Метрическое обучение на коннектомах

Аягоз Мусабаева[1,2], Юлия Додонова[1] и Дмитрий Петров[1,3]

[1] Институт проблем передачи информации им. А.А. Харкевича РАН
[2] Национальный исследовательский университет «Высшая школа экономики»
[3] Imaging Genetics Center, Университет Южной Калифорнии
amusabaeva@edu.hse.ru

Аннотация Очень важную роль в анализе данных играет метрика, с помощью которой измеряется расстояние между объектами. Активно развивающийся в последнее десятилетие подход метрического обучения (metric learning) предполагает автоматическое выучивание метрики на основе данных и некоторых наложенных ограничений, специфичных для задачи, которая решается в рамках конкретной предметной области. В данной работе мы проверяем гипотезу о том, что использование метрического обучения при оценивании расстояний между коннектомами – графами, представляющими сетевые структуры мозга человека – позволит получить более высокое качество классификации в задаче различения диагностических групп.

1 Введение

Методы машинного обучения всё чаще применяются в современных нейронауках. В частности, естественным образом в этой предметной области возникает задача классификации - автоматизированного различения групп здоровых людей и пациентов с различными патологиями, в частности нейродегенеративными и психиатрическими заболеваниями. В качестве входных данных в такого рода анализе используются тем или иным образом предварительно обработанные данные неинвазивной нейровизуализации.

Обработка данных нейровизуализации, поступающих на вход классификационного алгоритма, может различаться существенным образом. В простейшем случае классификация строится непосредственно на основе медицинских изображений мозга той или иной модальности. Однако в последнее десятилетие всё больший интерес вызывает так называемый коннектомный подход, когда на основе данных нейровизуализации реконструируется коннектом – граф, представляющий собой сеть связанных между собой регионов мозга. В качестве вершин такого графа рассматриваются макроуровневые регионы мозга, а ребра строятся на основе трактов, представляющих основные направления пучков миелинизированных волокон белого вещества, неконструированные на основе диффузно-тензорной магнитно-резонансной томографии (дМРТ). Сам термин «коннектом» используется, таким образом, для обозначения совокупности элементов центральной нервной системы и связей между ними, по аналогии с «геномом» в генетике и биоинформатике.

С точки зрения анализа данных коннектомы представляют собой довольно специфичные объекты. Так, задача различения диагностических групп теперь сводится к задаче классификации объектов, каждый из которых представляет собой неориентированный граф с фиксированным числом уникальным образом размеченных вершин. Одним из наиболее распространенных подходов к решению этой задачи является преобразование каждого такого входного объекта в вектор, представляющий либо всю последовательность ребер исходного графа (развернутый в вектор верхний треугольник исходной матрицы смежности), либо последовательность каких-либо локальных характеристик графа, в частности, степеней его вершин.

Однако и в этом случае не очевидно, каким образом должны оцениваться расстояния между полученными объектами. Для хорошей работы классификатора необходимо, чтобы семантически сходные объекты (например, представляющие пациентов с определенной патологией) оценивались как близкие, а семантически далекие (представляющие здоровых испытуемых vs носителей патологии) – как имеющие большие расстояния. Безусловно, существуют универсальные ответы на вопрос о том, как оценить расстояния между векторами – с помощью евклидова расстояния или cosine similarity, однако они не всегда оказываются способными отразить специфику объектов в конкретной задаче для специфичной предметной области.

Именно в такой ситуации – когда необходимо построить заранее неизвестную метрику для оценивания расстояний между объектами – оказывается применим активно развивающийся в последнее десятилетие подход метрического обучения (metric learning). Задача метрического обучения состоит в автоматическом выучивании метрики на основе данных и некоторых дополнительно наложенных содержательных ограничений (продолжая начатый выше пример, таким ограничением может быть информация о принадлежности объектов к одной или различным диагностическим группам). Обзор современных направлений развития области метрического обучения может быть найден в работе [1].

