Estimating Models' Uncertainty in Supervised and Semi-Supervised Classification Problems. Half-year Ph.D. review

Miquel Perelló-Nieto
miquel.perellonieto@bristol.ac.uk
University of Bristol

April 5, 2018

Abstract

In this review we explore the uncertainty of common supervised classification models. We explains some of their prior assumptions, how these assumptions bias their predicted probabilities, and how to interpret their confidence values in different situations. We also describe our proposed method to make common classifiers more reliable and versatile, and how these can be used in fluctuating scenarios in which unexpected classes and anomalies may appear during deployment. Furthermore, we show an extension of proper loss functions that allow classifiers; that minimize an empirical loss; to be trained with weak labels (labels that may be wrong). Finally, we discuss two future directions of our current work: (1) how to get better probability estimates in Deep Neural Networks, and (2) new methods to reuse old datasets whose labels may be outdated and weak.

keywords: Supervised learning, Semi-supervised learning, classifier calibration, classification with confidence, cautious classification, weak labels, proper losses, anomaly detection, outlier detection

1 Introduction

Machine learning can be largely divided into two subset of fields called unsupervised and supervised learning [Bishop, 2006].

In general, Unsupervised learning can be seen as a set of exploratory techniques that offer some insight into data being analysed. These methods can be grouped as Data Mining techniques like clustering, density estimation, signal de-noising, or anomaly detection to mention a few. Because these techniques do not need manual labelling or annotations, data to use these models is fairly cheap and abundant (eg. all the text on the web, photos and images or time series). However, with an unclear objective (but exploration) it is difficult to evaluate the performance of these models.

On the other hand, Supervised learning has a clear objective like the minimization of an expected loss in classification or regression problems. In Regression problems, the objective function to model has no finite range while in Classification problems the predictions consist on finite sets.

At the same time, some research has been done in order to combine both fields; known as Semi-supervised learning. In most of the cases, Semi-supervised learning focuses on using unsupervised techniques; which can be difficult to evaluate but has plenty of data; to improve Supervised methods; which have well defined performance measures but less data. However, it is still a question of debate in what circumstances unsupervised techniques are able to improve purely supervised methods. In general, it is a trade-off between the amount of labelled and unlabelled data available and how much information is contained in the unlabelled set that may help selecting better prior assumptions. Some research has been undergone in order to create hybrid models that join both approaches. For example Laserre et al. [Lasserre et al., 2006, Bishop and Lasserre, 2007] uses a linear combination of both methods by means of a weighting hyper-parameter. In this particular scenario the resulting model has purely generative and discriminative models on the extreme values 0 and 1 of the hyper-parameter, while the range in between behaves as an hybrid model.

2 Background and related work

This section will serve as a basis to understand the current 3 and future work 4 proposed in this Ph.D. Although the may seem to be unrelated, in our research we will understand what are the assumptions of the different methods and what can we interpret as common uncertainties and confidence levels.

First, in Section 2.1 we will understand some of the common assumptions that Supervised methods do and we will analyse their strengths and weaknesses. In order to simplify this task, we will focus on Classification problems as these can be interpreted as a subset of Regression problems.

Then, in Section2.2 we will describe a type of scenario in which the labels at hand are not fully reliable and occasionally missing. We refer to these type of labels as weak labels.

Finally, in Section 2.3 we give an interpretation to the classifiers output predictions in different scenarios: when we can assume that training and test samples are independent and identically distributed (Section 2.3.1) and when this can not be guaranteed (Section 2.3.2).

2.1 Supervised learning and proper losses

In Classification problems there exist at lest three types of models depending on their objective predictions. These are; in increasing order of training difficulty; models that try to predict only the correct class, a ranking of all the classes or proper posterior probabilities per class. The model and type of predictions that will need to be chosen for a particular problem will strongly depend on the amount of data and the complexity of the task. From the previous definition we can see that if a model is able to predict proper posterior probabilities per class we can obtain a ranking model as well. In the same manner, if our model can provide a ranking for all the classes it can predict the correct class as well [Cid-Sueiro, 2012].

