ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7. Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

A Comprehensive Introduction to Label Noise

Benoît Frénay[1] and Ata Kabán[2]

1- Université catholique de Louvain - ICTEAM/ELEN - Machine Learning Group
Place du Levant 3, 1348 Louvain-la-Neuve - Belgium

2- University of Birmingham - School of Computer Science
Edgbaston Birmingham B15 2TT - United Kingdom

Abstract. In classification, it is often difficult or expensive to obtain completely accurate and reliable labels. Indeed, labels may be polluted by label noise, due to e.g. insufficient information, expert mistakes, and encoding errors. The problem is that errors in training labels that are not properly handled may deteriorate the accuracy of subsequent predictions, among other effects. Many works have been devoted to label noise and this paper provides a concise and comprehensive introduction to this research topic. In particular, it reviews the types of label noise, their consequences and a number of state of the art approaches to deal with label noise.

1 Introduction

In classification, it is both expensive and difficult to obtain reliable labels, yet traditional classifiers assume and expect a perfectly labelled training set. This paper reviews the increasing literature devoted to label noise (i.e. errors in available labels), and it is largely based on [1, 2] and recent work.

Mislabelling may come from different sources. First, the available information may be insufficient to perform reliable labelling [3, 4], e.g. if the description language is too limited [5] or if data are of poor quality [6]. Second, even experts often make mistakes during labelling [4]. Third, classification is in some cases subjective [7, 8], which results in inter-expert variability [9]. For example, the pattern boundaries provided by two experts for the segmentation of electrocardiogram signals are often slightly different [10]. In addition, incorrect labels may come from communication or encoding problems [11, 3, 12]; real-word databases are estimated to contain around five percent of encoding errors [13, 14].

Figure 1 shows a taxonomy of label noise, proposed in [1] and inspired by the work of Schafer and Graham [15]. Three types of noise are distinguished here. First, label noise *completely at random (NCAR)* occurs independently of the true class and of the values of the instance features. Second, label noise that occurs *at random (NAR)* depends only on the true label. This can be used to model situations where some classes are more likely to be mislabelled than others. Third, label noise *not at random (NNAR)* is the more general case, where the mislabelling probability also depends on the feature values. This allows us to model labelling errors near the classification boundaries.

2 Consequences of Label Noise

Label noise is ubiquitous in real-word datasets, and has several consequences.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence
and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7.
Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

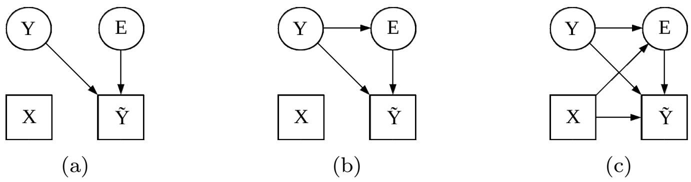

Figure 1: Statistical taxonomy of label noise proposed in [1]: (a) noisy completely at random (NCAR), (b) noisy at random (NAR) and (c) noisy not at random (NNAR). Squares and circles correspond to observed and unobserved variables respectively. Arrows represent statistical dependencies between the observed features X, the true class Y, the observed label $\tilde{Y}$ and E indicating whether a labelling error occurred. The complexity of dependencies in these models increase from left to right. The link between X and Y is not shown for clarity.

First, label noise decreases the prediction performances, which has been theoretically proved for simple models like linear classifiers [16, 17, 18], quadratic classifiers [19] or kNN classifiers [20, 21]. Many works [22, 23] have empirically confirmed this issue for other classifiers like decision trees induced by C4.5 and support vector machines. Boosting is also well known to be affected by label noise [24]. In particular, the AdaBoost algorithm tends to give too large weights to mislabelled instances [24, 25]. Most studies only deal with NCAR or NAR label noise, but NNAR label noise is also studied e.g. in [26].

Second, the number of necessary training instances [11, 27] may increase, as well as the complexity of inferred models, like e.g. the number of nodes of decision trees [22, 3] and the number of support vectors in SVMs [3].

