

ОБРАБОТКА И ПРОГНОЗ ДАННЫХ ОБРАЗОВАТЕЛЬНОГО ПРОЦЕССА НА ОСНОВЕ НЕЧЕТКОГО РЕГРЕССИОННОГО АНАЛИЗА

О.М. Полещук[✉], Е.Г. Комаров, Н.Г. Поярков

ФГБОУ ВО «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)», Россия, 105005, г. Москва, ул. 2-я Бауманская, д. 5, стр. 1

polshchuk@mgul.ac.ru

Представлены показатели качества нечетких регрессионных моделей, предназначенных для исследования зависимостей между качественными характеристиками образовательного процесса и для прогноза их значений, а также модель распознавания нечетких значений выходных характеристик регрессий. Приведен алгоритм выбора нечеткой регрессионной модели из линейной и нелинейной моделей на основе показателей их качества. Проведен анализ степеней влияния входных характеристик на выходную характеристику. Построена нечеткая регрессионная модель для прогноза успешности защиты диссертации при поступлении соискателя в аспирантуру и для исследования зависимостей между входными характеристиками соискателя и выходной характеристикой. Альтернативный подход к построению регрессионных моделей на основе нечисловых данных образовательного процесса позволяет не накладывать некорректные условия на исходные данные, считая их значениями случайных величин, и не использовать некорректные арифметические операции для элементов порядковых шкал.

Ключевые слова: образовательный процесс, нечеткая информация, нечеткая регрессионная модель, лингвистическая переменная

Ссылка для цитирования: Полещук О.М., Комаров Е.Г., Поярков Н.Г. Обработка и прогноз данных образовательного процесса на основе нечеткого регрессионного анализа // Лесной вестник / Forestry Bulletin, 2024. Т. 28. № 3. С. 133–140. DOI: 10.18698/2542-1468-2024-3-133-140

Многолетние исследования показали, что для обработки данных образовательного процесса, их анализа и прогноза недостаточно применения классических математических методов, поскольку они учитывают только неопределенность случайного характера и не учитывают нечеткость, которая появляется вследствие активного влияния человеческого фактора на процессы оценивания и управления [1–7].

Этап формализации полученной информации при моделировании процессов сферы образования имеет существенное значение и позволяет в дальнейшем применять адекватный математический аппарат для ее анализа. Если полученные данные объективно не являются значениями случайных величин, то очевидно, что исследователи будут искать альтернативные методы их формализации и обработки.

Соответственно, продолжая применять для анализа и прогноза характеристик образовательного процесса методы классического регрессионного анализа, мы априори закладываем ошибки, которые только накапливаются, начиная от этапа формализации и заканчивая этапом распознавания выходных данных.

Для того чтобы избежать накопления ошибок, исследователи на этапе формализации стали использовать нечеткие множества и лингвистические переменные, тем самым моделируя мыслительную деятельность лиц, выполняющих оценку характеристик, и учитывая субъективную составляющую в полученных данных [8–19].

Использование лингвистических переменных исключило некорректность операций с элементами порядковых шкал, которые традиционно использовались для оценки качественных (нечисловых) характеристик образовательного процесса. Поскольку большинство оцениваемых характеристик является качественным, то значение лингвистических переменных на этапе формализации достаточно велико. Альтернативный подход к формализации данных образовательного процесса обусловил развитие нечеткого регрессионного анализа и расширил его практическое использование не только в образовании, но и в других областях деятельности человека [20–28].

Разработка линейных и нелинейных нечетких регрессионных моделей не устранила пробелы, возникающие при выборе моделей, оценке их качества, распознавании выходной информации и практическом применении нечетких регрессионных моделей при решении задач образовательного процесса.

Цель работы

Цель работы — разработка показателей качества нечетких регрессионных моделей, алгоритма выбора вида регрессионной модели и модели распознавания выходных нечетких данных.

Основные понятия и определения

Нечетким числом \tilde{A} называется множество пар вида $\{(x, \mu_A(x)) : x \in X\}$, при этом значение функции принадлежности $\mu_A(x)$ называется степенью принадлежности x к \tilde{A} , R — подмножество действительной прямой, $\mu_A(x) : R \rightarrow [0, 1]$ [29].

