Сравнение эффективности логических методов в задачах анализа данных

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

1 сентября 2014 г.

Название. Сравнение эффективности комбинаторно-логических методов в задачах анализа данных.

Источник. Комбинаторно-логические методы являются одним из основных направлений исследований московской школы анализа. Им посвящена обширная литература. Постановка и решение задачи мотивированы последними результатами в этой области. Данные для экспериментов предоставлены группой анализа изображений INRIA Rhone-Alpes.

Задача. Состоит в исследовании возможностей комбинаторно-логических методов, и сравнении их эффективности в реальных (ImageNet: Fungus, Birds) и классических (UCI) задачах анализа данных. Особое внимание планируется уделять многоклассовой постановке.

Подходы к решению. Предполагается сравнить качество методов анализа данных, основанных на представлении данных дизъюнктивными нормальными формами. Перед началом работы настоятельно рекомендуется прочитать лекцию К.В. Воронцова о логических методах классификации.

Логические алгоритмы могут применяться, как непосредственно к данным, так и как дополнительная процедура для алгоритмической коррекции. В первом случае, следует снизить размерность исходной задачи(например провести PCA объясняющий $1-\varepsilon$ вариацию), затем построить градации на множестве признаков.

Во втором случае, предлагается использовать Weighted Majority Algorithm, над множеством ответов базовых классификаторов. Хорошее краткое введение приведено тут. В данном случае каждая конъюнкция ответов классификаторов (например SVM и Adaboost) может быть интерпретирована как эксперт, и Weighted Majority Algorithm как процедура коррекции с малым регретом. Таким образом, эффективность алгоритма зависит от исходного множества конъюнкций, построенного методами, эффективность которых мы хотим сравнить. Построение ДНФ, реализующей разделение классов, планируется проводить известными алгоритмами (см. список литературы ниже; все интересующие псевдокоды могу быть высланы), целью является сравнение их эффективности/качества(жадные алгоритмы, редукционный

алгоритм, аппроксимационный алгоритм, построение Хорновских ДНФ, построение ДНФ последовательным перемножением: см. раздел литература). Во (многие) стандартные алгоритмы в этой области не входят, опции тонкой настройки(например, выбор признаков для редукции). Так как в первую очередь, упор делался на "простоту" конечной комбинации конъюнкций. Тем самым сочетание стандартных алгоритмов (не ориентированных на практические задачи) и практических идей(например, вопрос выбора наиболее информативных признаков при построении ДНФ) может дать методы, интересные как специалистам-практикам, так и, в определенном смысле, теоретикам.

Конечной целью является сравнение результатов, полученных данными методами, с решениями полученными в рамках моделей решающих деревьев/списов (качество/скорость).

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.

Научная новизна и значимость. Список теоретической литературы по комбинаторно-логическим методам анализа данных, основанных на разделении классов ДНФ довольно обширен. В то же время, несмотря на достаточно глубокие наработки в теоретической области, практической составляющей в литературе уделяется не так много внимание. Определенным исключением является внимание к построению решающих деревьев/списков. В то же время процедуры синтеза, основанные на построении ДНФ, способны улавливать более тонкие закономерности в данных, оставаясь при этом достаточно робастными.

Системные требования. Для программной реализации части задачи рекомендуется использовать современные библиотеки анализа данных, такие как `VLFEAT` и `YAEL`, которые в настоящее время поддерживаются лишь для `Linux 64 bit` и `Mac OS X 64 bit`. Использование этих библиотек позволяет существенно ускорить и повысить качество решения задач. Однако оно не является обязательным, для реализации основных программных компонентов достаточно библиотек `LIBSVM`(от которой нам будет нужен модуль загрузки данных) и `SVMLIN`, предоставляющий относительно функционал для обучения линейных классификаторов в задачах полного и частичного обучения большой размерности. И `LIBSVM` и `SVMLIN` можно использовать под `Windows 32/64`, но это потребует время на сборку библиотеки. Ссылки на библиотеки, представлены в параграфе "Литература". Для решения выпуклых задач оптимизации рекомендуется использовать `CVX` в комбинации с солвером `MOSEK`. Все библиотеки являются свободно распространяемыми. В случае трудностей с их установкой и использованием, обращайтесь к консультанту.

1. Библиотека `VLFEAT`;

2. Библиотека `LIBSVM`;

3. Библиотека `SVMLIN`.

4. `CVX` является частю Матлаба. Вместе с `CVX` лучше всего использовать Солвер `MOSEK`, документация по которому доступна тут и тут.

Литература.

1. E. Boros, Y. Crama, P. L. Hammer, T. Ibaraki, A. Kogan, K. Makino. Logical Analysis of Data: Classification with Justification. // Annals of Operations Research. 2011. Vol. 188. Iss. 1. P. 33–61.

2. M. Marchand, J. Shawe-Taylor. The Set Covering Machine. Journal of Machine Learning Research 3 (2002) 723-746

3. Н.Н. Бондаренко, Ю.И. Журавлев. Алгоритм выбора конъюнкций для логических методов распознавания. Ж. вычисл. матем. и матем. физ., Т. 52:4 (2012), 746–749.

4. А.Г. Дьяконов. Построение ДНФ последовательным перемножением. Журнал вычислительной математики и математической физики, 2003, Т. 43, № 10, С. 1589–1600.

5. Ю.И. Журавлев, А.Ю. Коган. Реализация булевых функций с малым числом нулей дизъюнктивными нормальными формами и смежные задачи. // ДАН СССР. 1985. Т. 285. №4. С. 795-–799.

6. Ю.В. Максимов. Простые дизъюнктивные нормальные формы булевых функций с малым числом нулей // Доклады РАН. Серия «Математика». — 2012. Т. 445 №2 — С. 143–145.

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