ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

Impact of Biases in Big Data

Patrick Glauner[1], Petko Valtchev[2] and Radu State[1]

1- Interdisciplinary Centre for Security, Reliability and Trust,
University of Luxembourg
29, Avenue JF Kennedy, 1855 Luxembourg, Luxembourg

2- Department of Computer Science, University of Quebec in Montreal
201, av. President Kennedy, Montreal H2X 3Y7, Canada

Abstract. The underlying paradigm of big data-driven machine learning reflects the desire of deriving better conclusions from simply analyzing more data, without the necessity of looking at theory and models. Is having simply more data always helpful? In 1936, The Literary Digest collected 2.3M filled in questionnaires to predict the outcome of that year's US presidential election. The outcome of this big data prediction proved to be entirely wrong, whereas George Gallup only needed 3K handpicked people to make an accurate prediction. Generally, biases occur in machine learning whenever the distributions of training set and test set are different. In this work, we provide a review of different sorts of biases in (big) data sets in machine learning. We provide definitions and discussions of the most commonly appearing biases in machine learning: class imbalance and covariate shift. We also show how these biases can be quantified and corrected. This work is an introductory text for both researchers and practitioners to become more aware of this topic and thus to derive more reliable models for their learning problems.

1 Introduction

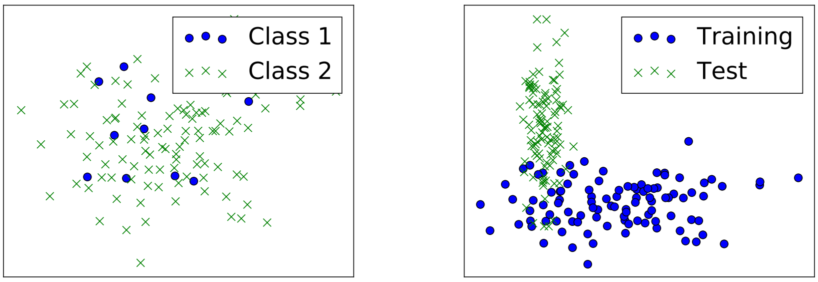
For about the last decade, the Big Data paradigm that has dominated research in machine learning can be summarized as follows: "It's not who has the best algorithm that wins. It's who has the most data." [1] In practice, however, most data sets are (systematically) biased.

Example 1. A spam filter is trained on a data set that consists of positive and negative examples. However, that training set was created a few years ago. Recent spam emails are different in two ways: the content of spam emails is different and the proportion of spam among all emails sent out has changed. As an outcome, the spam filter does not detect spam reliably and becomes even less reliable over time.

The appearance of biases in data sets imply a number of severe consequences including, but not limited to, the following: First, conclusions derived from biased - and therefore unrepresentative - data sets could simply be wrong due to lack of reproducibility and lack of generalizability. This is a common issue in research as a whole, as it has been argued that most research published may actually be wrong [2]. Second, these machine learning models may discriminate against subjects of under-represented categories [3, 4].

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6.
Available from http://www.i6doc.com/en/.

From a technical perspective, the most commonly appearing biases include *class imbalance* and *covariate shift*. Class imbalance is the case where classes are unequally represented in the data. An example is visualized in Fig. 1a. Covariate shift is the problem of drawing training and test data sets from different distributions. An example is visualized in Fig. 1b. These biases are often ignored in both research and practical applications. In part of the statistical literature, the phenomenon of biased data sets is called non-stationarity. In essence, this term indicates different statistics at a different time of collection of the training and test data sets, respectively [5].



(a) Class imbalance: Classes are unequally represented in the data.

(b) Covariate shift: Training and test data sets are drawn from different distributions.

Fig. 1: The most commonly appearing biases in data sets.

