

Мультимоделирование как универсальный способ описания выборки общего вида*

Логинев Р. А., Адуенко А. А., Стрижов В. В.

logipamar@yandex.ru

В случае неоднородных данных в машинном обучении использования одной модели недостаточно. Для выявления этого используют комбинации нескольких моделей - мультимодели. Работа нацелена на то, чтобы изучить по последовательности постепенно приходящих данных эволюцию представлений о модели. Рассмотреть, в какие моменты предпочтительнее разветвлять одну модель, а также какими критериями пользоваться для объединения ещё недообученного ответвления с имеющимися моделями. Исследования проводятся на синтетических данных из многоуровневой модели или смеси распределений.

Ключевые слова: мультимоделирование, бинарная классификация.

1 Введение

Работа посвящена исследованию решения задач бинарной классификации при помощи мультимоделей. Решение этой задачи может быть использовано в вопросах кредитного скоринга [], медицинской диагностики [], предсказания качества продукции и других областях. В некоторых задачах встречаются данные, для описания которых требуется вводить несколько моделей. Например, для задачи кредитного скоринга важность признаков в модели может отличаться в зависимости от региона заявителя. Как пример, многодетность может быть положительным параметром для более благополучных регионов и отрицательным для менее состоятельных. Тогда используют решающее правило о разделении выборки на кластеры, а затем на каждом из них строят отдельную модель. Такой подход называют многоуровневой моделью.

Один из алгоритмов построения и обучения оптимальной модели, основанный на байесовском выводе и ЕМ-алгоритме, описан в [Aduenko-main]. Более того, известна процедура выявления максимального числа необходимых моделей, а также построена функция, которая, в отличие от методов, основанных на дивергенциях Брегмана и KL-дивергенциях, позволяет оценить различимость двух моделей.

Однако этот алгоритм рассматривает выборку как статическую и известную заранее. В прикладных задачах появляются данные, имеющие временную структуру. Из-за этого на одном и том же объекте ответ с течением времени может различаться. Таким образом целью работы является исследование эволюции модели во времени. Более того, в статье приведены эксперименты, выявляющие необходимый размер выборки, которую возможно отделить для построения новой модели. Подобные результаты планируется получить на синтетических данных, где временная структура будет различной: случайный выбор модели и непрерывные отрезки во времени, на которых поступает каждая модель.

В упомянутой работе такая оптимизация алгоритма не приведена, и новый подход позволяет улучшить качество мультимоделей в задачах бинарной классификации.

Предлагается построить модификацию этого алгоритма, взяв его за основу. Один из методов заключается в том, чтобы для новой модели учитывать лишь ту часть выбор-

*Работа выполнена при финансовой поддержке РФФИ, проект №00-00-00000. Научный руководитель: Стрижов В. В. Задачу поставил: Стрижов В. В. Консультант: Адуенко А. А.

ки, которая поступила последней. Размер этой части является гиперпараметром и будет подобран на синтетических данных.

Затем полученный алгоритм предлагается сравнить с уже имеющимся как на построенных синтетических данных, так и на собранных в репозитории UCI данных о кредитном скоринге, бинаризованных данных о стоимости квартир и о качестве вина [1].

2 Постановка задачи

3 Введение(пример)

После аннотации, но перед первым разделом, располагается введение, включающее в себя описание предметной области, обоснование актуальности задачи, краткий обзор известных результатов, и т. п [1–4, 6, 7].

4 Название раздела

Данный документ демонстрирует оформление статьи, подаваемой в электронную систему подачи статей <http://jmla.org/papers> для публикации в журнале «Машинное обучение и анализ данных». Более подробные инструкции по стилевому файлу `jmla.sty` и использованию издательской системы \LaTeX находятся в документе `authors-guide.pdf`. Работу над статьёй удобно начинать с правки \TeX -файла данного документа.

4.1 Название параграфа.

Нет ограничений на количество разделов и параграфов в статье. Разделы и параграфы не нумеруются.

