Рассматривается возможность агрегации моделей зависимостей, содержащихся в данных, на примере агрегации линейных регрессионных моделей 2-х переменных и систем нечетких правил. Приведен необходимый и достаточный список статистик данных, позволяющих проводить однозначную аддитивную агрегацию без привлечения полного объема исходных данных. Предложено понятие «опыта» для обозначения списка характеристик, достаточного для проведения аддитивной агрегации.

УДК 519.8

С.Н. НЕЧУЙВИТЕР

АГРЕГАЦИЯ ЛИНЕЙНЫХ РЕГРЕССИОННЫХ МОДЕЛЕЙ И СИСТЕМ НЕЧЕТКИХ ПРАВИЛ

Введение. Последнее время развитие интел-лектуальных технологий и техники вышло на уровень, который позволяет проводить комплексное моделирование сложных сис-тем за приемлемое время. С другой стороны, к настоящему времени накоплены богатые экспериментальные данные во многих прикладных областях знаний (таких как медицина, финансовые рынки) и создано множество простых моделей, описывающих тот или иной аспект реальности. Поэтому актуальными становятся методы объе-динения уже наработанных моделей в рамках сложных комплексных.

Несмотря на наличие богатых исто-рических данных для многих сложных систем не существует точных моделей. Но было разработано множество нечетких методов, в основном опирающихся на экс-пертные оценки (это обусловлен тем, что, несмотря на отсутствие формальных моделей, человек умеет строить эффективные интуитивные модели).

Необходимость в объединении именно моделей обусловлена тем, что:

1) передача и последующая обработка полного объема данных может быть затруднена (особенно для многомерных случаев и нелинейных моделей);

2) может существовать необходимость повысить точность либо диапазон применимости модели.

В данной работе исследуется возможность агрегации линейных регрессионных моделей 2-х переменных и систем нечетких правил с гауссовскими функциями принадлежности.

Цель данной работы состоит в поиске расширения линейной регрессионной модели 2-х переменных и системы нечетких правил 2-х переменных, которое бы позволило проводить аддитивную агрегацию таких моделей.

Постановка задачи. Дано:

- моделируемая функциональная зависимость, заданная набором точек в 2-х мерном пространстве;
- стандартный метод построения линейной регрессионной модели 2-х переменных;
- метод Ванга моделирования функции системой нечетких уравнений с гауссовскими функциями принадлежности.

Необходимо найти необходимый и достаточный список дополнительных характеристик линейной регрессионной модели 2-х переменных и системы нечетких правил 2-х переменных, который бы позволил проводить аддитивную агрегацию таких моделей.

Достаточные и исчерпывающие характеристики регрессии

Пусть задано n точек данных в 2-х мерном пространстве $\left\{\left\langle x_i; y_i \right\rangle\right\}_{i=1}^{n}$. Предполагая, что переменная x задана точно, а переменная y – со случайным шумом, построим наилучшую линейную модель вида $y_i = a \cdot x_i + b + \varepsilon_i$, где a, b – параметры модели, $\{\varepsilon_i\}$ – множество независимых нормально распределенных случайных величин с одинаковым нулевым средним и дисперсией σ^2. Согласно методу наименьших квадратов под наилучшей моделью будем понимать модель с минимальной суммой квадратов невязок:

если $e_i = y_i - \left(a^* \cdot x_i - b^*\right)$ – невязка модели с данными, а $Q = \sum_i \left(y_i - a \cdot x_i - b\right)^2 = \sum_i e_i^2$ – оптимизируемая функция качества, то оптимальные параметры модели: $\left\langle a^*; b^* \right\rangle = \arg\min_{a,b \in R} \sum_i \left(y_i - a \cdot x_i - b\right)^2$.

Из необходимых условий экстремума функции:

$$\frac{\partial Q}{\partial a} = \sum_i e_i x_i = 0,$$

$$\frac{\partial Q}{\partial b} = \sum_i e_i = 0 \Rightarrow \overline{y} - a^* \overline{x} - b^* = 0 \Rightarrow y(x) = \overline{y} + a^* \cdot \left(x - \overline{x}\right),$$

$$a^* = \frac{\sum_i \left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sum_i \left(x_i - \overline{x}\right)^2} = \frac{\sum_i x_i y_i - n \cdot \overline{x} \cdot \overline{y}}{\sum_i x_i^2 - n \cdot \overline{x}^2},$$

где $\overline{x} = \frac{1}{n}\sum_i x_i$, $\overline{y} = \frac{1}{n}\sum_i y_i$ – средние значения переменных.

