УДК 519.688

АППРОКСИМАЦИЯ ВЕКТОРА ШЕПЛИ ДЛЯ ОЦЕНКИ ЗНАЧИМОСТИ ДАННЫХ

В.А. Ерофеева

Институт проблем машиноведения РАН

Россия, 199178, Санкт-Петербург, Большой проспект В.О., 61

E-mail: eva@ipme.ru

С.Э. Парсегов

Институт проблем управления им. В.А. Трапезникова РАН

Россия, 117997, Москва, Профсоюзная ул., 65

E-mail: s.e.parsegov@gmail.com

Ключевые слова: рынок данных, вектор Шепли, детерминированная аппроксимация, опознание со сжатием, агрегаторы данных.

Аннотация: На сегодняшний день многие практические приложения опираются на методы оптимизации и обучения, основанные на данных. С одной стороны, объем собранных данных постоянно растет. С другой стороны, владельцы данных не заинтересованы в их совместном использовании по соображениям конфиденциальности или коммерческим причинам. Правильные стимулы могут положительно повлиять на предпочтения конечных пользователей в отношении обмена данными. Таким образом, очень важно оценить вклад различных провайдеров («поставщиков») данных и распределить прибыль, полученную от использования общих данных, чтобы создать такие стимулы. Традиционным подходом к справедливому распределению является метод Шепли. Однако, несмотря на свои полезные свойства, этот метод требует огромного количества вычислений, поскольку число узлов совместного использования данных стремится к бесконечности. Вклад данной статьи заключается в оптимизированном подходе на основе метода опознания со сжатием (англ. compressed sensing).

1. Введение

В последние десятилетия мы наблюдаем тенденцию к сбору огромных объемов данных, генерируемых разнородными источниками, например, сотовыми телефонами, интеллектуальными счетчиками, электромобилями и т. д. В сочетании с достижениями в области машинного обучения эта тенденция превращает данные в ценный товар для большинства предприятий и отраслей. Обмен и анализ данных приносит огромные социальные и экономические выгоды во многих областях, таких как электронная коммерция, медицина, исследование операций, энергетические системы и т. д. Повсеместное использование анализа данных приводит к развитию рынков данных. В недавних обзорах освещается современное состояние дел в

области разработки рынков данных и подчеркивается важность исследований по оценке их качества (см., например, [5]).

Механизм распределения прибыли должен адекватно оценивать данные, предоставляемые различными продавцами данных, и вознаграждать их в соответствии с полезностью наборов данных для решения проблемы принятия решений. В литературе существует множество методов, посвященных проблеме распределения прибыли. Среди них широко используются концепции теории кооперативных игр. Большинство из них обеспечивает справедливость решения. Распространенные методы распределения учитывают предельные затраты (например, стоимость по Шепли), разделяемые и неразделяемые затраты (например, методы равных затрат, альтернативных избегаемых затрат и разрыва в затратах) или минимизируют наибольшую неудовлетворенность [6]. Метод Шепли – это широко используемый метод распределения прибыли. Он направлен на справедливое распределение прибыли на основе среднего маржинального вклада каждого участника (продавца данных). Маржинальные вклады оцениваются для всех уникальных комбинаций участников, называемых коалициями. Основное вычислительное узкое место метода связано с комбинаторным характером процесса создания коалиций и оценки каждой коалиции.

Среди описанных выше методов Шепли наиболее широко используется в работах, посвященных проектированию рынков данных [1]. Для преодоления его вычислительных проблем, исследователи предлагают аппроксимационные методы, которые позволяют получить приближенные решения за разумное время. Следуя этому направлению исследований, мы сосредоточимся на методах аппроксимации, подходящих для рынков данных.

2. Постановка задачи

Метод Шепли, предложенный в [9], стал классической концепцией в теории кооперативных игр для разделения общей прибыли. Он заключается в следующем. Рассмотрим множество игроков (узлов) $\mathcal{N} = \{1, 2, \ldots, n\}$. Определим кооперативную игру в виде пары, состоящей из множества игроков и характеристической функции $v(\cdot)$: $\mathcal{Q}_{\mathcal{N}} = \{v : 2^{\mathcal{N}} \to \mathbb{R} \mid v(\varnothing) = 0\}$.

