공학석사학위논문

Measuring the Economic Value of Data in Machine Learning: A Cooperative Game Approach

협력 게임이론을 활용한 기계학습에서의 데이터 경제적 가치 측정

2021 년 2 월

서울대학교 대학원

산업공학과

최 문 석

Measuring the Economic Value of Data in Machine Learning: A Cooperative Game Approach

협력 게임이론을 활용한 기계학습에서의 데이터 경제적 가치 측정

지도교수 이 덕 주

이 논문을 공학석사 학위논문으로 제출함

2021 년 2 월

서울대학교 대학원

산업공학과

최 문 석

최문석의 공학석사 학위논문을 인준함

2021 년 2 월

위 원 장 홍 성 필 (인)

부위원장 이 덕 주 (인)

위 원 장 우 진 (인)

Abstract

Measuring the Economic Value of Data in Machine Learning: A Cooperative Game Approach

Moonseok Choi
Department of Industrial Engineering
The Graduate School
Seoul National University

As machine learning thrives in both academia and industry at the moment, data plays a salient role in training and validating machines. Meanwhile, few works have been developed on the economic evaluation of the data in data exchange market. The contribution of our work is two-fold. First, we take advantage of semi-values from cooperative game theory to model revenue distribution problem. Second, we construct a model consisting of provider, firm, and market while considering the privacy and fairness of machine learning. We showed Banzhaf value could be a reliable alternative to Shapley value in calculating the contribution of each datum. Also, we formulate the firm's revenue maximization problem and present numerical analysis in the case of binary classifier with classical data examples. By assuming the firm only uses high quality data, we analyze its behavior in four different scenarios varying the data's fairness and compensating cost for data provider's privacy. It turned out that the Banzhaf value is more sensitive to the fairness of data than the Shapley value. We analyzed the maximum revenue proportion which the firm gives away to data providers, as well as the range of number of data the firm would acquire.

Keywords: Machine learning, Cooperative game theory, Semi-value, Data evaluation, Fairness, Privacy
Student Number: 2019-22469

Contents

Abstract i

Contents iii

List of Tables v

List of Figures vi

Chapter 1 Introduction 1

1.1 Research Background ····· 1

1.2 Problem Description ····· 2

1.3 Organization of the Thesis ····· 3

Chapter 2 Literature Review 4

2.1 Fair Machine Learning ····· 4

2.2 Private Machine Learning ····· 5

2.3 Data Valuation ····· 6

2.3.1 Dataset Price Estimation ····· 6

2.3.2 Equitable Price Estimation ····· 7

Chapter 3 Data Market Model 8

3.1 Basic Assumptions and Model Settings ····· 8

3.2 Firm's Profit Maximizing Problem ····· 10

3.3 Data Valuation ····· 12

3.4 Binary Classification Setting ····· 14

Chapter 4 Analysis **17**

4.1 Semi-value Approximation ····· 17

4.1.1 Convergence Analysis ····· 17

4.1.2 Group Data Calculation ····· 20

4.2 Binary Classification ····· 22

4.2.1 Parameter Analysis ····· 22

4.2.2 Scenario Analysis ····· 24

4.2.2.1 Description ····· 24

4.2.2.2 Synthetic Data ····· 25

4.2.2.3 Shapley Value Based Valuation ····· 26

4.2.2.4 Banzhaf Value Based Valuation ····· 28

4.2.2.5 Comparative Analysis ····· 30

4.3 Data Pricing ····· 33

Chapter 5 Conclusion **35**

Bibliography **38**

국문초록 **43**

List of Tables

Table 4.1 Semi-value convergence ····· 19

Table 4.2 Shapley value based scenarios ····· 27

Table 4.3 Banzhaf value based scenarios ····· 29

List of Figures

Figure 3.1 Data market model ····· 8

Figure 3.2 model accuracy with *n* highest semi-value data ····· 14

Figure 4.1 synthetic data ····· 17

Figure 4.2 semi-value approximation ····· 18

Figure 4.3 normal/fictitious data semi-value approximation ····· 18

Figure 4.4 group synthetic data ····· 20

Figure 4.5 semi-value calculation in group data ····· 21

Figure 4.6 unfair datasets ····· 25

Figure 4.7 fair datasets ····· 25

Figure 4.8 model accuracy with n highest Shapley value data in unfair dataset ····· 26

Figure 4.9 model accuracy with n highest Shapley value data in fair dataset ····· 26

Figure 4.10 model accuracy with n highest Banzhaf value data in unfair dataset ····· 28

Figure 4.11 model accuracy with n highest Shapley value data in fair dataset ····· 28

Figure 4.12 threshold α ····· 30

Figure 4.13 optimal N ····· 30

Figure 4.14 moons() dataset samples ····· 33

Figure 4.15 blobs() dataset samples ····· 33

Chapter 1

Introduction

1.1 Motivation

The development of machine learning over the last decade has been explosive at the same time as academic development and real-life application development. The evolution of machine learning, which has stagnated until very recently, has been aided by significant advances in computer hardware performance and the development of backpropagation theory, which can handle large datasets. The reason why deep learning has been able to surpass existing machine learning methodologies is that information from large datasets can be preserved and used as much as possible. Artificial intelligence is used in many real-life situations such as interpretation of medical device results and speech synthesis using medical, voice, and photographic data.

Despite the fact that everyone is aware of the importance of data, there is no standard measurement of data's economic value. Although domestic and foreign companies such as BDEX [19], Datastream Group [20], Info [21] and Selectstar [22] are operating the data market or providing necessary data, there are many differences in how data contributes to artificial intelligence market services depending on the machine learning model, learning method and model structure. However, since the size of the data market grows over time, specifying an economic value is essential to establishing the overall data market.

1.2 Problem Description

This work addresses the problem of modeling the data market and redistributing economic goods among the players that make up the market. Since none of the existing research deals with the economics of the data transaction market, we will briefly address only three entities: data providers, a firm providing artificial intelligence services and applicable markets. From a data market perspective, data providers and the firm correspond to the role of sellers and consumers, respectively.

Furthermore, we address two perspectives on the economic value of the data. The first is the economic value of big data, which considers how much the actual data set improves the service performance of a specific machine learning model. Second, we raise the question of value redistribution about how the firm, the entity that provides services and gains economic benefits, values the data to those who provided the actual data. Taking these into account, we will propose an overall data market structure while at the same time establishing the value of a particular dataset.

1.3 Organization of the Thesis

This paper consists of five chapters. Chapter 2 looks at the preceding studies, and chapter 3 presents a model of the data market and derives the optimal solution in a particular environment. In chapter 4, the conditions associated with data prices are analyzed and numerical examples identify the characteristics of the optimal solution. Finally, chapter 5 presents conclusions and future research directions.

Chapter 2

Literature Review

2.1 Fair Machine Learning

One of the representative problems that arises as artificial intelligence models replace existing technologies is fairness [1]. As a prime example, there are unfair results when the U.S. federal court predicts the possibility of recidivism with a machine learning model. Using demographic data from criminals, the probability of re-offending was calculated against blacks rather than whites. This has several causes; however, we emphasize that biased data was used for model learning, which was most responsible. Historically, the crime rate of black people was much higher, so the artificial intelligence model also judged race as the cause of crime without any filtering. Since then, academia has also been studying ways to create a fair artificial intelligence model for the real-life application.

There are three main ways to create a fair machine learning model. Pre-processing is the first way to find the bias of data before learning and reducing the impact [2,3]. There is a way to balance datasets by creating more data belonging to a particular group, and there is also a methodology that erases the specific attributes. Next, there is in-processing [4-6] to change the optimization method in the learning process, and post-processing is the method to correct the results after learning [7,8]. In this study, only pre-processing method will be addressed to measure the value of the data.

2.2 Private Machine Learning

One of the obstacles to the real-life application of artificial intelligence models is the problem of personal information leakage. For example, creating an artificial intelligence model applicable to the medical field requires patient information as learning data. The data of patients used includes personal information such as anthropometric values, medical history, etc., which causes problems in case of leakage. While data used in learning is not visible when using real-world artificial intelligence models, black-box attacks that use only learned models to infer data have also been developed.

Linkage attack is a method of restoring raw data by estimating common parts using multiple incomplete datasets [9]. This can be classified as part of the larger category of reconstruction attack and there are also attempts to restore the entire learning dataset using public data [14,15]. Membership inference attack is an attack that determines whether a particular data was used to learn an artificial intelligence model [10-12]. Property inference attack focuses on restoring certain properties of learning data rather than restoring the data itself [13,16].

In addition to the leakage of personal information of data, there is also a model training attack that attempts to emulate certain artificial intelligence models [17,18]. This not only lowers the economic value of a particular model but also causes problems that can reveal vulnerabilities in the model. Thus, in order for the actual artificial intelligence model to generate economic value in the market, the fairness and the privacy must be guaranteed prior to the performance of the model.