For classification models that are trained by the minimization of a loss function it is possible to define different losses that will optimize our model into one of the three aforementioned prediction scenarios. There is a specific type of loss for each of the cases. (1) Classification calibrated losses are only interested on the minimization of the miss-classification error (eg. accuracy). (2) Ranking calibrated losses minimize the expected loss of the full ranking. And by using (3) proper losses the model will obtain Bayesian a

posterior probabilities (eg. Brier score; also known as Mean Squared Error in regression problems; and Cross-entropy) [Buja et al., 2005].

In our research, we are interested on models that predict proper Bayesian a posterior probabilities as in this case we will be able to estimate the uncertanty levels on the models predictions. Also, as we mentioned before, these models can be applied to the other two simpler scenarios.

Notice that if a particular model is trained to minimize a proper loss, this will be theoretically able to obtain the Bayesian a posterior probabilities. However, this is only true in the limit with infinite number of samples and computational power and model complexity. Because in a finite amount of time is impossible to achieve these theoretical requirements, it is non-trivial to find an optimal regularisation for the models not to over-fit to the training data as it could reach a perfect performance in the training set, while obtaining poor predictions at test time. Two examples of theoretically calibrated models are Artificial Neural Networks [Hung et al., 1996, Zhang, 2000], and bagged trees [Niculescu-Mizil and Caruana, 2005].

On the other hand, there are classification models that because of their intrinsic assumptions tend to bias their predictions in one way or another. For example, some models push the posterior probabilities away from the extremes 0 and 1. This is the case of maximum margin methods like Support Vector Machines. Other methods are biased in the opposite direction and push the predicted posterior probabilities towards the boundaries 0 and 1. This is the case of Naive Bayes algorithm that assumes independence on the input features making every single feature contribution to make strong predictions.

For the cases in which the models do not predict good posterior probabilities there exist a set of post-processing tools called calibration methods whose objective is to find a mapping function between the model predictions and the current posterior probabilities. The most well known methods are Platt's Scaling [Platt, 1999] (that applies a Logistic Regression between the output scores and the true labels), binning methods; width binning, size binning, similarity binning [Bella et al., 2009] or Bayesian binning into quantiles [Naeini et al., 2015]; Isotonic regression [Zadrozny and Elkan, 2001, Zadrozny and Elkan, 2002] (a non-parametric method that used Pair-adjacent violators on the ROC curve to define the binning sizes per region) , or Beta calibration [Kull et al., 2017a, Kull et al., 2017b]. A good analysis of different calibration methods applied to ten clasification models can be seen in the work by Adrozny and Elkan [Zadrozny and Elkan, 2002] and Niculescu-Mizil and Caruana[Niculescu-Mizil and Caruana, 2005].

We will see in Section 2.3.1 how should we interpret the predicted posterior probabilities of classification models and the concept of confidence in their predictions. Also, we will show in our current work that this concept holds true in a set of different scenarios [Perello-Nieto et al., 2016].

2.2 Weak labels

We have seen problems in which the true classes are always available (Supervised learning). However, obtaining true; and objective; labels is commonly expensive.

For example, it is easier to obtain **multiple and subjective annotations from a crowd** of people instead of obtaining a reliable an objective label [Raykar et al., 2010]. In this case, we may assume that the true label is one of the selected ones, and some of the most extended approaches is to use the most voted label as the true label. However, we may expect annotators with different degrees of expertise on different samples. In this

situations other methods that try to model each annotator and by use of Expectation Maximization methods are able to obtain more accurate labels [Raykar et al., 2010].

In other scenarios, we may have a **supersets** of labels per sample for which we know that at least one of them is the true label [Hüllermeier and Cheng, 2015, Cour and Sapp, 2011]. This may be the case in a hierarchy of labels in which in some cases is easier to choose a parent class (that extends to several other subclasses) than the true leaf class. In this case Hüllermeier and Cheng [Hüllermeier and Cheng, 2015, Cour and Sapp, 2011] define an iterative method and losses that can be used with multiple annotated classes per sample.

Some times each sample has only one label but we accept that it may be wrong because of the inexperience of the annotator or ambiguous examples. This scenario is known as **label noise**.

We generalize all the previous scenarios with the concept of **weak labels**, in which we allow the true label to be in the superset of annotations, or not. Accepting the case of noisy labels and unlabelled cases as well.

