2012 11th International Conference on Machine Learning and Applications

One-Class versus Binary Classification: Which and When?

Colin Bellinger Shiven Sharma Nathalie Japkowicz
SITE, University of Ottawa
800 King Edward Avenue, Ottawa, Canada
{cbell059,sshar009,nat}@uottawa.ca

***Abstract*—Binary classifiers have typically been the norm for building classification models in the Machine Learning community. However, an alternate to binary classification is one-class classification, which aims to build models using only a single class of data. This is particularly useful when there is an over-abundance of data of a particular class. In such imbalanced cases, binary classifiers may not perform very well, and one-class classifiers then become the viable option. In this paper, we are interested in investigating the performance of binary and one-class classifiers as the level of imbalance increases, and, thus, uncertainty in the second class. Our objective is to gain insight into which classification paradigm becomes more suitable as imbalance and uncertainty increase. To this end, we conduct experiments on various datasets, both artificial and from the UCI repository, and monitor the performance of the binary and one-class classifiers as the size of the second class gradually decreases, thus increasing the level of imbalance. The results show that as the level of imbalance increases, the performance of binary classifiers decreases, whereas one-class classifiers stay relatively stable.**

***Keywords*-Machine learning, one-class classification, binary classification, imbalanced data.**

I. Introduction

The traditional methods of classification have always been those that use all data classes to build models. Such models are *discriminatory* in nature, since they learn to *discriminate* between classes. However, many real world situations are such that it is only possible to have data from one class, the *target class*; data from other classes, the *outlier classes*, is either very difficult or impossible to obtain. Examples of such domains include those in which there are almost an infinite number of instances from the outlier classes, such as in typist recognition [1], or those in which obtaining instances from the outlier classes is dependent upon the occurrence of a rare event[1], such as the detection of oil spills [2] or the inclusion of journal articles for systematic reviews [3]. Discriminatory methods cannot be used to their full potential in such situations, since by their very nature, they rely on data from all classes to build the discriminatory functions that separate the various classes. As a result, one-class learning methods become more appealing. These methods only use data from a single class to build a model, and are based on *recognition*, since their aim is to *recognize* data from a particular class, and reject data from all other classes.

One-class classification has seen a rise in application over the years, for example, in the use of document classification [4], typist recognition [1] and compliance verification of the CTBT [5]. However, to the best of our knowledge, the question of which classification paradigm, discriminatory or recognition-based, to apply, and when, has never explicitly been explored.

Since the use of either of the paradigms is dependant on the level of imbalance inherent in the dataset, a natural question to ask is: at what levels of imbalance does the use of binary classifiers becomes futile, and using one-class learning becomes the more suitable option? Clearly, if the data has a reasonable level of balance between the various classes, there is no reason why binary classifiers should not be used. It is at high levels of imbalance, and/or when there is a significant degree of uncertainty in the minority class, that the use of binary classifiers comes into scrutiny. We investigate this question by performing a series of experiments on both artificial datasets, and datasets from the UCI repository. The use of artificial datasets is purely for theoretical reasons, as it give us an opportunity to evaluate the learnt models using a large enough test set. The target class in each dataset has a fixed size, but the size of the outlier class is steadily decreased, thereby increasing the level of imbalance in the dataset. The performance of the classifiers is monitored over the increasingly imbalanced datasets. The results show a decreasing trend in binary classifier performance as the levels of imbalance increase. This result can be attributed to the binary classifier's failure to build a strong model of the second class. One-class classifiers, on the other hand, display fairly stable performance, offering support to their use in highly imbalanced cases.

The remainder of this paper is structured as follows; section II provides an overview of one-class classification (OCC). In Section III we elaborate on the issue of one-class versus binary classification methods in the context of imbalance. Section IV describes the various datasets in detail, and Section V describes the experimental framework. The results are presented in Section VI, and finally, concluding remarks are provided in Section VII.

II. One-Class Classification: An Overview

As we discussed in the preceding section, it is often the case that the data presented for inducing classifiers comes

[1]It is likely that the outlier class for classification is the target class in reality. However, we use the term *target class* to denote the *majority class*, while it may or may not be the intuitive target class.

