Обзор статьи “Что такое EM-алгоритм”

(“What is expectation maximization algorithm” by Chuong B Do & Serafim Batzoglou)

В предложенной статье рассматривается так называемый EM-алгоритм и его применение **в области** биоинформатики **для задач** оценки максимального правдоподобия параметров вероятностных моделей, зависящих от некоторых скрытых переменных.

В первой части своей работы авторы на примере конкретной задачи рассматривают метод оценки максимального правдоподобия, обобщением которого в случае неполноты исходных данных является EM-алгоритм.

EM-алгоритм состоит из последовательного чередования двух шагов, постепенно улучшающих данные значения оцениваемых параметров. Первый шаг (E-step) состоит из вычисления условного распределения скрытых переменных, исходя из значений наблюдаемых переменных и величины оцениваемых параметров с предыдущего шага. Для этого используется формула Байеса и формула полной вероятности. Второй шаг (M-step) необходим для пересчёта значений параметров с использованием результатов первого шага. Новые значения параметров должны максимизировать условное мат. ожидание логарифма правдоподобия.

Далее этот алгоритм также подробно рассматривается на конкретном примере.

Вторая часть статьи посвящена вопросу обоснования сходимости алгоритма. По сути, описанный выше процесс разбивает исходную сложную оптимизационную задачу, чья целевая функция может иметь множество локальных максимумов, на множество подзадач, гарантированно имеющих единственный глобальный максимум, который зачастую может быть вычислен аналитически. E-шаг обеспечивает соблюдение правильного порядка выбора подзадач, благодаря чему значение параметров сходится к некоторому локальному оптимуму логарифма исходной целевой функции.

Как и большинство алгоритмов оптимизации невыпуклых функций, EM-алгоритм гарантирует лишь сходимость к локальному оптимуму, что является существенным **недостатком** предложенного метода. Однако, благодаря таким **преимуществам,** как простота реализации и устойчивость, EM-алгоритм является достаточно распространённым.

В последней части работы рассматриваются приложения EM-алгоритма к задачам вычислительной биологии. Подробно рассматриваются следующие примеры:

● *Задача кластеризации данных экспрессии генов.* В данном случае, *наблюдаемыми переменными* является вектор значений измерений, *скрытыми переменными* - принадлежность каждого вектора к тому или иному классу, *оцениваемыми параметрами* - параметры многомерного распределения Гаусса для каждого из классов. В данном случае, EM-алгоритм может быть рассмотрен, как упрощенный вариант метода k-средних.

● *Задача обнаружения общего мотива заданной длины в нескольких нуклеотидных или аминокислотных последовательностях.* Здесь, *наблюдаемые переменные* - символы последовательностей, *скрытые переменные* - начальные позиции мотива в последовательностях, *оцениваемые параметры* - частота символов мотива в последовательностях в зависимости от позиции.

● *Задача определения гаплотипа для генотипа данной особи. Наблюдаемые переменные* - генотипы особей, *скрытые переменные* - принадлежность генотипов к одному из гаплотипов, *оцениваемые параметры* - частота встречаемости гаплотипов в популяции.

Подводя итоги, авторы относят EM-алгоритм к классу эффективных и простых методов построения моделей вычислительной биологии.