Third, the observed frequencies of the possible classes may be altered [28, 6, 29], which is of particular importance in medical contexts. Indeed, medical studies are often concerned about measuring the incidence of a given disease in a population, whose estimation may be biased by label noise. This is also important in model validation, since performance measures can be poorly estimated in the presence of label noise [30], like e.g. in the case of spam filters [31].

Eventually, other related tasks like feature selection [32, 33] or feature ranking [34] are also impacted by label noise.

3 State of the Art Methods to Deal with Label Noise

In light of Section 2, it is imperative to deal with label noise. There exist three types of approaches in the literature [35, 36, 37, 38, 39]: label noise-robust models, data cleansing methods and label noise-tolerant learning algorithms.

3.1 Label Noise-Robust Models

From a theoretical point of view, learning algorithms are seldom completely robust to label noise [38], except in some simple cases [40]. However, in practice,

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7. Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

some of them are more robust than others [41, 42]. For example, bagging achieves better results than boosting [24] and several boosting methods are known to be more robust than AdaBoost [43, 44, 45, 46]. For decision trees, the choice of the node splitting criterion can improve label noise-robustness [47]. In general, robust methods rely on overfitting avoidance to handle label noise [35].

3.2 Data Cleansing Methods

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 [48, 49] and anomaly detection [50], one can e.g. simply use methods based on ad hoc measures of anomaly and remove instances that are above a given threshold [51]. One can also remove instances that disproportionately increase the model complexity [52, 53].

Model predictions may also be used to filter instances [53, 54] – a simple heuristic is to remove training instances that are misclassified by a classifier [55], although this may remove too many instances [56, 57]. Iterative [58] and local model-based [59, 60] variants have been proposed, as well as voting filtering. With voting filtering [61, 3, 53, 62, 54], 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 [8, 63], or which seem suspicious [18]. Many kNN-based methods have also been proposed (see e.g. [64, 21, 65] for surveys and comparisons), which are mainly based on heuristics [66, 67, 64, 21]. For example, the reduced nearest neighbours [66] 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 [62, 68].

Hughes et al. [10] propose (i) to delete the label of the instances (and not the instances themselves) for which experts are less reliable and (ii) to use semi-supervised 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.

3.3 Label Noise-Tolerant Learning Algorithms

In the probabilistic community, some authors claim that detecting label noise is impossible without making assumptions [29, 69, 70]. For example, [29] 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 [70], prior information is necessary to break ties. Bayesian priors on the mislabelling probabilities [71, 69] can be used, but they should be chosen carefully, for *the results obtained depend on the quality of the prior distribution* [72]. Beta priors [71, 69, 73, 74, 75, 76] and Dirichlet priors [77, 78] are common choices; Bayesian methods exist for logistic regression [76, 79, 80, 32], hidden Markov

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7. Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

models [81] and graphical models [82]. Other approaches [75, 83, 84] 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 [85]. The latter is usually a uniform distribution on the instance domain, but other choices are possible. Lawrence et al. [86] 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. [86] derive an EM algorithm to learn a Fisher discriminant while inferring the true labels. This has been extended to non-Gaussian conditional class distributions [87], multi-class problems [88], sequential data [89] and mutual information estimation [33]. Discriminative classifiers equipped with label noise probabilities have also been devised in [90, 91]. The model-based treatment of label noise is quite intuitive, however a theoretical analysis of the resulting algorithms is still in its infancy [92]. Instead, guarantees for risk minimisation under random label noise [93] lead to different procedures to modify a given loss function and obtain new noise-tolerant algorithms.

Clustering can be used to detect mislabelled instances [94, 37], 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 [95, 96], 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 [95, 96, 97].

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 [98, 99, 100], support vector machines [101, 102] and ensembles obtained with boosting [103, 104, 105, 92]. Robust losses [106, 107, 108, 109, 110, 111] can also be used, and are theoretically shown to be less sensitive to outliers.

