Обучение представлению коллекций данных

Каримов П.Д.
Научный руководитель: Исаченко Р.В.

April 2024

Аннотация

В данной статье рассматривается задача сопоставления векторных представлений коллекций точек данных. Исходный набор данных состоит из размеченных пар — признаков и соответствующих им значений целевой переменной. Для достижения данной цели, рассматриваются оптимально обученные представления, которые позволяют преобразовать коллекции в векторном пространстве посредством аггрегации векторов-объектов, которые принадлежат соответствующим коллекциям. В качестве метрики качества применяется расстояние между векторами-представлениями групп.

Ключевые слова — обучение представлениям, глубокое обучение

1 Введение

Эффективность методов машинного обучения в значительной степени зависит от выбора представления данных (или признаков), на которых они применяются. По этой причине большая часть фактических усилий по внедрению алгоритмов машинного обучения направлена на разработку конвейеров предварительной обработки и преобразования данных [7, 10], которые в результате чего создается представление данных, способное поддерживать эффективное машинного обучения. Такая разработка характеристик важна, но трудоемкое и подчеркивает слабость нынешних алгоритмов обучения: их неспособность извлекать информацию из данных. неспособность извлечь и систематизировать дискриминационную информацию из данных.

Эта статья посвящена обучению представлениям, т.е. обучению представлений данных, которые облегчают извлечение полезной информации при построении классификаторов или других предсказателей. В случае с вероятностными моделями хорошим представлением часто является такое, которое отражает апостериорное распределение базовых объясняющих факторов для наблюдаемых входных данных. Хорошее представление также полезно в качестве входных данных для контролируемого предиктора [4]. Среди различных способов обучения представлениям в данной статье фокусируется на методах глубокого обучения [1]: тех, которые формируются путем композицией множества нелинейных преобразований, с целью получения более абстрактных представлений.

В рамках статьи решается задача сопоставления представлений коллекциям точек данных. Данная проблема в литературе обычно решается косвенно в рамках некоторой целевой задачи [6, 8, 11]. Можно выделить 2 способа её решения:

- Учёт векторов коллекций при обучении [11];
- Рассмотрение требуемых представлений после обучения на уровне объектов [6].

В работе рассматривается последний подход, вектор коллекции получается посредством аггрегации векторов объектов, которые этой коллекции принадлежат. Качество метода проверяется на основе расстояния между представлениями коллекций в результирующем пространстве.

2 Постановка задачи

Пусть дан датасет $\mathfrak{G} = \{(x_i, y_i)\}_{i=1}^n$, $x_i \in X$, $y_i \in \{1, ..., K\}$. Составим из этих точек данных множества:

$$G_{j,k} = \{x_i | (x_i, y_i) \in \mathfrak{G} \wedge y_i = k \forall i\} : \forall j_1, j_2 G_{j_1,k} \cap G_{j_2,k} = \emptyset.$$

Наша задача состоит в том, чтобы сопоставить каждой группе $G_{j,k}$ эмбеддинг $f_\theta(G_{j,k})$, представляющий собой информативное векторное представление $G_{j,k}$. Определение "информативного" в задаче обучения представлениям обычно формулируется под конкретную задачу [3], обычно полагают, что в рамках такого представления "близкие" в каком-то смысле объекты находятся в пространстве представлений близко, а "далёкие" — далеко.

3 Существующие работы

Данная задача обычно применяется к задаче объектного различения (instance discrimination [8]) для учёта групповой информации между объектами для улучшения качества результирующей модели. Ниже представлены общие описания рассмотренных статей.

3.1 Unsupervised Visual Representation Learning by Synchronous Momentum Grouping

Обычно это делается минимизацией функции потерь, которая учитывает взаимодействия между объектами в выборке:

$$L_i = -\log \frac{\exp(\text{sim}(f_\theta(x_i^a), f_\theta(x_i^b)))}{\sum_j \exp(\text{sim}(f_\theta(x_i^a), f_\theta(x_j)))}.$$

Можно попробовать обобщить представленную функцию до уровня коллекций, чтобы получить для них представления:

$$L_i = -\log \frac{\exp(\text{sim}(c_i^a, c_i^b))}{\sum_j \exp(\text{sim}(c_i^a, g_j))}.$$

Но если пытаться делать так — некоторые объекты могут быть близки одновременно нескольким группам, и качество полученных представлений будет неудовлетворительным.

Если же задача состоит в том, чтобы коллекции были таки существенно разные (с точки зрения объектов, которые хочется этим группам соотнести) — стоит рассматривать collection-item функцию:

$$L_i = -\log \frac{\exp(\text{sim}(f_\theta(x_i^a), c_i^b)}{\sum_j \exp(\text{sim}(f_\theta(x_i^a), g_j))}.$$

Здесь x_i^a - характеристики a-ого объекта, принадлежащего i-ому классу; c_i^b - групповая характеристика b-го объекта.