Также можно выписать в явном виде оценку ошибки моделирования значения y в произвольной точке x [1, с. 92 – 98]:

$$\sigma_y^{\,2} = \frac{\sigma^2}{n} + \sigma_a^{\,2} \cdot \left(x - \overline{x}\right)^2,$$

где σ^2 – истинная дисперсия случайного нормального шума, а $\sigma_a^{\,2} = \dfrac{\sigma^2}{n \cdot S_x^{\,2}}$ – дисперсия оценки параметра a, $S_x^{\,2} = \dfrac{1}{n}\sum_i \left(x_i - \overline{x}\right)^2$ – дисперсия переменной x.

Дисперсию шума можно оценить исходя из дисперсии невязок:

$$\hat{\sigma}_\varepsilon^{\,2} = \frac{\sum_i e_i^{\,2}}{n-2}.$$

Таким образом, оптимальные параметры модели и ее точность полностью определяются следующими статистиками исходных данных: n, $SX = \sum_{i=1}^{n} x_i$,

$$SY = \sum_{i=1}^{n} y_i,\ SXX = \sum_{i=1}^{n} x_i^{\,2},\ SXY = \sum_{i=1}^{n} x_i y_i,\ RSS = \sum_{i=1}^{n} e_i^{\,2} = \sum_{i=1}^{n} \left(y_i - a^* \cdot x_i - b^*\right)^2.$$

Тогда

$$a^* = \frac{SXY - n \cdot \dfrac{SX}{n} \cdot \dfrac{SY}{n}}{SXX - n \cdot \left(\dfrac{SX}{n}\right)^2} = \frac{n \cdot SXY - SX \cdot SY}{n \cdot SXX - SX \cdot SX},$$

$$y(x) = \frac{SY}{n} + a^* \cdot \left(x - \frac{SX}{n}\right) = \frac{n \cdot SXY - SX \cdot SY}{n \cdot SXX - SX \cdot SX} \cdot x + \frac{SY \cdot SXX - SX \cdot SXY}{n \cdot SXX - SX \cdot SX},$$

$$S_x^{\,2} = \frac{1}{n}\sum_i \left(x_i - \overline{x}\right)^2 = \frac{1}{n}\left(\sum_i x_i^{\,2} - n \cdot \overline{x}^2\right) = \frac{1}{n}\left(SXX - n \cdot \left(\frac{SX}{n}\right)^2\right) = \frac{SXX}{n} - \frac{SX^{\,2}}{n^2},$$

$\hat{\sigma}_\varepsilon^{\,2} = \dfrac{RSS}{n-2}$ – оценка дисперсии шума;

$\hat{\sigma}_a^{\,2} = \dfrac{\hat{\sigma}_\varepsilon^{\,2}}{n \cdot S_x^{\,2}}$ – оценка дисперсии параметра *a* модели;

$\hat{\sigma}_y^{\,2} = \dfrac{\hat{\sigma}_\varepsilon^{\,2}}{n} + \hat{\sigma}_a^{\,2} \cdot \left(x - \overline{x}\right)^2$ – оценка точности модели.

Агрегация линейных регрессионных моделей

Допустим, по данным несколько независимых серий измерений были построены две независимые линейные регрессионные модели согласно вышеописанной процедуре. Допустим, есть основания полагать, что обе выборки данных описываются одной и той же зависимостью. Для построения модели по совокупным данным достаточно вычислить перечисленные выше статистики данных. В отличие от классических параметров линейной регрессии они носят аддитивный характер:

$n_{aggr} = \sum_j n_j$, где j – индекс набора данных;

$$SX_{aggr} = \sum_j SX_j, \;\; SY_{aggr} = \sum_j SY_j,$$

$$SXX_{aggr} = \sum_j SXX_j, \;\; SXY_{aggr} = \sum_j SXY_j.$$

Пусть $\Delta_{ij} = \left(a_j^* x_i + b_j^*\right) - \left(a_{aggr} x_i + b_{aggr}\right) = a_\Delta x + b_\Delta$.