В нашей работе мы будем проверять гипотезу о том, что оценивание расстояний между объектами с использованием автоматического выучивания метрики на основе данных позволит повысить качество классификации в задаче различения диагностических групп. В качестве классификационной задачи мы будем рассматривать различение групп нормального развития и расстройств аутистического спектра. Каждый объект выборки будет представлять собой пару из двух элементов: диагноза и вектора, представляющего локальные характеристики (степени вершин) коннектома соответствующего пациента. Мы будем рассматривать только линейные подходы метрического обучения. Мы ожидаем, что, выучивая метрику Махаланобиса на основе данных, мы сможем обнаружить специфические для коннектомов паттерны и улучшить качество классификации.

2 Метрическое обучение

Главная идея метрического обучения заключается в том, чтобы, используя какую либо действительно-значную функцию, для которой выполняются свойства метрики, приспособить ее для решения задачи машинного обучения. По сути, выучивая эту функцию по обучающей выборке, мы хотим, чтобы она говорила, насколько похожи или не похожи наши объекты в том смысле похожести, который определяется нашей задачей. Самым простым случаем такой функции является метрика Махаланобиса:

$$d_M(x,y) = \sqrt{(x-y)^T M(x-y)},$$

где M – это какая-то положительно определенная матрица.

В случае бинарной классификации задача машинного обучения формулируется как задача поиска матрицы M с минимальным следом на множестве всех положительно определенных матриц, такой, что, например, будут выполнятся граничные условия вида: $d_M(x_i, x_j) < d_M(x_i, x_l)$, где объекты x_i, x_j принадлежат одному классу, а x_l- другому.

Существуют разнообразные способы постановки задачи поиска матрицы M в случае, когда используется метрика Махаланобиса как основной параметр модели машинного обучения. В нашей работе мы используется два линейных метода, которые ищут метрику махаланобиса.

Первый из них, Sparse Determinant Metric Learning (SDML) [2], заключается в том, что в нем ставится задача добиться как можно более разреженной матрицы M. Такая постановка существенна в случае, когда размерность признаков очень высока. Поиск матрицы M осуществляется следующим образом:

$$\min_M D_g(M\|M_0) + \lambda\|\mathbf{M}\|_{1,off} + \eta\mathbf{L}(S,D),$$

где

$$D_g(M\|M_0) = trace(M_0^{-1}M) - \log\det M$$

– это лог-детерминантное отклонение матрицы M от M_0, M_0 – это некоторая постоянная матрица, что помогает контролировать ее разреженность,

$$\|\mathbf{M}\|_{1,off} = \sum_{i\neq j} M_{i,j}$$

– аналог l_1 – регуляризации и

$$\mathbf{L}(S,D) = \sum_{i,j=1}^{n} (x_i^T M x_i - x_j^T M x_j) K_{i,j}$$

– граничные условия, $K_{i,j} = 1$, если x_i, x_j – похожи, и $K_{i,j} = -1$, если x_i, x_j – различны.

Для такой постановки задачи оптимизации был предложен блочный итеративный алгоритм [2], который и использовался в данной работе.

Второй метод, Large Margin Nearest Neighbor (LMNN) [3], основан на минимизации расстояния между ближайшими соседями в смысле метрики Махаланобиса. Поскольку любую положительно определенную матрицу можно представить в виде $M = L^T L$, то расстояние можно переписать в виде:

$$d_M(x,y) = \sqrt{(x-y)^T M(x-y)} = \sqrt{(x-y)^T L^T L(x-y)} = \|\mathbf{L}(\mathbf{x}-\mathbf{y})\|_2.$$