We will see in our current 3.2 and future work 4.2 how in particular cases it is possible to use weak labels combined with proper loss functions to train models that find accurate Bayesian a posteriori probabilities.

2.3 Safe probabilities

One of the most important aspects of this Ph.D. is to be able to evaluate the confidence of classifiers in their predictions. However, there is no unique way to interpret the predicted probabilities for all the models, as each model is build upon different set of assumptions [Grünwald, 2017].

We will differentiate here two possible scenarios, in Section 2.3.1 we will explain models that make the strong assumption that the training and test data are independent and identically distributed from the same set. And in Section 2.3.2 we will see scenarios in which the training data is a partial; and poor; representation of the full domain.

In each of these cases, the posterior probabilities will have a different meaning and we will joint both concepts in our current work [Perello-Nieto et al., 2016].

2.3.1 IID training and test sets

A common assumption in all the aforementioned scenarios is that training and test sets are identically and independently distributed from the same data. In this case, the only source of uncertainty during predictions are the regions near the decision boundary in which most of the errors may happen. Or regions with basically same probabilities for more than one class. This type of uncertainty reflects intrinsic ambiguities (or overlap) between different classes.

In this scenarios, occasionally the misclassification cost may be higher than abstaining. With **cautious classification** it is possible to define optimal thresholds (or windows) around different ranges of probabilities in which the classifier should abstain [Ferri and Hernández-Orallo, 2004, Chow, 1970].

Also, given the limited size of training data it is possible that outliers are not properly captured during training. This is not because the iid. assumption being false, but because the finite nature of the samples makes difficult to capture all these possible values. However, given the sparse nature of outliers we will consider the common solutions applied to

outliers detection in the set of techniques used in the non iid. scenario as its possible to find similarities to anomaly detection and on-line learning methods.

2.3.2 Non iid training and test sets

As we mentioned before, there are multiple scenarios in which the iid. assumption does not hold. Some examples are: on-line learning, in which a set of categories are known, but these may shift and change over time, or new classes may appear; one-class classification, where during training we have samples from only one class and we are interested on detecting samples not belonging to this class; novelty detection, and anomaly detection, that generalize the concept of one class to the general setting in which multiple classes may be known but we are still interested on detecting non-expected situations or patterns.

In al these examples, it is expected that new data from regions of the input space not known during training may appear. Most of the proposed solutions for these kind of problems involve the use of Semi-supervised techniques as there is usually an unsupervised part that tries to estimate densities or create good summaries of the training data while using supervised learning for tasks such as classification. The absence of knowledge about the test distribution is the clear source of uncertainty in these scenarios. The inability of getting samples from novel regions makes these kind of problems difficult, and most of the solutions rely on the imposition of strong assumptions.

One-class classifiers are trained with samples of one class and knowing that in reality there are other classes that will show up during test. Multiple methods are known, and Khan et al. makes a review [Khan and Madden, 2014] with a good comparison between the best known methods evaluated in different domains. One of these methods designed by Hemstalk et al. [Hempstalk et al., 2008] makes use of density estimator methods in order to generate artificial data. Then, it is possible to train a classifier to discern between original and synthetic examples. A similar method has been used recently to train Artificial Neural Networks with Generative Adversarial nets [Goodfellow et al., 2014a, Goodfellow et al., 2014b].

In **anomaly detection** problems the known classes are well defined but there may be a small number of samples of not well represented (and unknown) classes. Very similar to the one-class classifier, the methods usually consist on two phases of unsupervised and supervised steps in which supervised classifiers are trained to discern between the known classes, while unsupervised density estimators try to reject non-familiar cases [Landgrebe et al., 2006, Landgrebe et al.,].

Novelty detection techniques are mostly unsupervised techniques. Markou et al. made two reviews one with statistical approaches [Markou and Singh, 2003a] and one for Artificial Neural Networks [Markou and Singh, 2003b] that analyse multiple models. The statistical and parametric methods use Gaussian mixture models, Hidden Markov models (HMM) and hypothesis testing, while some of the non-parametric approaches are k-nearest neighbour (k-nn), parzen density estimation, string matching approaches and clustering. Some fo the methods based on ANNs are Support Vector Machines (SVMs), Radial Basis Functions (RBFs), Hopefield networks and Self-organizing maps (SOM). Although the review by Markou et al. is really wide in the amount of models that are described, there is a lack of comparison between some of the models and it is left to the reader to further investigate their performance in common problems.

