Fair Samples, Fair Predictions: KNN's Battle Against Gender Bias

Jonathan Ohop
University of Massachusetts Amherst
300 Massachusetts Ave, Amherst, MA 01003
JOhop@uMass.edu

Abstract

Since machine learning impacts many areas of our lives, it's essential for these algorithms to be both fair and accurate. This work uses feature weighting and datasets like the 1994 Adult Census and German Credit Data to address an important problem. By investigating various distance metrics in k-NN classifiers(k-Nearest Neighbors), the research reveals that implementing feature weighting can significantly reduce bias while maintaining accuracy, ultimately contributing to the development of more equitable machine learning systems..

1. Introduction

Machine learning algorithms, like k-Nearest Neighbors, are increasingly used in important decision-making roles, raising concerns about fairness and bias. Many datasets used to train these algorithms can contain biases that lead to discriminatory outcomes, particularly affecting groups based on race, gender, and socioeconomic status. While it's common to remove sensitive features from training data to promote fairness, this approach often falls short, as bias can still emerge from other features. Additionally, the reliance on kNN and similar models can perpetuate these biases if not carefully managed, underscoring the need for ongoing evaluation and improvement of fairness in machine learning applications.

In a job hiring situation, a k-Nearest Neighbors (kNN) model might unintentionally favor candidates from certain zip codes linked to specific racial groups, causing qualified individuals to be passed over simply because of where they live. Similarly, if a kNN algorithm evaluates resumes that mostly include graduates from top schools, it might favor those candidates and overlook equally qualified applicants from less well-known schools or diverse backgrounds. Additionally, if most successful applicants in the data are men, kNN might unfairly favor men because they appear more often in the nearest neighbors. This bias can be further amplified if features like experience are unevenly distributed between men and women, making this a specific case we will address in this project.

This topic is important because machine learning makes our daily lives more efficient, especially in areas where getting things right is essential and there's little room for mistakes. As these technologies continue to advance, it's crucial to understand their impact on society, making responsible development and use more important than ever.

This project aims to tackle these issues by investigating and applying preprocessing techniques to minimize bias in kNN classifiers. The goal is to promote an equitable decision-making process while maintaining performance, ultimately contributing to the development of fairer machine learning systems.

According to Hort et al. (2021):

> "The most general baseline is to compare the fairness achieved by classification models after applying a bias mitigation method with the fairness of a fairness-agnostic Original Model."[5]

In other words, the "most general baseline" is a basic comparison to determine whether bias mitigation methods actually make models fairer. It compares models that have been adjusted for fairness with an original model that hasn't been adjusted. This helps researchers understand how much fairness improves when using these methods by showing the difference between a standard model and one that's been modified to be fairer. It's a straightforward way to measure the effectiveness of these methods in reducing bias in models. To further this approach, my two baselines focus on race and gender equality, providing insights into how bias mitigation techniques can enhance fairness across different demographic groups. By evaluating these factors, I will demonstrate the practical impact of fairness adjustments in machine learning.

2. Related work

The importance of fairness in machine learning, especially in algorithms like k-NN, is highlighted by definitions that

consider both individual and group perspectives. "Fairness Through Awareness" [3] defines fairness as treating individuals equitably, ensuring that decisions are informed by sensitive attributes without bias. Their approach emphasizes the importance of treating individuals fairly and suggests using algorithms that consider personal circumstances to help reduce biases. "Fairness and Machine Learning" [1] defines group fairness as ensuring that no demographic group is unfairly disadvantaged due to sensitive attributes like race, gender, or religion. They emphasize the importance of equal treatment and outcomes for all groups, underlining the need for algorithms to avoid favoring or disadvantaging any specific demographic. Together, these definitions highlight the essential role of fairness in creating equitable machine learning systems, ensuring that both individual and group perspectives are taken into account in algorithms like k-NN.

The application of preprocessing techniques in machine learning to guarantee equitable treatment is greatly influenced by concepts of fairness, addressing both individual and group-level equity. According to Fairness Through Awareness, fairness should be tailored to each individual, taking sensitive characteristics into consideration to lessen bias. This aligns with preprocessing techniques such as relabeling. To minimize accuracy loss and ensure fairness—especially by achieving equal outcomes for protected groups—labeling techniques like "massaging" modify labels, often focusing on individuals near the threshold of classification.