4 Experimental Considerations to Assess Algorithms

There exist only a few real-world datasets where mislabelled instances have been identified [112, 8, 83, 113, 31]. In most experiments, label noise is artificially introduced in datasets. NCAR label noise is introduced by picking instances at random and flipping their label [36]. Several works use asymmetric flipping strategies to consider NAR label noise [4, 114, 12, 41, 42, 23], in order to simulate situations where some classes are more likely to be polluted than others. Eventually, a few works deal with NNAR label noise which is introduced in ambiguous regions [26, 40]. Open research questions include how to obtain more real-world datasets where mislabelled instances are clearly identified and what the characteristics of real-world label noise are. In the literature, it is not yet

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7. Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

clear if and when NCAR, NAR or NNAR label noise is the most realistic.

Criteria to assess algorithms which deal with label noise include the classification accuracy [58, 61, 3, 53, 62, 114], the model complexity [61, 3, 62], the accuracy of the estimation of true frequencies from observed frequencies [28, 29, 30] and the filter precision for data cleansing methods [3, 62, 114, 115].

References

[1] Benoît Frénay and Michel Verleysen. Classification in the presence of label noise: a survey. *IEEE Trans. Neural Netw. Learn. Syst.*, 2014, in press, 25 pages.

[2] Benoît Frénay. *Uncertainty and Label Noise in Machine Learning.* PhD thesis, Université catholique de Louvain, 2013.

[3] Carla E. Brodley and Mark A. Friedl. Identifying mislabeled training data. *J. Artif. Intell. Res.*, 11:131–167, 1999.

[4] Ray J. Hickey. Noise modelling and evaluating learning from examples. *Artif. Intell.*, 82(1-2):157–179, 1996.

[5] P. B. Brazdil and P. Clark. Learning from imperfect data. In P. B. Brazdil and K. Konolige, editors, *Machine Learning, Meta-Reasoning and Logics*, pages 207–232. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1990.

[6] A P Dawid and A M Skene. Maximum likelihood estimation of observer error-rates using the em algorithm. *J. Roy. Stat. Soc. C. - App.*, 28(1):20–28, 1979.

[7] Padhraic Smyth, Usama M. Fayyad, Michael C. Burl, Pietro Perona, and Pierre Baldi. Inferring ground truth from subjective labelling of venus images. In *Advances in Neural Information Processing Systems 7*, pages 1085–1092, Denver, CO, November–December 1994.

[8] Andrea Malossini, Enrico Blanzieri, and Raymond T. Ng. Detecting potential labeling errors in microarrays by data perturbation. *Bioinformatics*, 22(17):2114–2121, 2006.

[9] Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. Building a large annotated corpus of english: the penn treebank. *Comput. Linguist.*, 19(2):313–330, 1993.

[10] Nicholas P Hughes, Stephen J Roberts, and Lionel Tarassenko. Semi-supervised learning of probabilistic models for ecg segmentation. In *Ann. Int. Conf. IEEE Engineering in Medicine and Biology Society*, pages 434–437, San Francisco, CA, September 2004.

[11] Dana Angluin and Philip Laird. Learning from noisy examples. *Mach. Learn.*, 2:343–370, 1988.

[12] Xingquan Zhu and Xindong Wu. Class noise vs. attribute noise: A quantitative study. *Artif. Intell. Rev.*, 22:177–210, 2004.

[13] Ken Orr. Data quality and systems theory. *Commun. ACM*, 41(2):66–71, 1998.

[14] Thomas Redman. The impact of poor data quality on the typical enterprise. *Commun. ACM*, 2(2):79–82, 1998.

[15] Joseph L. Schafer and John W. Graham. Missing data: our view of the state of the art. *Psychol. methods*, 7(2):147–177, 2002.

[16] Peter A. Lachenbruch. Discriminant analysis when the initial samples are misclassified. *Technometrics*, 8(4):657–662, 1966.

[17] G. J. McLachlan. Asymptotic results for discriminant analysis when the initial samples are misclassified. *Technometrics*, 14(2):415–422, 1972.