978-0-7695-4913-2/12 $26.00 © 2012 IEEE
DOI 10.1109/ICMLA.2012.212

with either an overabundance of a single class, or the complete absence of all other classes (apart from a single, target class), thereby necessitating the use of OCC. One approach to OCC is to use density estimation [6]. This is performed by attempting to fit a statistical distribution to the data from the single class (the target data), and using the learnt density function to classify instances as belonging either to the target class (high density values), or to the set of outlier class (low density values). Parametric approaches rely on reliably estimating the distribution of the data beforehand, a challenging and impractical task given that most real-world data have complex distributions. An alternative approach to parametric techniques would be to use non-parametric techniques, such as Parzen Windows [6]. But, as the dimensionality of the data increases, these methods suffer from the well known curse-of-dimensionality problem, whereby the computational complexity for density estimation increases drastically.

There are algorithms designed specifically for one-class (OC) classification. An example of a OC classifier is the autoassociator (AA) [7], which can be thought of as a compression neural network, where the aim is to try to recreate the input at the output, with the compression taking place at the hidden layers. Hempstalk *et al*., in [1], describe a method, PDEN, for estimating the probability density function of a single class by first obtaining a rough estimate of the density of target class, generating an artificial class based on it and then performing binary learning. Yet another example of a OC classifier is the OC Support Vector Machine (OCSVM) [8]. OCSVMs assume the origin in the kernel space to be the second class, and, subsequently, learn a boundary that separates the target class from the origin.

III. One-Class versus Binary Classification

In this section, we discuss in detail the performance of binary classifiers in the context of levels of imbalance in the dataset from a Bayesian perspective.

The Bayes Rule for classification, assuming a zero/one loss function, given two classes ω_1 and ω_2, is: Classify as ω_1 if $p(x|\omega_1)P(\omega_1) > p(x|\omega_2)P(\omega_2)$, else classify as ω_2.

If instances from class ω_1 are much more abundant than those from class ω_2, we will get, for the prior probabilities, $P(\omega_1) >> P(\omega_2)$. Also, given the rarity of instances from class ω_2, the probability density functions (PDF) will be related as $p(x|\omega_1) >> p(x|\omega_2)$. Only in extremely rare, exceptional cases will this inequality be reversed. Given these relationships between the priors and the PDF, using the rule mentioned previously, we observe that we will almost always classify an instance as belonging to class ω_1. Clearly, the resulting classifier will be extremely biased towards the majority class, and will thus not be suitable for use in an imbalanced domain.

This analysis shows, from a Bayesian perspective, the effects of imbalance on binary classifiers; they almost always become biased towards the majority class, effectively ignoring the minority class. In contrast, OCC ignores prior probabilities, since, given a single class ω, the notion of using prior probabilities becomes moot. What we are interested in OCC is estimating the PDF of the given target class; once we have that, we can perform classification by imposing a threshold τ on the value given by the PDF for a given test instance:

$$\text{Classification}(x) = \begin{cases} \text{target}, & \text{if } p(x|\omega) \geq \tau \\ \text{outlier}, & \text{otherwise} \end{cases} \quad (1)$$

As we only use information from a single class to build a model (in this case, estimating the PDF), there is no bias present in it. Therefore, OCC, from the analysis shown here, becomes the better choice for building classification models when extreme levels of imbalance are present in the data.

The discussion here is from a purely theoretical perspective, and does not necessarily relate to any particular classifier. What we are interested in is empirically verifying the analysis presented in this section by running different classifiers on various datasets, and seeing at what point the use of binary classifiers becomes detrimental to the problem at hand.

IV. Description of the Data Sets

This section provides a description the various data sets used in the experiments. We begin by describing the artificial datasets that we create, followed by the UCI datasets.

A. Artificial Data

The purpose of using artificial data is to create an idealized data distribution on which we can concretely test the trends of classifier performance as class imbalance increases. It provides us with an ample test set, eliminating the need for using cross-validation. Having a very small outlier class causes the resulting test sets in cross-validation to be small, and the performance metric value may not represent the true classifier performance.