More generally, however, the term *bias* is multifaceted in the field of machine learning and describes different matters: The inductive bias of a learning algorithm refers to the set of assumptions a learner makes [6]. For example, logistic regression assumes that the training data is linearly separable. In contrast, the term bias is often used as a synonym for underfitting in the literature [7]. Moreover, the parameter w_0 of a hypothesis

$$h(x) = w_0 + w_1 x_1 + ... + w_n x_n$$

is sometimes called bias as it allows to shift a hypothesis by a fixed offset [7].

Our core contribution in this work is that we provide a systematic review of the research on biased data sets in the field of machine learning. We show that Big Data is useful, but not a silver bullet that could simply always be used without correcting the biases in it. The results of this review allow researchers to pay attention to this topic in their own research and thus leading to more reliable models. The rest of this paper is organized as follows: Section 2 provides a selection of popularized issues caused by applications of machine learning models trained on biased data sets. Section 3 provides a general introduction of how biases in data sets are defined. Sections 4 and 5 provide reviews of class imbalance and covariate shift - the most commonly appearing biases in data sets,

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

respectively. Section 6 provides references to other biases in data sets studied in the literature. Section 7 summarizes this work.

2 The more data, the better?

Historically, biased data sets have been a long-standing issue in statistics. The following example describes the failed prediction of the outcome of the 1936 US presidential election. It is often cited in the statistics literature in order to illustrate the impact of biases in data. This example is discussed in detail in [8].

Example 2. The Democratic candidate Franklin D. Roosevelt was elected President in 1932 and ran for a second term in 1936. Roosevelt's Republican opponent was Kansas Governor Alfred Landon. *The Literary Digest*, a general interest weekly magazine, had correctly predicted the outcomes of the elections in 1916, 1920, 1924, 1928 and 1932 based on straw polls. In 1936, The Literary Digest sent out 10M questionnaires in order to predict the outcome of the presidential election. The Literary Digest received 2.3M returns and predicted Landon to win by a landslide. However, the predicted result proved to be wrong, as quite the opposite happened: Roosevelt won by a landslide. This leads to the following questions:

1. How could the prediction turn out to be completely wrong despite the 2.3M participants?
2. How could The Literary Digest actually collect 10M addresses in 1936?

The Literary Digest compiled their data set of 10M recipients mainly from car registrations and phone directories. In that time, the households that had a car or a phone represented a disproportionally rich, and thus biased, sample of the overall population that particularly favored the Republican candidate Landon. In contrast, George Gallup only interviewed 3K handpicked people, which were an unbiased sample of the population. As a consequence, Gallup could predict the outcome of the election very accurately [9].

Even though this historic example is well understood in statistics nowadays, similar or related issues happen every day dozens of times in modern Big Data-oriented machine learning. We now discuss selected examples that result from biases in modern applications of Big Data-driven machine learning.

Example 3. It has been argued that most data on humans may be on white people and thus may not represent the overall population [10]. As a consequence, the predictions of models trained on such biased data may cause infamous news. For example, in 2015, Google added an auto-tagging feature to its Photos app. This new feature automatically assigned tags to photos, such as bicycle, dog, etc. However, some black users reported that they were tagged as "gorillas", which led to major criticism of Google [3]. Most likely, this mishap was caused by a biased training set, in which black people were largely underrepresented.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

The examples provided in this Section show that having simply more data is not always helpful in training reliable models, as the data sets used may be biased. In the following Sections, we discuss the most commonly appearing biases in data sets. We also present different strategies for assessing biased models and how to correct biases. These techniques include weighting training examples as well as subsampling methods. As a consequence, having data that is more representative is favorable, even if the amount of data used is less than just using the examples from a strongly biased data set.