2.3 Data Valuation

2.3.1 Dataset Price Estimation

Studies estimating the value of a dataset differ as metrics vary [25]. The most extensive survey turned out to be a query-based data pricing. QueryMarket model first presented in [23] overcomes the disadvantages of the inflexible market for buying and selling existing datasets and the difficulty of providing a fixed price to consumers. In this system, the data seller discloses the dataset and the consumer sends queries regarding the data they need. Using those queries, the data seller selects data that only has the information needed by the consumer and delivers it to the consumer. [24] presents a market structure for sellers who buy data multiple times, such as reducing the price of data they already have. Furthermore, [25] presented an additional characteristic of the structure that would prevent consumers from trading profits in the market.

To address the challenges of data sets that depend on the machine learning model, [26] presents a model-based packing (MBP) model. This was designed to add a broker between data sellers and consumers to a market that trades the model itself. We want to solve the optimization problem of maximizing the seller's revenue while also increasing the accessibility of consumers by using consumer demand curves. In this study, the revenue of sellers were further analyzed, presenting a model in which consumers (firms) of the preceding studies could buy the optimal dataset.

2.3.2 Equitable Price Estimation

In artificial intelligence-related fields other than economics, research is actively conducted to measure the contributions of data providers (sellers) rather than market modeling. As a tool measuring the contribution of training datum to boost the performance of a particular artificial intelligence model, [27-29] utilizes the Shapley value from the cooperative game theory. Due to the nature of the Shapley value, time complexity of the approximation algorithm is $O(n!)$ because every permutation of the training data contributed must be considered. Therefore, the purpose of [27-29] is to create an efficient Shapley value approximating algorithm. As in 2.3.1, it is difficult to consider the value of data for a general artificial intelligence model so studies that apply Shapley value to specific models are also being conducted. [30] expresses the reward function as the contribution of each data under the multi-armed bandit problem in reinforcement learning. [31,32] focuses on securing model robustness, efficiency, etc. by applying Shapley value to federated learning which uses multiple machine learning models at once. [33] demonstrates that the exact Shapley value can be found in $O(n \log n)$ time when applied to k-nearest-neighbor algorithm. In this work, we further present and analyze Banzhaf value as an alternative to Shapley value, and correlate the fairness problem of 2.1 with semi-value to analyze how the data contribute to model fairness.

Chapter 3

Data Market Model

3.1 Basic Assumptions and Model Setting

We propose a data market model, in which the whole process from data collection to economic value redistribution will be analyzed. The model consists of three entities: data providers, the firm, and the market. The model is shown in figure 3.1. The model consists of two steps. The first stage is shown in black for the data collection stage and the second stage is in red for the revenue distribution phase.

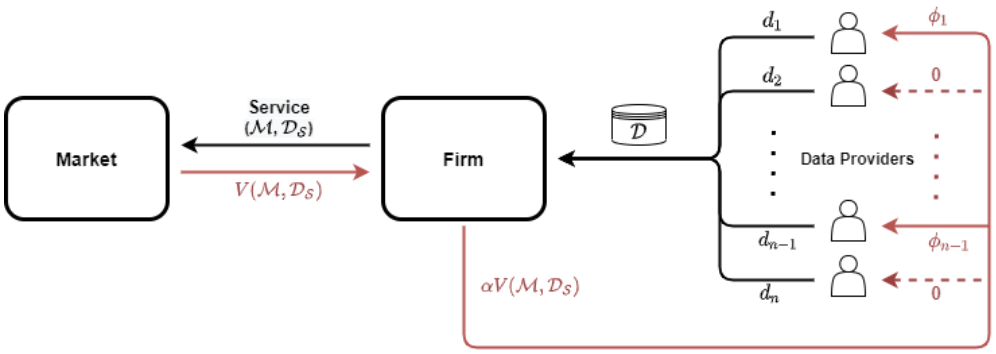

Figure 3.1 | Data market model

In the first stage of data collection, data providers will provide the firm with the dataset $\mathcal{D} := d_1, ..., d_n$. The company selects the optimal dataset $\mathcal{D}_S$ from the dataset $\mathcal{D}$ and

then trains the machine learning model $\mathcal{M}$ with $\mathcal{D}_S$. The firm then provides service to the market with the trained model. The types of services may vary depending on the purpose of the machine learning model. For instance, the firm can provide a language translation model to an IT platform or a model that reads MRI results to a hospital. Currently, most machine learning models are free but GPT-3, a natural language processing model made by openAI, has started to be paid for. In addition, there are many paid services that predict the likelihood of getting a disease by using medical data such as heart attack and obesity [45,46]. This is covered in detail in 3.2.

The second step represents the process of generating and distributing revenue. The firm earns profit $V(\mathcal{M}, \mathcal{D}_S)$ in return for providing service to the market and keeps $1-\alpha$ ($\alpha \in [0,1]$) of the total revenue for itself. The rest, $\alpha V(\mathcal{M}, \mathcal{D}_S)$, is distributed to data providers based on their contribution. The firm also compensates selected data providers consistently for the cost of personal information leakage. This is covered in detail in 3.3.

3.2 Firm's Profit Maximizing Problem

The firm wants to maximize its profit margin by taking the data as input and determining the optimal dataset. If this can be determined, the return on data providers can be calculated, which in turn translates into profitability of the data, or economic value. (3.2) represents an economic model. The company's net profit is defined as Π, the return is defined as $V(\mathcal{M}, \mathcal{D})$, and the cost is defined as $c(\mathcal{M}, \mathcal{D})$.

$$\mathcal{D}_S := \arg\max_{d \subseteq \mathcal{D}} \Pi(\mathcal{M}, \mathcal{D}) \tag{3.1}$$

$$\Pi(\mathcal{M}, \mathcal{D}) = V(\mathcal{M}, \mathcal{D}) - c(\mathcal{M}, \mathcal{D}) \tag{3.2}$$

$$V(\mathcal{M}, \mathcal{D}) = \mathcal{M}_A(\mathcal{D}) + \mathcal{M}_F(\mathcal{D}) \tag{3.3}$$

$$c(\mathcal{M}, \mathcal{D}) = c_F(\mathcal{M}, \mathcal{D}) + c_v(\mathcal{M}, \mathcal{D}) \tag{3.4}$$

$$c_F(\mathcal{M}, \mathcal{D}) = |\mathcal{D}| \cdot k_P \tag{3.5}$$

$$c_v(\mathcal{M}, \mathcal{D}) = \alpha V(\mathcal{M}, \mathcal{D}) \tag{3.6}$$

Performance is the most important factor when the machine learning model service creates economic value from the market [27,28]. The corresponding revenue function for the model performance is $\mathcal{M}_A : D \to \Re$. In order for a machine learning model to succeed in the marketplace, the fairness covered in 2.1 must be met as well as the performance. The fairer the service, the more widely it can be used in the market. Thus, we define an additional profit function regarding fairness, $\mathcal{M}_F : D \to \Re$. This is expressed in (3.3).

The firm's costs are represented by the sum of fixed cost, $c_F(\mathcal{M}, \mathcal{D})$, and variable cost,

$c_V(\mathcal{M}, \mathcal{D})$. 2.2 addresses the privacy of the training data. Therefore, this model assumes that the firm pays k_p to the providers of the optimal datasets for the personal information leakage. It is also assumed that the firm pays the differential cost for the use of the training dataset according to data quality. The sum of these two costs is expressed in (3.4). Finally, (3.1) defines the optimal dataset of the firm profit maximization problem as $\mathcal{D}_S$.

3.3 Data Valuation

This paper evaluates the economic value of data through an analysis of the revenue distribution problem. Data can be adjusted not only for the performance of machine learning models but also for fairness and privacy. In this model, the training data can only improve the performance of the model. Therefore, the more performance-enhancing data for fixed machine learning models and environments, the more economic benefits the data provider must receive. This can be modeled as a transferable utility (TU) game $\Gamma(\mathcal{M}, \mathcal{D})$. Let's define the data, $\mathcal{D} = \{d_1, \ ... \ , \ d_n\}$, as the game participant and $\mathcal{M}: \ 2^{|\mathcal{D}|} \rightarrow \mathbb{R}$ as a characteristic function on coalitions such that $S \subseteq \mathcal{D}$ [34-36]. The characteristic function corresponds to the economic gain from the market. In addition, when $\Gamma_{\mathcal{D}}$ defined as a power set for the entire dataset, the data provider's revenue distribution function $\psi : \Gamma_{\mathcal{D}} \rightarrow \mathbb{R}^{|\mathcal{D}|}$ must satisfy three characteristics [36-41].