Kamiran and Calders (2012) explain that

> "Massaging uses a ranker to determine the best candidates for relabeling" [6]

Perturbation also aims to reduce bias, but instead of focusing on labels, it focuses on modifying features. Perturbation adjusts the data to make the sensitive characteristics more similar across groups, while keeping the order of individuals within each group the same.

As noted by Feldman et al. (2015), "perturbation modifies non-protected attributes such that their values for privileged and unprivileged groups are comparable" [4].

To ensure that no demographic group is unfairly disadvantaged from a group fairness perspective, methods such as reweighing and sampling seek to adjust the training data to better represent underrepresented groups. Reflecting the concept of group fairness, reweighing assigns higher weights to instances from underrepresented or misclassified groups.

Calders et al. (2009) explain that

> "Instances in the unprivileged group and positive label receive a higher weight, as this is less likely." [2]

Sampling, in contrast, modifies the dataset by either down-sampling, udersampling or up-sampling, oversampling. Down-sampling reduces the number of points in overrepresented groups, while up-sampling increases the number of points in underrepresented groups. Both techniques help balance the distribution of the data. While sampling changes the structure of the dataset, reweighing only adjusts how much each data point impacts the model, without altering the dataset itself. In contrast to sampling, which modifies the data, reweighing affects the importance of the data points without changing the composition of the data.

3. Method

Expanding on this idea, I developed two baselines with preprocessing methods to evaluate the fairness of classification models. Scikit-learn's k-Nearest Neighbors was used for the main experimentation throughout this project. It operates by determining which data points in the training set are closest to a brand-new, unobserved data point. I used sample datasets in an attempt to mitigate bias between the two gender groups. One of the datasets used is the Adult Income dataset, also called the "Census Income" dataset.

	age	workclass	fnlwgt	education	education.num	marital.status	occupation	relationship	race	sex	capital.gain	capital.
0	90	?	77053	HS-grad	9	Widowed	?	Not-in-family	White	Female	0	4356
1	82	Private	132870	HS-grad	9	Widowed	Exec-managerial	Not-in-family	White	Female	0	4356
2	66	?	186061	Some-college	10	Widowed	?	Unmarried	Black	Female	0	4356
3	54	Private	140359	7th-8th	4	Divorced	Machine-op-inspct	Unmarried	White	Female	0	3900
4	41	Private	264663	Some-college	10	Separated	Prof-specialty	Own-child	White	Female	0	3900
...	...	...	...	...	...	...	...	...	...	...	...	...
32556	22	Private	310152	Some-college	10	Never-married	Protective-serv	Not-in-family	White	Male	0	0
32557	27	Private	257302	Assoc-acdm	12	Married-civ-spouse	Tech-support	Wife	White	Female	0	0
32558	40	Private	154374	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	0	0
32559	58	Private	151910	HS-grad	9	Widowed	Adm-clerical	Unmarried	White	Female	0	0
32560	22	Private	201490	HS-grad	9	Never-married	Adm-clerical	Own-child	White	Male	0	0

Figure 1. A visualization of the Adult data set and some of its features.

This dataset is designed to predict whether a person's annual income is over $50,000 based on factors like age, education, occupation, and marital status, helping classify individuals as either earning above or below $50,000. The other dataset used is the German Credit Data dataset, which aids in predicting credit risk by categorizing individuals based on their financial stability. It contains features such as credit history, loan purpose, and employment status. The dataset helps in building models to assess individuals' ability to repay loans.

To assess how preprocessing influences bias mitigation in k-Nearest Neighbors (k-NN), we compare a preprocessed-aware model with an unaware model without preprocessing. This comparison reveals whether preprocessing reduces bias and improves the fairness of the

model's decisions, providing insight into how it affects both accuracy and fairness. The fairness-unaware model trains a k-NN classifier on the entire dataset without incorporating fairness adjustments, such as balancing the representation of different groups like male and female.

