Рецензия

на доклад по теме «Ленивые методы классификации в машинном обучении».

В данном докладе описывается метод классификации, который не требует построения модели для выполнения классификации тестируемого примера. Изначально стоит уточнить понимание наличия модели: что есть модель? Как канонический пример был приведен метод классификации k-NN. Далее отмечен факт, что подобная классификация не является устойчивой к выбросам. Таким образом, для реального применения метода требуется качественный выбор базы примеров, который в данном случае можно считать предобработкой и самой моделью. Как основная задача была выбрана цель улучшить результаты, показываемые решающим лесом. В этом случае стоит ожидать, что докладчик будет хорошо разбираться в том, какие плюсы и минусы есть у последнего метода и как он устроен — однако, в данном случае это не так. Из того, что я увидел, была использована готовая реализация метода, и дальнейшее знакомство было ограничено (причины могут различаться); при этом докладчик запутался в используемых подходах — в попытке объяснения того, как работает решающий лес для ленивого случая (разреженные данные, во избежание построения большого количества различных классификаторов под разные наборы доступных признаков предлагается делать это для каждого примера отдельно) была сделана попытка объяснения метода работы схемы, использующей построение ассоциативных правил по прецедентам, у которых значения имеющихся в наличии признаков совпадают. Далее описывается метод, пытающийся построить минимальное правило, которое не нарушает некоторые ограничения по критериям вроде множества поддержки.

Суммарно доклад производит впечатление того, что первая попытка рассказа была произведена именно на семинаре. В любом случае, не выглядит как доклад аспиранта 3ьего года (предыдущий доклад на порядок лучше, но, может быть, про свою работу рассказывать сложнее).

Плюсы:

- видно, что была проделана работа, и докладчик имеет представление о том, что было сделано.

Минусы (пожелания):

- требуется более качественная проработка доклада (но начало было хорошее);
- две версии слайдов с разными материалами — плохая готовность к докладу;
- слишком много лишней информации, рассказ не имеет понятной структуры и последовательности (и эта структура явно не оговорена), что приводит к утере понимания, о чем рассказывается в каждый конкретный момент;
- большое количество лишней информации — показатель того, что докладчик имеет трудности с выделением значимых идеи и лишних для понимания этих идей деталей — это может быть воспринято как плохое понимание материала;
- алгоритм позиционируется для решения крупных задач (банки, медицина), но метод проверяется на базе, где признаковое пространство имеет размерность порядка 30, а примеров около 1000; даже при этом оценки времени работы примерно 10 минут; то есть, алгоритм не пригоден для целевой задачи, где объёмы данных имеют на порядки большие размеры;
- алгоритм проверяется «чисто теоретически»; заявлено, что подготовка базы для тестирования — work in progress, что странно для аспиранта, который скоро планирует защищаться (а насколько хорошо и быстро метод сработает на собранной базе?); а если весь акцент именно на теоретическом обосновании, то зачем нужна база для тестирования?
- есть проблемы с «честностью» сравнения методов, так как вводятся ограничения на оба метода (например, глубина 2 для решающего леса) — нет объяснения того, что эти ограничения являются равноценными с точки зрения потери качества.

Автор рецензии: Александр Сергеев (aesergeev@hse.ru)