Finally, **online learning** methods are train in a continuous manner in which past data may not reflect the full distribution of new data; new classes may appear. One of the methods to deal with these situations is proposed by De Faria et al. [de Faria et al., 2015]

and consists on annotating in real time all the samples for which the classifiers are not confident about. After several iterations the set of annotated samples may start generating clusters which are evaluated using clustering algorithms. Finally, by the definition of certain thresholds these clusters are incorporated as new classes and classification models are trained with the new data.

A common denominator of all these methods is a necessity to model the training distribution in order to detect future deviations from it. In order to have an opportunity to detect these deviations, it is strictly necessary for the new samples not to follow the same distribution as the original training, as in that case, it would be impossible to differentiate them. We will see in our current work 3.1 that by modifying our assumptions we can solve different sets of problems in the same framework.

3 Current work

In this section, we describe how we made use of the theory and publications explained on the background Section 2 in order to publish two papers. We start by interpreting the estimated probabilities of classification models and understanding what we can call confidence in different scenarios in order to create a common framework (Section 3.2). In a separated but related publication, we evaluate some proper loss functions in order to train common classifiers in weak label scenarios (Section 3.2).

3.1 Classifiers with confidence

As we have seen in the Section Safe probabilities 2.3 most of the classification models assume that the training and test data are independent and identically distributed. However, we were interested on adapting these classifiers to consider that other unknown classes may appear during test.

With that in mind, in our publication "Background Check: A General Technique to Build More Reliable and Versatile Classifiers" [Perello-Nieto et al., 2016] we modified the Bayesian equation in order to incorporate an additional class for the unknown classes. We called this one the *background class*, while all the known classes where bounded together in the *foreground class*. Because these two classes share a common normalization factor it was possible to define the new posterior probabilities just from the ratio between the *foreground* and the *background*. Then, as the knowledge about the *background class* is unknown during training, we defined its distribution as a functional of the *foreground* density with two degrees of freedom. These two degrees of freedom simply adjusted their ratios and created an hybrid method between classification with confidence and anomaly detection problems in multi-class classification problems.

We demonstrated empirically with 41 datasets from the UCI machine learning repository [Lichman, 2013] that our method achieved and in many cases surpassed specialised state-of-the-art approaches.

3.2 Classifiers in weak label scenarios

We already mentioned that classification models that are trained with proper losses are theoretically capable of achieving Bayesian a posteriori probabilities. However, these losses were originally designed to be trained with true labels. The work by Cid-Suerio [Cid-Sueiro, 2012] demonstrates theoretically that under certain mild assumptions it is

possible to adapt proper loss functions to be applied in weak label scenarios. The method relies in the assumption that there exists a mixing matrix M that contains all the possible probabilities between the true and the weak labels. If this matrix is known, it is straight forward to adapt a proper loss function to the weak scenario. Furthermore, there exist a set of mixing matrices for which it is not necessary to know its values in order to create proper weak loss functions [Cid-Sueiro et al., 2014].

In our article "Adapting Supervised Classification Algorithms to Arbitrary Weak Label Scenarios" [Perelló-Nieto et al., 2017] we performed an empirical analysis to evaluate the validity of the theoretical results in real world datasets. The analysis was performed in 31 real datasets from the UCI machine learning repository [Lichman, 2013]. In order to generate the weak labels we had to simulate different mixing process by generating random mixing matrices M with different conditions. In our experiments we showed that the theoretical results agree with finite datasets when the prior assumptions about the mixing process are correct. And in some cases the assumptions do not hurt the performance of the models trained with the weak labels when comparing to other state-of-the-art methods.

4 Future work

In this section we discuss two problems in which we are currently working for which initial experiments have been developed but further analysis is required. We start in Section 4.1 by an analysis the performance of a new calibration method (Beta calibration) when this is applied to modern artificial neural networks. Then in Section 4.2 we intend to extend methods to train with proper weak loss functions where the estimation of the mixing process is possible given that a join set of weak and true labels is available.