По факту, кооперативная игра – это функция стоимости, определенная на подмножествах $\mathcal{N}$. Каждое подмножество $S \subseteq \mathbb{N}$ представляет собой коалицию игроков, а $v(S)$ обозначает ценность, приписываемую коалиции S при условии, что нулевая коалиция $\varnothing$ получает нулевую ценность.

Таким образом, значение Шепли для узла i – это его вклад в достижение общей прибыли, взвешенный и просуммированный по всем возможным комбинациям:

$$\phi_i(v) = \sum_{S \subseteq \mathcal{N} \setminus \{i\}} \frac{|S|!(n - |S| - 1)!}{n!} \Big[v(S \cup \{i\}) - v(S) \Big]. \tag{1}$$

Несмотря на то, что метод Шепли обеспечивает справедливое распределение прибыли, он не обладает масштабируемостью и является вычислительно трудоемким. В связи с этим, были предложены различные техники аппроксимации, среди которых опознание со сжатием, используемое в рассматриваемой статье [3, 7]. Учитывая область применения, требуется построить детерминированный алгоритм

с возможностью воспроизводимости найденных значений при одних и тех же наборах данных, что не всегда возможно при использовании существующих рандомизированных подходов.

3. Аппроксимация вектора Шепли

Введение в опознание со сжатием.

Рассмотрим сжимаемый сигнал $\mathbf{x} \in \mathbb{R}^d$, который имеет разреженное представление $\mathbf{s} \in \mathbb{R}^d$ (содержащее в основном нули) в разреживающем базисе $\Psi \in \mathbb{R}^{d\times d}$:

$$\mathbf{x} = \Psi\mathbf{s}. \tag{2}$$

Вектор $\mathbf{s}$ называется K-разреженным в Ψ, если он содержит ровно K ненулевых элементов. Когда процесс измерения вычислительно затратен и/или частично невозможен, целью является разработка матрицы измерений, которая уменьшает их количество и дает возможность восстановить $\mathbf{x}$. В этом контексте, измерения $\mathbf{y} \in \mathbb{R}^p$ определяются как

$$\mathbf{y} = C\mathbf{x}, \tag{3}$$

где $K < p << d$, $C \in \mathbb{R}^{p\times d}$ – матрица измерений.

Знание Ψ и $\mathbf{s}$ дает возможность восстановить исходный вектор $\mathbf{x}$. Таким образом, комбинируя (2) и (3), мы формулируем постановку задачи опознания со сжатием: найти разреженный вектор $\mathbf{s}$, который согласуется с измерениями

$$\mathbf{y} = C\Psi\mathbf{s} = \Theta\mathbf{s}. \tag{4}$$

Искомый разреженный вектор можно найти решив задачу ℓ_1-оптимизации.

Аппроксимация посредством опознания со сжатием и QR разложения.

Преобразуем (1) в матричную форму. Пусть $\mathcal{S}_i = \{S_{i,1}, \ldots, S_{i,m}\}$ – множество, состоящее из всех возможных комбинаций (коалиций) для игрока i таких, что $\forall j \in \{1, \ldots, m\}$: $S_{i,j} \subseteq \mathcal{N} \setminus \{i\}$. Тогда, для каждого игрока и коалиции мы определяем вес $w_{i,j}$ и маржинальный вклад $u_{i,j}$ игрока i в коалицию $S_{i,j}$ как

$$w_{i,j} = \frac{|S_{i,j}|!(n - |S_{i,j}| - 1)!}{n!}, \;\; u_{i,j} = v(S_{i,j} \cup \{i\}) - v(S_{i,j}). \tag{5}$$

Объединяя $w_{i,j}$ и $u_{i,j}$, получаем

$$\phi_i(v, \mathcal{S}_i) = \mathbf{w}^{\mathrm{T}}\mathbf{u}_i, \tag{6}$$

где $\mathbf{w} = [w_1, \ldots, w_m]^{\mathrm{T}}$, $\mathbf{u}_i = [u_{i,1}, \ldots, u_{i,m}]^{\mathrm{T}}$. Предположим, что $\mathbf{u}_i$ сжимаема в некотором базисе Ψ. Аналогично (2), получим $\mathbf{u}_i = \Psi\mathbf{s}_i$.