Непрерывным нечетким числом называется нечеткое число с непрерывной функцией принадлежности. Дискретным нечетким числом называется нечеткое число с дискретной функцией принадлежности.

Лингвистической переменной называется множество $\{X, T(X), U, V, S\}$, где X — название переменной; $T(X) = \{X_l, l = 1, m\}$ — терм-множество переменной X или множество термов или названий лингвистических значений переменной X ; V — синтаксическое правило, порождающее названия значений лингвистической переменной X ; S — семантическое правило, которое ставит в соответствие каждому терму из множества $T(X)$ нечеткое подмножество универсального множества U [1].

В работе [30] дано определение агрегирующего отрезка $[c_1, c_2]$ для нечеткого числа $\tilde{A} = (a_1, a_2, a_L, a_R)$:

$$c_1 = \int_0^1 \frac{2a_1 - (1-\alpha)a_L}{2} 2\alpha d\alpha = a_1 - \frac{1}{6}a_L,$$

$$c_2 = \int_0^1 \frac{2a_1 - (1-\alpha)a_R}{2} 2\alpha d\alpha = a_1 - \frac{1}{6}a_R.$$

В работе [10] определено расстояние $\rho(\tilde{A}, \tilde{B})$ между нечеткими числами $\tilde{A} = (a_1, a_2, a_L, a_R)$, $\tilde{B} = (b_1, b_2, b_L, b_R)$ на основе их агрегирующих отрезков $[c_1, c_2]$, $[d_1, d_2]$

$$\rho(\tilde{A}, \tilde{B}) = \sqrt{(c_1 - d_1)^2 + (c_2 - d_2)^2}.$$

Наиболее известными и широко применяемыми нечеткими регрессионными моделями являются линейная и нелинейная регрессионные модели.

Линейная нечеткая регрессионная модель для нечетких множеств имеет следующий вид [31]

$$\tilde{Y} = \tilde{a}_0 + \tilde{a}_1 \tilde{X}_1 + \dots + \tilde{a}_n \tilde{X}_n.$$

Входными и выходными данными модели являются нечеткие числа. Предполагается, что неизвестные коэффициенты регрессионной модели являются треугольными нечеткими числами, которые определяются тремя параметрами. Первый параметр — абсцисса вершины треугольника,

второй и третий параметры соответственно длины левого и правого крыльев треугольника.

Нелинейная нечеткая регрессионная модель для нечетких множеств имеет следующий вид [27]

$$\begin{aligned} \tilde{Y} = & \tilde{a}_0 + \tilde{a}_1 \tilde{X}_1^2 + \dots + \tilde{a}_m \tilde{X}_m^2 + \\ & + \tilde{a}_{m+1} \tilde{X}_1 \tilde{X}_2 + \dots + \tilde{a}_{mm} \tilde{X}_{m-1} \tilde{X}_m + \\ & + \tilde{a}_{m(m+1)} \tilde{X}_1 + \dots + \tilde{a}_{m(m+1)} \tilde{X}_m. \end{aligned}$$

Предполагается, что неизвестные коэффициенты нелинейной регрессионной модели являются треугольными нечеткими числами.

Постановка задачи и ее решение

При прогнозе значений нечисловых характеристик образовательного процесса на основе нечеткого регрессионного анализа существенное значение имеет качество построенной модели. В классическом регрессионном анализе числовыми показателями качества регрессионной модели служат коэффициент детерминации R^2 и стандартная ошибка S [32]

$$R^2 = \frac{\sum_{j=1}^n (y_{jm} - \bar{y})^2}{\sum_{j=1}^n (y_j - \bar{y})^2},$$

$$S = \sqrt{\frac{1}{n-m-1} \sum_{j=1}^n (y_j - y_{jm})^2},$$

где $y_j, j = \overline{1, n}$ — исходные выходные данные;

$y_{jm}, j = \overline{1, n}$ — модельные выходные данные,

$$\bar{y} = \frac{1}{n} \sum_{j=1}^n y_j.$$

Для нечетких регрессионных моделей определим аналоги коэффициента детерминации и стандартной ошибки, используя агрегирующие отрезки нечетких данных [33, 34].