The ideal number of neighbors, k, is selected through K-Fold cross-validation, which evaluates multiple values and calculates the average accuracy. Next, the model is evaluated on the training data and then trained on the full dataset using the chosen optimal k. This model serves as a starting point for comparison with fairness-aware models, but it overlooks fairness, potentially leading to biased performance across groups and favoring one gender over another.

Unlike the fairness-unaware model, the fairness-aware model incorporates fairness during preprocessing, ensuring that the model performs equally well for different groups—in this case, male and female. Biases or imbalances in the data can be addressed by preprocessing procedures, such as ensuring that different groups are fairly represented. These strategies can help create a more balanced dataset, improving the model's ability to make fair predictions and promoting more equitable outcomes across different groups.

This study's findings highlight the significant impact of gender bias, particularly on the prediction accuracy for male and female groups, in machine learning models. At first, there was a notable disparity in the Adult dataset, with 10,771 female records and 21,790 male entries. I also simulated disparity within the German dataset by extracting 700 males and 300 females. The raw, unbalanced data used to train the bias-unaware model led to its performance being affected by the gender gap. In this approach, the model learned predominantly from the male data due to its greater representation, leading to a skewed learning process. This difference highlights the inherent bias in the data, as the model was more inclined to correctly predict male outcomes due to the predominance of male records in the dataset.

3.1. UnderSampling

To address this bias, a bias-aware model was introduced. This model employed a preprocessing method called undersampling to reduce the size of the male group to match that of the female group.

> "Sampling methods change the training data by changing the distribution of samples (e.g., adding, removing samples) or adapting their impact on training. Similarly, the impact of training data instances can be adjusted by reweighing their importance."[6]

Within the Adult dataset, the model created a more balanced dataset by randomly selecting 10,771 male records from the larger pool, ensuring equal representation of both genders. Under-sampling may slightly reduce the information from the male group, but it was necessary to prevent the gender imbalance from disproportionately affecting the model. Once this technique was applied, the model was retrained, and its performance reassessed. The fairness-aware model achieved an accuracy of 80.025%, slightly lower than the 80.05% accuracy of the bias-unaware model, suggesting that the fairness-aware model may have sacrificed some accuracy in favor of reducing bias or increasing fairness, particularly for the minority class. This difference reflects the common effect of balancing datasets, with the reduction in accuracy being minimal and expected due to the reduced male dataset. It's interesting to note that, in this case, the fairness-aware model could be considered slightly less accurate than the fairness-unaware model, even if the difference is minimal. Can we really call it "accuracy" if it lacks fairness? It's like forming an opinion based on misleading or incomplete information. In that case, it seems it was never a true measure of accuracy to begin with.

Although the female accuracy was significantly lower in the fairness-unaware model, at 68.96%, the male accuracy in the fairness-unaware model remained high at 85.71%. By comparison, the accuracy for females in the bias-aware model improved to 72.51% after the dataset was balanced. This rise illustrates how bias mitigation works. The algorithm improved its ability to forecast female outcomes by ensuring equal representation of both genders. Significantly, the accuracy for males stayed consistent at 85.71%, suggesting that the under-sampling did not affect the prediction of male outcomes and instead contributed to creating a more equitable situation for females.

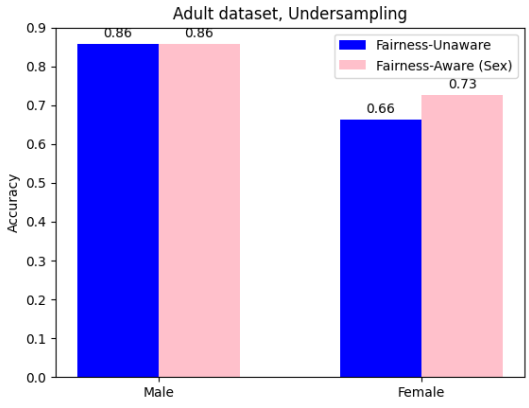

Figure 2. Shows an increase in female accuracy after undersampling.

Although the bias-unaware model demonstrated significant overall accuracy at first, its results were influenced by the gender disparity present in the dataset. Due to a

slight decrease in overall accuracy, the bias-aware model enhanced predictive fairness for females by under-sampling the male group. This improvement in fairness, evident in the higher accuracy for females, highlights the critical importance of addressing gender bias in machine learning. In addition to yielding more equitable results, this approach ensures that the model is fairer and more representative of all gender groups, leading to a more just and balanced prediction system while maintaining similar performance overall.

