Применение анализа формальных понятий для исследования нейросетей

A Preprint

Черемискин Егор Андреевич
ММП ВМК МГУ
egorcheremiskin@yandex.ru

Гуров Сергей Исаевич
ММП ВМК МГУ
sgur@cs.msu.ru

2023

Abstract

Нейронные сети достигают высоких результатов в разных задачах машинного обучения, при этом зачастую возникает вопрос их интерпретируемости. В данной работе мы постараемся попробовать исследовать нейронные сети, используя анализ формальных понятий. Мы рассмотрим "схожесть"нейросетей и попробуем сравнить их, а также попробуем построить иерархию классов в задаче классификации. Это поможет нам лучше понимать работу нейросетей, а также правильно выбирать подходящую архитектуру нейросети для той или иной задачи.

Keywords Анализ формальных понятий

1 Введение

Применить анализ формальных понятий для исследования или улучшения работы алгоритмов машинного обучения уже пробовали другие исследователи. В работе [Hanika and Hirth, 2020] анализ формальных понятий используется для масштабирования данных. В работе [Bernhard Ganter and Hirth, 2023] исследуется вопрос масштабирования данных с помощью АФП и применение масштабирования для улучшения работы древовидных классификаторов. В работах [Hanika and Hirth, 2023] и [Amirata Ghorbani and Kim, 2019] авторы используют АФП для исследования интепретируемости работы древовидных классификаторов. В работе [Hanika and Hirth, 2021] АФП используется для уменьшения размерности данных. В работе [Hirth and Hanika, 2022] и [Fong and Vedaldi, 2018] авторы применяют АФП для исследования нейросетей и их интерпретируемости. В работе [Julius Adebayo and Kim, 2020] АФП используется для изучения изображений и влияние входных изображений на выход нейросети. В работе [Jiayuan Mao and Wu, 2019] АФП используется для изучения нейросетей, который используются для распознавания изображений.

2 Постановка задачи

Рассмотрим нейронные сети как сложные функции от объектов обучающей выборки. Тогда, вводя некоторую метрику и используя АФП, появляется возможность сравнивать нейронные сети. Для сравнения нейронных сетей были взяты веса последнего линейного выходного слоя нейронной сети и активации нейронной сети на объектах обучающей выборки. В качестве метрики можно взять расстояние Громова-Вассерштейна. Для построения иерархии классов также будем использовать веса линейного выходного слоя и некоторый порог, в нашем случае - нулевой.

Рассмотрим задачу классификации, где каждый признак объектов является бинарным. Любой небинарный признак можно свести к нескольким бинарным признакам: для категориального признака достаточно завести для каждой категории или для некоторых множеств категорий новый признак, для

вещественного признака достаточно ввести порог, или же разбить вещественную прямую на отрезки, где каждому отрезку будет соответствовать свой новый бинарный признак.

Пусть G - множество объектов, M - множество бинарных признаков, а $I \subseteq G \times M$ - отношение, которое показывает, какие объекты какими признаками обладают. Тогда назовем тройку (G, M, I) контекстом. Введем операторы Галуа для подмножеств этих множеств:

$$A' = \{\ m \in M \mid gIm\ \forall g \in A\}$$

$$B' = \{\ g \in G \mid gIm\ \forall m \in B\}$$

Если $A' = B$ и $B' = A$, то (A, B) называется формальным понятием, где $A \in G, B \in M$. Множество A называется формальным объемом, а множество B - формальным содержанием.

Если для двух понятий (A_1, B_1) и (A_2, B_2) выполняется $A_1 \subseteq A_2$ (или $B_2 \subseteq B_1$, что аналогично), то (A_1, B_1) является подпонятием (A_2, B_2). Тогда на множестве понятий можно ввести отношение $(A_1, B_1) \leq (A_2, B_2)$. Упорядоченное множество всех формальных понятий является решеткой формальных понятий [Ganter and Wille, 1999].

Пусть теперь у нас признаки являются не бинарными, а категориальными, то есть могут принимать несколько отдельных значений. Пусть W - множество всевозможных значений, которые могут принимать признаки из M. Тогда назовем четверку (G, M, W, I) многозначным контекстом, где $I \subseteq G \times M \times W$, причем $(g, m, w) \in I$ и $(g, m, v) \in I$ тогда и только тогда, когда $w = v$.

Для того, чтобы построить формальные понятия для многозначного контекста, необходимо применить масштабирование: по некоторым правилам построить новый контекст $S_m = (G_m, M_m, I_m)$ для некоторого признака m. В качестве правила может служить, например, пороговое значение для признака m, тем самым признак m станет бинарным и по нему можно будет построить новый контекст [Ganter and Wille, 1989].

Рассмотрим некоторую нейронную сеть, решающую задачу бинарной классификации. У такой нейронной сети на выходе появляется вектор $[0, 1]^k$, где k - число классов. На ее вход подается объект g из обучающей выборки G размера n. Признаковое описание объекта имеет размер m: $g \in \mathbb{R}^m$. Представим всю нейронную сеть, кроме последнего выходного слоя, как вектор функций размера h $N = \{n_1, ..., n_h\}$, где $n_i : \mathbb{R}^m \to \mathbb{R}$ для каждой функции из вектора N. Последний выходной слой нейронной сети представим как вектор функций $C = \{c_1, ..., c_k\}$, где $c_i : \mathbb{R}^h \to \mathbb{R}$, то есть за каждый класс отвечает своя отдельная функция c_i, которая получает на вход все выходные значения скрытых слоев нейросети и возвращает вероятность объекта принадлежать данному классу.