Обозначим через $[y_{1j}, y_{2j}], j = \overline{1, n}$ агрегирующие отрезки исходных выходных нечетких чисел $\tilde{Y}_j, j = \overline{1, n}$, через $[y_{1mj}, y_{2mj}], j = \overline{1, n}$ агрегирующие отрезки модельных выходных нечетких чисел $\tilde{Y}_{jm}, j = \overline{1, n}$, через $[y_1, y_2]$ агрегирующий отрезок среднего исходного значения \tilde{Y} , через $\rho(\tilde{Y}_j, \tilde{Y}_{jm}) = \sqrt{(y_{1j} - y_{1mj})^2 + (y_{2j} - y_{2mj})^2}$ расстояние между исходными и модельными выходными нечеткими числами, через $\rho(\tilde{Y}_i, \tilde{Y}) = \sqrt{(y_{1j} - y_1)^2 + (y_{2j} - y_2)^2}$ расстояние между исходным выходным нечетким числом и средним исходным выходным нечетким числом, через $\rho(\tilde{Y}_{jm}, \tilde{Y}) = \sqrt{(y_{1mj} - y_1)^2 + (y_{2mj} - y_2)^2}$

расстояние между модельным выходным и средним исходным выходным нечеткими числами. Тогда аналогами коэффициента детерминации и стандартной ошибки соответственно назовем следующие величины

$$\hat{R}^2 = \frac{\sum_{j=1}^n \rho^2(\tilde{Y}_{jm}, \tilde{Y})}{\sum_{j=1}^n \rho^2(\tilde{Y}_j, \tilde{Y})},$$

$$\hat{S} = \sqrt{\frac{1}{n-m-1} \sum_{j=1}^n \rho^2(\tilde{Y}_j, \tilde{Y}_{jm})}.$$

Определенные показатели качества нечетких регрессионных моделей позволяют выбирать между линейной и нелинейной моделью в пользу той, у которой показатели лучше.

Пусть Y — выходная качественная характеристика, формализованная в виде лингвистической переменной \tilde{Y} со значениями $\tilde{Y}_l = (a_{l1}, a_{l2}, a_{lL}, a_{lR})$, $l = 1, p$, соответствующими уровням Y_l , $l = 1, p$ шкалы, используемой для ее оценки, \tilde{X}_i , $i = 1, m$ — формализованные с помощью лингвистических переменных входные характеристики.

При получении выходного нечеткого числа $\tilde{Y}_m = (y_1, y_2, y_L, y_R)$ регрессионной модели возникает проблема его распознавания [35].

Поскольку выходная характеристика является качественной (нечисловой), то необходимо выходную информацию идентифицировать с одним из лингвистических значений Y_l , $l = 1, p$, формализованных с помощью нечетких чисел $\tilde{Y}_l = (a_{l1}, a_{l2}, a_{lL}, a_{lR})$, $l = 1, p$. Таким образом, нечеткое число $\tilde{Y}_m = (y_1, y_2, y_L, y_R)$ необходимо идентифицировать с одним из нечетких чисел $\tilde{Y}_l = (a_{l1}, a_{l2}, a_{lL}, a_{lR})$, $l = 1, p$. Для этого найдем агрегирующие отрезки $[y_{1m}, y_{2m}]$, $[y_{1l}, y_{2l}]$, $l = 1, p$ нечетких чисел \tilde{Y}_m , \tilde{Y}_l , $l = 1, p$ и расстояния

$$\rho(\tilde{Y}_m, \tilde{Y}_l) = \sqrt{(y_{1m} - y_{1l})^2 + (y_{2m} - y_{2l})^2}, l = 1, p.$$

Если $\rho(\tilde{Y}_m, \tilde{Y}_k) = \min \rho(\tilde{Y}_m, \tilde{Y}_l)$, $l = 1, p$, то выходное значение регрессионной модели \tilde{Y}_m идентифицируется со значением Y_k характеристики Y .

Примеры

Рассмотрим применение нечетких регрессионных моделей для исследования влияния аудиторного и внеаудиторного видов контроля знаний на итоговую оценку по учебной дисциплине.