The model developed for the German Credit dataset assesses and enhances the fairness of a k-Nearest Neighbors (k-NN) classifier in predicting credit risk, utilizing the German credit data. The main objective is to evaluate the model's performance both before and after applying under-sampling to correct the gender imbalance in the dataset. To replicate the results of the previous experiment with the Adult dataset on a smaller scale, 700 males and 300 females were selected from the dataset, which might lead the model to favor the larger group. The overall accuracy of the model's performance is assessed first, followed by separate assessments for the male and female groups.

To convert the personal_status column into a binary sex column (1 for men and 0 for women), the code first loads and preprocesses the data. It then applies one-hot encoding to categorical features to prepare the data for machine learning. The k-NN classifier is trained and evaluated on the imbalanced dataset, achieving an overall accuracy of 79.67%. The accuracy rates were 83.33% for males and 84.44% for females, with females slightly outperforming males.

As seen in the previous Adult dataset example, machine learning algorithms often exhibit bias toward the majority class in imbalanced datasets because it occurs more frequently during training. This can lead the model to focus more on predicting the majority class, especially when using class-based evaluation metrics. The accuracy results before sampling showed higher accuracy for females, despite the imbalance, which suggests that the model might be performing better for females. This could be because the model is more careful when predicting the minority class, trying to reduce false negatives. These results indicate that the model isn't biased toward men. It might also mean the model is better at identifying patterns for the minority class, or that the features are more useful for women than for men.

Under-sampling is still necessary to reduce bias caused by class imbalance, even though the model performed better for females before under-sampling. Even if the accuracy for the minority class is decent, models in imbalanced datasets often become biased toward the majority class. This bias can lead to overfitting on the majority class and cause the model to miss important patterns in the minority class, hurting its ability to generalize. Under-sampling helps by balancing the dataset, allowing the model to learn more equally from both classes. As a result, both male and female performance improves, especially in terms of metrics like precision.

Under-sampling is used to address the gender imbalance by randomly reducing the male data to 300 records, creating a balanced dataset with an equal number of males and females. Following the application of under-sampling, the model undergoes retraining, resulting in a slight improvement in overall accuracy to 82.22%. More significantly, both the male and female accuracies are now equal at 82.22%, suggesting that the model's performance is now gender-neutral while maintaining similar overall performance.

The findings highlight how crucial it is to address class imbalance in machine learning, as the dataset's imbalance led the model to favor males prior to under-sampling. The model's fairness was enhanced by under-sampling the male group, which resulted in more evenly distributed performance for the sexes. This method shows how under-sampling can improve fairness and reduce gender bias in predictive models, especially in sensitive applications like credit scoring.

3.2. Oversampling

This is content in the subsection.

Diving deeper into sampling, there exists another subgenre of sampling called SMOTE or Synthetic Minority Over-sampling Technique.

Hort describes SMOTE as a popular method and also defines characteristics of this mitigation technique:

> "SMOTE does not duplicate instances but generates synthetic ones in the neighborhood of the minority group"[5]

Instead of increasing the number of existing data points like oversampling or reducing the data as in undersampling, this method generates new, synthetic samples for the smaller class. This technique can enhance model performance, especially when dealing with imbalanced datasets.

I chose upsampling as the second preprocessing method because, after exploring downsampling, I wanted to gain a deeper understanding of its counterpart.

Preprocessing the data by turning the sex and occupation variables into numerical features is the first step in the upsampling experiment for the Adult dataset. To handle missing values, one-hot encoding and imputation are used. The dataset initially contains 10,771 females and 21,790 males before sampling. After the minority class is upsampled to balance the gender distribution, the dataset contains 21,790 males and 21,790 females, ensuring equal representation of both genders during training. In order to find the optimal value of k for both fairness-unaware and fairness-aware

models, the code performs cross-validation on both unbalanced and balanced datasets, evaluating k values from 1 to 20 for the k-NN model.

