Data valuation

UNIVERSIDAD COMPLUTENSE DE MADRID
UNIVERSIDAD POLITÉCNICA DE MADRID

TRABAJO FINAL DE MÁSTER

14 de julio de 2023

Autor: Francisco AGUILAR MARTÍNEZ
Tutores: Carlos GREGORIO RODRÍGUEZ
Miguel DE BENITO DELGADO

UNIVERSIDAD
COMPLUTENSE
MADRID

Data valuation

UNIVERSIDAD COMPLUTENSE DE MADRID
UNIVERSIDAD POLITÉCNICA DE MADRID

TRABAJO FINAL DE MÁSTER

14 de julio de 2023

Autor: Francisco AGUILAR MARTÍNEZ
Tutores: Carlos GREGORIO RODRÍGUEZ
Miguel DE BENITO DELGADO

Índice general

CAPÍTULO 1

Estado del arte

Desde la aparición del valor de Shapley como un método de reparto justo de recompensas en juegos cooperativos [1], este concepto se ha utilizado en diversos campos como la economía [2], estudio de sistemas multiagente [3] e incluso en áreas en las que su aplicación puede resultar menos evidente como la genética [4]. Esta versatilidad se debe, en parte, a su sólida base matemática y a sus intuitivas interpretaciónes, entre las que podemos resaltar:

- Pago justo: El valor de Shapley de un jugador es la cantidad que este debería recibir si las recompensas se distribuyesen de manera que los jugadores fueran recompensados en función de su contribución a la recompensa total.
- Poder de negociación: El valor de Shapley puede interpretarse como una medida del poder de negociación de un jugador. Un jugador tiene más poder de negociación si su ausencia causa una mayor disminución en la recompensa total que se puede obtener.

El valor de Shapley se basa en una serie de axiomas fundamentales, que garantizan propiedades como su equidad, eficiencia y simetría. En economía relajar estos axiomas, con el objetivo de dar lugar a nuevas formas de reparto de recomenpensa, ha sido uno de los principales temas de estudio. Ejemplos de esto serían el concepto de semivalor, el cual se obtiene al eliminar el axioma de eficiencia [5, 6]. Este axioma asegura que la suma de los valores de todos los jugadores sea igual a la recompensa total disponible. Por tanto, al suprimirlo, los semivalores permiten cierta flexibilidad en este aspecto, lo que puede ser útil en situaciones en las que no todos los beneficios se pueden distribuir. Del mismo modo, eliminar el axioma de simetría, que establece que dos jugadores con igual contribución deben recibir igual recompensa, lleva al valor de Banzhaf [7], que proporciona una medida de poder de un jugador basada en cuánto puede cambiar el resultado de un juego al unirse o abandonar una coalición. Cabe destacar que tanto el valor de Shapley como el valor de Banzhaf pueden obtenerse como particularizaciones del concepto de semivalor.

Debido a la naturaleza combinatoria del valor de Shapley, el cálculo de este es altamente costoso a nivel computacional y resulta en una tarea cuya complejidad crece exponencialmente al aumentar el número de jugadores. Es por esto que surgen métodos de estimación del mismo, la mayoría de estos métodos se basan en técnias de Montecarlo. En 1960, Irwin Mann y el propio Shapley mencionan las estimaciones basada en muestreo de permutaciones [8]. Pero no es hasta 2015 que se lleva a cabo un análisis de la complejidad a la hora de muestrear usando dicha técnica [9]. Tras esto, en 2019, Covert propone un nuevo método de estimación basado en la técnica de muestreo por importancia que mejora los actuales [10]. Cabe destacar el *ApproShapley* propuesto en [11] desarrollado por los profesores J.Castro, D.Gómez y J.Tejada de la UCM.

Una de las primeras apariciones del valor de Shapley en el campo del aprendizaje computacional data del año 2005 como un método de selección de variables [12]. Más tarde, en 2017, se utiliza en el diseño del marco SHAP [13], enfocado en la evaluación de la importancia de variables en modelos de predicción. Sin embargo, no es asta el año 2019 en el que se introduce como una alternativa a los métodos de valoración de datos del momento [14], acuñándose así el concepto *Data Shapley*.

Al igual que el cálculo del valor de Shapley, el cálculo de *Data Shapley* es altamente costoso a nivel computacional, por lo que surgen también varios métodos de aproximación, entre los que podemos destacar *Group Testing* [15], métodos de aproximación y cálculo exacto para problemas en los que se aplican métodos como KNN o derivados de este [16] y diversas técnicas basadas en métodos de Montecarlo como las vistas en [14].

Cuando se prueba la eficacia de *Data Shapley* en problemas de aprendizaje computacional como la detección de outliers o datos corruptos [14], la investigación sigue el mismo camino que años antes en teoría de juegos y se empiezan a reciclar conceptos como los semivalores o el valor de Banzhaf. Es en la línea de los semivalores que en 2022 surge *Beta Shapley* [17], una generalización de *data Shapley* que supera los resultados de los métodos más actuales de valoración de datos en varias tareas como son detección de muestras mal etiquetadas y selección de puntos problemáticos a la hora de entrenar un modelo.

En 2023, aparece *data BanzHaf* [18], un nuevo método de valoración de datos derivado del valor de Banzhaf. Este nuevo método surge como una solución a la falta de robustez de las herramientas de valoración de datos existentes, falta de robustez causada en parte por factores difíciles de controlar como la aleatoriedad del método del descenso del gradiente estocástico, el cual es ampliamente usado hoy en día. Para solventar esta falta de robustez se apoya en el concepto de *Safety Margin*, y demuestra que el valor de Banzhaf es el semivalor con mayor *Safety Margin*. *Data Banzhaf* supera a los existentes métodos de valoración basados en semivalores en varias tareas de aprendizaje automático.