В качестве примера было исследовано влияние аудиторного и внеаудиторного видов контроля знаний на итоговую оценку по математике. Как аудиторный контроль знаний рассмотрено выполнение контрольной работы по теме «Пределы и дифференциальное исчисление». Как внеаудиторный контроль знаний рассмотрено

самостоятельное выполнение расчетно-графической работы по этой же теме. Формализация исходных данных была выполнена на основе лингвистических переменных, значениями которых являются нечеткие переменные. Оценки знаний студентов, во-первых, объективно не являются значениями случайных величин, а, во-вторых, все арифметические операции с оценками, измеренными в порядковой шкале, некорректны.

Итоговые оценки знаний по математике, измеренные по шкале «неудовлетворительно», «удовлетворительно», «хорошо», «отлично», формализованы на основе лингвистической переменной. Аналогично формализованы оценки за контрольную и расчетно-графическую работы.

В качестве значений входных переменных нечеткой регрессионной модели \tilde{X}_1, \tilde{X}_2 рассмотрены формализованные оценки за контрольную и расчетно-графическую работы по математике. Как значения выходной переменной \tilde{Y} рассмотрены формализованные итоговые оценки по математике.

Построены две нечеткие регрессионные модели:

$$\tilde{Y} = (-0,11, 0,06, 0,05) + (1,21, 0,02, 0,01)\tilde{X}_1 + (0,05, 0,004, 0), \tilde{X}_2;$$

$$\tilde{Y} = (-0,29, 0,03, 0,07) + (1,08, 0,04, 0,03)\tilde{X}_1^2 + (0,03, 0,02, 0,02), \tilde{X}_2^2 + (-0,63, 0,05, 0,01)\tilde{X}_1\tilde{X}_2 + (0,9, 0,01, 0)\tilde{X}_1 + (0,07, 0, 0,002)\tilde{X}_2.$$

Для каждой модели найдены аналоги коэффициента детерминации и стандартной ошибки. Для линейной модели — $\hat{R}^2 = 0,636$, $\hat{S} = 0,183$, для нелинейной модели — $\hat{R}^2 = 0,429$, $\hat{S} = 0,217$.

Для дальнейшего анализа была выбрана нечеткая линейная модель, имеющая лучшие показатели качества.

Поскольку все коэффициенты регрессионной модели являются нечеткими числами, то для сравнительного анализа коэффициенты при входных переменных были дефаззифицированы по методу центра тяжести [10]

$$a_i = \frac{\int_{-\infty}^{+\infty} x \mu_i(x) dx}{\int_{-\infty}^{+\infty} \mu_i(x) dx}, i = 1, 2,$$

где $\mu_i(x)$, $i = 1, 2$ — функции принадлежности нечетких коэффициентов \tilde{a}_i , $i = 1, 2$ регрессионной модели.

Полученные результаты $a_1 = 1,196$, $a_2 = 0,049$ свидетельствуют о существенном влиянии аудиторного контроля знаний студентов на итоговую оценку по математике, подтверждают существенную объективность аудиторного контроля знаний, поскольку, находясь под контролем преподавателя, студенты не пользуются дополнительными источниками, не списывают решения друг у друга, не находят похожие или в точности такие же задачи в интернете, после чего бездумно копируют их.

Рассмотрим применение нечетких регрессионных моделей для прогноза успешности защиты диссертации аспирантом.

Итоговой аттестацией аспиранта согласно приказу Минобрнауки России № 951 от 20.10.2021 (с изм. от 06.05.2022) является оценка диссертации на предмет ее соответствия критериям, установленным согласно Федеральному закону Российской Федерации от 23 августа 1996 г. № 127-ФЗ «О науке и государственной научно-технической политике». Переход обучения в аспирантуре из образовательного формата в научный обеспечил переход с ФГОС (федеральные государственные образовательные стандарты) на ФГТ (федеральные государственные требования). Это изменение направлено на повышение качества подготовки аспирантов, поскольку фактом успешного завершения их обучения является не сдача образовательных дисциплин и получения удостоверения, а защита диссертации или практическая готовность диссертации к защите.

В связи с этими изменениями большое значение имеет оценка научно-исследовательской деятельности поступающего в аспирантуру и прогноз защиты диссертации в положенные сроки.