4.1 Beta calibration for Artificial Neural Networks

We mentioned in the background section 2 that certain models need to be calibrated in order to predict proper posterior probabilities. In [Niculescu-Mizil and Caruana, 2005] the authors investigate the biases on the posterior probabilities of ten different classification models and compare the performance of well known calibration methods (Platt's Scaling and Isotonic regression). The authors explain how maximum margin methods push the probabilities away from 0 and 1. This fact usually creates distributions for positive an negative classes that resemble Gaussian distributions (eg. Support Vector Machines). On the other hand, methods that assume independence between features; such as Naive Bayes; tend to be over-confident pushing the probabilities towards 0 and 1. Finally, the authors show that Artificial neural networks (ANN) and bagged trees do not suffer from these biases and predict well calibrated probabilities.

Although in theory Artificial Neural Networks are well calibrated and can achieve good estimates of Bayesian a posterior probabilities [Richard and Lippmann, 1991, Hung et al., 1996] recent work [Guo et al., 2017] shows empirically that the training complexity of modern neural networks prevents the networks to be properly calibrated. They show that the number of layers, number of units per layer, weight decay, or batch normalisation all affect into the models obtaining good posterior probabilities. Then, the authors evaluate several calibration methods (including new ones proposed for ANNs) with various classification problems and several state-of-the-art Deep Neural Networks (eg. ResNet [He et al., 2015], ResNets with stochastic depth [Huang et al., 2016] and DenseNets [Huang

et al., 2017]). The results show that by applying a simple calibration methods proposed by the authors the these Deep neural networks are able to reduce the calibration error on the validation and test sets.

Recent work on Calibration [Kull et al., 2017a], proposes a new method to calibrate models whose posterior probability distributions doesn't resemble Gaussian but Beta distributions. This is the case of Naive Bayes or Adabost which predictions are pushed to the 0 and 1 extremes. The authors show that in this situation Platt's scalling can worsen the predictions of the original models. In general, it is a better fit when the output domain of the original model is already in the interval (0, 1) as the assumption of Beta distributions in an closed range is more realistic than the assumption of truncated normal distributions.

In this work we are analysing how the new Beta calibration can be applied into Deep neural networks and if this method can be incorporated as a layer of the networks and trained directly with back-propagation.

Also, we are considering a new extension of the Beta calibration to the multi-class scenario using in this case Dirichlet distributions.

4.2 Recycling data with weak (and old) labels

We already demonstrated in [Perelló-Nieto et al., 2017] that it is possible to train classification models with weak labels by using proper weak loss functions. Although the results were obtained in 31 real world datasets, the weak labels were synthetically generated in order to test different assumptions.

In our current work, we are interested on a real dataset in which 65.000 samples have weak (and old) labels, while about 1.000 samples have the old weak labels plus new annotated true labels. We are now investigating if it is possible to reuse the 65.000 old samples in order to improve the performance of fully supervised classification methods trained only on the small set of 1.000 samples. Our proposed method uses the Expectation Maximization algorithm in order to estimate the most plausible mixing matrix M from the small set with weak and true labels, while training a model that minimizes a proper weak loss in the full set of samples (65.000 + 1.000). Our current results seems to tell that when the number of true labels is reduced, our method surpasses the validation accuracy of a purely supervised method. However, at this moment the results do not seem statistically significant given the number of executions, sample size and variance in the performance.

We are now trying with additional datasets in order to get statistically significant results with less variance.

5 Discussion

In this review, we have described the work undergone during the first part of my Ph.D. With an introduction to the related topics and some context to understand the direction of our work.

First we have given a brief introduction to the topics of supervised classification models, calibration methods, proper losses, safe probabilities, and weak labels. All with a focus on understanding the posterior probabilities of classification models and obtaining confidence values from their predictions.

Then, we have shown how our understanding of these topics was translated into the publication of our articles in two well known conferences: the International Conference

on Data Mining and the Advances in Intelligent Data Analysis. We have explained the main contributions of both publications demonstrating that our methods achieved and many times surpassed state-of-the-art approaches.

Finally, we described our future work and what is its current state, one in the direction of obtaining better calibrated deep neural networks. And the other for training classifiers when only weak labels are available. Our current results seem to go into the right direction and we hope to be able to publish them in the forthcoming months.