We use two artificial datasets which are various combinations of multimodal and unimodal target and outlier distributions. These are comprised of unimodal bivariate Gaussian distributions. The standard deviations for both dimensions are the same in both the target and outlier classes; only the means vary. We completely specify these distributions using six parameters (U: Unimodal, M: Multimodal, B: Bimodal):

Data 1: U target and M outlier distributions:

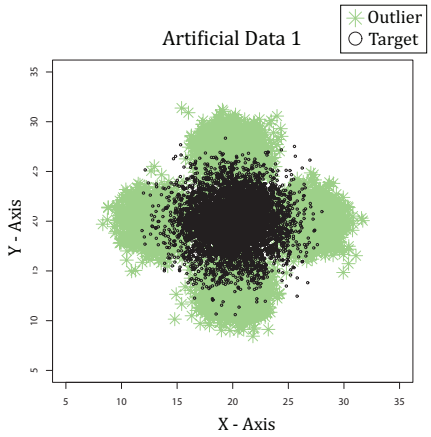
$$\text{Target}: N([\mu_{u1}, \mu_{u2}], \sigma_t)$$
$$\text{Outlier}: N([\mu_{u1}, \mu_{u2}+2.75\sigma_t], \sigma_o) \cup N([\mu_{u1}, \mu_{u2}-2.75\sigma_t], \sigma_o) \cup N([\mu_{u1}+2.75\sigma_t, \mu_{u2}], \sigma_o) \cup N([\mu_{u1}-2.75\sigma_t, \mu_{u2}], \sigma_o)$$

Data 2: B target and M outlier distributions:

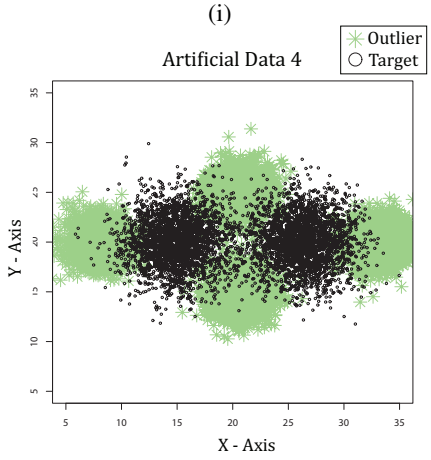
$$\text{Target}: N([\mu_{m1}, \mu_{m2}], \sigma_t) \cup N([\mu_{m1}, \mu_{m2}+4.5\sigma_t], \sigma_t)$$
$$\text{Outlier}: N([\mu_{m1}+2.25\sigma_t, \mu_{m2}+1.984\sigma_t], \sigma_o) \cup N([\mu_{m1}+2.25\sigma_t, \mu_{m2}-1.984\sigma_t], \sigma_o) \cup N([\mu_{m1}-2.75\sigma_t, \mu_{m2}], \sigma_o) \cup N([\mu_{m1}+7.25\sigma_t, \mu_{m2}], \sigma_o)$$

The values for the parameters are as follows: μ_{u1}: 20, μ_{u2}: 20, μ_{m1}: 15, μ_{m1}: 20, σ_t: 2.5 and σ_o: 1.5.

Figure 1 shows the plot of the datasets generated by the aforementioned distributions.



(i)



(ii)

Fig. 1. The artificial datasets.

TABLE I
DESCRIPTION OF THE UCI DATASETS USED IN THIS PAPER. THE WBCD DATASET IS THE WISCONSIN BREAST CANCER DATASET.

Dataset	Number of Targets	Number of Outliers
Diabetes	500	268
Heart Disease	150	120
Hepatitis	123	32
Ionosphere	225	126
Thyroid Disease	3541	231
Sonar	111	97
WBCD	357	212

B. UCI Datasets

Table I lists the datasets used, along with the number of target instances and outlier instances in each dataset. All the datasets are binary problems, with numeric attributes, and no missing values. As can be seen, the initial ratios indicate that there is not nearly enough imbalance to warrant using OCC.

V. EXPERIMENTAL FRAMEWORK

We use the Autoassociator (AA) and the Probability Density Estimator (PDEN) for one-class classification. The binary classifiers use are: Multilayer Perception (MLP), Decision Trees (DTree), Support Vector Machines (SVM), Nearest Neighbour (IBK)

Apart from AA, all classifiers have been implemented in WEKA [9], and run with their default settings. This is done so as to prevent any bias resulting from the fine tuning the parameters in order to obtain optimal results from specific datasets. For PDEN, we use the Gaussian Estimator as the density estimator, and AdaBoost with Decision Stumps as the class probability estimator. Both of these were used with default settings. The experiments with AA were implemented using the AMORE[2] R package, and run in R[3]. One hidden layer was used for the AA in all the experiments, and the number of training iterations was set to 50. The momentum value was set to 0.99, and the learning rate to 0.01. The number of hidden units for the artificial datasets were set to 4. For all other datasets, they varied from 1 to the number of dimensions of the particular dataset, and the number of units giving the best results were chosen.