Для построения регрессионной модели в целях прогноза успешности защиты диссертации аспирантом необходимо определить систему входных характеристик, которая зависит от научной специальности и формируется на начальном этапе анализа. Сложность этого анализа состоит в исследовании разнородных характеристик, подавляющее большинство которых являются качественными (нечисловыми). Поскольку, как было сказано ранее, для значений качественных характеристик некорректны все арифметические операции, то формализация их значений осуществляется на основе лингвистических переменных. Для выявления системы входных характеристик, влияющих на успешность защиты диссертации, применяется модель, разработанная в работе [18].

Предположим, что в систему входных характеристик включены характеристики X_i , $i = 1, m$, которыми, как правило, являются: успеваемость

по основным дисциплинам, рейтинговая оценка в рамках этих дисциплин, успешность студенческой научно-исследовательской работы, логичность мышления, дисциплинированность, исполнительность, лидерство. Каждая из характеристик X_i , $i = 1, m$ формализуется с помощью лингвистической переменной, термы которой соответствуют уровням X_{il} , $l = 1, m_l$, $i = 1, m$ шкалы, используемой для ее оценки. Таким образом, входным характеристикам X_i , $i = 1, m$ поставлены в соответствие лингвистические переменные \tilde{X}_i , $i = 1, m$, значениями которых являются нечеткие множества \tilde{X}_{il} , $l = 1, m_l$, $i = 1, m$. Для примера были выбраны три входные характеристики X_i , $i = 1, 3$: X_1 — «Рейтинговая оценка по основным дисциплинам», X_2 — «Успешность студенческой научно-исследовательской работы», X_3 — «Логичность мышления». Для оценки характеристик X_i , $i = 1, 3$ использовались шкалы соответственно с уровнями X_{11} = «Удовлетворительно», X_{12} = «Хорошо», X_{13} = «Отлично», X_{21} = «Не успешно», X_{22} = «Достаточно успешно», X_{23} = «Успешно», X_{31} = «Низкая», X_{32} = «Средняя», X_{33} = «Высокая», X_{34} = «Очень высокая». Все шкалы формализованы с помощью лингвистических переменных \tilde{X}_i , $i = 1, 3$, а уровни с помощью нечетких множеств \tilde{X}_{1l} , $l = 1, 3$, \tilde{X}_{2l} , $l = 1, 3$, \tilde{X}_{3l} , $l = 1, 4$.

Выходная характеристика Y измеряется в шкале с уровнями Y_k , $k = 1, 3$: Y_1 = «Не успешно» (диссертация не представлена), Y_2 = «Достаточно успешно» (диссертация завершена на 75 %), Y_3 = «Успешно» (диссертация готова к защите). Уровни шкалы Y_k , $k = 1, 3$ формализованы с помощью значений Y_k , $k = 1, 3$ лингвистической переменной

$$\tilde{Y}_1 = (0, 0, 3, 0, 0, 2), \tilde{Y}_2 = (0, 5, 0, 8, 0, 2, 0, 1), \\ \tilde{Y}_3 = (0, 9, 1, 0, 1, 0).$$

Для линейной и нелинейной моделей найдены аналоги коэффициента детерминации и стандартной ошибки. Для линейной модели — $\hat{R}^2 = 0,782$, $\hat{S} = 0,129$, для нелинейной модели — $\hat{R}^2 = 0,365$, $\hat{S} = 0,253$, поэтому для анализа была выбрана линейная модель

$$\tilde{Y} = (-1,34, 0,15, 0,07) + (0,83, 0,04, 0,08) \tilde{X}_1 + \\ + (1,36, 0,02, 0,05) \tilde{X}_2 + (0,91, 0,17, 0,09) \tilde{X}_3.$$

Как видно из модели, наибольший коэффициент имеет характеристика успешности студенческой научно-исследовательской работы, поэтому при отборе кандидатов в аспирантуру следует обращать на эту характеристику пристальное внимание, как на оказывающую существенное влияние на успешность защиты диссертации аспирантом.