4.2 Теоретическую часть работы

желательно структурировать с помощью окружений `Def`, `Axiom`, `Hypothesis`, `Problem`, `Lemma`, `Theorem`, `Corollary`, `State`, `Example`, `Remark`.

Определение 1. Математический текст хорошо структурирован, если в нём выделены определения, теоремы, утверждения, примеры, и т. д., а неформальные рассуждения (мотивации, интерпретации) вынесены в отдельные параграфы.

Утверждение 1. Мотивации и интерпретации наиболее важны для понимания сути работы.

Теорема 1. Не менее 90% коллег, заинтересовавшихся Вашей статьёй, прочитают в ней не более 10% текста.

Доказательство. Причём это будут именно те разделы, которые не содержат формул. ■

Замечание 1. Выше показано применение окружений `Def`, `Theorem`, `State`, `Remark`, `Proof`.

5 Некоторые формулы

Образец формулы: $f(x_i, \alpha^\gamma)$.

Образец выключной формулы без номера:

$$y(x, \alpha) = \begin{cases} -1, & \text{если } f(x, \alpha) < 0; \\ +1, & \text{если } f(x, \alpha) \geq 0. \end{cases}$$

Таблица 1 Подпись размещается над таблицей.

Задача	CCEL	boosting
Cancer	3.46 \pm 0.37 (3.16)	4.14 \pm 1.48
German	25.78 \pm 0.65 (1.74)	29.48 \pm 0.93
Hepatitis	18.38 \pm 1.43 (2.87)	19.90 \pm 1.80

Образец выключной формулы с номером:

$$y(x, \alpha) = \begin{cases} -1, & \text{если } f(x, \alpha) < 0; \\ +1, & \text{если } f(x, \alpha) \geq 0. \end{cases} \quad (1)$$

Образец выключной формулы, разбитой на две строки с помощью окружения align:

$$R'_N(F) = \frac{1}{N} \sum_{i=1}^N \left(P(+1 | x_i) C(+1, F(x_i)) + \right. \\ \left. + P(-1 | x_i) C(-1, F(x_i)) \right). \quad (2)$$

Образцы ссылок: формулы (1) и (2).

6 Пример иллюстрации

Рисунки вставляются командой `\includegraphics`, желательно с выравниванием по ширине колонки: `[width=\linewidth]`.

Практически все популярные пакеты рисуют графики с подписями, которые трудно читать на бумаге и на слайдах из-за малого размера шрифта. Шрифт на графиках (подписи осей и цифры на осях) должны быть такого же размера, что и основной текст.

При значительном количестве рисунков рекомендуется группировать их в одном окружении `{figure}`, как это сделано на рис. ??.

7 Пример таблицы

Подпись делается *над таблицей*, см. таблицу 1.

8 Заключение

Желательно, чтобы этот раздел был, причём он не должен дословно повторять аннотацию. Обычно здесь отмечают, каких результатов удалось добиться, какие проблемы остались открытыми.

Литература

Литература

- [1] *Author N.* Paper title // 10-th Int'l. Conf. on Anyscience, 2009. Vol. 11, No. 1. Pp. 111–122.

-
- [2] *Автор И. О.* Название книги. Город: Издательство, 2009. 314 с.
 - [3] *Автор И. О.* Название статьи // Название конференции или сборника, Город: Изд-во, 2009. С. 5–6.
 - [4] *Автор И. О., Соавтор И. О.* Название статьи // Название журнала. 2007. Т. 38, № 5. С. 54–62.
 - [5] `www.site.ru` — Название сайта. 2007.
 - [6] *Воронцов К. В.* $\text{\LaTeX} 2_{\epsilon}$ в примерах. 2006. <http://www.ccas.ru/voron/latex.html>.
 - [7] *Львовский С. М.* Набор и вёрстка в пакете \LaTeX . 3-е издание. Москва: МЦНМО, 2003. 448 с.