Тогда $RSS_j^{\,aggr} = \sum_i \left(y_i - a_{aggr} \cdot x_i - b_{aggr}\right)^2 = \sum_i \left(y_i - a_j^* \cdot x_i - b_j^* + \Delta_{ij}\right)^2 =$

$$= \sum_i \left(e_i^{\,2} + 2e_i\left(a_\Delta x + b_\Delta\right) + \Delta_i^{\,2}\right) = \sum_i e_i^{\,2} + \sum_i \Delta_{ij}^{\,2},$$

так как $\sum_i e_i x_i = 0$ и $\sum_i e_i = 0$ по условию минимизации суммы квадратов невязок.

$$RSS = \sum_j \left(RSS_j + \sum_i \Delta_{ij}^{\,2} \right).$$

По статистикам агрегированных данных вычисляем параметры модели.

Агрегация линейной регрессионной модели и системы нечетких правил

Под системой нечетких правил будем понимать совокупность утверждений вида:

«Если $x \in X$, то $y \in Y$»,

где A и B задаются гауссовскими функциями принадлежности:

$$p_A(x) = \frac{1}{\sqrt{2\pi}\sigma_x} \exp\left(-\frac{\left(x - \mu_x\right)^2}{2\sigma_x^{\,2}} \right),$$

$$p_B(y) = \frac{1}{\sqrt{2\pi}\sigma_y} \exp\left(-\frac{\left(y - \mu_y\right)^2}{2\sigma_y^{\,2}} \right).$$

В общем случае система нечетких правил соответствует нелинейной зависимости между переменными [2], но в данном случае будем рассматривать ее как особую форму записи списка опорных точек. Для агрегации этих точек и линейной регрессионной модели вычислим вышеуказанные статистики:

$$n = \sum_i w_i,$$

где w_i – эффективное количество точек данных, соответствующих этому правилу;

$$SX = \sum_i w_i \int_{-\infty}^{\infty} x p_A(x) dx = \sum_i w_i \mu_{xi}, \quad SY = \sum_i w_i \mu_{yi},$$

$$SXX = \sum_i w_i \left(\mu_{xi}{}^2 + \sigma_{xi}{}^2 \right),$$

$$SXY = \sum_i w_i \mu_{xi} \mu_{yi},$$

$$RSS = \sum_i w_i \left(\left(\mu_{yi} - a^* \mu_{xi} - b^* \right)^2 + \sigma_{xi}{}^2 + \sigma_{yi}{}^2 \right).$$

Параметры модели агрегированных данных вычисляются по формулам предыдущего раздела.

Таким образом, для агрегации системы нечетких правил с линейной регрессионной моделью необходимо знать статистический вес правил – эффективное количество точек данных, на которых оно было образовано.

Выводы. Вышеизложенные рассуждения приводят к следующим выводам.

Теорема. Знание статистик данных *n*, *SX*, *SY*, *SXX*, *SXY* и *RSS* является необходимым и достаточным для агрегации линейных регрессионных моделей 2-х переменных.

Доказательство. Данный список статистик является достаточным, так как через эти статистики выражаются все параметры линейной регрессионной модели. Он является необходимым, так как эти статистики – независимые линейные комбинации моментов 1-го и 2-го порядков набора данных, а эти моменты являются независимыми.

Следствие 1. Классических параметров *a* и *b* линейной регрессионной модели вида $y(x) = a \cdot x + b$ недостаточно для однозначной агрегации таких моделей.

Доказательство. Так как вышеперечисленные статистики необходимы для агрегации и являются независимыми, то двух параметров недостаточно для передачи всей информации о данных, содержащейся в модели.

Следствие 2. Возможна аддитивная агрегация линейных регрессионных моделей 2-х переменных в отсутствие полного объема данных, но при наличии вышеперечисленных статистик. Правила ее проведения указаны в соответствующем разделе статьи.

Следствие 3. Возможно аддитивное обновление модели новыми данными, путем вычисления по этим данным вышеуказанного перечня статистик и добавления их к статистикам модели.