Для того чтобы спрогнозировать успешность защиты диссертации, необходимо выходное модельное число идентифицировать с одним из лингвистических значений Y_k , $k = 1, 3$: $Y_1 = \langle \text{«Не успешно»}$, $Y_2 = \langle \text{«Достаточно успешно»}$, $Y_3 = \langle \text{«Успешно»}$. Пусть выходным значением является нечеткое число $\tilde{Y}_m = (0,6, 0,75, 0,1, 0,1)$.

Найдены агрегирующие отрезки $[y_{1k}, y_{2k}]$, $k = 1, 3$ для нечетких чисел $\tilde{Y}_1 = (0, 0,3, 0,02)$, $\tilde{Y}_2 = (0,5, 0,8, 0,2, 0,1)$, $\tilde{Y}_3 = (0,9, 1,0, 1,0)$ и агрегирующий отрезок $[y_{1m}, y_{2m}]$ для нечеткого числа $\tilde{Y}_m = (0,6, 0,75, 0,1, 0,1)$: $[y_{11}, y_{21}] = [0, 0,33]$, $[y_{12}, y_{22}] = [0,47, 0,82]$, $[y_{13}, y_{23}] = [0,88, 1]$, $[y_{1m}, y_{2m}] = [0,58, 0,77]$.

Найдены квадраты расстояний

$$\rho^2(\tilde{Y}_m, \tilde{Y}_k) = \sqrt{(y_{1m} - y_{1k})^2 + (y_{2m} - y_{2k})^2},$$

$$k = \overline{1, 3}: \rho^2(\tilde{Y}_m, \tilde{Y}_1) = 0,53,$$

$$\rho^2(\tilde{Y}_m, \tilde{Y}_2) = 0,015, \rho^2(\tilde{Y}_m, \tilde{Y}_3) = 0,14.$$

Прогнозом успешности защиты диссертации является «Достаточно успешно» (диссертация завершена на 75 %).

Выводы

Определены показатели качества нечетких регрессионных моделей, позволяющие осуществлять выбор лучшей модели для анализа качественных характеристик образовательного процесса и прогноза их значений.

Разработана модель идентификации выходного нечеткого значения модели с одним из уровней лингвистической шкалы, используемой для оценки выходной характеристики.

Значения входных и выходных характеристик нечетких регрессионных моделей являются значениями лингвистических переменных.

На примере показан алгоритм выбора нечеткой регрессионной модели из двух на основе показателей их качества. Проанализированы степени влияния входных характеристик на выходную характеристику.

Построена линейная нечеткая регрессионная модель для прогноза успешности защиты соискателем диссертации и для исследования зависимостей между входными характеристиками соискателя и выходной характеристикой.


Развитие нечеткого регрессионного анализа и его применение для обработки информации образовательного процесса позволяет учитывать нечеткость исходных данных, не накладывая на них некорректные условия, считая значениями случайных величин, корректно оперировать с ними, тем самым избегая накопления ошибок и снижая риски принятия ошибочных управляющих решений по результатам итоговых данных.

