters the literature in three approaches to deal with noisy labels. The first set of literature concerns the ML algorithms that have some form of robustness against the presence of noisy labels. Such algorithms can still give good predictions also in the presence of label noise. Examples of such studies are Biggio et al. (2011), Bootkrajang & Kaban (2012), Bouveyron et al. (2009), Li et al. (2007), Stempfel & Ralaivola (2009), Sigurdsson et al. (2002) and Sukhbaatar & Fergus (2014).

There is a separate set of literature that treats the topic of 'outlier-robust' machine learning, see Baher et al. (2010), Kim & Ghahramani (2008), Larsen et al. (1998) and Talak et al. (2024)

as examples. From these papers it becomes clear that this literature refers misclassified labels that occur with a small probability. In official statistics however the term ’outlier’ is used for when the value of a variable deviates strongly from the value of similar units, but it is a *correct* value. In the AI literature it is apparently used in as a special case of label noise, namely when it occurs with a small probability.

In the second set of literature uses the strategy to try to predict which cases are likely to be error-free. This can for instance be done by predicting the labeled units with multiple machine learning models. One then assumes that the cases where the predictions of the machine learning models agree are more likely to be error-free than the other cases. An example of this approach is given by Brodley & Friedl (1999).

The third set of papers builds an explicit model for label noise is used as part of the learning process. Examples of this approach are given in Eskin (2002), Garg et al. (2021), Lawrence & Schölkopf (2001), Rantalainen & Holmes (2011) and Sigurdsson et al. (2002). Literature that uses a label noise model usually make the assumption that throughout the population (trainingset) there is a fixed (unknown) proportion of the labels that is incorrect, see for instance Eskin (2002), Mansour & Parnas (1998), Rantalainen & Holmes (2011) and Sigurdsson et al. (2002).

Frénay et al. (2011) mentions a fourth approach, which they refer to as ’plausibilistic approaches’, which refers to experts that have given their opinion about uncertainties with respect to the labels. Next, specific algorithms can deal with those uncertainties, see for instance Côme et al. (2008).

References

Baher, H. L., Lemaire, V., & Trinquart, R. (2010). On the intrinsic robustness to noise of some leading classifiers and symmetric loss function - an empirical evaluation. *arXiv:2010.13570v5 [cs.LG] 21 Jun 2021*.

Biggio, B., Nelson, B., & Laskov, P. (2011). Support vector machines under adversarial label noise. *Journal of Machine Learning Research - Proceedings Track*, *20*, 97-112.

Bootkrajang, J., & Kaban, A. (2012). Label-noise robust logistic regression and its applications. *Joint European conference on machine learning and knowledge discovery in databases*, 143-158.

Bouveyron, C., Girard, S., & Olteanu, M. (2009). Supervised classification of categorical data withuncertain labels for dna barcoding. In *Advances in computationalintelligence and learning* (p. 29-34). ESANN'2009 proceedings, European Symposium on Artificial Neural Networks.

Brodley, C. E., & Friedl, M. A. (1999). Identifying mislabeled training data. *Journal of artificial intelligence research*, *11*(1), 131-167. Retrieved from http://dl.acm.org/citation.cfm?id=3013545.3013548

Côme, E., Oukhellou, L., Denœux, T., & Aknin, P. (2008). Mixture model estimation with soft labels. In *Fourth international workshop on soft methods in probabilities and statistics* (p. 282–293). doi: 10.1007/978-3-540-85027-4_21

Eskin, E. (2002). Detecting errors within a corpus using anomaly detection. *Proceedings of the 1st North American chaper of the Association for Computational Linguistics conference*, 148-153.

Frénay, B., de Lannoy, G., & Verleysen, M. (2011). Label noise-tolerant hidden markov models for segmentation: Application to ecgs. In D. Gunopulos, T. Hofmann, D. Malerba, & M. Vazirgiannis (Eds.), *Machine learning and knowledge discovery in databases* (pp. 455–470). Berlin, Heidelberg: Springer Berlin Heidelberg.

Frénay, B., & Verleysen, M. (2014). Classification in the presence of label noise: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, *25*, 845-869. doi: 10.1109/TNNLS.2013.2292894

Garg, S., Ramakrishnan, G., & Thumbe, V. (2021). Towards robustness to label noise in text classification via noise modeling. In *Proceedings of the 30th acm international conference on information & knowledge management* (p. 3024–3028). New York, NY, USA: Association for Computing Machinery. Retrieved from https://doi.org/10.1145/3459637.3482204 doi: 10.1145/3459637.3482204

Kim, H.-C., & Ghahramani, Z. (2008). Outlier robust gaussian process classification. In N. da Vitoria Lobo et al. (Eds.), *Structural, syntactic, and statistical pattern recognition* (pp. 896–905). Berlin, Heidelberg: Springer Berlin Heidelberg.

Kolcz, A., & Cormack, G. V. (2009). Genre-based decomposition of email class noise. In *Proceedings of the 15th acm sigkdd international conference on knowledge discovery and data mining* (pp. 427–436).

Larsen, J., Nonboe, L., Hintz-Madsen, M., & Hansen, L. (1998). Design of robust neural network classifiers. In *Proceedings of the 1998 ieee international conference on acoustics, speech and signal processing, icassp '98 (cat. no.98ch36181)* (Vol. 2, p. 1205-1208 vol.2). doi: 10.1109/ICASSP.1998.675487

Lawrence, N. D., & Schölkopf, B. (2001). Estimating a kernel fisher discriminant in the presence of label noise. In *Proceedings of the eighteenth international conference on machine learning* (p. 306–313). San Francisco, CA, USA: Morgan Kaufmann Publishers Inc.

Li, Y., Wessels, L. F., de Ridder, D., & Reinders, M. J. (2007). Classification in the presence of class noise using a probabilistic kernel fisher method. *Pattern Recognition*, *40*(12), 3349-3357. Retrieved from https://www.sciencedirect.com/science/article/pii/S0031320307002166 doi: https://doi.org/10.1016/j.patcog.2007.05.006

Mansour, Y., & Parnas, M. (1998). Learning conjunctions with noise under product distributions. *Information Processing Letters*, *68*, 189–196.

Rantalainen, M., & Holmes, C. (2011). Accounting for control mislabeling in case–control biomarker studies. *Journal of proteome research*, 5562-5567.

Sarma, A., & Palmer, D. D. (2004). Context-based speech recognition error detection and correction. In *Proceedings of hlt-naacl 2004: Short papers* (pp. 85–88). Retrieved from https://www.aclweb.org/anthology/N04-4022

Sigurdsson, S., Larsen, J., Hansen, L., Philipsen, P., & Wulf, H. (2002). Outlier estimation and detection application to skin lesion classification. *IEEE International Conference on Acoustics Speech and Signal Processing*, *1*. doi: 10.1109/ICASSP.2002.5743975

Stempfel, G., & Ralaivola, L. (2009). Learning svms from sloppily labeled data. In C. Alippi, M. Polycarpou, C. Panayiotou, & G. Ellinas (Eds.), *Artificial neural networks – icann 2009* (pp. 884–893). Berlin, Heidelberg: Springer Berlin Heidelberg.

Sukhbaatar, S., & Fergus, R. (2014). Learning from noisy labels with deep neural networks. *arXiv preprint arXiv:1406.2080*, *2*(3), 4.

Talak, R., Georgiou, C., Shi, J., & Carlone, L. (2024). Outlier-robust training of machine learning models. *arXiv:2501.00265v1 [cs.LG] 31 Dec 2024*.