[18] Tom Heskes. The use of being stubborn and introspective. In *Proc. ZiF Conf. Adaptative Behavior and Learning*, pages 55–65, Bielefeld, Germany, April 1994.

[19] Peter A. Lachenbruch. Note on initial misclassification effects on the quadratic discriminant function. *Technometrics*, 21(1):129–132, 1979.

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7. Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

[20] Seishi Okamoto and Yugami Nobuhiro. An average-case analysis of the k-nearest neighbor classifier for noisy domains. In *Proc. 15th Int. Joint Conf. Artifical intelligence - Vol. 1*, pages 238–243, Nagoya, Aichi, Japan, August 1997.

[21] D Randall Wilson and Tony R Martinez. Reduction techniques for instance-based learning algorithms. *Mach. Learn.*, 38(3):257–286, 2000.

[22] J. R. Quinlan. Induction of decision trees. *Mach. Learn.*, 1(1):81–106, 1986.

[23] David Nettleton, Albert Orriols-Puig, and Albert Fornells. A study of the effect of different types of noise on the precision of supervised learning techniques. *Artif. Intell. Rev.*, 33(4):275–306, 2010.

[24] Thomas G. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization. *Mach. Learn.*, 40(2):139–157, 2000.

[25] Wenxin Jiang. Some theoretical aspects of boosting in the presence of noisy data. In *Proc. 18th Int. Conf. Machine Learning*, pages 234–241, Williamstown, MA, June–July 2001.

[26] Peter A. Lachenbruch. Discriminant analysis when the initial samples are misclassified ii: Non-random misclassification models. *Technometrics*, 16(3):419–424, 1974.

[27] Javed A. Aslam. On the sample complexity of noise-tolerant learning. *Inform. Process. Lett.*, 57(4):189–195, 1996.

[28] Irwin Bross. Misclassification in 2 x 2 tables. *Biometrics*, 10(4):478–486, 1954.

[29] Anil Gaba and Robert L. Winkler. Implications of errors in survey data: A bayesian model. *Manage. Sci.*, 38(7):913–925, 1992.

[30] Chuck P. Lam and David G. Stork. Evaluating classifiers by means of test data with noisy labels. In *Proc. 18th Int. Joint Conf. Artificial intelligence*, pages 513–518, Acapulco, Mexico, August 2003.

[31] Gordon V. Cormack and Aleksander Kolcz. Spam filter evaluation with imprecise ground truth. In *Proc. 32nd Int. ACM SIGIR Conf. Research and Development In Information Retrieval*, pages 604–611, Boston, MA, July 2009.

[32] Richard Gerlach and James Stamey. Bayesian model selection for logistic regression with misclassified outcomes. *Stat. Model.*, 7(3):255–273, 2007.

[33] Benoît Frénay, Gauthier Doquire, and Michel Verleysen. Feature selection with imprecise labels: Estimating mutual information in the presence of label noise. *Comput. Stat. Data An.*, submitted for publication.

[34] Ahmad Abu Shanab, Taghi M. Khoshgoftaar, and Randall Wald. Robustness of threshold-based feature rankers with data sampling on noisy and imbalanced data. In *Proc. 25th Int. Florida Artificial Intelligence Research Society Conf.*, Marco Island, FL, May 2012.

[35] Choh-Man Teng. A comparison of noise handling techniques. In *Proc. 14th Int. Florida Artificial Intelligence Research Society Conf.*, pages 269–273, Key West, FL, May 2001.

[36] Shahram Golzari, Shyamala Doraisamy, Md Nasir Sulaiman, and Nur Izura Udzir. The effect of noise on rwtsairs classifier. *Eur. J. Sci. Res.*, 31(4):632–641, 2009.

[37] Charles Bouveyron and Stéphane Girard. Robust supervised classification with mixture models: Learning from data with uncertain labels. *Pattern Recogn.*, 42(11):2649–2658, 2009.

[38] Naresh Manwani and P. S. Sastry. Noise tolerance under risk minimization. *IEEE Trans. on Syst., Man, Cybern.*, in press.