References

[Bella et al., 2009] Bella, A., Ferri, C., Hernández-Orallo, J., and Ramírez-Quintana, M. J. (2009). Similarity-binning averaging: A generalisation of binning calibration. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 5788 LNCS, pages 341–349. Springer, Berlin, Heidelberg.

[Bishop and Lasserre, 2007] Bishop, C. and Lasserre, J. (2007). Generative or Discriminative? Getting the Best of Both Worlds. In Bernardo, J., Bayarri, M., Berger, J., Dawid, A., Heckerman, D., Smith, A., and West, M., editors, *Bayesian Statistics*, volume 8, pages 3 – 24.

[Bishop, 2006] Bishop, C. M. (2006). *Pattern recognition and machine learning.*, volume 1. New York: springer, 2006.

[Buja et al., 2005] Buja, A., Stuetzle, W., and Shen, Y. (2005). Loss functions for binary class probability estimation and classification: Structure and applications. *Working draft, November*.

[Chow, 1970] Chow, C. K. (1970). On Optimum Recognition Error and Reject Tradeoff. *IEEE Transactions on Information Theory*, 16(1):41–46.

[Cid-Sueiro, 2012] Cid-Sueiro, J. (2012). Proper losses for learning from partial labels. In *Advances in Neural Information Processing Systems*, pages 1565–1573.

[Cid-Sueiro et al., 2014] Cid-Sueiro, J., García-García, D., and Santos-Rodríguez, R. (2014). Consistency of losses for learning from weak labels. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, volume 8724 LNAI, pages 197–210. Springer Berlin Heidelberg.

[Cour and Sapp, 2011] Cour, T. and Sapp, B. (2011). Learning from Partial Labels. *Journal of Machine Learning Research*, 12(2):1501–1536.

[de Faria et al., 2015] de Faria, E. R., Ponce de Leon Ferreira Carvalho, A. C., and Gama, J. (2015). MINAS: multiclass learning algorithm for novelty detection in data streams. *Data Mining and Knowledge Discovery*, 30(3):640–680.

[Ferri and Hernández-Orallo, 2004] Ferri, C. and Hernández-Orallo, J. (2004). Cautious Classifiers. In *Proceedings of ROC Analysis in Artificial Intelligence, 1st International Workshop (ROCAI-2004)*, pages 27–36.

[Goodfellow et al., 2014a] Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014a). Generative Adversarial Nets. *Advances in Neural Information Processing Systems 27*, pages 2672–2680.

[Goodfellow et al., 2014b] Goodfellow, I. J., Shlens, J., and Szegedy, C. (2014b). Explaining and Harnessing Adversarial Examples. *arXiv:1412.6572 [cs, stat]*.

[Grünwald, 2017] Grünwald, P. (2017). Safe probability. *Journal of Statistical Planning and Inference*.

[Guo et al., 2017] Guo, C., Pleiss, G., Sun, Y., and Weinberger, K. Q. (2017). On Calibration of Modern Neural Networks.

[He et al., 2015] He, K., Zhang, X., Ren, S., and Sun, J. (2015). Deep Residual Learning for Image Recognition.

[Hempstalk et al., 2008] Hempstalk, K., Frank, E., and Witten, I. H. (2008). One-Class Classification by Combining Density and Class Probability Estimation. In Daelemans, W., Goethals, B., and Morik, K., editors, *Machine Learning and Knowledge Discovery in Databases*, volume 5211 of *Lecture Notes in Computer Science*, pages 505–519. Springer Berlin Heidelberg, Berlin, Heidelberg.

[Huang et al., 2017] Huang, G., Liu, Z., Weinberger, K. Q., and van der Maaten, L. (2017). Densely connected convolutional networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, volume 1, page 3.

[Huang et al., 2016] Huang, G., Sun, Y., Liu, Z., Sedra, D., and Weinberger, K. Q. (2016). Deep networks with stochastic depth. In *European Conference on Computer Vision*, pages 646–661. Springer.

[Hüllermeier and Cheng, 2015] Hüllermeier, E. and Cheng, W. (2015). Superset Learning Based on Generalized Loss Minimization. In *Machine learning and Knowledge Discovery in Databases*, pages 260–275. Springer, Cham.