Таким образом, мы пытаемся приблизить объекты коллекции к ней самой, и отдалить эти объекты от других групп.

Коллекции инициализируются, например, алгоритмом кластеризации. Обновляются они следующим образом:

$$c_i = \arg\min_{g_k} \text{sim}(f_\theta(x_i), g_k),$$
$$\text{g}_k \leftarrow \beta g_k + (1-\beta)\text{mean}_{c_t = g_k} f_\theta(x_t).$$

Такой способ позволяет через дифференцируемым образом обновлять представления коллекций, и этим отличается от аналогов, которые приводятся в указанной статье.

Поскольку составленные таким образом группы могут сколлапсировать из-за того, что на каждом шаге мы движемся по батчу — авторы предлагают периодически перегруппировывать центроиды.

3.2 GroupFace: Learning Latent Groups and Constructing Group-based Representations for Face Recognition

В этой статье авторы предлагают к эмбедам инстансным добавлять прямо групповые фичи (явно), представляя специфическую архитектуру сети.

В качестве функции потерь рассматривается сумма классификационной (CE-функция потерь) и репрезентационной (ArcFace-функция потерь) с определёнными весами.

Специфичным образом определяется принадлежность к коллекции, пытаясь адресовать неравномерность распределения по ним - не $\arg\max_k p(G_k|x)$, а $\arg\max_k \frac{1}{K}(p(G_k|x) - \mathbb{E}[p(G_k|x)]) + \frac{1}{K}$. Перескоринг интуитивно обосновывается тем, что матожидание представленной величины равно $\frac{1}{K}$.

3.3 The Group Loss for Deep Metric Learning

В этой статье авторы ставят в противовес контрастивной или триплетной функции потерь — классическому выбору в обучении представлениям, свою, которая каждому объекту в батче сопоставляет группу. Делается это доставанием характеристик из нейронной сети, подачей софтмакс-вероятностей как начальному приближению их классов и запуском определённого итерационного процесса. По окончании результат вместе с меткой подаются в кроссэнтропийную функции потерь, и осуществляется обновление весов посредством алгоритма обратного распространения.

4 Теоретические результаты

В рамках представленной задачи было решено посмотреть в сторону того, что будет, если мы оптимально потренируем представления на уровне объектов, например:

Теорема 1 *Пусть мы имеем оптимально обученную функцию представления объектов $f_\theta(x)$ с точки зрения Triplet loss-a, то есть для любого айтема x_a, его позитива x_p и негатива x_n верно, что*

$$\exists m : ||f_\theta(x_a) - f_\theta(x_p)|| - ||f_\theta(x_a) - f_\theta(x_n)|| \leq m \quad \forall (a, p, n).$$

Рассмотрим группы $G_{j_1,k_1}, G_{j_2,k_1}, G_{p_1,k_2}$, в качестве эмбеддинга группы возьмём $f_\theta(G_{j,k}) = \frac{1}{|G_{j,k}|} \sum_{x \in G_{j,k}} f_\theta(x)$. Тогда

$$||f_\theta(G_{j_1,k_1}) - f_\theta(G_{j_2,k_1})|| \leq 2 \max\{m, \max_{s_1 \in G_{j_1,k_1}, s_2 \in G_{p_1,k_2}} ||f_\theta(s_1) - f_\theta(s_2)||\}.$$

Таким образом, выбор представления группы как среднего арифметического поайтемных эмбедов в случае, если $m < 0$, является в какой-то степени оправданным. Отметим, однако, что в теореме ничего не упоминается про расстояние с центроидом отрицательного примера ещё, в этом направлении ещё предстоят некоторые исследования.

5 Эксперимент

Целью эксперимента является проверка качества представлений коллекций на основе расстояния между ними.

Датасет Omniglot [9] содержит 1623 различных рукописных символа из 50 различных алфавитов. Каждый из 1623 символов был нарисован в режиме онлайн через Amazon's Mechanical Turk 20 разными людьми. Для этого датасета обычно в качестве аугментации данных применяют повороты на 90,180,270 градусов. Каждый алфавит в представленной постановке можно использовать как группу G_k, в качестве групп $G_{j,k}$ взяв подмножество элементов, принадлежащих одной букве. Таким образом, мы получим необходимые нам группы.

В качестве функции представления возьмём нейронную сеть (предварительно обучим её), после чего проанализируем значения $||f_\theta(G_{j_1,k_1}) - f_\theta(G_{j_2,k_1})||$ и $\max_{s_1 \in G_{j_1,k_1}, s_2 \in G_{p_1,k_2}} ||f_\theta(s_1) - f_\theta(s_2)||$.