[Property 1] Linearity

$$\psi(\mathcal{M} + \mathcal{M}') = \psi(\mathcal{M}) + \psi(\mathcal{M}'), \ \psi(\lambda\mathcal{M}) = \lambda\psi(\mathcal{M}), \ \forall \mathcal{M}, \mathcal{M}' \in \Gamma_{\mathcal{D}}, \ \lambda \in \mathbb{R}$$

[Property 2] Symmetricity

$$\psi_i(\mathcal{M}) = \psi_j(\mathcal{M}), \text{ if } \mathcal{M}(S \cup \{d_i\}) = \mathcal{M}(S \cup \{d_j\}), \ \forall S \subseteq \mathcal{D} \backslash \{d_i, d_j\}$$

[Property 3] Dummy Datum

$$\psi_i(\mathcal{M}) = 0, \text{ if } \mathcal{M}(S \cup \{d_i\}) = \mathcal{M}(S), \ \forall S \subseteq \mathcal{D} \backslash \{d_i\}$$

Property 1 states that the distribution of profits is equitable only when there is linearity for two different characteristic functions. Property 2 guarantees that data i, j should receive the same amount of revenue if the contribution of two data is same when added to any subset of the entire dataset without data i, j. Last property 3 means that data i receives no revenue if there is no contribution of data i when added to any subset of the entire dataset without data i.

The values that meet the above three conditions are defined as semi-value; there are two examples [42-44]. The first is Shapley value and the value for data i is defined as (3.7). Second is Banzhaf value and the value for data i is defined as (3.8).

$$\phi_i(\mathcal{M}) = \sum_{S \subseteq \mathcal{D} \setminus \{d_i\}} \frac{|S|!\ (n-|S|-1)!}{n!} \cdot \left(\mathcal{M}(S \cup \{d_i\}) - \mathcal{M}(S)\right) \tag{3.7}$$

$$\beta_i(\mathcal{M}) = \sum_{S \subseteq \mathcal{D} \setminus \{d_i\}} \frac{1}{2^{n-1}} \cdot \left(\mathcal{M}(S \cup \{d_i\}) - \mathcal{M}(S)\right) \tag{3.8}$$

In this work, we implement both Shapley and Banzhaf value when measuring the contribution of data to model performance. However, Banzhaf value does not consider the order in which data is added unlike Shapley value. It is not necessary to consider all permutations because a datum is used multiple times when training machine learning models. Therefore, an experiment to propose and compare Banzhaf value instead of the Shapley value considered by the studies in 2.3.2 is subsequently conducted. This is described in detail in 4.1.

3.4 Binary Classification Setting

In this section, to reduce the time complexity of the semi-value approximation in 3.3, we will limit it to the case of the binary classification to obtain and analyze the optimal solution. In this case, the performance of the model is simply defined as classification accuracy. In addition, machine learning fairness can be easily defined. On the other hand, solving the firm's profit optimization problem in 3.2 requires the computation over all subsets of a set of n data providers. Therefore, it becomes an NP-hard problem with the complexity of $O(2^n)$ so that we add several assumptions to obtain the explicit solution of this optimization problem.

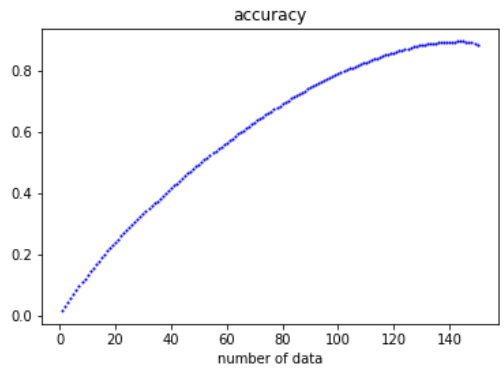

Figure 3.2 | model accuracy with *n* highest semi-value data

First, to avoid the NP-hard problem, assume that the firm selects *n* data in the order in which it has the highest semi-value contributions, *i.e.,* $|\mathcal{D}_S| = n$. In this case, as shown in figure 3.2, the more data is extracted, the less accuracy is improved. If we use the whole dataset, the accuracy begins to diminish, which suggests that the use of faulty data is damaging the

model's performance. These data have negative semi-value and the company has no reason to use them. Consequently, the revenue from the model performance follows diminishing marginal utility law.

Furthermore, the more diverse the contributions of the data, the curve will be concave and if all data contributions are the same, it would be a straight line. Therefore, after estimating this concave monotonic increasing function, assume that it is the firm's model performance revenue function $\mathcal{M}_A(\mathcal{D})$. k_A was set below 0.1 under the assumption which we need more than 20 data and wants minimum accuracy of 0.9.

$$\mathcal{M}_A(D) = 1 - \exp(-k_A n), \quad k_A \in [0, 0.1] \tag{3.9}$$

Moreover, assume that the proportion of the two classes of data is $q \in [0, 0.5]$ to define the revenue function for machine learning fairness. For the rest of the cases, symmetry can be used. Then, when *n* data is selected, the number of data for each class is each qn, (1-q)n . Assuming that the less data difference between classes results in the fairer model, the fairness function is defined as follows. The range of k_F is assumed to be less than $\frac{1}{n}$ in accordance with $\mathcal{M}_A(\mathcal{D})$ having the maximum value of 1. That is, the firm's revenue is assumed to be positive.

$$\mathcal{M}_F(D) = -k_F(1-2q)n, \quad k_F \in \left[0, \frac{1}{n}\right] \tag{3.10}$$

Thus, the optimization problem in 3.2 is defined as follows, which the explicit optimal solution can be easily found. In addition, suppose that if the firm keeps all the revenue from the market itself ($\mathcal{M}_A(\mathcal{D}) = 1,\ \alpha = 0$) and use perfectly fair data (q=0.5), the firm's revenue is positive. In other words, assume that $k_P \in \left[0, \frac{1}{n}\right]$. In this case, explicit optimization solutions are derived in (3.11).

$$\begin{aligned} N^* &= \text{argmax}_n\ (1\text{-}\alpha)\left(\mathcal{M}_A(D) + \mathcal{M}_F(D)\right) - k_P n \\ &= \frac{1}{k_A} log \frac{(1-\alpha)k_A}{k_P + (1-\alpha)(1-2q)k_F} \end{aligned} \tag{3.11}$$

Chapter 4

Analysis

4.1 Semi-value Approximation

4.1.1 Convergence Analysis

This section will compare Shapley and Banzhaf value, the most widely used semi-values covered in 3.3, using Monte-Carlo approximation method. While there are many papers studying the approximation method of Shapley value, there is no study of whether Banzhaf value is well approximated in vast machine learning data. First, we create 150 cluster data represented in figure 4.1 below and artificially insert 10 incorrect data. These are purple dots in a yellow cluster or vice versa.