[39] Hua Yin and Hongbin Dong. The problem of noise in classification: Past, current and future work. In *IEEE 3rd Int. Conf. Communication Software and Networks*, pages 412–416, Xi'an, China, May 2011.

[40] P. S. Sastry, G. D. Nagendra, and Naresh Manwani. A team of continuous-action learning automata for noise-tolerant learning of half-spaces. *IEEE Trans. on Syst., Man, Cybern. B, Cybern.*, 40:19–28, February 2010.

[41] Andres Folleco, Taghi M. Khoshgoftaar, Jason Van Hulse, and Amri Napolitano. Identifying learners robust to low quality data. *Informatica*, 33:245–259, 2009.

[42] Taghi M. Khoshgoftaar, Jason Van Hulse, and Amri Napolitano. Supervised neural network modeling: an empirical investigation into learning from imbalanced data with labeling errors. *IEEE Trans. Neural Netw.*, 21:813–830, May 2010.

[43] Gunnar Rätsch, Takashi Onoda, and Klaus-Robert Müller. Regularizing adaboost. In *Advances in Neural Information Processing Systems 11*, pages 564–570, Denver, CO, November–December 1998.

[44] Gunnar Rätsch, Takashi Onoda, and Klaus Robert Müller. An improvement of adaboost to avoid overfitting. In *Proc. 5th Int. Conf. Neural Information Processing*, pages 506–509, Kitakyushu, Japan, October 1998.

[45] Gunnar Rätsch, Bernhard Schölkopf, Alex J. Smola, Sebastian Mika, Takashi Onoda, and Klaus-Robert Müller. Robust ensemble learning for data mining. In *Proc. 4th Pacific-Asia Conf. Knowledge Discovery and Data Mining, Current Issues and New Applications*, pages 341–344, Kyoto, Japan, April 2000.

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

[47] Joaquín Abellán and Serafín Moral. Building classification trees using the total uncertainty criterion. *Int. J. Intell. Syst.*, 18(12):1215–1225, 2003.

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

[49] Victoria Hodge and Jim Austin. A survey of outlier detection methodologies. *Artif. Intell. Rev.*, 22(2):85–126, 2004.

[50] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Comput.Surv.*, 41(3):15:1–15:58, 2009.

[51] Jiang-wen Sun, Feng-ying Zhao, Chong-jun Wang, and Shi-fu Chen. Identifying and correcting mislabeled training instances. In *Proc. Future Generation Communication and Networking - Vol. 1*, pages 244–250, Jeju-Island, Korea, December 2007.

[52] Dragan Gamberger and Nada Lavrač. Conditions for occam's razor applicability and noise elimination. In *Proc. 9th Eur. Conf. Machine Learning*, pages 108–123, Prague, Czech Republic, April 1997.

[53] Dragan Gamberger, Rudjer Boskovic, Nada Lavrac, and Ciril Groselj. Experiments with noise filtering in a medical domain. In *Proc. 16th Int. Conf. Machine Learning*, pages 143–151, Bled, Slovenia, June 1999.

[54] Taghi M. Khoshgoftaar and Pierre Rebours. Generating multiple noise elimination filters with the ensemble-partitioning filter. In *Proc. 2004 IEEE Int. Conf. Information Reuse and Integration*, pages 369–375, Las Vegas, NV, November 2004.

[55] Piyasak Jeatrakul, Kok Wai Wong, and Chun Che Fung. Data cleaning for classification using misclassification analysis. *J. Adv. Comput. Intell. and Intell. Informatics*, 14(3):297–302, 2010.

[56] N. Matic, I. Guyon, L. Bottou, J. Denker, and V. Vapnik. Computer aided cleaning of large databases for character recognition. In *Proc. 11th IAPR Int. Conf. Pattern Recognition, Conf. B: Pattern Recognition Methodology and Systems*, pages 330–333, The Hague, Netherlands, August–September 1992.