3 Biases in data sets

In supervised learning, training examples $(x^{(i)}, y^{(i)})$ are drawn from a training distribution $P_{train}(X, Y)$, where X denotes the data and Y the label, respectively. The training set is biased if the following inequality holds true:

$$P_{train}(X, Y) \neq P_{test}(X, Y). \tag{1}$$

Different biases are visualized in Fig. 1. In order to reduce a bias, it has been shown that example $(x^{(i)}, y^{(i)})$ can be weighted during training as follows [11]:

$$w_i = \frac{P_{test}(x^{(i)}, y^{(i)})}{P_{train}(x^{(i)}, y^{(i)})}. \tag{2}$$

However, computing $P_{train}(x^{(i)}, y^{(i)})$ may be impractical in many cases because of the limited amount of data in the training domain. In the following sections, we discuss different biases for which specific assumptions about $P_{train}(X, Y)$ and $P_{test}(X, Y)$ are made.

4 Class imbalance

Class imbalance refers to the case where classes are unequally represented in the data. When comparing training set and test set, respectively, we assume [12]:

$$P_{train}(Y) \neq P_{test}(Y), \tag{3}$$
$$P_{train}(X|Y) = P_{test}(X|Y). \tag{4}$$

An example is depicted in Fig. 1a. Imbalanced classes appear frequently in machine learning. Machine learning models trained on imbalanced data sets often tend to predict the majority class. The appearance of imbalanced classes also affects the choice of evaluation metric. Accuracy and recall are the most commonly used metric in contemporary research works in machine learning [11, 13]. However, both metrics are affected by class imbalance. As a consequence, in many machine learning works, overly high accuracies or recalls are reported [14].

Example 4. Anomaly detection problems often work on particularly imbalanced data sets. A test set containing 1K customers of which 999 have regular

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

behavior and 1 has irregular behavior, (1) a classifier always predicting regular behavior has an accuracy of 99.9%, whereas in contrast, (2) a classifier always predicting irregular behavior has a recall of 100%. While the classifier of the first example has a very high accuracy and intuitively seems to perform very well, it will never predict any irregular behavior. In contrast, the classifier of the second example will find all customers that have irregular behavior, but may potentially trigger many costly and unnecessary interventions for customers that have a regular behavior [14].

Example 5. The Modified National Institute of Standards and Technology (MNIST) database consists of 60K training images and 10K test images used for recognition of hand-written digits [15], for which examples are depicted in Fig. 2a. MNIST has been used in the fields of computer vision and machine learning for the last 20 years. The test accuracies reported in recent research are above 99.6% [16, 17]. The distribution of test labels is depicted in Fig. 2b. We notice that this data set is mainly imbalanced between the different classes. As a consequence, the accuracy is not the right metric for MNIST, as an increase of this metric does not necessarily imply an increased predictive power of a model. We would like to add that the distribution of labels is nearly the same for the training set. Furthermore, there is another imbalance between the training set and test set, respectively. However, that imbalance is less noticeable and we have therefore focused on the imbalance between the labels in each set.

(a) Example images. (b) Distribution of test labels.

Fig. 2: MNIST data set.

A number of metrics that are insensitive to class imbalance can be found in the literature. One common metric is to use a receiver operating characteristic (ROC) curve, which plots the true positive rate against the false positive rate for varying decision threshold values. An example is depicted in Figure 3. The area under the curve (AUC) is a performance measure between 0 and 1, where any binary classifier with an AUC > 0.5 performs better than random guessing [18]. Another metric that is insensitive to class imbalance is the Matthews correlation coefficient (MCC):

$$\frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}, \tag{5}$$

which measures the accuracy of binary classifiers taking into account the imbalance of both classes, ranging from -1 to $+1$ [19]. Furthermore, for multi-class

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

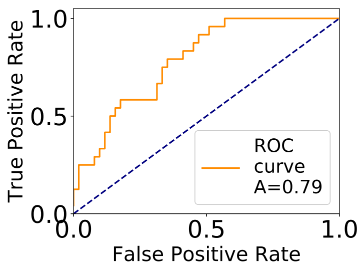


Fig. 3: Example of receiver operating characteristic (ROC) curve.

problems the intraclass correlation coefficient (ICC) has been proposed [20]. It can be interpreted as the fraction of the total variance that is between the different classes. It has been successfully applied to imbalanced multi-class learning problems [21].