Aunque en este trabajo nos centramos en métodos de valoración de datos que se derivan directamente de conceptos de la teoría de juegos, existen otros métodos que, aún utilizando teoría de juegos, siguen enfoques relativamente distintos. Podemos destacar algunas obras como [19] en la que sugieren un método de valoración de datos para modelos generativos que utiliza la discrepancia media máxima (MMD) entre la fuente de datos y la distribución real de datos. En [20] proponen una medida de diversidad, llamada volumen robusto (RV), para valorar las fuentes de datos. La robustez de RV se discute en términos de la estabilidad frente a la replicación de datos. Finalmente en [21] utilizan diferencias estadísticas entre los datos de origen y un conjunto de datos de referencia como la métrica de valoración. Estas diferencias estadísticas se miden mediante el uso de los conceptos de diversidad y relevancia de los datos previamente comentados.

Como algo casi anecdótico hay literatura en la que se lleva a cabo valoración de datos mediante aprendizaje por refuerzo [22]. En dicho trabajo, se utiliza una red neuronal profunda para obtener un estimador de la propabiliadad de cada dato de ser usado en el entrenamiendo del modelo de predicción. Este estimador se obtiene mediante aprendizaje por refuerzo.

Bibliografía.

[1] Lloyd S Shapley «A Value for n-Person Games», 1953 ed. por Harold W. Kuhn y Albert W. Tucker, págs. 307-317.

[2] Alvin E Roth «*The Shapley value: essays in honor of Lloyd S. Shapley*» Cambridge University Press, 1988.

[3] Shaheen S. Fatima, Michael Wooldridge y Nicholas R. Jennings «A linear approximation method for the Shapley value» *Artificial Intelligence* 172.14, 2008, págs. 1673-1699.

[4] Stefano Moretti, Vito Fragnelli, Fioravante Patrone y Stefano Bonassi «Using coalitional games on biological networks to measure centrality and power of genes» *Bioinformatics (Oxford, England)* 26, 2010, págs. 2721-30.

[5] Pradeep Dubey y Robert J. Weber *Probabilistic Values for Games* Cowles Foundation Discussion Papers 471 Cowles Foundation for Research in Economics, Yale University, 1977.

[6] Pradeep Dubey, Abraham Neyman y Robert James Weber «Value Theory without Efficiency» *Mathematics of Operations Research* 6.1, 1981, págs. 122-128.

[7] J.F. Banzhaf «Weighted voting doesn't work: A mathematical analysis» *Rutgers Law Review* 19.2, 1965, págs. 317-343.

[8] Irwin Mann y Lloyd S. Shapley «Values of Large Games, IV: Evaluating the Electoral College by Montecarlo Techniques», 1960.

[9] Sasan Maleki «Addressing the computational issues of the Shapley value with applications in the smart grid» Tesis doct. University of Southampton, ago. de 2015.

[10] Ian C Covert, Scott M Lundberg y Su-In Lee «Shapley feature utility» *IEEE Transactions on Information Theory*, 2019.

[11] Javier Castro, Daniel Gómez y Juan Tejada «Polynomial calculation of the Shapley value based on sampling» *Computers & Operations Research* 36.5, 2009 Selected papers presented at the Tenth International Symposium on Locational Decisions (ISOLDE X), págs. 1726-1730.

[12] Shay Cohen, Eytan Ruppin y Gideon Dror «Feature selection based on the shapley value» *other words* 1.98Eqr, 2005, pág. 155.

[13] Scott Lundberg y Su-In Lee «A Unified Approach to Interpreting Model Predictions», 2017 eprint: arXiv:1705.07874.

[14] Amirata Ghorbani y James Zou «Data Shapley: Equitable Valuation of Data for Machine Learning», 2019 eprint: arXiv:1904.02868.

[15] Ruoxi Jia et al. «Scalability vs. Utility: Do We Have to Sacrifice One for the Other in Data Importance Quantification?», 2019.

[16] Ruoxi Jia et al. «Efficient Task-Specific Data Valuation for Nearest Neighbor Algorithms» *Proc. VLDB Endow.* 12.11, 2019, págs. 1610-1623.

[17] Yongchan Kwon y James Zou «Beta Shapley: a Unified and Noise-reduced Data Valuation Framework for Machine Learning», 2021 eprint: arXiv:2110.14049.

[18] Jiachen T. Wang y Ruoxi Jia «Data Banzhaf: A Robust Data Valuation Framework for Machine Learning», 2022 eprint: arXiv:2205.15466.

[19] Sebastian Shenghong Tay, Xinyi Xu, Chuan Sheng Foo y Bryan Kian Hsiang Low «Incentivizing Collaboration in Machine Learning via Synthetic Data Rewards», 2021 arXiv: 2112.09327 [cs.LG].

[20] Xinyi Xu, Zhaoxuan Wu, Chuan Sheng Foo y Bryan Kian Hsiang Low «Validation Free and Replication Robust Volume-based Data Valuation» 34, 2021 ed. por M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang y J. Wortman Vaughan, págs. 10837-10848.

[21] Mohammad Mohammadi Amiri, Frederic Berdoz y Ramesh Raskar «Fundamentals of Task-Agnostic Data Valuation», 2022 arXiv: 2208.12354 [cs.LG].

[22] Jinsung Yoon, Sercan Arik y Tomas Pfister «Data Valuation using Reinforcement Learning» Proceedings of Machine Learning Research 119, 2020 ed. por Hal Daumé III y Aarti Singh, págs. 10842-10851.