Обучение метрик в задачах полного и частичного обучения

Ю.В. Максимов,
ИППИ РАН, ПреМоЛаб МФТИ, INRIA Rhone-Alpes
yury.maximov@phystech.edu, yuri.maximov@inria.fr

4 сентября 2014 г.

Название. Обучение метрик в задачах полного и частичного обучения. Задача оптимизации метрики является одним из популярных исследований в области машинного обучения. Ссылка на краткий обзор результатов в этой области, который содержит описание десятков методов, приведен в параграфе "Литература".

Как правило, задача поиска оптимальной метрики формулируется в виде NP-hard проблемы, допускающей однако простую выпуклую релаксацию. Например

$$\sum_{i,j} \phi((x_i - x_j)^T M (x_i - x_j)) \cdot 1_{y(x_i) \neq y(x_j)} \to \min_M$$

$$\text{s.t.}: \quad M \succcurlyeq 0$$

$$M = M^T$$

$$\text{diag}(M) = 1_n,$$

где функция ϕ выпукла(в нашем случае релаксация будет называться полуопределенной). Смысл написанной формулы состоит в том, что ближайшие соседи для каждой точки, должны иметь тот же класс, что и сама точка. При этом, полученная в ходе решения матрица M представима в виде $M = L \cdot L^T$, где матрица L в данном случае будет отвечать линейному проектору в пространство с требуемым свойством.

Цель работы : опробовать методы выпуклой оптимизации на задачах данного класса, сравнить полученные результаты с существующими методами по скорости/качеству решения.

Нюансы, научная новизна и значимость. В решении данной задачи предполагается поступательное движение от пункта к пункту. В зависимости от силы студента, все или не все пункты могут быть пройдены.

1. Приведенная релаксация не является единственной. Цель: опробовать разные релаксации.

2. Решение задачи даже в исходной, довольно частной постановке, требует больших вычислительных затрат, что неприемлемо для больших задач. Цель: съэкономить в скорости. Для этого есть некоторый набор стандартных приемов, о которых консультантом будет рассказано в свой срок.

3. Обобщение же указанной постановки на задачи частичного обучения приводит к задаче DC-оптимизации и изучена достаточно слабо, а потому представляет первичный интерес работы (далеко не все методы, любимые специалистами по невыпуклой оптимизации, были изучены специалистами по анализа данных). Тут много методовв. В первую очередь хотелось бы перенести на задачу частичного обучения метрик техники, развитые для SVM. Ссылки будут даны дополнительно при достижении этого пункта.

Алгоритмы и псевдокод. Псевдо-код разрабатываемых методов, равно как и ссылки на известные техники будут выдаваться консультантом по достижении очередного этапа в решении задачи.

Данные. Планируется опробовать методы на данных репозитория UCI, LIBSVM и данные ImageNet, из которых извлечены deep features.

1. Репозиторий LIBSVM
2. Репозиторий UCI
3. Репозиторий ImageNet

Подготовленные данные по ImageNet предоставляются консультантом задачи. В первую очередь важны результаты по ImageNet.

Литература. По обучению метрик:

1. Auruelien Bellet. Tutorial on Metric Learning.
2. A. Bellet, A. Habrard and M. Sebban. A Survey on Metric Learning for Feature Vectors and Structured Data.

По выпуклой оптимизации:

1. Ю.Е. Нестеров. Введение в выпуклую оптимизацию. М: МНЦМО, 2010.

По невыклой оптимизации, возникающей в задачах частичного(трансдуктивного обучения) и собственно по трансдуктивному обучению, рекомендуется ознакомиться с обзорами:

1. Xiaojin Zhu. Semi-Supervised Learning Tutorial. ICML tutorial.
2. O. Chapelle, B. Schцlkopf and A. Zien. Semi-Supervised Learning. MIT Press.

Дополнительные теоретические материалы, для подготовки к выполнению проекта будут высланы студенту по e-mail, при необходимости.