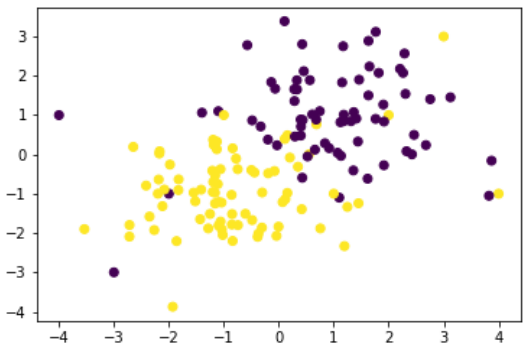

Figure 4.1 | synthetic data

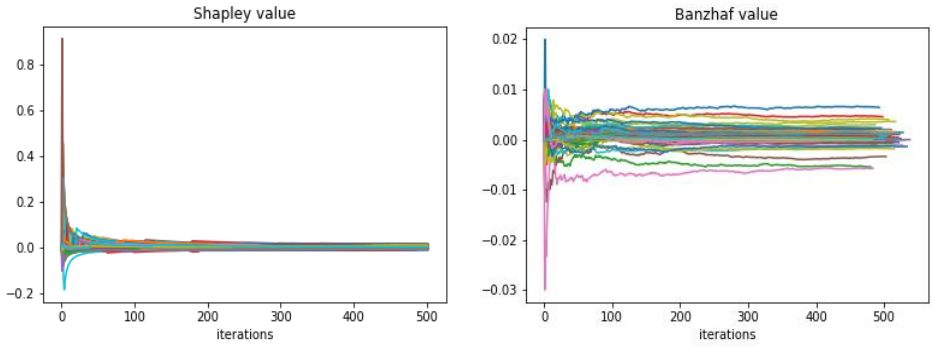

Figure 4.2 | semi-value approximation

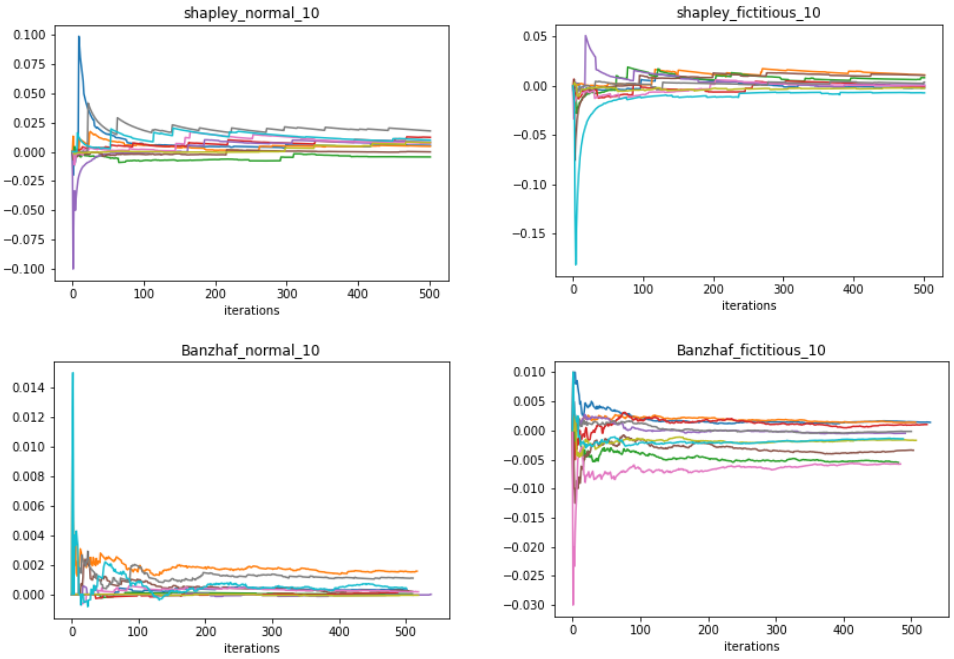

Figure 4.3 | normal/fictitious data semi-value approximation

Figure 4.2 shows the approximation results of Shapley and Banzhaf value under binary classification problem setting using RBF kernel SVM(support vector machine) on 150 data after 500 iterations. Both values seem to converge well due to the large variability in the beginning but in figure 4.3, where only 10 data are actually drawn, we can see that Banzhaf value is more convergent. In addition, abnormal data exhibits lower semi-value than the normal data and most of them are negative. In the same way, the results of several experiments were summarized in table 4.1 below, examining both cases of regression and classification. Using the given data sets available in the sklearn package, the semi-value values were considered to converge when there was no error greater than 0.0005. Results that are not significantly different from the previous analysis can be identified. In particular, in the case of regression, the Shapley value did not converge.

		Time(s)		Iteration	
		Shapley	Banzhaf	Shapley	Banzhaf
Classification	Data_blobs	12.523	2.518	153	61
	Data_moons	22.112	1.096	265	27
	Make_hastie_10_2	19.409	4.528	161	82
Regression	Load_boston	-	4.514	-	98
	Load_iris	-	0.475	-	21

Table 4.1 | semi-value convergence

4.1.2 Group Data Calculation

4.1.1 approximates the semi-value of each datum and analyzes its convergence. This is a time-consuming task even with simple regression and classification problems with low-dimension data. Therefore, studies such as [27,31,32] also measure semi-value for data groups. Figure 4.4 assumes a situation in which four people have combined their own datasets forming data_blobs(), data_moons() provided by the Sklearn package. In other words, we identified the extent to which a dataset, not single datum, contributed to machine learning performance, and the results are shown in figure 4.5.

For each dataset, we solved the kernel SVM classification problem and utilized both linear and RBF kernel. In these cases, the exact semi-value be calculated. The two semi-values exhibited similar behavior which the four data providers were ranked in the same order. For the blobs() dataset, green data ranked from 1st to 4th as the kernel changed and the moons() dataset showed that the order of importance was randomly reversed depending on the kernel. That is to say that both values are highly sensitive. It can also be seen that the two values make little difference. It can be interpreted that the approximation of the Shapley value is due to poor convergence in 4.1.1 and that Banzhaf value can also be used instead of Shapley value for data groups.

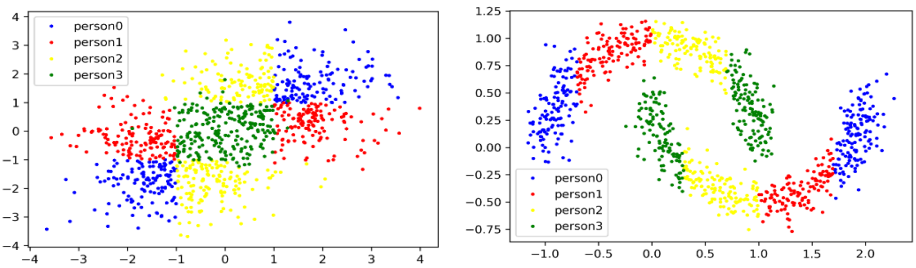

Figure 4.4 | group synthetic data

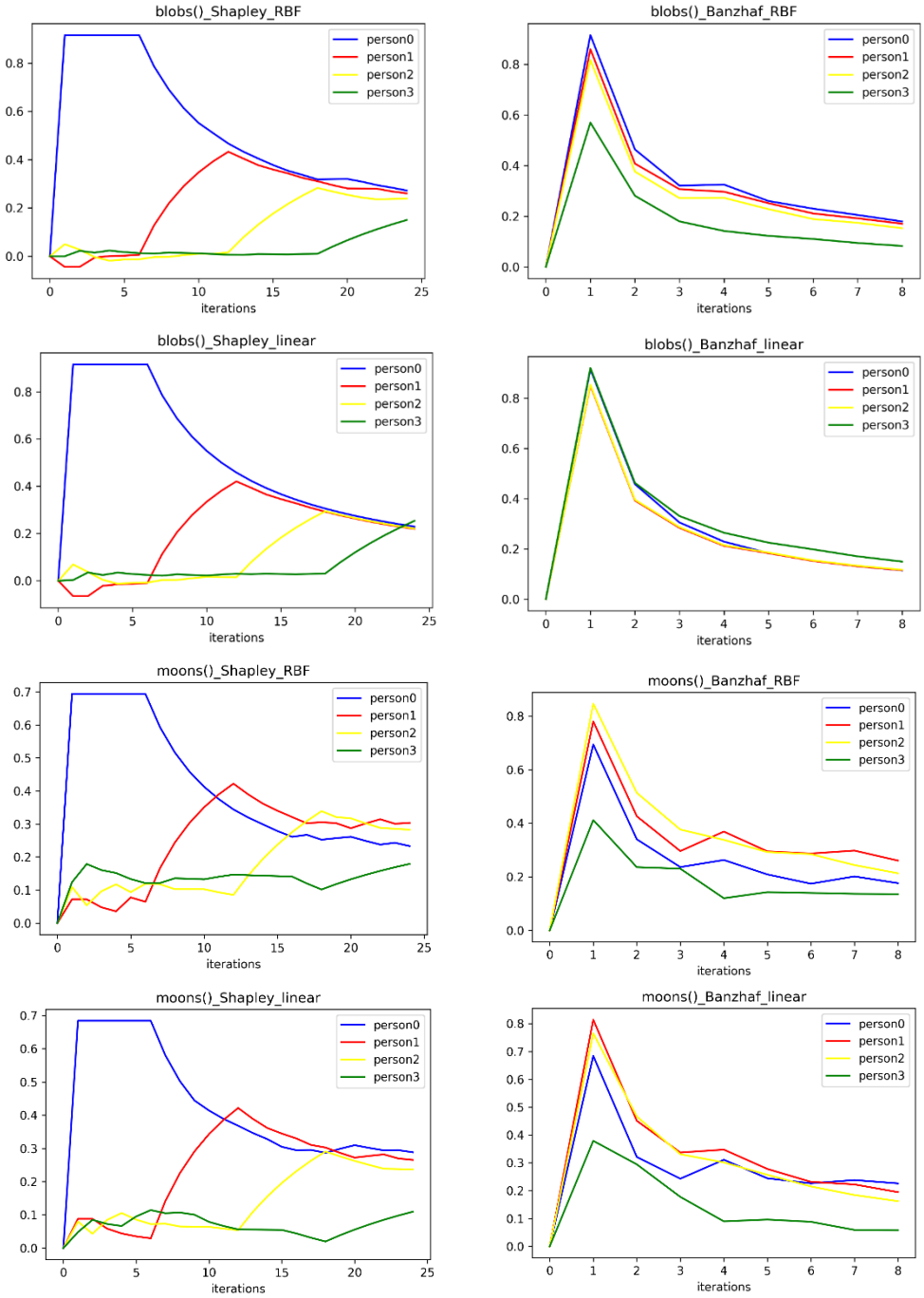

Figure 4.5 | semi-value calculation in group data

4.2 Binary Classification

4.2.1 Parameter Analysis

Under the assumption in 3.4, we will analyze changes in the optimal solution for the four variables using (3.8).

Proposition 1. For fixed k_A, k_P, k_F, q, if $\alpha \le 1 - \frac{k_P}{k_A - (1-2q)k_F}$, then $\frac{\partial N^*}{\partial \alpha} \le 0$.

Otherwise, $N^* = 0$.