[Hung et al., 1996] Hung, M., Hu, M., Shanker, M., and Patuwo, B. (1996). Estimating posterior probabilities in classification problems with neural networks. *International Journal of Computational Intelligence and Organizations*, 1(1):49–60.

[Khan and Madden, 2014] Khan, S. S. and Madden, M. G. (2014). One-class classification: taxonomy of study and review of techniques. *The Knowledge Engineering Review*, 29(03):345–374.

[Kull et al., 2017a] Kull, M., De Menezes, T., Filho, S., Flach, P., Filho, T. S., and Flach, P. (2017a). Beta calibration: a well-founded and easily implemented improvement on logistic calibration for binary classifiers. *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, 54:623–631.

[Kull et al., 2017b] Kull, M., Silva Filho, T. M., and Flach, P. (2017b). Beyond sigmoids: How to obtain well-calibrated probabilities from binary classifiers with beta calibration. *Electronic Journal of Statistics*, 11(2):5052–5080.

[Landgrebe et al., 2006] Landgrebe, T. C., Tax, D. M., Paclík, P., and Duin, R. P. (2006). The interaction between classification and reject performance for distance-based reject-option classifiers. *Pattern Recognition Letters*, 27(8):908–917.

[Landgrebe et al.,] Landgrebe, T. C. W., Tax, D. M. J., Paclík, P., Duin, R. P. W., and Andrew, C. A combining strategy for ill-defined problems.

[Lasserre et al., 2006] Lasserre, J., Bishop, C., and Minka, T. (2006). Principled Hybrids of Generative and Discriminative Models. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Volume 1 (CVPR'06)*, volume 1, pages 87–94. IEEE.

[Lichman, 2013] Lichman, M. (2013). UCI machine learning repository.

[Markou and Singh, 2003a] Markou, M. and Singh, S. (2003a). Novelty detection: a review - part 1: statistical approaches. *Signal Processing*, 83(12):2481–2497.

[Markou and Singh, 2003b] Markou, M. and Singh, S. (2003b). Novelty detection: a review- part 2:. *Signal Processing*, 83(12):2499–2521.

[Naeini et al., 2015] Naeini, M. P., Cooper, G. F., and Hauskrecht, M. (2015). Obtaining Well Calibrated Probabilities Using Bayesian Binning. *Proceedings of the ... AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence*, 2015:2901–2907.

[Niculescu-Mizil and Caruana, 2005] Niculescu-Mizil, A. and Caruana, R. (2005). Predicting good probabilities with supervised learning. In *Proceedings of the 22nd international conference on Machine learning - ICML '05*, pages 625–632, New York, New York, USA. ACM Press.

[Perelló-Nieto et al., 2017] Perelló-Nieto, M., Santos-Rodríguez, R., and Cid-Sueiro, J. (2017). Adapting Supervised Classification Algorithms to Arbitrary Weak Label Scenarios. In *International Symposium on Intelligent Data Analysis*, pages 247–259, London, UK. Springer, Cham.

[Perello-Nieto et al., 2016] Perello-Nieto, M., Telmo De Menezes Filho, E. S., Kull, M., and Flach, P. (2016). Background check: A general technique to build more reliable and versatile classifiers. In *Data Mining (ICDM), 2016 IEEE 16th International Conference on*, pages 1143–1148. IEEE.

[Platt, 1999] Platt, J. (1999). Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. *Advances in large margin classifiers*, 10(3):61–74.

[Raykar et al., 2010] Raykar, V. C., Yu, S., Zhao, L. H., Hermosillo Valadez, G., Florin, C., Bogoni, L., Moy, L., and Org, L. M. (2010). Learning From Crowds. *Journal of Machine Learning Research*, 11(Apr):1297–1322.

[Richard and Lippmann, 1991] Richard, M. D. and Lippmann, R. P. (1991). Neural Network Classifiers Estimate Bayesian a posteriori Probabilities. *Neural Computation*, 3(4):461–483.

[Zadrozny and Elkan, 2001] Zadrozny, B. and Elkan, C. (2001). Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. *ICML*.

[Zadrozny and Elkan, 2002] Zadrozny, B. and Elkan, C. (2002). Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '02*, page 694, New York, New York, USA. ACM Press.

[Zhang, 2000] Zhang, G. P. (2000). Neural networks for classification: a survey. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 30(4):451–462.