The fairness-unaware model, trained on an unbalanced dataset, has an accuracy of 79.93%, while the fairness-aware model, trained on a balanced dataset, achieves an accuracy of 81.15%. Although the fairness-aware model shows a slight overall improvement, it performs worse for females (76.54%) compared to the fairness-unaware model (73.46%). Both models have similar accuracy for males (83.33%). This highlights a trade-off: balancing the dataset to reduce bias may improve fairness but can slightly hurt performance for certain groups, like females. However, the fairness-aware model still helps mitigate bias by improving overall gender balance in performance.

Although the drop in female accuracy with the fairness-aware model is disappointing compared to the fairness-unaware model, it is a normal part of bias mitigation. The dataset doesn't always behave as expected when attempting to reduce bias. While balancing the dataset can help prevent the model from favoring one group over another, it can also make the underrepresented group less accurate. The dataset is balanced in the fairness-aware approach by including an equal number of female and male examples. This ensures equal representation and trains the model on repeated female data, which can lead to overfitting if the examples aren't fully representative of the broader population. Instead of learning general trends, upsampling the female data could cause the model to learn and apply patterns from the repeated examples. The model performs worse on female test data as a result of its inability to generalize to fresh, unseen female data. This suggests that, to ensure fairness, the disadvantaged group might sometimes lose accuracy. The lower female accuracy suggests that while the fairness-aware model reduces gender bias, it may not be the best choice if high accuracy, especially for females, is the main goal.

To sum up, the code demonstrates how to mitigate bias using fairness-aware models and clearly illustrates the trade-off between accuracy and fairness. The challenge of balancing fairness and performance is shown by the fact that the fairness-aware model ensures equal representation of males and females, while the fairness-unaware model performs better for females. This suggests that other methods, like post-processing or more advanced models, are needed to better balance accuracy and fairness without sacrificing individual performance.

Applying SMOTE to the German dataset yielded some issues. The majority class was males at 700, while the female class had 300. The goal was to create enough synthetic samples for the female class to match the majority class of males at 700. However, SMOTE was unable to place the female class at exactly 700, as needed. Since SMOTE generates new examples by blending existing samples from the minority class, it requires a diverse set of data to work effectively. If the minority class has very few unique examples or if the examples are too similar, SMOTE may struggle to create enough diverse new samples. This problem is especially noticeable when the minority class is small or quite different from the majority class. In such cases, SMOTE might not produce enough synthetic data to balance the class sizes effectively.

In order to maintain the integrity of the experiment, upsampling was chosen instead. This model uses the German Credit dataset to train a kNN classifier for predicting credit risk. The dataset, which contains information like checking account status, credit history, and employment, is first loaded. The personal_status column is used to identify gender, with males mapped to 1 and females mapped to 0. Then, one-hot encoding is applied to categorical features to prepare the dataset for the model. The dataset initially contains 700 male and 300 female samples, creating an imbalance in the gender distribution.

The dataset is split into training and test sets, and the model is trained on the data. The accuracy is measured before any adjustments are made. Before upsampling, the model achieves an accuracy of 79.67% on the unbalanced data, with 700 males and 300 females. This imbalance suggests the model may be biased toward predicting the majority class (males).

The code uses random oversampling to increase the number of female samples to match the number of male samples, addressing the gender imbalance. The code duplicates female samples at random until the target number of 700 females is reached. The k-NN model is then run again to assess the improved accuracy after the new dataset is balanced with 700 males and 700 females.

Not only does the model's accuracy increase to 84.05% after oversampling, but more importantly, both males and females now share an accuracy of 83.3%, illustrating that both genders are now treated equally in the model. A balanced dataset prevents the model from favoring the majority class (males), making it less biased and more equitable. This shows that oversampling reduces bias and enhances model performance, producing predictions that are more equitable for both genders. In situations with notable class disparity, this method demonstrates how oversampling can reduce bias in machine learning models and produce predictions that are fairer to all groups.

3.3. Novel Approach

In the new approach, the k-Nearest Neighbors (KNN) classifier finds the data points that are nearest to a specified point for classification using a distance measure. To ensure that all features are on the same scale, StandardScaler is used to scale them. This prevents the distance computa-

tion from being dominated by larger features. Furthermore, some attributes are given weights, which increases their significance in calculating the separation between points. KNN computes the distances automatically using Euclidean distance by default.

