ПРОГРАММНЫЙ КОМПЛЕКС АППРОКСИМАЦИЯ ДВУМЕРНЫХ ПЛОТНОСТЕЙ ВЕРОЯТНОСТИ

Д.А. Кудрявцев

Самарский государственный аэрокосмический университет им. академика С.П. Королёва (национально исследовательский университет)

В докладе рассматривается применение параметрической модели для определения вида и параметров двумерного распределения по имеющейся дискретной двумерной выборке. Предлагаемая методика основана на классификации законов распределения и определении параметров распределения с использованием нейронных сетей. Приводится описание программного комплекса, реализующего аппроксимацию двумерных плотностей вероятности.

Доклад посвящен исследованию двумерных плотностей вероятности через построение параметрической модели неизвестного непрерывного закона распределения и описанию программного комплекса, реализующего данный подход. При определении выражения неизвестной двумерной плотности вероятности с помощью параметрической модели по имеющейся выборке необходимо решить 2 задачи:

1. Определить вид параметрической модели, адекватно описывающей представленную выборку
2. Определить неизвестные параметры модели

Учитывая, что нейронные сети являются универсальным и широко используемым механизмом при решении задач аппроксимации и классификации, в рамках дипломного проекта рассматривается возможность использования нейронных сетей при построении параметрических моделей плотностей вероятности. Таким образом, поставленные задачи решаются в 4 этапа:

1. Преобразовать исследуемую дискретную двумерную выборку в вид, удобный для работы с нейронной сетью;
2. Определенить тип распределения исследуемой выборки с помощью многослойного персептрона;
3. Определенить параметры распределения по методу моментов с учетом известного типа распределения;
4. Определить параметры распределения методом нейронных сетей с использованием алгоритма обучения радиально-базисной сети.

Данный подход исследования двумерных распределений можно использовать для различных видов распределений, но для упрощения генерации исходных данных и отладки алгоритмов рассматривается параметрическая модель, заданная следующим способом:

$$\begin{cases} X = X_1, \\ Y = \sqrt{1-\rho^2} \cdot X_2 + \rho^2 \cdot X_1, \end{cases} \tag{1}$$

где X_1 и X_2 – независимые величины, ρ – коэффициент корреляции.

Ограничимся рассмотрением нормального и экспоненциального одномерных законов, а также закона Вейбулла.

В качестве сети определяющей вид распределения был выбран многослойный персептрон, так как данная сеть является одной из самых универсальных, позволяющей решать задачу классификации. Для удобства формализации задачи классификации будем представлять случайную последовательность в виде частотной двумерной гистограммы с разбиением на M_x коридоров по оси x и M_y коридоров по оси y, построенной по исследуемой выборке. Для восстановления аналитического выражения функции плотности вероятности из набора узловых точек используется градиентный алгоритм.

Для восстановления аналитического выражения функции плотности вероятности из набора узловых точек используется алгоритм нейросетевой аппроксимации. Для примера, в качестве базовой модели для определения коэффициентов нормального-нормального закона распределения вероятности, который был установлен на этапе классификации, берется *RBF*-сеть, нейроны скрытого слоя которой являются двумерными функциями Гаусса вида:

$$G(x,y)=e^{-\frac{(\nu_x x-c_x)^2}{2}-\frac{(\nu_y y-c_y)^2}{2}}. \tag{2}$$

Для нормального-экспоненциального закона распределения вероятности:

$$G(x,y)=e^{-\frac{(\nu_x x-c_x)^2}{2}-\lambda_y y}. \tag{3}$$

Для нормального-Вейбулла закона распределения вероятности:

$$G(x,y)=\frac{y^{k_y}}{\beta_y}\cdot e^{-\frac{(\nu_x x-c_x)^2}{2}-\frac{y^{k_y}}{\beta_y}}. \tag{4}$$

Эти функции используются для построения аппроксимирующей модели, которая выглядит следующим образом:

$$\hat{f}(x,y)=w\cdot G(x,y). \tag{5}$$

Неизвестные коэффициенты w, ν_x, ν_y, c_x, c_y, β_y, k_y, λ_y в выражениях (2-4) являются настраиваемыми, а их значения определяются в процессе обучения нейронной сети.

Таким образом, целевая функция определяется по формуле:

$$E=\frac{1}{2}\sum_{i=0}^{M_x-1}\sum_{j=0}^{M_y-1}(\hat{f}(x_i,y_j)-f(x_i,y_j))^2\to\min. \tag{6}$$

Приведем формулы вычисления параметров для нормального-нормального закона распределения вероятности:

$$\begin{aligned}
\nu_x&=\nu_x+\eta\cdot w\cdot G(x,y)\cdot(G(x,y)-f(x_i,y_j))(\nu_x x_i-c_x)x_i,\\
\nu_y&=\nu_y+\eta\cdot w\cdot G(x,y)\cdot(G(x,y)-f(x_i,y_j))(\nu_y y_j-c_y)y_j,\\
c_x&=c_x-\eta\cdot w\cdot G(x,y)\cdot(G(x,y)-f(x_i,y_j))(\nu_x x_i-c_x),\\
c_y&=c_y-\eta\cdot w\cdot G(x,y)\cdot(G(x,y)-f(x_i,y_j))(\nu_y y_j-c_y).
\end{aligned} \tag{7}$$