The performance measure we use is the geometric mean of the per-class accuracies [10]. It is given by $gmean = \sqrt{acc_1 \times acc_2}$, where acc_i is the accuracy of the classifier on instances belonging to class i. By definition, the metric is immune to class imbalance. Evaluation is done using stratified 10-fold cross validation for the UCI datasets. For the artificial datasets, we use a dedicated training and test set.

To simulate the effect of imbalance, we fix the size of the target set, and steadily decrease the size of the outlier set. Let r be the ratio of outlier instances to target instances, *i.e.*, $r = \frac{|outliers|}{|targets|}$. We divide the range from r to 0.001 into 20 intervals of a fixed width, $w = \frac{r-0.001}{20}$. We then get size of each new outlier set, s as $s = (r - (w \times i)) \times |target|, \forall i \in \{0, 19\}$.

VI. EXPERIMENTAL RESULTS

We begin with the results over the artificial data, followed by the UCI dataset. It is worth noting that what we are interested in seeing are the trends; consequently, actual values are unimportant. Therefore, we only present graphs which plot the LOWESS curve [11] for the actual values.

A. Results on Artificial Data

The results of the binary and one-class classifiers are shown in Figure 2. The performance trends of the binary classifiers are clear: there is a steady decline as the levels of imbalance increase, starting at around imbalance levels of 1:2.8. The one-class classifiers on the other hand remain stable throughout. The slight deviations in the AA are due to it using the outlier class from the training set for setting the threshold. The stability of the one-class classifiers can be attributed to the fact that they only use information from the target class

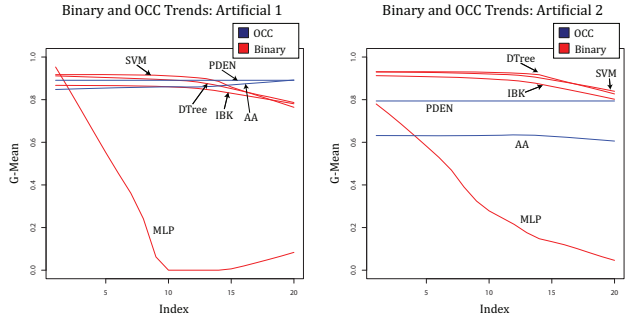
[2]AMORE: A MORE flexible neural network package, http://cran.r-project.org/web/packages/AMORE/index.html

[3]The R Project for Statistical Computing, http://www.r-project.org/

TABLE II
INITIAL RATIOS OF THE UCI DATASET, ALONG WITH THE RATIO AT WHICH BINARY LEARNING STARTS TO DETERIORATE.

Dataset	Initial Ratio	Deterioration Ratio
Diabetes	1:1.86	1:3.73
Heart Disease	1:1.25	1:2.5
Hepatitis	1:3.84	1:6.47
Ionosphere	1:1.78	1:3
Thyroid Disease	1:15.32	1:37
Sonar	1:1.14	1:2.92
WBCD	1:1.68	1:4.82

to build the models. As a result, they are not affected by the lack of information from the outlier class. Amongst the binary classifiers, Dtree, SVM and IBK all perform relatively similarly, whereas the MLP has the worst performance, with a sharp decline in performance at higher levels of imbalance.



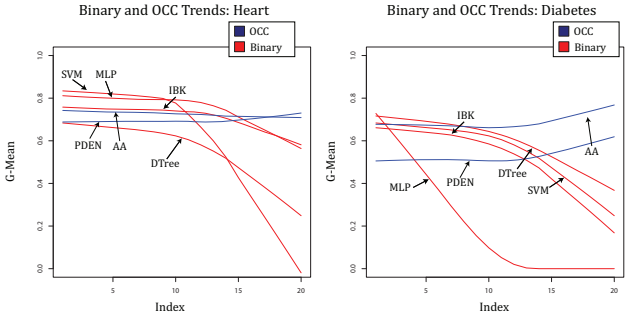
(i) (ii)

Fig. 2. Performance trends of the binary and one-class classifiers over the artificial datasets. A greater value of Index implies a greater imbalance.