Если матрица C плотная (заполненная), то она все равно захватывает большое количество векторных компонент и не позволяет избежать вычислительной проблемы, связанной с вычислением маржинальных вкладов. Введем новую матрицу $B \in \mathbb{R}^{l\times m}$ и выразим измерения в виде

$$\mathbf{y}_i = B\mathbf{u}_i = B\Psi\mathbf{s}_i, \tag{7}$$

где $B = [\mathbf{e}_{\gamma_1}, \ldots, \mathbf{e}_{\gamma_l}]$. Здесь и далее $\mathbf{e}_i \in \mathbb{R}^m$ – это канонический базисный вектор с единицей на выбранном индексе i и нулями в других местах, $\gamma = \{\gamma_1, \ldots, \gamma_l\} \subset \{1, \ldots, m\}$. Матрица B дает разреженное представление и задает компоненты $\mathbf{u}_i$, которые мы должны измерить. Авторы [2] предложили использовать QR-разложение для получения оптимальных компонент вектора γ. Объединяя подход к аппроксимации посредством опознания со сжатием и построения матрицы B посредством QR-разложения, мы приходим к предлагаемому алгоритму.

Мы получаем форму измерения (7) из QR-разложения с перестановками:

$$(VP)^{\mathrm{T}} = R^{\mathrm{T}}Q^{\mathrm{T}} = P^{\mathrm{T}}\Psi = B\Psi, \quad B = P^{\mathrm{T}}_{l\times m}, \tag{8}$$

где Q – ортогональная матрица, R – верхнетреугольная матрица, а P – матрица перестановки столбцов.

QR-CS аппроксимация вектора Шепли.

Предлагаемый алгоритм сводится к выполнению следующих шагов для каждого $i \in \mathcal{N}$:

1. Входные данные: Ψ, число измерений l, множество игроков $\mathcal{N}$, точность ϵ
2. Сгенерировать $\mathcal{S}_i$ и вычислить $\mathbf{w}$ используя (5)
3. Сформировать матрицу B используя (8)
4. Вычислить l маржинальных вкладов: $\mathbf{y}_i = B\mathbf{u}_i$
5. Решить задачу оптимизации: $\widehat{\mathbf{s}}_i \leftarrow \arg\min_{\mathbf{s}_i} \|\mathbf{s}_i\|_1 \quad \text{s.t. } \|\Theta\mathbf{s}_i - \mathbf{y}_i\|^2 < \epsilon$
6. Вычислить искомое значение: $\widehat{\mathbf{u}}_i = \Psi Q\widehat{\mathbf{s}}_i$
7. Оценить значение Шепли для игрока i используя (6): $\widehat{v}_i = \mathbf{w}^{\mathrm{T}}\widehat{\mathbf{u}}_i$
8. Сформировать вектор распределения прибыли $\widehat{\mathbf{v}} = [\widehat{v}_1, \ldots, \widehat{v}_n]^{\mathrm{T}}$

4. Моделирование

Секция демонстрирует эффективность предложенного алгоритма в соответствии с существующей рандомизированной техникой, опубликованной в [8]. Для сравнения используется следующая метрика. Распределение игрока i определяется как доля в процентах, которая должна быть взята из общей суммы. Доли всех игроков в сумме составляют 100 процентов. В качестве референсного значения выбрано распределение, рассчитанное точным аналитическим методом Шепли.