Corollary 1.1. If $k_A \ge \frac{k_P}{1-\alpha} + (1-2q)k_F$, $\frac{\partial N^*}{\partial k_P} \le 0$, $\frac{\partial N^*}{\partial k_F} \le 0$, $\frac{\partial N^*}{\partial q} \ge 0$ are satisfied. Otherwise, $N^* = 0$.

The proposition 1 states that if the firm does not take more than a certain proportion of the revenue, the company has no incentive to collect data and train the machine learning model. That is, the company can define an upper bound on the percentage of revenue it can distribute to data providers. Also, data providers can require the maximum amount α of revenue from the firm. When the firm takes more than the threshold α, the larger the ratio, the more data it selects.

Furthermore, the greater the compensation for data providers' personal information, the smaller the number of data is drawn. Similarly, the fairer the data, or the less market revenue for data fairness, the more data the firm tends to select.

Proposition 2. For fixed α, k_P, k_F, q, if $k_A^* = e \cdot \left(\frac{k_P}{1-\alpha} + (1-2q)k_F \right)$,

$N^* = \frac{1}{k_A^*}$ is the optimal solution.

The greater the proportion of important data, *i.e.*, high semi-value data, k_A increases. Proposition 2 states that the firm picks more data as the contribution of the data are indifferent, *i.e.*, k_A is small. The firm can conclude that it is always optimal to collect only datasets that can produce a constant model performance of $\mathcal{M}_A^*(\mathcal{D}) = 1 - \frac{1}{e}$.

4.2.2 Parameter Analysis

4.2.2.1 Description

This section uses numerical examples to establish scenario analysis. Scenarios considered in this study can be summarized in two ways. The first is the fairness of the data. The second is whether the data provider requires high privacy compensation or not. In this scenario, the market assumes that services are impossible unless fairness precedes them. Based on these two criteria, we construct the following four scenarios. For each scenario, the firm obtains an upper bound α on the revenue proportion for the data provider and a range of *n* which the number of data to be selected by the firm.

Scenario 1: Fair data and a high privacy compensation level

$$\left(q = 0.5,\ k_P = \frac{1}{N},\ k_F = \frac{1}{N} \right)$$

Scenario 2: Unfair data and a high privacy compensation level

$$\left(q = 0.4,\ k_P = \frac{1}{N},\ k_F = \frac{1}{N} \right)$$

Scenario 3: Fair data and a low privacy compensation level

$$\left(q = 0.5,\ k_P = \frac{1}{2N},\ k_F = \frac{1}{N} \right)$$

Scenario 4: Unfair data and a low privacy compensation level

$$\left(q = 0.4,\ k_P = \frac{1}{2N},\ k_F = \frac{1}{N} \right)$$

4.2.2.1 Synthetic Data

This section uses the blobs(), moons() dataset provided in the sklearn package. Two datasets are shown in figure 4.6 and 4.7. Each class was represented by different shape, a circle and a triangle, and the color represents fairness. Figure 4.6 and 4.7 each exhibits unfair and fair dataset. The training and test dataset each have 1000 points. Classification is performed with RBF kernel SVM. Shapley and Banzhaf value approximation were done with 500 and 1,000 iterations, respectively.

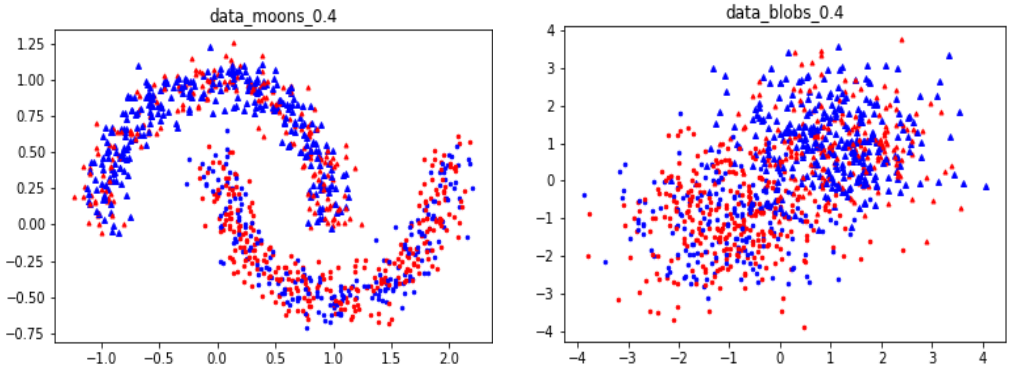

Figure 4.6 | unfair datasets

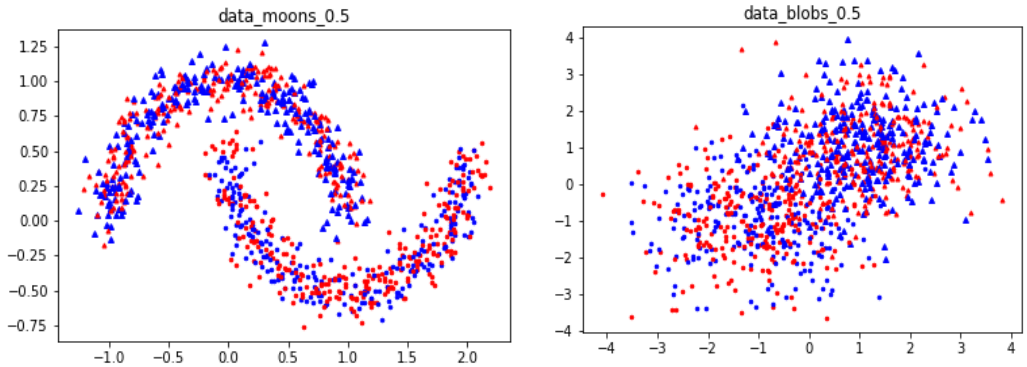

Figure 4.7 | fair datasets

4.2.2.3 Shapley value based valuation

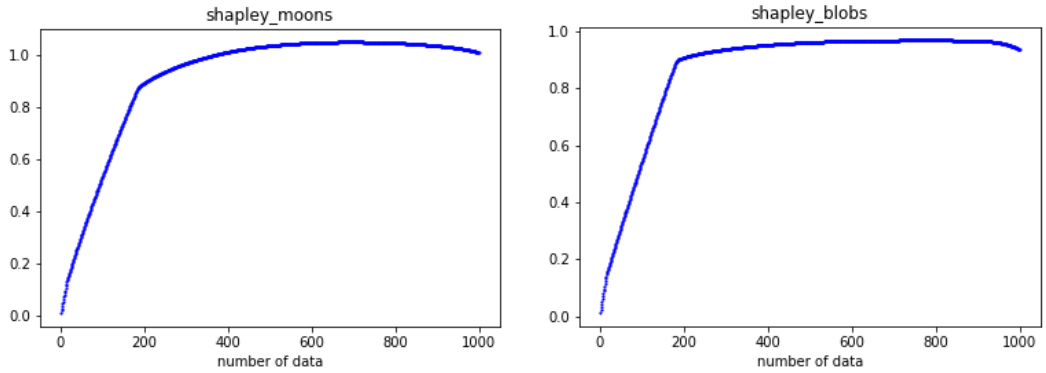

Figure 4.8 | model accuracy with *n* highest Shapley value data in unfair dataset

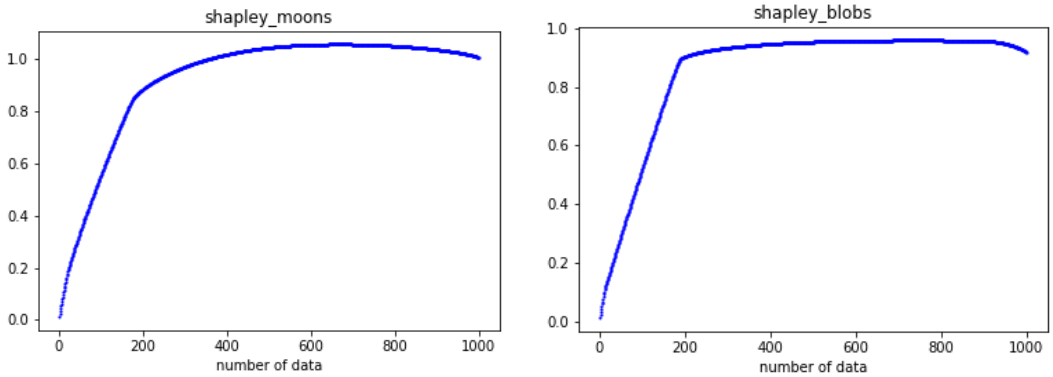

Figure 4.9 | model accuracy with *n* highest Shapley value data in fair dataset

Cumulative accuracy graphs according to Shapley value for each of the two datasets are shown in figure 4.8 and 4.9. Figure 4.8 shows that we only need 200 data out of the 1000 data points in order to achieve the accuracy of 0.9 in both datasets. The cumulative accuracy

value of the fair dataset shown in figure 4.9 is almost the same as the unfair dataset. Therefore, we set k_A to 0.0102 and 0.0115, respectively. The summarized scenario-specific result is in table 4.2.