B. Results on UCI Datasets

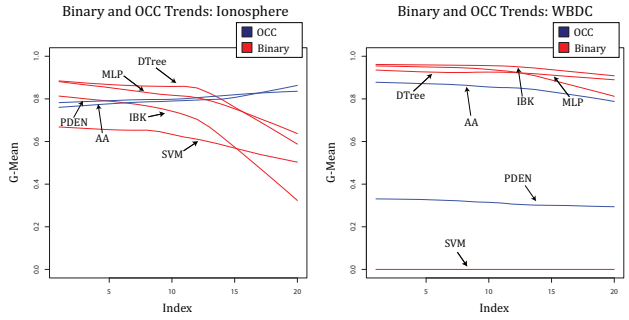
Figures. 3, 4, 5 and 6 show the performance trends of the binary and one-class classifiers over the various UCI datasets. These datasets originally come with levels of imbalance that are not extreme, thus making them conducive for binary classification. However, as we decrease the size of the outlier class and increase imbalance between the targets and outliers, there is a clear declining trend in performance in all the binary classifiers. This offers support to the fact that an increasing imbalance will cause deterioration in performance of binary classifiers. Table II displays the approximate ratios at which binary learning starts to deteriorate. The ratios indicate the number of target instances for each outlier instance. Thus, a ratio of $1:t$ indicates that there are t target instances for a single outlier instance in the dataset.

For one-class classifiers, one would expect their performance to remain stable, regardless of the level of imbalance. However, since we use 10-fold cross validation, the number of outliers in the test set changes, and as a result, we get different performance values for different sizes of the outlier set. But in all cases, the trends in one-class classifier performance and not even close to as pronounced as those in the binary case.



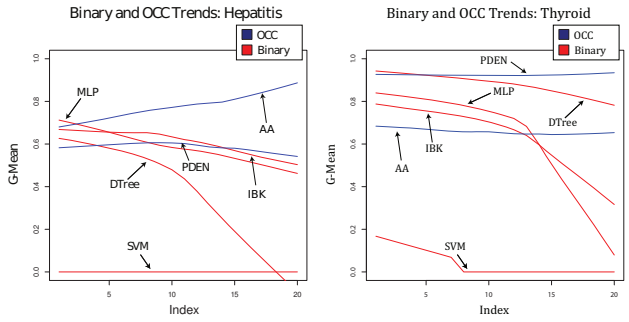
(i) (ii)

Fig. 3. Performance trends of the binary and one-class classifiers over the (i) Heart and (ii) Diabetes UCI datasets.



(i) (ii)

Fig. 4. Performance trends of the binary and one-class classifiers over the (i) Ionosphere and (ii) WBDC UCI datasets.



(i) (ii)

Fig. 5. Performance trends of the binary and one-class classifiers over the (i) Hepatitis and (ii) Thyroid UCI datasets.

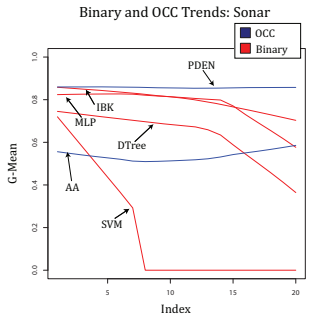


Fig. 6. Performance trends of the binary and one-class classifiers over the Sonar UCI dataset.

VII. Conclusions and Future Work

Given the inherent imbalance present in most real world datasets, it is natural to wonder which classification paradigm would be suitable, i.e., one based on discrimination (binary classification), or one based on recognition (one-class classification). We investigate the performance of binary and one-class classifiers over datasets in which we purposely decrease the size of the outlier (or second) class, thereby increasing the level of imbalance between classes. The results show that in all cases, the performance of the binary classifiers decreases as the imbalance increases. The one-class classifiers on the other hand, remain relatively stable in performance. More importantly, the performance of the binary classifiers is generally found to degrade below that of the OCC when the balance is taken to the extreme.