In order to correct the class imbalance during training, a number of methods are proposed in the literature. First, weighting examples by the inverse proportion of examples per class using Eq. 2 is proposed in the literature [12]. On the one hand, one intuitive method is undersampling the majority classes by dropping training examples, either randomly or by specific criteria [22, 23]. This approach leads to smaller data sets, but may lack variation, as important examples could have been dropped. On the other hand, oversampling the minority classes by creating more training examples is proposed in the literature. Most trivially, training examples can simply be randomly copied. However, there are also more sophisticated algorithms, such as the synthetic minority over-sampling technique (SMOTE), which attempts to create synthetic examples representing the minority class by interpolating between neighboring data points [24]. Generally, adding more examples leads to larger training sets, which, in turn, leads to increased training time. Therefore, combinations of oversampling and undersampling were proposed [25, 26].

5 Covariate shift

Covariate shift refers to the case where the training data and test data are distributed differently. We assume [12]:

$$P_{train}(X) \neq P_{test}(X), \tag{6}$$

$$P_{train}(Y|X) = P_{test}(Y|X). \tag{7}$$

An example is depicted in Fig. 1b. Covariate shift appears frequently in machine learning as discussed in Section 2. Machine learning models trained on biased training sets tend not to generalize on test data that is from the true underlying distribution of the population.

Example 6. Non-technical losses (NTL) appear in power grids during distribution and describe irregular power usage, in particular electricity theft. NTL

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

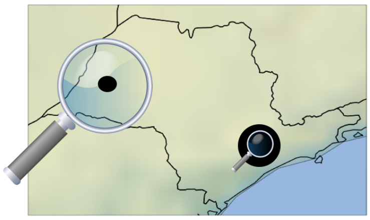


Fig. 4: Example of spatial bias: Most inspections are carried out in the small city due to hidden selection criteria. This sample of customers inspected thus does not represent the overall population of customers [28].

are reported to range up to 40% of the total electricity distributed in countries such as Brazil, India, Malaysia or Pakistan [14, 27]. Recent research on NTL detection mainly uses machine learning models that learn anomalous behavior from customer data and known irregular behavior that was reported through on-site inspection results. We have previously shown that in many cases, the set of inspected customers is biased as depicted in Fig. 4 [28]. A reason for this bias is that past inspections have been largely focused on certain criteria and were not sufficiently spread across the population. As a consequence, when learning from the inspection results, a bias is learned, making predictions less reliable.

The literature distinguishes classifiers into local learners and global learners, respectively [29]. For a local learner, the prediction of the learner depends only on $P(Y|X)$ for an increasing number of training examples. Previous research addresses this behavior with the term "asymptotically". We assume $P_{train}(Y|X) = P_{test}(Y|X)$ in Eq. 7. Hence, a local learner is not affected by covariate shift. Examples include logistic regression and hard-margin support vector machine (SVM). In contrast, the prediction of a global learner depends asymptotically on both, $P(Y|X)$ and $P(X)$. We assume $P_{train}(X) \neq P_{test}(X)$ in Eq. 6. Hence, a global learner is affected by covariate shift. Examples include decision tree learners such as ID3 or C4.5, naive Bayes and soft-margin SVM [28]. The terms "global" and "local", respectively, have been established as follows: A global learner also uses $P(X)$, which is a (global) distribution over the entire input data. In contrast, a local learner uses $P(Y|X)$, which refers for every $x^{(i)} \in X$ to a local distribution $P(Y|x^{(i)})$.