Scenario	Threshold α		Maximum N^*	
	moons()	blobs()	moons()	blobs()
1	0.902	0.913	228	212
2	0.9	0.912	210	197
3	0.951	0.957	296	273
4	0.95	0.956	263	243

Table 4.2 | Shapley value based scenarios

4.2.2.4 Banzhaf value based valuation

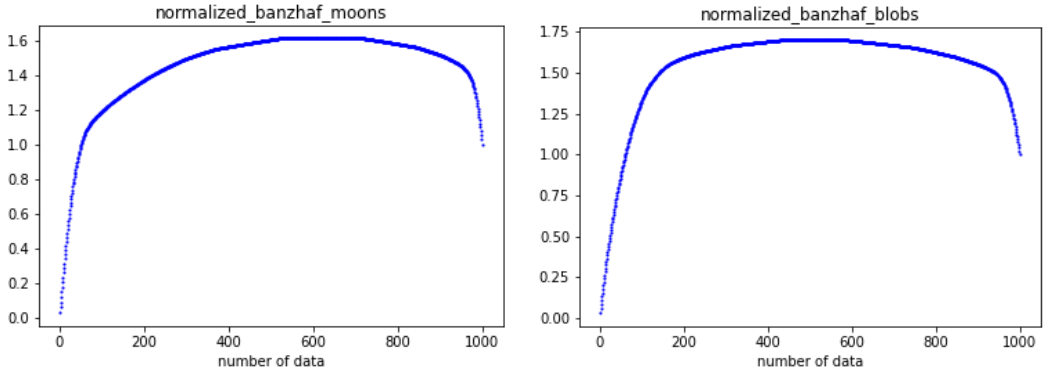

Figure 4.10 | model accuracy with *n* highest Banzhaf value data in unfair dataset

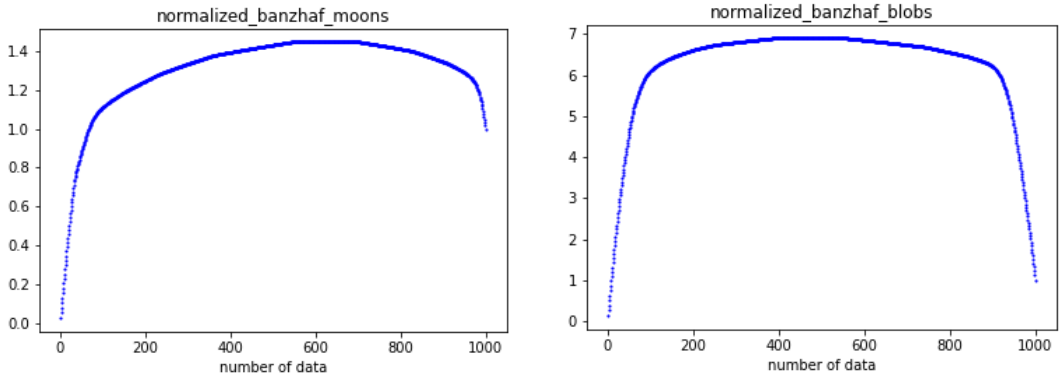

Figure 4.11 | model accuracy with *n* highest Banzhaf value data in fair dataset

The cumulative accuracy graph according to the Banzhaf value for each of the two data sets can be found in figure 4.10 and 4.11. We normalized Banzhaf value in order to make a fair comparison with Shapley value. Consequently, k_A is set to 0.0145, 0.0136 for the unfair moons(), blobs() dataset and 0.0171, 0.0195, respectively, for the fair dataset. The summarized scenario-specific result is in table 4.3.

Scenario	Threshold α		Maximum N^*	
	moons()	blobs()	moons()	blobs()
1	0.942	0.949	166	152
2	0.93	0.925	172	179
3	0.971	0.974	207	188
4	0.965	0.963	209	218

Table 4.3 | Banzhaf value based scenarios

4.2.2.5 Comparative Analysis

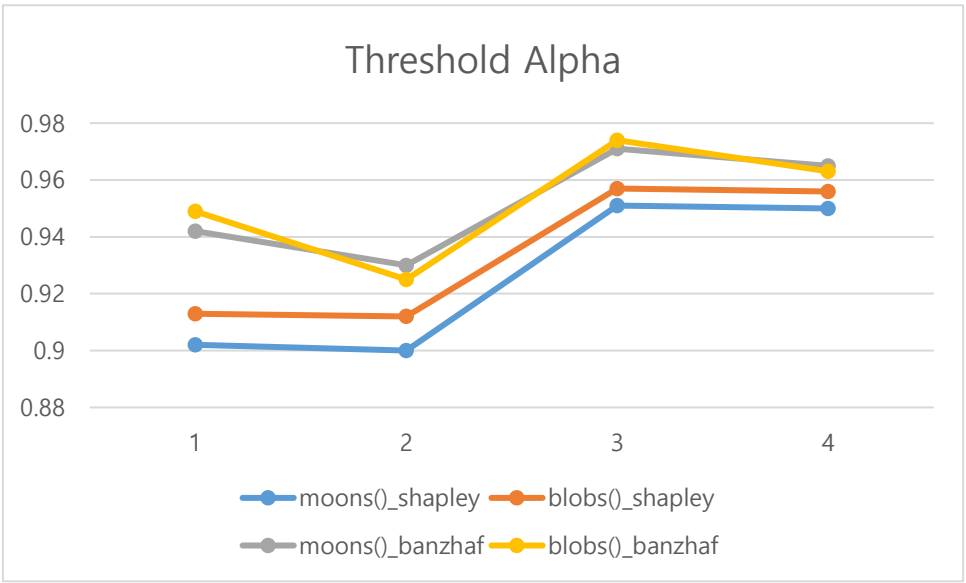

Figure 4.12 | threshold α

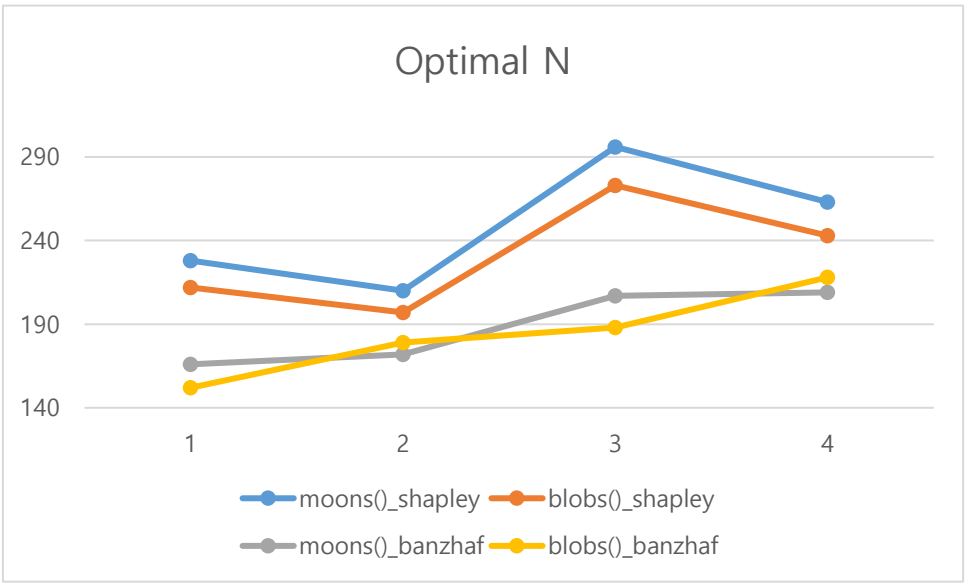

Figure 4.13 | optimal N^*

Scenario-specific results are shown in figure 4.12 and 4.13. If both datasets are fair, the firm distributes a large proportion of revenue to the data provider. That is, the fairer their dataset, data providers can also request greater allocation in terms of revenue distribution. Furthermore, companies tend to select fair data, which can be interpreted as the unfairer data they select, unfairer the artificial intelligence model used in the market becomes. However, the results using Banzhaf value showed a tendency to extract more data if it is unfair, although the data similarly distributes less revenue to the data provider. This can be interpreted as a reason that Banzhaf value puts emphasis on certain data so even if the firm selects more data, it does not have a significant impact on the firm profit.

Moreover, scenarios 1 and 2 pay twice the data provider's privacy fixed costs compared to scenario 3 and 4. As a result, the firm can distribute more revenue to its data providers in scenario 3 and 4 and utilize more data. This brings in the same results for both Shapley and Banzhaf value. However, Shapley value was more affected by the fairness of the data than by fixed costs while Banzhaf value showed similar results between using unfair data and reducing fixed costs.