Задача будет состоять в том, чтобы отыскать соответствующую матрицу L, которая бы минимизировала функцию вида:

$$\min_L \sum_{i,j} \eta_{i,j} \|\mathbf{L}(\mathbf{x_i}-\mathbf{x_j})\|^2 + \lambda \sum_{i,j,l} \eta_{i,j}(1-y_{i,l})[1+\|\mathbf{L}(\mathbf{x_i}-\mathbf{x_j})\|^2 - \|\mathbf{L}(\mathbf{x_i}-\mathbf{x_l})\|^2]_+,$$

где $\eta_{i,j}$ – это индикатор того, что x_j является одним из k ближайших соседей x_i, то есть первое слагаемое отвечает за то, чтобы минимизировать расстояние между ближайшими соседями, $y_{i,j}$ – это индикатор принадлежности одному классу, и, соответственно, второе слагаемое можно представить как максимизацию $1+\|\mathbf{L}(\mathbf{x_i}-\mathbf{x_j})\|^2 \geq \|\mathbf{L}(\mathbf{x_i}-\mathbf{x_l})\|^2$, где для x_i x_j является ближайшим соседом, а x_l – нет, $[z]_+ = \max(z, 0)$.

Поскольку очень часто мы можем предположить наличие нелинейных закономерностей в данных, то естественным продолжением идеи метрического обучения является поиск функции похожести/различия среди нелинейных функций. Имеются работы, посвященные именно поиску нелинейных функций, задающих метрику. Однако в данной работе мы не будем рассматривать этот подход и ограничимся только первым, более простым шагом – будем рассматривать только линейные функции.

3 Постановка задачи

Цель работы - оценить возможность использования линейных методов метрического обучения для улучшения качества классификации при решении задачи различения диагностических групп.

Логика наших рассуждений такова. Предполагая, что расстояние между нашими объектами измеряется по метрике Махаланобиса, мы ожидаем, что при «правильном» подсчете расстояния классификатор, работающий на основе этих расстояний, сможет лучше разделить объекты, принадлежащие различным классам. Говоря о более высоком качестве работы классификатора, мы ожидаем получить не только лучшее качество различения объектов разных классов (более высокое значение площади под ROC-кривой, ROC AUC), но и меньшую вариативность этой оценки (меньшее стандартное отклонение ROC AUC по различным разбиениям на обучающую и тестовую подвыборки).

В качестве классификационной задачи мы будем рассматривать задачу различения групп здоровых испытуемых и пациентов с расстройствами аутистического спектра. Каждому испытуемому в используемом наборе

данных соответствует пара: коннектом – граф, репрезентирующий сетевую структуру мозга – и метка класса, к которому относится данный испытуемый.

4 Методы

В качестве входных данных мы используем не собственно граф, а вектор взвешенных степеней его вершин. На этих векторах мы ищем метрику Махаланобиса. В качестве основных моделей метрического обучения мы будем рассматривать описанные выше методы SDML [2] и LMNN [3], поскольку эти методы устойчивы и в большинстве случаев сходятся.

На следующем шаге мы получаем попарные расстояния относительно полученной метрики и решаем задачу классификации (различения нормального развития и патологии) на основе этих расстояний. В качестве базового классификатора мы используем метод опорных векторов.

Методу опорных векторов может подаваться на вход ядро – матрицу, каждый элемент которой в своем роде представляет меру похожести объектов. Мы использовали два способа получить ядро на основе матрицы попарных расстояний.

Во-первых, мы считали экспоненциальное ядро на попарных расстояниях (полученных с помощью выученной метрики), то есть матрица расстояний $d = (x_i - x_j)^T M (x_i - x_j)_{i,j=1}^n$ преобразовывалась следующим образом:

$$\kappa_{i,j}^{exp} = \exp\left(-\alpha d_{i,j}\right),$$

где $\kappa_{i,j}^{exp}$ - элемент получаемого экспоненциального ядра, α - параметр, оптимизируемый при обучении.

Во-вторых, в качестве меры сходства мы считали ковариацию между объектами, то есть

$$\kappa_{i,j}^{inn} = x_i^T M x_j, i,j = 1, \ldots, n$$

.