In almost all cases presented here, for both UCI and artificial datasets, there appears to be at least an imbalance ratio of $1:2.5$ before binary classifier performance starts to deteriorate. In other words, when there are at least two and a half times as many target instances as outlier instances, the performance of the binary classifiers may not be as strong as that of a recognition-based method. Indeed, in some cases, the imbalance ratio is well over $1:3.5$ before a decline in binary classifier performance happens. Different binary classifiers have different points of imbalance after which their performance starts to decline, making any generalizations over the discriminatory paradigm in the presence of imbalance difficult. In addition, each classification problem is unique. Specifically, some problems are easier to model than others, thus, fewer instances are required. Indeed, this appears to be the case in the WBDC problem. And, as a result, the majority of the binary learners are capable of effectively model the problem without succumbing to the class imbalance.

The answer to the question posed in the title of the paper, *Which and When?*, is not a simple one. The ultimate choice of which classification paradigm to use will depend on the problem at hand. While the work presented does show that performance of binary classifiers decreases with increasing imbalance, this does not imply that binary classifiers should not be used if there is any form of imbalance present. Indeed, even with a decent level of imbalance, certain binary classifiers can come up with effective decision boundaries. It is only when the imbalance is extreme to the point that the minority class is not providing nearly enough information does the value of using a recognition based paradigm becomes apparent.

Furthermore, one must be cautious when empirically evaluating binary and one-class classifiers over datasets that are inherently suited for binary classification. Examples of such datasets are the UCI datasets; the initial imbalance ratios presented in Table II illustrate this point. Both paradigms are complementary to each other; the success of one is usually dependent on the failure of the other, and as a result, comparisons between them can never be absolute, but only relative to the context of their application.

We conducted experiments using two popular one-class classifiers, the autoassociator and PDEN, and four well known binary classifiers. Continuing this study, we will explore the performance trends for other one-class and binary classifiers for increasing levels of imbalance. Furthermore, it would be worth taking this work further by examining each classifier individually, and discovering the various nuances present in them that contribute to their performance over imbalanced datasets. A deeper understanding into the workings of both binary and one-class classifiers over different datasets, with varying levels of imbalance, can help facilitate the selection of the appropriate classifier for the task at hand. Indeed, the results here show that the performance of any classifier is highly dependent on both the nature of the dataset and the degree of imbalance inherent in it.

References

[1] K. Hempstalk, E. Frank, and I. H. Witten, "One-class classification by combining density and class probability estimation," in *Machine Learning and Knowledge Discovery in Databases*, ser. Lecture Notes in Computer Science, vol. 5211. Berlin: Springer, 2008, pp. 505–519.

[2] M. Kubat, R. C. Holte, and S. Matwin, "Machine learning for the detection of oil spills in satellite radar images," in *Machine Learning*, 1998, pp. 195–215.

[3] S. Matwin, A. Kouznetsov, D. Inkpen, O. Frunza, and P. O'Blenis, "A new algorithm for reducing the workload of experts in performing systematic reviews," *Journal of the American Medical Informatics Association*, vol. 17, pp. 446–453, 2010.

[4] L. M. Manevitz and M. Yousef, "One-class svms for document classification," *The Journal of Machine Learning Research*, vol. 2, pp. 139–154, 2002.

[5] C. Bellinger and B. J. Oommen, "On simulating episodic events against a background of noise-like non-episodic events," in *42nd Summer Computer Simulation Conference, SCSC 2010, Ottawa, Canada, July 11-14, 2010. Proceedings*, 2010.

[6] R. O. Duda, P. Hart, and D. Stork, *Pattern Classification*. Wiley-Interscience, 2000.

[7] N. Japkowicz, "Supervised versus unsupervised binary-learning by feedforward neural networks," *Machine Learning Volume 42, Issue 1/2*, vol. 42, pp. 97–122, 2001.

[8] B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson, "Estimating the support of a high-dimensional distribution," *Neural computation*, vol. 13, no. 7, pp. 1443–1471, 2001.

[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: An update," *ACM SIGKDD Explorations Newsletter*, vol. 11, no. 1, pp. 10–18, 2009.

[10] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: One-sided selection," in *In Proceedings of the Fourteenth International Conference on Machine Learning*. Morgan Kaufmann, 1997, pp. 179–186.

[11] W. Cleveland, "Robust locally weighted regression and smoothing scatterplots," *Journal of the American Statistical Association*, pp. 829–836, 1979.