Следствие 4. Возможна аддитивная агрегация линейных регрессионных моделей 2-х и системы нечетких правил, если возможно узнать значения вышеперечисленных статистик. Она проводится по тем же правилам, что и агрегация линейных регрессионных моделей.

Следствие 5. Необходимость вышеперечисленных статистик для агрегации некоторых моделей и возможность их записи в *аддитивном виде* указывают на возможность введения специального понятия «опыта».

Опыт (об исходных данных, лежащий в основе модели) – аддитивный объект, несущий в себе весь объем информации о данных необходимый для построения указанного типа моделей.

Переформулируем выводы с использованием этого понятия.

В работе показана недостаточность внутреннего содержания таких форм знания о данных как линейная регрессионная модель и система нечетких правил для дальнейшей агрегации этих знаний. Также продемонстрирована возможность построения объектов, представляющих «опыт» об исходных данных, лежащий в основе этих моделей. Наличие соответствующего «опыта» позволяет проводить аддитивную агрегацию таких моделей путем сложения «опыта», лежащего в их основе. А также аддитивно обновлять модель, будем простого добавления к имеющемуся в модели «опыту» новый «опыт», рассчитанный по новым данным.

Заключение. В данной работе рассматривается возможность агрегации моделей зависимостей, содержащихся в данных (т. е., по сути, возможность агрегации знаний об этих зависимостях) на примере агрегации линейных регрессионных моделей 2-х переменных и систем нечетких правил.

Была показана недостаточность внутренних параметров линейной регрессионной модели для однозначной агрегации таких моделей. Приведен необходимый и достаточный список статистик данных, позволяющих проводить однозначную агрегацию линейных регрессионных моделей 2-х переменных и систем нечетких правил без привлечения полного объема исходных данных. Более того, предложенный список статистик позволяет проводить аддитивную агрегацию этих моделей.

Для обозначения таких объектов как предложенный список статистик, а именно позволяющих проводить аддитивную агрегацию моделей предложено понятие «опыта» (об исходных данных, лежащих в основе модели).

Таким образом, знание опыта, лежащего в основе модели, позволяет проводить аддитивную агрегацию моделей данного типа, аддитивную агрегацию с моделями с аналогичной структурой опыта, аддитивное обновление модели новыми данными, путем извлечения из них опыта.

Перспективным представляется продолжить изложенные рассуждения на случай многомерной линейной регрессии и метода главных компонент. Для более сложных случаев, когда параметры оптимальной модели не выражаются в явном виде из данных, перспективным представляется анализ поведения оптимизируемой функции качества в окрестности оптимума.

С.М. Нечуйвітер

АГРЕГАЦІЯ ЛІНІЙНИХ РЕГРЕСІЙНИХ МОДЕЛЕЙ ТА СИСТЕМ НЕЧІТКИХ ПРАВИЛ

Розглядається можливість агрегації моделей залежностей, що містяться у даних, на прикладі агрегації лінійних регресійних моделей 2-х змінних та систем нечітких правил 2-х змінних. Наведено необхідний та достатній перелік характеристик даних, що дозволяє проводити однозначну адитивну агрегацію таких моделей без використання повного обсягу вихідних даних. Запропоновано поняття «досвіду» для позначення переліку характеристик, достатнього для проведення адитивної агрегації.

S.M. Nechuiviter

AGGREGATION OF LINEAR REGRESSION MODELS AND FUZZY RULE SYSTEMS

Possibility of aggregation of dependency models contained in data is considered on the example of aggregation of linear regression models of two variables and fuzzy rule systems with two variables. Necessary and sufficient set of data features is presented that allows to carry out a unique additive aggregation without full amount of output data. A concept of “experience” is proposed for designation of the set of features sufficient for additive aggregation description.

1. *Шор Я.Б.* Статистические методы анализа и контроля качества и надежности. – М.: Госэнергоиздат, 1962. – 552 с.
2. *Борисов В.В., Круглов В.В., Федулов А.С.* Нечеткие модели и сети. – М.: Горячая линия-Телеком, 2007. – 284 с.

Получено 25.10.2011

Об авторе:

Нечуйвитер Сергей Николаевич,
аспирант Физико-технического учебно-научного центра НАН Украины.
e-mail snechuiviter@gmail.com