На этих ядрах обучался метод опорных векторов. Чтобы получить базовый уровень классификации, относительно которого оценивать эффективность предлагаемого нами подхода, мы использовали метод опорных векторов с линейным ядром на тех же входных данных (взвешенных степенях вершин).

Таким образом, сочетание двух методов метрического обучения и двух способов построения ядра, а также добавление базового классификатора дало нам пять классификаторов, поведение которых мы будем сравнивать на реальных данных.

Параметры для метрического обучения и базового классификатора подбирались внутри обучающей выборки в рамках 10-фолдовой кросс-валидации. После этого параметры фиксировались и на 100 других разбиениях на обучающую и тестовую выборку с использованием 10-фолдовой кросс-валидации оценивалось качество работы классификатора. В качестве метрики качества

рассматривался показатель ROC AUC. Мы приводим среднее значение и стандартное отклонение этого показателя для разных разбиений выборки на десять частей

5 Эксперименты

5.1 Данные

В данной работе использовался набор данных UCLA Autism (UCLA - Университет Калифорнии в Лос-Анжелесе) [4].

В базе данных UCLA Autism присутствуют испытуемые контрольной группы и пациенты с расстройствами аутистического спектра. Всего в базе содержится 94 пациента, 51 человек с патологией и 43 пациента без выявленых отклонений. Набор данных включает готовые коннектомы - графы, восстановленные на основе данных нейровизуализации и представляющие сетевые структуры мозга, по одному графу для каждого испытуемого.

Все коннектомы имеют размерность 264×264, то есть содержат 264 вершины, соответствующие различным регионам мозга. Кроме того, в данных имеются пространственные координаты для вершин графа – центры соответствующих регионов в едином стандартизированном атласе мозга.

В качестве предварительной обработки данных проводилась их нормализация в соответствии с [5]. На первом шаге исходные элементы $a_{i,j}$ взвешенной матрицы смежности нормировались следующим образом:

$$a_{ij}^{weighted} = \frac{a_{ij}}{l_{ij}^2},$$

где l_{ij} рассчитывалось как евклидово расстояние между центрами регионов i и j. После этого над элементами полученной матрицы производилось следующее преобразование:

$$w_{ij}^{normed} = \frac{w_{ij}}{\sqrt{\delta_i \delta_j}},$$

где w_{ij} - элемент предварительно нормированной на квадрат евклидова расстояния матрицы, δ_i - взвешенная степень вершины i этой матрицы. Преимущества сочетания такой последовательности преобразования исходных коннектомов обсуждаются в работе [5].

Наконец, для каждого из полученных нормализованных графов был получен вектор взвешенных степеней его вершин.

5.2 Результаты

Результаты классификации с использованием расстояний на основе выученной метрики представлены в Таблице 1. Рисунок 1 визуализирует полученные результаты, представляя не только средние, но и медианные значения ROC AUC для каждого из классификаторов.

Таблица 1. Результаты классификации с использованием расстояний на основе выученной метрики

Метод метрического обучения	Тип ядра для SVM	ROC AUC ± std
LMNN	ковариационное	0.752 ± 0.144
SDML	ковариационное	0.736 ± 0.138
SDML	экспоненциальное	0.670± 0.160
LMNN	экспоненциальное	0.668 ± 0.140
–	линейное	0.700± 0.128