В качестве примера взята эталонная игра, представленная в [4]. Пусть $\mathcal{Q}_{\mathbb{N}}$ – игра «Аэропорт», где $\mathbb{N} = \{1, \ldots, 20\}$ и характеристическая функция $v(\cdot)$ определена следующим образом:

$$v(S) = \max_{i\in S} c_i, \quad \forall S \subseteq \mathbb{N}, \quad \mathbf{c} = [c_1, \ldots, c_{20}]^{\top},$$
$$\mathbf{c} = [\underbrace{1}_{2\text{ раза}}, \underbrace{2}_{3\text{ раза}}, 3, \underbrace{4}_{4\text{ раза}}, \underbrace{5}_{2\text{ раза}}, \underbrace{6}_{2\text{ раза}}, \underbrace{7}_{3\text{ раза}}, \underbrace{8}_{4\text{ раза}}]^{\top}.$$

На рисунке 1 показана ошибка распределения методами QR-CS и Random CS в игре «Аэропорт». Здесь мы видим, что ошибка распределения QR-CS остается в пределах $\pm 2,5$ процентов для каждого игрока. Усредненная ошибка распределения метода Random CS лежит в интервале $[-10\%, 5\%]$.

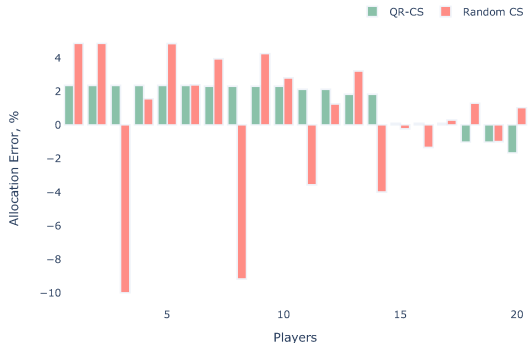

Рис. 1. Сравнение методов в игре «Аэропорт»

5. Заключение

В этой статье изучена проблема распределения прибыли между узлами совместного доступа к данным. Традиционный подход, основанный на значении Шепли, дает справедливое распределение, но требует большого количества вычислений. Предложенный в статье метод обеспечивает более точную аппроксимацию и однородный уровень ошибок при использовании детерминированной техники, которая необходима для воспроизводимости результатов на рынках данных.

Исследование выполнено при поддержке Российского научного фонда, грант №22-71-00072, https://rscf.ru/project/22-71-00072/.

Список литературы

1. Agarwal A., Dahleh M., Sarkar T. A marketplace for data: An algorithmic solution // Proceedings of the 2019 ACM Conference on Economics and Computation. 2019. P. 701–726.
2. Brunton S., Kutz N. Data-driven science and engineering: Machine learning, dynamical systems, and control. Cambridge University Press, 2019.
3. Candes E., Wakin M. An introduction to compressive sampling // IEEE signal processing magazine. 2008. Vol. 25, No. 2. P. 21–30.
4. Castro J., Gomez D., Tejada J. Polynomial calculation of the shapley value based on sampling // Computers & Operations Research. 2009. Vol. 36, No. 5. P. 1726–1730.
5. Driessen S., Monsieur G., Van Den Heuvel W. Data market design: a systematic literature review // IEEE access. 2022. Vol. 10. P. 33123–33153.
6. Gao E., Sowlati T., Akhtari S. Profit allocation in collaborative bioenergy and biofuel supply chains // Energy. 2019. Vol. 188. P. 116013.
7. Granichin O., Pavlenko D. Randomization of data acquisition and l1-optimization (recognition with compression) // Automation and Remote Control. 2010. Vol. 71. P. 2259–2282.
8. Jia R., Dao D., Wang B., Hubis F., Hynes N., Gurel N., Li B., Zhang C., Song D., Spanos C. Towards efficient data valuation based on the shapley value // 22nd International Conference on Artificial Intelligence and Statistics. 2019. P. 1167–1176.
9. Shapley L. A value for n-person games. Princeton University Press, 2016.