Пусть для построенной сложной функции нам дана матрица $\mathbb{W}$ размера $h \times k$, где каждый элемент $w_{i,j}$ отвечает за вес между нейронами n_i и c_j. Пусть также нам дана матрица $\mathbb{O}$ размера $n \times h$, где каждый элемент $o_{i,j}$ равен $n_j(g_i)$, то есть равен активации нейрона n_j для объекта g_i обучающей выборки. Тогда назовем пару $(\mathbb{O}, \mathbb{W})$ многозначным представлением понятий.

Таким образом, мы можем представить выход нейронной сети как

$$\mathbb{O}(g) \cdot \mathbb{W}(c) + b,$$

где b - некоторое смещение. Перепишем эту формулу в следующем виде:

$$|\mathbb{O}(g)| \cdot |\mathbb{W}(c)| \cdot \cos\left(\mathbb{O}(g), \mathbb{W}(c)\right) + b,$$

для нее нам необходимо посчитать косинус угла между $\mathbb{O}(g)$ и $\mathbb{W}(c)$.

Таким образом, мы можем рассматривать объекты обучающей выборки и классы в одном пространстве, и рассматривать задачу классификации как нахождение минимального расстояния между объектом и классами в этом пространстве. То есть мы можем ввести некоторую функцию расстояния между объектами и классами $d : G \times C \to \mathbb{R}$. В качестве такой функции может быть евклидово или косинусное расстояния. Задачу классификации в такое случае можно решать с помощью алгоритма KNN с одним ближайшим соседом, то есть для каждого объекта тестовой выборки будет находиться ближайший сосед, который является классом, и объекту тестовой выборки будет приписываться этот класс [Cunningham and Delany, 2020].

Более того, используя многозначные представления понятий нейросетей, можно сравнивать сами нейросети. Для сравнения удобно использовать расстояние Громова-Вассерштейна. Оно считается для

двух матриц. В качествве этих матриц можно взять как две матрицы $\mathbb{O}$ двух нейросетей, так и две матрицы $\mathbb{W}$ этих нейросетей. Для того, чтобы посчитать расстояние Громова-Вассерштейна, строятся две матрицы, полученные из двух исходных матриц подсчетом попарного евклидово расстония для каждой матрицы и нормированные на максимальный элемент построенных матриц. Обозначим эти матрицы C_1 и C_2 для исходных двух матриц. Тогда расстояние Громова-Вассерштейна считается по следующей формуле [Memoli, 2011]:

$$GW = \min_T \sum_{i,j,k,l} L(C_{1,i,k}, C_{2,j,l}) T_{i,j} T_{k,l},$$

где L - квадратичная функция потерь, T - обучаемый параметр в оптимизационной задаче, по которому находится минимум функционала. Нижняя индексация для матриц обозначает элемент матрицы, то есть $T_{i,j}$ обозначает элемент матрицы T в строке i и столбце j.

Тем самым для каждой пары нейросетей строится их многозначные представление понятий и считается их расстрояние Громова-Вассерштейна по описанной выше формуле.

Сравнивать нейросети можно и по-другому. Пусть у нас есть многозначное представление понятий $(\mathbb{O}, \mathbb{W})$ и некоторые пороговые значения $\delta_{\mathbb{O}}$ и $\delta_{\mathbb{W}}$. Тогда построим формальные контексты $\mathbb{O}_\delta$ и $\mathbb{W}_\delta$:

$$\mathbb{O}_\delta = (G, N, I_{\mathbb{O}}) : \ (g_i, n_j) \in I_{\mathbb{O}} \iff n_j(g_i) > \delta_{\mathbb{O}}$$

$$\mathbb{W}_\delta = (C, N, I_{\mathbb{W}}) : \ (c_i, n_j) \in I_{\mathbb{W}} \iff w_{i,j} > \delta_{\mathbb{W}}$$

Эти контексты можно использовать как матрицы, тогда способ сравнения будет аналогичен предыдущему, за тем исключением, что матрицы теперь бинарные, либо же использовать их для построения решетки формальных понятий, так как теперь матрица является масштабированной и по ней можно построить контекст. Используя решетку формальных понятий, можно построить иерархию классов, или же использовать решетки каким-либо другим способом для сравнения классов или самих нейросетей.

эксперимент? Вывод? Анализ ошибок?

Список литературы

Tom Hanika and Johannes Hirth. On the lattice of conceptual measurements. 2020.

Tom Hanika Bernhard Ganter and Johannes Hirth. Scaling dimension. 2023.

Tom Hanika and Johannes Hirth. Conceptual views on tree ensemble classifiers. 2023.

James Y. Zou Amirata Ghorbani, James Wexler and Been Kim. Towards automatic concept-based explanations. 2019.

Tom Hanika and Johannes Hirth. Quantifying the conceptual error in dimensionality reduction. 2021.

Johannes Hirth and Tom Hanika. Formal conceptual views in neural networks. 2022.

Ruth Fong and Andrea Vedaldi. Net2vec: Quantifying and explaining how concepts are encoded by filters in deep neural networks. 2018.

Michael Muelly Ian Goodfellow Moritz Hardt Julius Adebayo, Justin Gilmer and Been Kim. Sanity checks for saliency maps. 2020.

Pushmeet Kohli Joshua B. Tenenbaum Jiayuan Mao, Chuang Gan and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. 2019.

Bernhard Ganter and Rudolf Wille. Formal Concept Analysis: Mathematical Foundations. Springer-Verlag, 1999.

Bernhard Ganter and Rudolf Wille. Conceptual scaling. Springer-Verlag, 1989.

Padraig Cunningham and Sarah Jane Delany. k-nearest neighbour classifiers: 2nd edition (with python examples). 2020.

Facundo Memoli. Gromov-wasserstein distances and the metric approach to object matching. 2011.