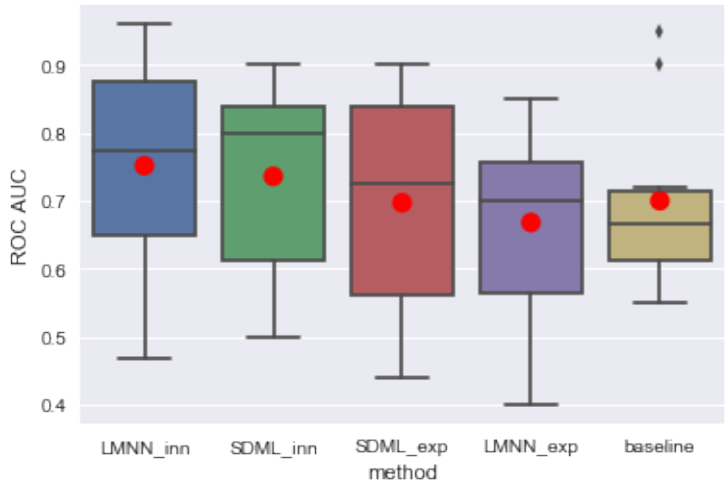

Рис. 1. Боксплоты значений ROC AUC для анализируемых классификаторов. В подписи горизонтальной оси LMNN и SDML соответствуют двум методам, использованным для метрического обучения, inn и exp обозначают ковариационное и экспоненциальное ядро SVM, соответственно, и baseline представляет базовый классификатор - SVM. Горизонтальные линии боксплотов соответствуют медианным значениям, красные точки - средним значениям для тех же данных.

Поскольку средние значения в данном случае оказываются сильно смещенными, при интерпретации результатов мы ориентируемся скорее на медианные значения и распределения получаемых значений ROC AUC. На основании полученных результатов можно говорить о том, что использование метрического обучения и SVM с ядром на основе расстояний, посчитанных с помощью выученной метрики, позволяет в среднем получать более высокое качество классификации, чем при использовании стандартного линейного SVM. Из двух типов ядер, построенных нами на основе попарных расстояний, ковариационное ядро, по-видимому, работает несколько лучше, чем экспоненциальное. Систематических различий между двумя проанализиро-

ванными подходами к выучиванию метрики в контексте данной задачи мы не обнаружили.

6 Заключение

В данной работе мы использовали метрическое обучение для того, чтобы на основе автоматически выученной метрики оценивать расстояния между векторами, репрезентирующими локальные характеристики коннектомов. Мы использовали полученные расстояния, чтобы построить ядро SVM-классификатора для различения объектов из групп нормального развития и расстройств аутистического спектра. Мы продемонстрировали, что автоматическое выучивание метрики позволяет улучшить качество классификации, в сравнении с использованием стандартного линейного ядра на тех же объектах.

Наша работа требует продолжения, в первую очередь – верификации полученных результатах на других наборах данных и в других классификационных задачах из данной предметной области. Кроме того, мы планируем сравнить эффективность для данных задач не только линейных, но и нелинейных методов метрического обучения.

Благодарности

Исследование проведено в Институте проблем передачи информации им. А.А. Харкевича РАН при поддержке Российского научного фонда, проект 17-11-01390.

Список литературы

1. Aurélien Bellet, Amaury Habrard, Marc Sebban:
   A Survey on Metric Learning for Feature Vectors and Structured Data
   arXiv e-preprints (2003) https://arxiv.org/abs/1306.6709
2. Guo-Jun Qi, Jinhui Tang, Zheng-Jun Zha, Tat-Seng Chua,Hong-Jiang Zhang :
   An Efficient Sparse Metric Learning in High-Dimensional Space via l_1-Penalized Log-Determinant Regularization
   ICML 26, 841–848 (2009)
3. Kilian Q. Weinberger, John Blitzer, Lawrence K. Saul. Distance Metric Learning for Large Margin Nearest Neighbor Classification
   JMLR, 10: 207-244, (2009)
4. Brown, J. A., Rudie, J. D., Bandrowski, A., Van Horn, J. D., Bookheimer, S. Y. (2012). The UCLA multimodal connectivity database: a web-based platform for brain connectivity matrix sharing and analysis. Frontiers in neuroinformatics, 6, 28.
5. Petrov, D., Dodonova, Y., Zhukov, L., Belyaev, M. (2016) Boosting Connectome Classification via Combination of Geometric and Topo- logical Normalizations. Proc. of Pattern Recognition in Neuroimaging.