[57] Isabelle Guyon, Nada Matic, and Vladimir Vapnik. Discovering informative patterns and data cleaning. In U.M. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, editors, *Advances in knowledge discovery and data mining*, pages 181–203. AAAI/MIT Press, Cambridge, MA, 1996.

[58] George H. John. Robust decision trees: Removing outliers from databases. In *Proc. 1st Int. Conf. Knowledge Discovery and Data Mining*, pages 174–179, Montreal, Quebec, Canada, August 1995.

[59] Léon Bottou and Vladimir Vapnik. Local learning algorithms. *Neural Comput.*, 4(6):888–900, 1992.

[60] Nicola Segata, Enrico Blanzieri, Sarah Delany, and Pàdraig Cunningham. Noise reduction for instance-based learning with a local maximal margin approach. *J. Intell. Inf. Syst.*, 35(2):301–331, 2010.

[61] Carla E. Brodley and Mark A. Friedl. Identifying and eliminating mislabeled training instances. In *Proc. 13th Nat. Conf. Artificial intelligence*, pages 799–805, Portland, Oregon, August 1996.

[62] Sofie Verbaeten and Anneleen Van Assche. Ensemble methods for noise elimination in classification problems. In *Proc. 4th Int. Conf. Multiple Classifier Systems*, pages 317–325, Guildford, UK, June 2003.

[63] Chen Zhang, Chunguo Wu, Enrico Blanzieri, You Zhou, Yan Wang, Wei Du, and Yanchun 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.

[64] D. Randall Wilson and Tony R. Martinez. Instance pruning techniques. In *Proc. Int. Conf. Machine Learning*, pages 403–411, Nashville, TN, July 1997.

[65] Sarah Jane Delany, Nicola Segata, and Brian Mac Namee. Profiling instances in noise reduction. *Knowl.-Based Syst.*, 31:28–40, 2012.

[66] Geoffrey W. Gates. The reduced nearest neighbor rule. *IEEE Trans. Inf. Theory*, 18:431–433, May 1972.

[67] Dennis L Wilson. Asymptotic properties of nearest neighbor rules using edited data. *IEEE Trans. on Syst., Man, Cybern.*, 2:408–421, July 1972.

[68] Amitava Karmaker and Stephen Kwek. A boosting approach to remove class label noise. *Int. J. Hybrid Intell. Syst.*, 3(3):169–177, 2006.

[69] Lawrence Joseph, Theresa W. Gyorkos, and Louis Coupal. Bayesian estimation of disease prevalence and the parameters of diagnostic tests in the absence of a gold standard. *Am. J. Epidemiol.*, 141(3):263–272, 1995.

[70] Tim Swartz, Yoel Haitovsky, Albert Vexler, and Tae Yang. Bayesian identifiability and misclassification in multinomial data. *Can. J. Stat.*, 32(3):285–302, 2004.

[71] Anil Gaba. Inferences with an unknown noise level in a bernoulli process. *Manage. Sci.*, 39(10):1227–1237, 1993.

[72] Martin Ladouceur, Elham Rahme, Christian A. Pineau, and Lawrence Joseph. Robustness of prevalence estimates derived from misclassified data from administrative databases. *Biometrics*, 63(1):272–279, 2007.

[73] Lawrence Joseph and Theresa W. Gyorkos. Inferences for likelihood ratios in the absence of a "gold standard". *Med. Decis. Making*, 16(4):412–417, 1996.

[74] Paul Gustafson, Nhu D. Le, and Refik Saskin. Case-control analysis with partial knowledge of exposure misclassification probabilities. *Biometrics*, 57(2):598–609, 2001.

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

[76] Carlos Daniel Paulino, Paulo Soares, and John Neuhaus. Binomial regression with misclassification. *Biometrics*, 59(3):670–675, 2003.

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

[78] Juxin Liu, Paul Gustafson, Nicola Cherry, and Igor Burstyn. Bayesian analysis of a matched case-control study with expert prior information on both the misclassification of exposure and the exposure-disease association. *Stat. Med.*, 28(27):3411–3423, 2009.