Для нормального-экспоненциального:

$$\begin{aligned}
\nu_x&=\nu_x+\eta\cdot w\cdot G(x,y)\cdot(G(x,y)-f(x_i,y_j))(\nu_x x_i-c_x)x_i,\\
c_x&=c_x-\eta\cdot w\cdot G(x,y)\cdot(G(x,y)-f(x_i,y_j))(\nu_x x_i-c_x),\\
\lambda_y&=\lambda_y+\eta\cdot w\cdot G(x,y)\cdot(G(x,y)-f(x_i,y_j))y_j.
\end{aligned} \tag{8}$$

Для нормального-Вейбулла:

$$\begin{aligned}
\nu_x&=\nu_x+\eta\cdot w\cdot G(x,y)\cdot(G(x,y)-f(x_i,y_j))(\nu_x x_i-c_x)x_i,\\
c_x&=c_x-\eta\cdot w\cdot G(x,y)\cdot(G(x,y)-f(x_i,y_j))(\nu_x x_i-c_x),\\
\beta_y&=\beta_y-\eta\cdot w\cdot G(x,y)\cdot(G(x,y)-f(x_i,y_j))\frac{y_j^{k_y}}{\beta_y^{2}},\\
k_y&=k_y-\eta\cdot w\cdot G(x,y)\cdot(G(x,y)-f(x_i,y_j))\ln(y_j)(1-\frac{y_j^{k_y}}{\beta_y}).
\end{aligned} \tag{9}$$

Оценка погрешности аппроксимации вычисляется по формуле:

$$\Delta=\iint_D(f(x,y)-\hat{f}(x,y))^2dxdy \tag{10}$$

Программный комплекс позволяет моделировать случайные величины, задавая различные параметры распределения, создавать обучающую выборку, а так же отображать гистограмму. Исследование выборки состоит в том, что необходимо загрузить обучающую выборку, задать исследуемую выборку, классифицировать выборку. После определения типа распределения, необходимо определить параметры с использованием метода моментов или градиентного алгоритма.

Для исследования погрешности было проведено 29 испытаний [2]. Также, полагая величину погрешности случайной величиной, распределенной по нормальному закону, величина погрешности оценивается по правилу «трёх сигма». В каждом из испытаний строилась гистограмма 15 на 15 столбцов и моделировалась случайная величина, состоящая из 10000 отсчетов и распределенная по двумерному нормальному закону. Результаты исследования представлены в таблице 1.

Первая строка – погрешность, полученная с использованием методов описанных в статье, вторая строка – результаты исследования погрешности аппроксимации RBF-сетью [3].

Таблица 1 – Результаты исследования методов.

	СКП			
	m_x	σ_x	$m_x + 3\sigma_x$	*max*
1	0,0137	0,0038	0,0243	0,0246
2	0,0144	0,0008	0,0169	0,0160
3	0,0116	0,0035	0,0221	0,0224
4	0,0231	0,0029	0,0318	0,0309

Третья строка – результаты, описанные в [4], последняя строка – результат использования метода моментов.

В таблице m_x представляет выборочное среднее погрешности аппроксимации, рассчитанное по 29 значениям, σ_x – корень выборочной дисперсии, *max* – максимальная величина погрешности среди проведенных испытаний.

Таким образом, по результатам испытаний, которые приведены в таблице 1 можно сделать вывод о том, что метод, описанный в докладе имеет меньшую погрешность, чем метод моментов, однако большую, чем метод, основанный на аппроксимации RBF-сетью [3] и метод используемый для двумерной независимой выборки [4].

Литература

1. Осовский, С. Нейронные сети для обработки информации [Текст] / Осовский С.: Пер. с польского И.Д. Рудинского. – М.: Финансы и статистика, 2002. – 344 с.
2. Лёзин И.А. «Автоматизированный комплекс аппроксимативного анализа двумерных законов распределения ортогональными полиномами и нейронными сетями» // Информационные технологии в высшем профессиональном образовании: Сборник докладов II межрегиональной научно-практической конференции / Под.ред. О.А. Тарабрина, А.В. Очеповского / Самарский государственный аэрокосмический университет – Тольятти-Самара, 2007. – С.84-87. – Библиогр.: с.87.
3. Лёзин И.А., Прохоров С.А., Лёзина И.В. «Аппроксимация двумерной плотности вероятности ортогональными полиномами»// Радиотехника и связь. Материалы четвертой международной научно-технической конференции / Саратовский государственный технический университет – Саратов, 2007. – С. 17-22. – Библиогр.: с.22.
4. Методы нормирования метрологических характеристик, оценки и контроля характеристик погрешностей средств статистических измерений. РТМ 25139-74 Текст. // Минприбор, 1974. 76 с.
5. Лёзин И.А. «Анализ погрешности измерения геометрических размеров лопаток газотурбинных двигателей» // Труды Международной конференции с элементами научной школы для молодежи. – Самара, 2010. – С. 95-99. – Библиогр.: с.99. - ISBN 978-5-7883-0851-7