When using a data set, we need to assess whether the training set has actually a covariate shift. The Kullback-Leibler divergence [30] is a measure of the difference of two probability distributions. However, it is challenging (1) to adapt this measure to multi-dimensional data that is a combination of discrete and continuous features, which is common in machine learning, and (2) to define criteria from what values on a distance is an indicator for a covariate shift. We have recently proposed decision tree learning for finding a model that is able to distinguish between training and test distributions. The rationale behind our methodology is as follows: First, we add a feature s and assign the values 1 or 0 to the training data ($s = 1$) or test/production data ($s = 0$), respectively.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

Second, decision trees are global learners and thus sensitive to covariate shift. Third, optimizing a decision tree to predict the label s and thus maximizing this distinction between the two sets is equivalent to finding the best binary classification between test/production data and original training data. Next, the performance of the classifier is quantified using the Matthews correlation coefficient (MCC), which is defined in Eq. 5. As a result, the MCC value is the magnitude of the covariate shift in the data set [28].

Instance weighting using density estimation has been proposed for correcting covariate shift [31]. Examples can either be weighted during training [32] or the weights can be used for rejection sampling [29]. Historically, the Heckman method has been proposed to correct covariate shift by estimating the probability of an example being selected into the training sample [33]. However, the Heckman method only applies to linear regression models.

6 Other biases

Below we list other types of biases that have been investigated. Without any pretension for exhaustivity, we define those biases and refer the reader to the corresponding literature for further details. For instance, a change of functional relations can create a new bias and thus lead to $P_{train}(Y|X) \neq P_{test}(Y|X)$ [12]. Also, it has been shown that biases can be created by transforming the feature space [34]. Furthermore, a bias specific to neural networks has been reported: During training, a change of the weights in one layer may alter the distribution of the input to the following layer. This so-called internal covariate shift slows down convergence of training a neural network and may result in a neural network that overfits [35]. Internal covariate shift can be compensated by normalizing the input of every layer. By doing so, it has been reported that the training can be significantly accelerated. The resulting neural network is also less likely to overfit. This approach is radically different to regularization [36], as it addresses the cause of overfitting rather than trying to improve a model that overfits.

7 Conclusions

In this work, we first have presented a number of historic and modern examples of biased data sets that resulted in unreliable models. Biases occur in machine learning whenever training sets are not representative for the test data. Even though biases have been recognized as an issue in statistics since the mid-20th century, they only recently started to get more attention in machine learning. We then provided an extensive review of biases in machine learning, with a (special) focus on the most common ones: class imbalance and covariate shift. We have shown how these biases can be quantified and corrected. As a consequence, in many cases it may not be helpful to simply have more data, but rather to have (possibly less) data that is more representative.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

Acknowledgement

The present project is supported by the National Research Fund, Luxembourg under grant agreement number 11508593.

References

[1] Michele Banko and Eric Brill. Scaling to very very large corpora for natural language disambiguation. In *Proceedings of the 39th annual meeting on association for computational linguistics*, pages 26–33. Association for Computational Linguistics, 2001.

[2] John PA Ioannidis. Why most published research findings are false. *PLoS medicine*, 2(8):e124, 2005.

[3] Sophie Curtis. Google photos labels black people as gorillas. the telegraph. http://www.telegraph.co.uk/technology/google/11710136/Google-Photos-assigns-gorilla-tag-to-photos-of-black-people.html, 2015. [Online; accessed December 28, 2017].

[4] Yilun Wang and Michal Kosinski. Deep neural networks are more accurate than humans at detecting sexual orientation from facial images. *Journal of Personality and Social Psychology*, 2017.

[5] Moamar Sayed-Mouchaweh and Edwin Lughofer. *Learning in non-stationary environments: methods and applications*. Springer Science & Business Media, 2012.

[6] Tom M Mitchell. Machine learning. 1997. *Burr Ridge, IL: McGraw Hill*, 45(37):870–877, 1997.

[7] Christopher Bishop. Pattern recognition and machine learning. *Springer, New York*, 2006.

[8] Maurice C Bryson. The literary digest poll: Making of a statistical myth. *The American Statistician*, 30(4):184–185, 1976.

[9] Tim Harford. Big data: are we making a big mistake? ft magazine. http://www.ft.com/intl/cms/s/2/21a6e7d8-b479-11e3-a09a-00144feabdc0.html, 2014. [Online; accessed January 15, 2016].