[79] Jorge Alberto Achcar, E. Z Martinez, and F Louzada-Neto. Binary data in the presence of misclassifications. In *16th Symp. Int. Association for Statistical Computing*, pages 581–587, Praga, Czech Republic, August 2004.

[80] Pat McInturff, Wesley O Johnson, David Cowling, and Ian A Gardner. Modelling risk when binary outcomes are subject to error. *Stat. Med.*, 23(7):1095–1109, 2004.

[81] M. J. García-Zattera, T. Mutsvari, A. Jara, D. Declerckc, and E. Lesaffrea. Correcting for misclassification for a monotone disease process with an application in dental research. *Stat. Med.*, 29(30):3103–3117, 2010.

[82] Frederik O. Kaster, Bjoern H. Menze, Marc-André Weber, and Fred A. Hamprecht. Comparative validation of graphical models for learning tumor segmentations from noisy manual annotations. In *Proc. 2010 Int. MICCAI Conf. Medical Computer Vision: Recognition Techniques and Applications in Medical Imaging*, pages 74–85, Beijing, China, September 2011.

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

[84] Daniel Hernandez-Lobato, José Miguel Hernandez-Lobato, and Pierre Dupont. Robust multi-class gaussian process classification. In *Advances in Neural Information Processing Systems 24*, pages 280–288, Granada, Spain, December 2011.

[85] Yishay Mansour and Michal Parnas. Learning conjunctions with noise under product distributions. *Inform. Process. Lett.*, 68(4):189–196, 1998.

[86] Neil D. Lawrence and Bernhard Schölkopf. Estimating a kernel fisher discriminant in the presence of label noise. In *Proc. of the 18th Int. Conf. Machine Learning*, pages 306–313, Williamstown, MA, June–July 2001.

[87] Yunlei Li, Lodewyk F.A. Wessels, Dick de Ridder, and Marcel J.T. Reinders. Classification in the presence of class noise using a probabilistic kernel fisher method. *Pattern Recogn.*, 40(12):3349–3357, 2007.

[88] Jakramate Bootkrajang and Ata Kabán. Multi-class classification in the presence of labelling errors. In *Proc. 19th Eur. Symp. Artificial Neural Networks*, pages 345–350, Bruges, Belgium, April 2011.

[89] Benoît Frénay, Gaël de Lannoy, and Michel Verleysen. Label noise-tolerant hidden markov models for segmentation: application to ecgs. In *Proc. 2011 Eur. Conf. Machine learning and Knowledge Discovery in Databases - Vol. I*, pages 455–470, Athens, Greece, September 2011.

[90] Jakramate Bootkrajang and Ata Kabán. Label-noise robust logistic regression and its applications. In *Proc. ECML-PKDD(1)*, pages 143–158, Bristol, UK, 2012.

[91] Jakramate Bootkrajang and Ata Kabán. Classification of mislabelled microarrays using robust sparse logistic regression. *Bioinformatics*, 29(7):870–877, 2013.

[92] Ata Kabán Jakramate Bootkrajang. Learning a label-noise robust logistic regression: Analysis and experiments. In *IDEAL*, pages 569–576, Hefei, China, October 2013.

[93] Nagarajan Natarajan, Inderjit S. Dhillon, Pradeep D. Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *NIPS*, pages 1196–1204, Lake Tahoe, NV, December 2013.

[94] Umaa Rebbapragada and Carla E. Brodley. Class noise mitigation through instance weighting. In *Proc. 18th Eur. Conf. Machine Learning*, pages 708–715, Warsaw, Poland, September 2007.

[95] T. Denœux. A k-nearest neighbor classification rule based on dempster-shafer theory. *IEEE Trans. Syst., Man, Cybern.*, 25:804–813, May 1995.

[96] Thierry Denœux. A neural network classifier based on dempster-shafer theory. *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans*, 30:131–150, March 2000.

[97] Zoulficar Younes, Fahed abdallah, and Thierry Denœux. Evidential multi-label classification approach to learning from data with imprecise labels. In *Proc. 13th Int. Conf. Information Processing and Management of Uncertainty*, pages 119–128, Dortmund, Germany, June–July 2010.

