Секция «Теория вероятностей и математическая статистика»

Об индуктивном порождении и выборе существенно нелинейных регрессионных моделей с учетом погрешности как зависимых, так и независимых измеряемых переменных

Рудой Георгий Игоревич

Аспирант

Московский физико-технический институт, Москва, Россия

E-mail: 0xd34df00d@gmail.com

В большинстве естественнонаучных приложений в ходе анализа результатов эксперимента требуется восстановить функциональную зависимость между измеряемыми величинами. При этом необходима возможность экспертной интерпретации полученной зависимости. Кроме того, зачастую требуется выбрать единственную модели из некоторого множества, например, заранее предложенных экспертом или индуктивно порожденных альтернатив.

Для построения структурно сложных моделей предлагается использовать метод символьной регрессии [1, 2], позволяющий строить существенно нелинейные параметрические регрессионные модели, представляющие собой суперпозиции аналитических функций. Параметры каждой индуктивно порожденной модели минимизируют среднеквадратичное отклонение в предположении о нормальности распределения регрессионных остатков, например, с помощью алгоритма Левенберга-Марквардта (АЛМ). Различные модели обычно сравниваются согласно ошибке на обучающей выборке [1, 2].

Однако измеряемые величины, как независимые, так и зависимые, известны не абсолютно, а лишь с некоторой конечной точностью. Соответственно, требуется выбрать не только регрессионную модель, минимизирующую сумму квадратов регрессионных остатков, но и оценить зависимость ее параметров от вариации входных данных в рамках некоторых экспертно заданных предположений о погрешностях измерения.

Кроме того, непосредственно в процессе построения модели необходимо учитывать погрешности независимых переменных, для чего общепринятые алгоритмы оптимизации существенно нелинейных регрессионных моделей не подходят, так как они измеряют расстояние от каждой экспериментально снятой точки до кривой, соответствующей оптимизируемой модели, по вертикали, что соответствует предположению об отсутствии ошибки измерения независимых переменных.

В настоящей работе предлагается алгоритм оптимизации, основанный на алгоритме Левенберга-Марквардта, позволяющий учитывать погрешности измерения измеряемых величин. Предложенный алгоритм также может быть применен в случае различия погрешности определения каждой из физических величин в различных точках.

Для этого квадрат каждого регрессионного остатка для каждой точки обучающей выборки нормируется на величину, зависящую от погрешности измерения и коэффициентов линеаризованной в окрестности соответствующей точки модели, оптимальные параметры которой требуется найти. На полученной сумме нормированных остатков выполняется некоторое число итераций стандартного АЛМ, после чего процесс повторяется снова с обновленными параметрами линеаризованной модели согласно новому полученному приближению.

Показывается, что данный алгоритм эквивалентен оптимизации стандартным немодифицированным АЛМ некоторого функционала, получающегося из искомого. Таким образом, доказываются свойства корректности и сходимости предлагаемого алгоритма.

В вычислительном эксперименте предложенный метод применен для восстановления зависимости показателя преломления полимера n от длины волны λ.

Предложенный алгоритм дополняет и уточняет ранее описанный метод индуктивного порождения и выбора регрессионных моделей [3] (который, кроме того, учитывает и структурную сложность порожденных моделей), дополняя его предложенным в настоящей работе критерием устойчивости моделей.

Источники и литература

1) Sammut C., Web.. G. I. Symbolic Regression. - Berling: Springer, 2010.
2) Zelinka I., Oplatkova Z., Nolle L. Analytic Programming- Symbolic Regression by means of arbitrary evolutionary algorithms. - UK Sym. Journal, 2005, Vol 6, № 9, Paper 5
3) Рудой Г. И., Стрижов В. В. Алгоритмы индуктивного порождения суперпозиций для аппроксимации измеряемых данных. - Информатика и ее применения, 2013 - №7, С. 44-53.