[10] United States. Executive Office of the President and John Podesta. *Big data: Seizing opportunities, preserving values*. White House, Executive Office of the President, 2014.

[11] Nathalie Japkowicz and Shaju Stephen. The class imbalance problem: A systematic study. *Intelligent data analysis*, 6(5):429–449, 2002.

[12] Jing Jiang. A literature survey on domain adaptation of statistical classifiers, 2008.

[13] Yuchun Tang, Yan-Qing Zhang, Nitesh V Chawla, and Sven Krasser. Svms modeling for highly imbalanced classification. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(1):281–288, 2009.

[14] Patrick Glauner, Jorge Augusto Meira, Petko Valtchev, et al. The challenge of non-technical loss detection using artificial intelligence: A survey. *International Journal of Computational Intelligence Systems*, 10(1):760–775, 2017.

[15] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[16] Li Wan, Matthew Zeiler, Sixin Zhang, Yann Le Cun, and Rob Fergus. Regularization of neural networks using dropconnect. In *International Conference on Machine Learning*, pages 1058–1066, 2013.

[17] Sara Sabour, Nicholas Frosst, and Geoffrey E Hinton. Dynamic routing between capsules. In *Advances in Neural Information Processing Systems*, pages 3859–3869, 2017.

[18] Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.

ESANN 2018 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25-27 April 2018, i6doc.com publ., ISBN 978-287587047-6. Available from http://www.i6doc.com/en/.

[19] Brian W Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451, 1975.

[20] William M Stanish and Noel Taylor. Estimation of the intraclass correlation coefficient for the analysis of covariance model. *The American Statistician*, 37(3):221–224, 1983.

[21] Philipp Werner, Frerk Saxen, and Ayoub Al-Hamadi. Handling data imbalance in automatic facial action intensity estimation. *FERA*, page 26, 2015.

[22] Ivan Tomek. Two modifications of cnn. *IEEE Trans. Systems, Man and Cybernetics*, 6:769–772, 1976.

[23] Inderjeet Mani and I Zhang. knn approach to unbalanced data distributions: a case study involving information extraction. In *Proceedings of workshop on learning from imbalanced datasets*, volume 126, 2003.

[24] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[25] Gustavo EAPA Batista, Ana LC Bazzan, and Maria Carolina Monard. Balancing training data for automated annotation of keywords: a case study. In *WOB*, pages 10–18, 2003.

[26] Xu-Ying Liu, Jianxin Wu, and Zhi-Hua Zhou. Exploratory undersampling for class-imbalance learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 39(2):539–550, 2009.

[27] Joaquim L Viegas, Paulo R Esteves, R Melício, et al. Solutions for detection of non-technical losses in the electricity grid: A review. *Renewable and Sustainable Energy Reviews*, 80:1256–1268, 2017.

[28] Patrick Glauner, Angelo Migliosi, Jorge Augusto Meira, et al. Is big data sufficient for a reliable detection of non-technical losses? In *2017 19th International Conference on Intelligent System Application to Power Systems (ISAP)*, pages 1–6, Sept 2017.

[29] Bianca Zadrozny. Learning and evaluating classifiers under sample selection bias. In *Proceedings of the twenty-first international conference on Machine learning*, page 114. ACM, 2004.

[30] Solomon Kullback. Letter to the editor: The kullback-leibler distance. *The American Statistician*, 1987.

[31] Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

[32] Corinna Cortes and Mehryar Mohri. Domain adaptation and sample bias correction theory and algorithm for regression. *Theoretical Computer Science*, 519:103–126, 2014.

[33] James J. Heckman. Sample selection bias as a specification error. *Econometrica*, 47(1):153–161, 1979.

[34] Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Advances in neural information processing systems*, pages 137–144, 2007.

[35] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *International Conference on Machine Learning*, pages 448–456, 2015.

[36] Christopher M Bishop. *Neural networks for pattern recognition.* Oxford university press, 1995.