[98] W. Krauth and Mézard. Learning algorithms with optimal stability in neural networks. *J. Phys. A: Math. Gen.*, 20:L745–L752, 1987.

[99] Yi Li and Philip M. Long. The relaxed online maximum margin algorithm. *Mach. Learn.*, 46(1-3):361–387, 2002.

[100] Roni Khardon and Gabriel Wachman. Noise tolerant variants of the perceptron algorithm. *J. Mach. Learn. Res.*, 8:227–248, 2007.

[101] Aravind Ganapathiraju, Joseph Picone, and Mississippi State. Support vector machines for automatic data cleanup. In *Proc. 6th Int. Conf. Spoken Language Processing*, pages 210–213, Beijing, China, October 2000.

[102] Chun-fu Lin and Sheng-de Wang. Training algorithms for fuzzy support vector machines with noisy data. *Pattern Recog. Lett.*, 25(14):1647–1656, 2004.

[103] Carlos Domingo and Osamu Watanabe. Madaboost: A modification of adaboost. In *Proc. 13th Ann. Conf. Computational Learning Theory*, pages 180–189, San Francisco, CA, June 2000.

[104] Nikunj C. Oza. Aveboost2: Boosting for noisy data. In *Proc. 5th Int. Conf. Multiple Classifier Systems*, pages 31–40, Cagliari, Italy, June 2004.

[105] Vanessa Gómez-Verdejo, Manuel Ortega-Moral, Jerónimo Arenas-García, and Aníbal R. Figueiras-Vidal. Boosting by weighting critical and erroneous samples. *Neurocomputing*, 69(7-9):679–685, 2006.

[106] L Mason, J Baxter, P Bartlett, and M Frean. Functional gradient techniques for combining hypotheses. In A Smola, P Bartlett, B Schölkopf, and D Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 221–246. MIT Press, Cambridge, MA, 2000.

[107] Nir Krause and Yoram Singer. Leveraging the margin more carefully. In *Proc. 21st Int. Conf. Machine learning*, pages 63–70, Banff, Alberta, Canada, July 2004.

[108] Linli Xu, Koby Crammer, and Dale Schuurmans. Robust support vector machine training via convex outlier ablation. In *Proc. 21st Nat. Conf. Artificial intelligence - Vol. 1*, pages 536–542, Boston, MA, July 2006.

[109] Hamed Masnadi-Shirazi and Nuno Vasconcelos. On the design of loss functions for classification: theory, robustness to outliers, and savageboost. In *Advances in Neural Information Processing Systems 21*, pages 1049–1056, December 2008.

[110] Guillaume Stempfel and Liva Ralaivola. Learning svms from sloppily labeled data. In *Proc. 19th Int. Conf. Artificial Neural Networks: Part I*, pages 884–893, Limassol, Cyprus, September 2009.

[111] Hamed Masnadi-Shirazi, Vijay Mahadevan, and Nuno Vasconcelos. On the design of robust classifiers for computer vision. In *IEEE Computer Society Conf. Computer Vision and Pattern Recognition*, pages 779–786, San Francisco, CA, June 2010.

[112] David D. Lewis, Yiming Yang, Tony G. Rose, and Fan Li. Rcv1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.*, 5:361–397, 2004.

[113] Shuiwang Ji and Jieping Ye. Generalized linear discriminant analysis: A unified framework and efficient model selection. *IEEE Trans. Neural Netw.*, 19:1768–1782, October 2008.

[114] Xingquan Zhu, Xindong Wu, and Qijun Chen. Eliminating class noise in large datasets. In *Proc. 20th Int. Conf. Machine Learning*, pages 920–927, Washington, DC, August 2003.

[115] Luis Daza and Edgar Acuna. An algorithm for detecting noise on supervised classification. In *Proc. World Cong. Engineering and Computer Science 2007*, pages 701–706, San Francisco, CA, October 2007.