Список литературы

- [1] Заде Л.А. Понятие лингвистической переменной и его применение к принятию приближенных решений. М.: Мир, 1976. 165 с.
- [2] Voskoglou M.G. Fuzzy Logic as a Tool for Assessing Students' Knowledge and Skills // *Education Science*, 2013, v. 3(2), pp. 208–221.
- [3] Lin C.T., Lee C.S. Neural Network based Fuzzy Logic Control and Decision System // *IEEE Transactions on Comput*, 1991, v. 40, no. 12, pp. 1320–1336.
- [4] Mendes R. R., Voznika F. D., Freitas, A. A., Nievola J. C. Discovering fuzzy classification rules with genetic programming and co-evolution // *Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery*, Freiburg, Germany, September 3–5, 2001, pp. 314–325.
- [5] Poleshchuk O., Komarov E. The determination of students' fuzzy rating points and qualification levels // *1st International Fuzzy Systems Symposium — FUZZYSS'2009*, 1–2 October, 2009, Ankara, Turkey, 2009, pp. 218–224.
- [6] Ilahi R., Widiaty I., Gafar A. Abdullah Fuzzy system application in education // *IOP Conference Series Materials Science and Engineering*, 2018, v. 434 (1), p. 012308.
- [7] Darwish A., Poleshchuk O. Fuzzy Models for Educational Data Mining // *J. of Telecommunications*, 2012, v. 15, no. 2, pp. 8–22.
- [8] Ruspini E.H. A new approach to clustering // *Information and Control*, 1969, v. 15, pp. 22–32.
- [9] Tamura S., Higuchi S., Tanaka K. Pattern classification based on fuzzy relations // *IEEE Transactions on Systems Man and Cybernetics SMC1*, 1971, no.1, pp. 61–66. DOI:10.1109/TSMC.1971.5408605
- [10] Zadeh L.A. Similarity relations and fuzzy orderings // *Information Sciences*, 1971, v. 3, pp. 177–200.
- [11] Hwang C.L., Lin N.J. Group decision making under multiple criteria. Berlin: Springer, 1987. 400 p.
- [12] Dunn J.C. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters // *J. of Cybernetics*, 1973, v. 3, pp. 32–57.
- [13] Ruspini E.H. Recent developments in fuzzy clustering // *Fuzzy Set and Possibility Theory*. N.Y.: Pergamon Press, 1982, pp. 133–146.
- [14] Ryjov A.P. The Concept of a Full Orthogonal Semantic Scope and the Measuring of Semantic Uncertainty // *Fifth International Conference Information Processing and Management of Uncertainty in Knowledge-Based Systems*, Iran, 4–7 December, 1994, pp. 33–34.
- [15] Bezdek J.C. Selected applications in classifier design // *Pattern recognition with fuzzy objective function algorithms*, 1981, v. 2, pp. 203–239.
- [16] Dubois D., Prade H. Ranking Fuzzy Numbers in Setting of Possibility Theory // *Information Science*, 1983, v. 30, pp. 183–224.
- [17] Borisov A.N., Krumberg O.A., Fedorov I.P. Decision making on the basis of fuzzy models: Examples of use. Riga: Zinatne, 1990. 184 p.
- [18] Ryjov A. Fuzzy Linguistic Scales: Definition, Properties and Applications. *Soft Computing in Measurement and Information Acquisition. Studies in Fuzziness and Soft Computing*, 2003, v. 127, pp. 2–12.
- [19] Полещук О.М. Кластерный анализ экспертной информации на основе Z-чисел // *Лесной вестник / Forestry Bulletin*, 2022. Т. 26. № 1. С. 143–148. DOI: 10.18698/2542-1468-2022-1-143-148
- [20] Sabic D.A., Pedrycz W. Evaluation on fuzzy linear regression models // *Fuzzy Sets and Systems*, 1991, v. 39, pp. 51–63.

- [21] Tanaka H., Ishibuchi H., Yoshikawa S. Exponential possibility regression analysis // *Fuzzy Sets and Systems*, 1995, v. 69, pp. 305–318.
- [22] Chang Y.-H.O. Hybrid fuzzy least-squares regression analysis and its reliability measures // *Fuzzy Sets and Systems*, 2001, v. 119, pp. 225–246.
- [23] Domrachev V.G., Poleshuk O.M. A regression model for fuzzy initial data // *Automation and Remote Control*, 2003, v. 64, no. 11, pp. 1715–1724.
- [24] Yager R.R., Filev D.P. On the issue of defuzzification and selection based on a fuzzy set // *Fuzzy Sets and Systems*, 1993, v. 55, pp. 255–272.
- [25] Алтунин А.Е., Семухин М.В. Модели и алгоритмы принятия решений в нечетких условиях. Тюмень.: Изд-во Тюменского государственного университета, 2002. 268 с.
- [26] Liu F., Mendel J.M. Encoding words into interval Type-2 fuzzy sets using an interval approach // *IEEE Trans. Fuzzy Systems*, 2008, v. 16, no. 6, pp. 187–201.
- [27] Runkler T.A., Katz C. Fuzzy clustering by particle swarm optimization // *Proceedings of the IEEE International Conference on Fuzzy Systems*, Oslo, 1–4 November, 2006. Berlin, 2006, pp. 34–41.
- [28] Phyo O., Chaw E. Comparative Study of Fuzzy PSO (FPSO) Clustering Algorithm and Fuzzy C-Means (FCM) Clustering Algorithm // *National J. of Parallel and Soft Computing*, 2019, v. 1, no. 1, pp. 62–67.
- [29] Zadeh L.A. Fuzzy logic and approximate reasoning // *Synthese*, 1975, v. 80, pp. 407–428.
- [30] Aliev R., Guirimov B.: Z-number clustering based on general Type-II fuzzy sets // *Advances in Intelligent Systems and Computing*, 2018, v. 896, pp. 270–278.
- [31] Jamal M., Khalif K., Mohamad S. The implementation of Z-numbers in fuzzy clustering algorithm for wellness of chronic kidney disease patients // *J. of Physics: Conference Series*, 2018, v. 1366, pp. 201–209.
- [32] Zadeh L.A. A note on Z-numbers // *Inf. Sci.*, 2011, v. 14(181), pp. 2923–2932. DOI: 10.1016/j.ins.2011.02.022
- [33] Комаров Е.Г., Полещук О.М., Поярков Н.Г. Изучение взаимосвязей между качественными признаками при нечеткой исходной информации // *Обозрение прикладной и промышленной математики*, 2005. Т. 12. Вып. 4. С. 992–993.
- [34] Darwish A., Poleshchuk O., Komarov E. A new fuzzy linear regression model for a special case of interval type-2 fuzzy sets // *Applied Mathematics & Information Sciences*, 2016, v. 10, no. 3, pp. 1209–1214.
- [35] Poleshchuk O.M. Formalization, Prediction and Recognition of Expert Evaluations of Telemetric Data of Artificial Satellites Based on Type-II Fuzzy Sets // *Machine Learning and Data Mining in Aerospace Technology. Studies in Computational Intelligence*, 2020, v. 836, pp. 39–64.