With the use of Shapley value, the fairness of the data did not have a significant impact on the firm's decisions. This is because Shapley value has not changed much in figure 4.8 and 4.9. However, for Banzhaf value, there was a noticeable change between fair and unfair dataset. Comparison of figure 4.10 and 4.11 shows that the number of data that is not used increases in fair data, *i.e.*, only small portion of the whole dataset can boost the performance of machine learning models. This can be interpreted as accelerating the process of eliminating the impact of sensitive attributes on fair datasets. Thus, as shown in figure 4.12 and 4.13, Banzhaf value-based results tend to pick less data and instead reward data providers more than in the case of Shapley value. This can be interpreted as the use of

Banzhaf value allows data providers to demand a higher rate of revenue sharing. In other words, using Banzhaf value rather than Shapley value is a favorable condition for data providers.

4.3 Data Pricing

This section will produce the final price of dataset by selecting representative data points from two datasets under the assumptions of the binary classification model. We will use the fair moons() and blobs() datasets in figure 4.7 and draw five sample data each to produce the relative price of 10 data. Figure 4.14 and 4.15 show the samples taken from the two datasets in red.

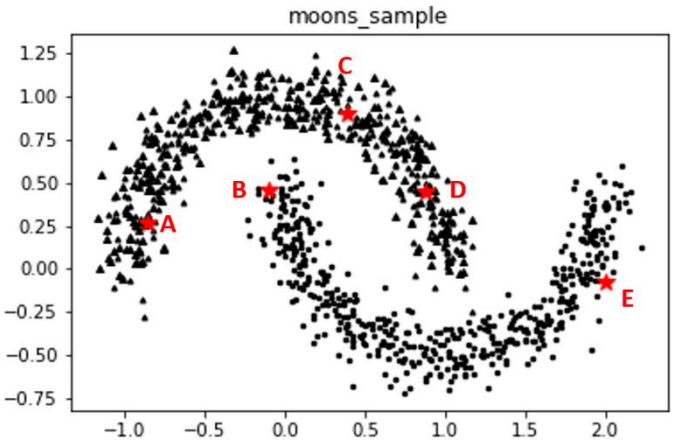

Figure 4.14 | moons() dataset samples

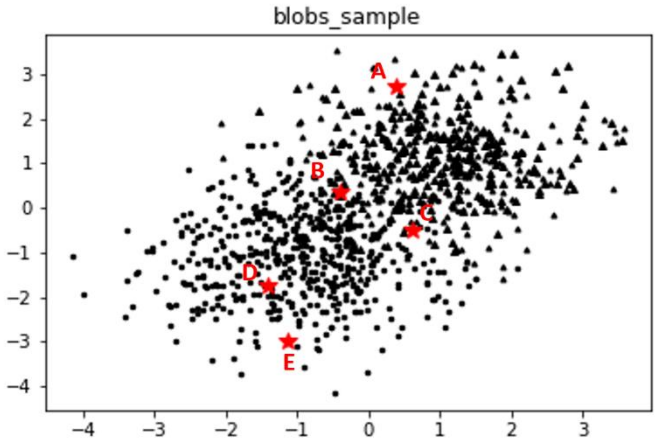

Figure 4.15 | blobs() dataset samples

Additional assumptions are needed to determine the exact relative price through numerical analysis. First, we use the Shapley value. The samples A, B, C, D, E in figure 4.14 have Shapley values of -1.681e-4, -6.283e-5, 8.495e-3, 5.816e-4, 8.366e-3, respectively. Figure 4.15 shows values of 6.939e-5, 9.798e-3, 7.037e-4, -9.925e-6 and 1.022e-5. Second, set the alpha as 0.5. This was established under the assumption that both the firm and data providers share equitably because the threshold alpha analyzed in 4.2.2 exceeded 0.9. Finally, suppose that it is a scenario 1 situation in which data providers want a high level of privacy.

The relative price of 10 data can be explicitly obtained under the above assumptions. First, among the data A, B, C, D and E in figure 4.14, only C and E are selected and each data can be converted into economic values of 5.031e-3, 4.97e-3. It can be interpreted that this is about 0.5% of the market's revenue and that it is highly valuable because the firm uses only about 200 data of the 1000 data. For the five sample data in figure 4.15, only B and C are selected by the firm and the remaining data are not selected, resulting in a value of zero. B and C has economic values of 5.46e-3 and 1.32e-3, respectively, 0.5% and 0.1% of the total market revenue.

Chapter 5

Conclusion

In this paper, we model two markets at once to estimate the economic value of artificial intelligence related data. The first is the data market where the firm acts as a consumer and data providers as a seller. Next, the market and the firm play the role of consumers and sellers in the market where the service from trained machine learning model is traded. Together, the entire market was modeled using three entities: data providers, the firm and the market. There are three main contributions to this study.

First, we showed that the Banzhaf value is a promising alternative to the Shapley value by performing the convergence analysis. The reason why Shapley value required a heavy computation is that all of the permutations were considered. Due to the nature of the data there is no need to implement such a scheme because single datum is used many times in the training of a machine learning model. Therefore, we propose and analyze Banzhaf value with less computational complexity while maintaining important properties. In addition, Banzhaf value is not normalized so we used regularized Banzhaf value while performing comparative analysis with Shapley value. The analysis shows that Banzhaf value has better convergence. In both regression and classification tasks, Banzhaf value converged faster than Shapley value. Furthermore, we present the possibility of Banzhaf value replacing Shapley value in the case of finding the exact values of the actual grouped data, not only Monte Carlo approximation.

Secondly, we proposed an overall data market model that takes fairness and privacy of data

into account. This was modeled as a structure that guarantees greater fairness in the data, increases total profit in the market and rewards data providers with the risk of personal information leakage from the data. We also found the explicit optimal solution in the case of binary classification task with numerical examples and constructed four scenarios for further analysis. Scenario-specific analysis result show that Banzhaf value is more sensitive to Shapley value with respect to data fairness and assigning data value less evenly than Shapley value. We can infer that the instability of Shapley value method is the cause for the observed phenomenon.

In addition, while compensation for fair dataset tended to be greater than unfair dataset, there was no correlation with the number of data that the firm actually selects. In other words, as a result of adding the factors which induce the firm to collect fair data, the firm reduced the compensation for the data providers but continued to increase the number of data used until a specific performance was achieved. From the data provider's point of view, it may be better to collect high-quality data, even if it is unfair, rather than fair but low-quality data.

Finally, through the explicit optimization under the assumption of a binary classification model, we can assign monetary value of each datum. Only four out of ten samples were selected by the firm and were valued at approximately 0.5% of the total market revenue, according to the analysis. That is, it can be interpreted that under this research model, the firm selects high-quality data and provides a large reward for those data. Therefore, this model shows that data providers have no choice but to make efforts to obtain better quality data.

In this study, many assumptions were used to model the market for general artificial intelligence models. There are many ways in which this research can be developed. First,

we can analyze the convergence of semi-value and behavior of the model by using real-world large datasets such as CIFAR and COMPAS. Moreover, we can develop data fairness guarantee in the model if existing fairness measures, such as equalizedods, were used. Lastly, using game and market theory to further analyze the optimization of the data provider's perspective and the market consumer curve will provide a better equilibrium. This will be reflected in further research to improve the practicality of this model and ensure theoretical verification.

Bibliography

[1] Hardt, Moritz., *How Big Data Is Unfair*, https://medium.com/@mrtz/how-big-data-is-unfair-9aa544d739de, 2014.

[2] d'Alessandro, Brian, Cathy O'Neil, and Tom LaGatta, *Conscientious classification: A data scientist's guide to discrimination-aware classification*, Big data, 5 (2017), pp. 120-134.

[3] Zemel, Rich, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork, *Learning fair representations,* In International Conference on Machine Learning, 2013, pp. 325-333.

[4] Dwork, Cynthia, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel, *Fairness through awareness,* In Proceedings of the 3rd innovations in theoretical computer science conference, 2012, pp. 214-226.

[5] Agarwal, Alekh, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach, *A reductions approach to fair classification,* arXiv preprint arXiv:1803.02453 (2018).

[6] Madras, David, Elliot Creager, Toniann Pitassi, and Richard Zemel, *Learning adversarially fair and transferable representations*, arXiv preprint arXiv:1802.06309 (2018).

[7] Kim, Michael P., Amirata Ghorbani, and James Zou, *Multiaccuracy: Black-box post-processing for fairness in classification*, In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, 2019, pp. 247-254.

[8] Awasthi, Pranjal, Matthäus Kleindessner, and Jamie Morgenstern, *Equalized odds postprocessing under imperfect group information*, In International Conference on Artificial Intelligence and Statistics, 2020, pp. 1770-1780.

[9] Vatsalan, Dinusha, Peter Christen, Christine M. O'Keefe, and Vassilios S. Verykios, *An evaluation framework for privacy-preserving record linkage*, Journal of Privacy and Confidentiality 6, 1 (2014).

[10] Shokri, Reza, Marco Stronati, Congzheng Song, and Vitaly Shmatikov, *Membership inference attacks against machine learning models*, In 2017 IEEE Symposium on Security and Privacy, 2017, pp. 3-18.