В качестве архитектуры нейронной сети взята модель EfficientNet_b2 [12], для триплетной функции потерь взят отступ (margin [2]) равный 1000. Нейронная сеть обучалась в течении X итераций, во время обучения применяется внутрипачечный hard-negative mining [13].

Метрикой качества задачи будет среднее нормы разности между эмбеддингами групп. В случае рассмотрения групп в рамках одного алфавита следует добавить так же верхнюю оценку на норму разности между ними (см. теорему (1)).

В ходе эксперимента было обнаружено, что при увеличении отступа в триплетной функции потерь соответствующим образом растут масштабы расстояния между представлениями объектов одной коллекции (что никак не противоречит сути рассматриваемой функции потерь — достаточно, чтобы разница расстояния между объектами одной коллекции и расстояния между объектами разных коллекций была близка к значению отступа, взятого со знаком минус). Результаты

эксперимента продемонстрированы несколькими примерами оценок, которые предлагалось получить выше.

$\|\|f_\theta(G_{j_1,k_1}) - f_\theta(G_{j_2,k_1})\|\|$	$\max \|\|f_\theta(s_1) - f_\theta(s_2)\|\|$
734.82	1750.37
280.37	1907.42
254.03	3338.06

По этим данным видно, что полученная теоретическая оценка оказалась довольно грубой, и требуется больше исследований в этой области.

6 Дальнейшие планы

В дальнейшем планируется получить более точную верхнюю границу для разницы между представлениями коллекций, а также получение нижней границы между ними. Кроме того, планируется получить результаты с выбором других функций представления коллекций (например, медоид) и связанных с ними теоретических результатов.

Список литературы

[1] Kourosh T Baghaei и др. “Deep representation learning: Fundamentals, perspectives, applications, and open challenges”. В: 2022.

[2] Vassileios Balntas и др. “Learning local feature descriptors with triplets and shallow convolutional neural networks”. В: янв. 2016, с. 119.1—119.11. DOI: 10.5244/C.30.119.

[3] Yoshua Bengio, Aaron C. Courville и Pascal Vincent. “Representation Learning: A Review and New Perspectives”. В: т. 35. 2012, с. 1798—1828. URL: https://api.semanticscholar.org/CorpusID:393948.

[4] Jacob Devlin и др. “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”. В: *North American Chapter of the Association for Computational Linguistics*. 2019. URL: https://api.semanticscholar.org/CorpusID:52967399.

[5] Ismail Elezi и др. “The Group Loss for Deep Metric Learning”. В: *European Conference on Computer Vision*. 2019. URL: https://api.semanticscholar.org/CorpusID:208527171.

[6] Mohsen Heidari и Kazim Fouladi. “Using Siamese Networks with Transfer Learning for Face Recognition on Small-Samples Datasets”. В: февр. 2020, с. 1—4. DOI: 10.1109/MVIP49855.2020.9116915.

[7] Alexander Isenko и др. “Where Is My Training Bottleneck? Hidden Trade-Offs in Deep Learning Preprocessing Pipelines”. В: *Proceedings of the 2022 International Conference on Management of Data*. SIGMOD ’22. Philadelphia, PA, USA: Association for Computing Machinery, 2022, с. 1825—1839. ISBN: 9781450392495. DOI: 10.1145/3514221.3517848. URL: https://doi.org/10.1145/3514221.3517848.

[8] Yonghyun Kim и др. “GroupFace: Learning Latent Groups and Constructing Group-Based Representations for Face Recognition”. В: июнь 2020, с. 5620—5629. DOI: 10.1109/CVPR42600.2020.00566.

[9] Brenden M. Lake, Ruslan Salakhutdinov и Joshua B. Tenenbaum. “Human-level concept learning through probabilistic program induction”. В: т. 350. 6266. 2015, с. 1332—1338. DOI: 10.1126/science.aab3050. URL: https://www.science.org/doi/abs/10.1126/science.aab3050.

[10] Cancheng Li. “Preprocessing Methods and Pipelines of Data Mining: An Overview”. В: т. abs/1906.08510. 2019. URL: https://api.semanticscholar.org/CorpusID:195218459.

[11] Bo Pang и др. “Unsupervised Visual Representation Learning by Synchronous Momentum Grouping”. В: *European Conference on Computer Vision*. 2022. URL: https://api.semanticscholar.org/CorpusID:250490993.

[12] Mingxing Tan и Quoc V. Le. “EfficientNet: Rethinking Model Scaling for Convolutional Neural Networks”. В: 2020. arXiv: 1905.11946 [cs.LG].

[13] Shaohua Wan и др. “Bootstrapping Face Detection with Hard Negative Examples”. В: авг. 2016.