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Scaling and weighting have an impact on the algorithm's determination of which neighbors are most crucial for predictions.

The Adult dataset was chosen due to its larger size, with over 30,000 samples, making it better suited for advanced techniques like sampling and feature weighting. Larger datasets like this help prevent issues such as overfitting and make it easier to address fairness. In contrast, smaller datasets like the German Credit dataset, with only 1,000 samples, are more sensitive to these techniques and can quickly lead to overfitting or errors. As a result, larger datasets offer greater stability and are more effective for handling fairness issues.

A new method to reduce gender bias was applied to the Adult dataset using several preprocessing steps. First, the data is prepared by converting categorical features into numbers with one-hot encoding, fixing missing values, and scaling the features so they all contribute equally. The main focus is on reducing gender bias through two techniques: sampling and adjusting feature weights.

There are 10,771 females and 21,790 males in the dataset. Because there are more men than women, this leads to an imbalance. Models may become biased and perform better on the majority group (males, in this case) when there is such an imbalance.

This imbalance is addressed by balancing the dataset and lowering the proportion of men. To ensure that there are an equal number of males and females, we randomly select 10,771 males, matching the number of females. This procedure is called downsampling. To obtain a balanced dataset, we combine the selected males and females.

Lastly, we shuffle the data to ensure that the model is not impacted by any patterns or order in the data.

In the dataset, some features like education, occupation, hours worked, capital gain/loss, and age are more important for predicting income than others. To make sure the model pays extra attention to these important features, we give them higher weights. We set the weight for these features to 10. This means that the model will treat them as more important when making predictions. By doing this, the model can make more accurate predictions because it's focusing more on the features that matter more for predicting income. It also helps make the predictions more fair because the important features are properly considered.

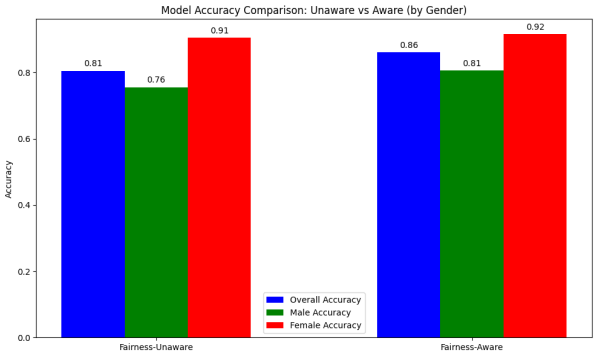

Figure 3. Results of the novel approach mitigating bias.

The Fairness-Unaware model is trained on the original unbalanced dataset, where there are more males than females, leading to better performance for females (90.5%) than males (75.6%). This imbalance creates bias in the model. In contrast, the Fairness-Aware model is trained on a balanced dataset, ensuring an equal representation of males and females. Additionally, feature weighting is applied to prioritize certain features, enhancing the model's fairness and performance, such as education and occupation. As a result, the Fairness-Aware model maintains high accuracy for females at 91.6% but also improves accuracy for males at 80.6%, achieving an overall accuracy of 86.1%. This shows that by balancing the data and adjusting feature importance, the model reduces bias and ensures more equitable performance across gender groups.

References

[1] S. Barocas, M. Hardt, and A. Narayanan. *Fairness and Machine Learning*. fairmlbook.org, 2019. Available online: https://fairmlbook.org/. 2

[2] T. Calders, F. Kamiran, and M. Pechenizkiy. Building classifiers with biased data. In *Proceedings of the 2009 IEEE International Conference on Data Mining*, pages 319–328. IEEE, 2009. 2

[3] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pages 214–226. ACM, 2012. 2

[4] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM, 2015. 2

[5] M. Hort, Z. Chen, J. M. Zhang, M. Harman, and F. Sarro. Bias mitigation for machine learning classifiers: A comprehensive survey. *ACM Computing Surveys*, 2021. 1, 4

[6] F. Kamiran and T. Calders. Massaging: A data preprocessing technique to reduce bias. In *Proceedings of the 2012 European Conference on Machine Learning and Knowl-*

edge Discovery in Databases (ECML PKDD), pages 183–196.
Springer, 2012. 2, 3