[11] Nasr, Milad, Reza Shokri, and Amir Houmansadr, *Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning*, In 2019 IEEE Symposium on Security and Privacy, 2019, pp. 739-753.

[12] Salem, Ahmed, Yang Zhang, Mathias Humbert, Pascal Berrang, Mario Fritz, and Michael Backes, *Ml-leaks: Model and data independent membership inference attacks and defenses on machine learning models*, arXiv preprint arXiv:1806.01246 (2018).

[13] Fire, Michael, Gilad Katz, Lior Rokach, and Yuval Elovici, *Links reconstruction attack*, In Security and Privacy in Social Networks, (2013), pp. 181-196.

[14] Lacharité, Marie-Sarah, Brice Minaud, and Kenneth G. Paterson, *Improved reconstruction attacks on encrypted data using range query leakage*, In 2018 IEEE Symposium on Security and Privacy, 2018, pp. 297-314.

[15] Ganju, Karan, Qi Wang, Wei Yang, Carl A. Gunter, and Nikita Borisov, *Property inference attacks on fully connected neural networks using permutation invariant representations*, In Proceedings of the 2018 ACM SIGSAC Conference on Computer and Communications Security, 2018, pp. 619-633.

[16] Naveed, Muhammad, Seny Kamara, and Charles V. Wright, *Inference attacks on property-preserving encrypted databases,* In Proceedings of the 22nd ACM SIGSAC

Conference on Computer and Communications Security, 2015, pp. 644-655.

[17] Tramèr, Florian, Fan Zhang, Ari Juels, Michael K. Reiter, and Thomas Ristenpart, *Stealing machine learning models via prediction apis*, In 25th {USENIX} Security Symposium ({USENIX} Security 16), 2016, pp. 601-618.

[18] Juuti, Mika, Sebastian Szyller, Samuel Marchal, and N. Asokan, *PRADA: protecting against DNN model stealing attacks*, In 2019 IEEE European Symposium on Security and Privacy (EuroS&P), 2019, pp. 512-527.

[19] https://www.bdex.com/

[20] https://datastreamgroup.com/

[21] https://infutor.com/

[22] https://selectstar.ai/

[23] Koutris, Paraschos, Prasang Upadhyaya, Magdalena Balazinska, Bill Howe, and Dan Suciu, *Querymarket demonstration: Pricing for online data markets*, Proceedings of the VLDB Endowment, 5 (2012), pp. 1962-1965.

[24] Koutris, Paraschos, Prasang Upadhyaya, Magdalena Balazinska, Bill Howe, and Dan Suciu, *Toward practical query pricing with QueryMarket*, In proceedings of the 2013 ACM SIGMOD international conference on management of data, (2013), pp. 613-624.

[25] Lin, Bing-Rong, and Daniel Kifer, *On arbitrage-free pricing for general data queries,* Proceedings of the VLDB Endowment, 7 (2014), pp. 757-768.

[25] Roth, Aaron, *Technical Perspective: Pricing information (and its implications*), Communications of the ACM 60, 12 (2017), pp. 78.

[26] Chen, Lingjiao, Paraschos Koutris, and Arun Kumar, *Towards model-based pricing for machine learning in a data marketplace*, In Proceedings of the 2019 International Conference on Management of Data, 2019, pp. 1535-1552.

[27] Ghorbani, Amirata, and James Zou, *Data shapley: Equitable valuation of data for machine learning*, arXiv preprint arXiv:1904.02868 (2019).

[28] Jia, Ruoxi, David Dao, Boxin Wang, Frances Ann Hubis, Nick Hynes, Nezihe Merve Gurel, Bo Li, Ce Zhang, Dawn Song, and Costas Spanos, *Towards efficient data valuation based on the shapley value*, arXiv preprint arXiv:1902.10275 (2019).

[29] Kwon, Yongchan, Manuel A. Rivas, and James Zou, *Efficient computation and analysis of distributional Shapley values*, arXiv preprint arXiv:2007.01357 (2020).

[30] Yoon, Jinsung, Sercan O. Arik, and Tomas Pfister, *Data Valuation using Reinforcement Learning*, arXiv preprint arXiv:1909.11671 (2019).

[31] Wang, Tianhao, Johannes Rausch, Ce Zhang, Ruoxi Jia, and Dawn Song, *A Principled Approach to Data Valuation for Federated Learning*, arXiv preprint arXiv:2009.06192 (2020).

[32] Yu, Han, Zelei Liu, Yang Liu, Tianjian Chen, Mingshu Cong, Xi Weng, Dusit Niyato, and Qiang Yang, *A fairness-aware incentive scheme for federated learning*, In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, 2020, pp. 393-399.

[33] Jia, Ruoxi, David Dao, Boxin Wang, Frances Ann Hubis, Nezihe Merve Gurel, Bo Li, Ce Zhang, Costas J. Spanos, and Dawn Song, *Efficient task-specific data valuation for nearest neighbor algorithms*, arXiv preprint arXiv:1908.08619 (2019).

[34] Nash, John, *Non-cooperative games*, Annals of mathematics, (1951), pp. 286-295.

[35] Peleg, Bezalel, and Peter Sudhölter, *Introduction to the theory of cooperative games*, vol. 34, Springer Science & Business Media, 2007

[36] Bachrach, Yoram, Evangelos Markakis, Ariel D. Procaccia, Jeffrey S. Rosenschein, and Amin Saberi, *Approximating power indices*, In Proceedings of the 7th international joint conference on Autonomous agents and multiagent systems, 2

(2008), pp. 943-950.

[37] Packel, Edward W., and John Deegan, *An axiomated family of power indices for simple n-person games*, In Power, Voting, and Voting Power, Physica, Heidelberg, 1981, pp. 213-221.

[38] Amer, Rafael, and José Miguel Giménez, *An axiomatic characterization for regular semivalues*, Mathematical Social Sciences, 51 (2006), pp. 217-226.

[39] Peters, Hans, *Game theory: a multi-leveled approach*, Springer, 2015.

[40] Kurz, Sascha, *The power of the largest player*, Economics Letters, 168 (2018), pp. 123-126.

[41] Matsui, Tomomi, and Yasuko Matsui, *A survey of algorithms for calculating power indices of weighted majority games*, Journal of the Operations Research Society of Japan, 43 (2000), pp. 71-86.

[42] Haimanko, Ori, *The axiom of equivalence to individual power and the Banzhaf index*, Games and Economic Behavior, 108 (2018), pp. 391-400.

[43] Dubey, Pradeep, and Lloyd S. Shapley, *Mathematical properties of the Banzhaf power index*, Mathematics of Operations Research, 4 (1979), pp. 99-131.

[44] Deegan, John, and Edward W. Packel, *A new index of power for simplen-person games*, International Journal of Game Theory, 7 (1978), pp. 113-123.

[45] Lundervold, Alexander Selvikvåg, and Arvid Lundervold, *An overview of deep learning in medical imaging focusing on MRI*, Zeitschrift für Medizinische Physik, 29 (2019), pp. 102-127.

[46] Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan et al, *Language models are few-shot learners*, arXiv preprint arXiv:2005.14165 (2020).

국문초록

기계학습이 현재 이론과 실생활 적용 모두에서 발전함에 따라 데이터는 인공지능 모델을 훈련하고 검증하는 데 중요한 역할을 하고 있다. 한편, 데이터 교환 시장에서 데이터의 경제성 평가에 대한 연구는 초기 단계이다. 본 논문의 기여는 두 가지 관점에서 접근할 수 있다. 첫째, 협동 게임 이론의 개념인 semi-value를 모델 수익 분배 문제에 활용한다. 둘째, 인공지능 모델의 공정성과 개인정보보호성을 고려한 데이터 제공자, 기업, 시장으로 구성된 모델을 제안한다. 본 연구에서 Banzhaf 값은 각 데이터의 기여도를 계산할 때 Shapley 값의 대안이 될 수 있음을 확인하였다. 또한 회사의 수익 극대화 문제를 모델링하였고, 추가적으로 데이터 예제를 사용하여 이진 분류 모델의 경우 수치 분석을 제시하였다. 이를 통해, Banzhaf 값은 Shapley 값보다 데이터의 공정성에 더 민감하다는 것을 확인하였다. 나아가 기업이 고품질 데이터만을 사용한다는 가정하에 데이터의 공정성과 데이터 제공자의 개인정보에 대한 보상비용을 달리하는 네 가지 시나리오에서 기업의 행동을 분석하였다. 기업은 데이터가 공정할수록 데이터 제공자에게 더 큰 수익을 보장해주었고, 고정비용이 작아질수록 가변비용을 통해서 데이터 제공자에게 수익을 나눠주는 것을 확인하였다.

주요어: 기계학습, 협동게임이론, semi-value, 데이터 가치, 공정성, 개인정보보호성

학번: 2019-22469