Сведения об авторах

Полещук Ольга Митрофановна  — д-р. техн. наук, профессор, зав. кафедрой «Высшая математика и физика», ФГБОУ ВО «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)» (Мытищинский филиал), poleshchuk@mgul.ac.ru

Комаров Евгений Геннадиевич — д-р. техн. наук, профессор, зав. кафедрой «Информационно-измерительные системы и технологии приборостроения», ФГБОУ ВО «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)» (Мытищинский филиал), komarov@mgul.ac.ru

Поярков Николай Геннадьевич — канд. техн. наук, доцент, декан Космического факультета, ФГБОУ ВО «Московский государственный технический университет имени Н.Э. Баумана (национальный исследовательский университет)» (Мытищинский филиал), royarkov@mgul.ac.ru

Поступила в редакцию 13.11.2023.

Одобрено после рецензирования 22.03.2024.

Принята к публикации 27.03.2024.

PROCESSING AND PREDICTION OF EDUCATIONAL PROCESS DATA BASED ON FUZZY REGRESSION ANALYSIS

O.M. Poleshchuk✉, E.G. Komarov, N.G. Poyarkov

BMSTU (Mytishchi branch), 1, 1st Institutskaya st., 141005, Mytishchi, Moscow reg., Russia

poleshchuk@mgul.ac.ru

Quality indicators of fuzzy regression models designed to study the dependencies between the qualitative characteristics of the educational process and to predict their values, as well as a model for recognizing fuzzy values of the output characteristics of regressions are presented. An algorithm for selecting a fuzzy regression model from linear and nonlinear models based on their quality indicators is given. An analysis of the degree of influence of input characteristics on the output characteristic is carried out. A fuzzy regression model has been constructed to predict the success of the dissertation defense when the applicant enters the PhD program and to study the dependencies between the applicant's input characteristics and the output characteristic. An alternative approach to the construction of regression models based on non-numerical data of the educational process allows not to impose incorrect conditions on the initial data, considering them to be the values of random variables, and not to use incorrect arithmetic operations for the elements of ordinal scales.

Keywords: educational process, fuzzy information, fuzzy regression model, linguistic variable.

Suggested citation: Poleshchuk O.M., Komarov E.G., Poyarkov N.G. *Obrabotka i prognoz dannykh obrazovatel'nogo protsessa na osnove nechetkogo regressionnogo analiza* [Processing and prediction of educational process data based on fuzzy regression analysis]. *Lesnoy vestnik / Forestry Bulletin*, 2024, vol. 28, no. 3, pp. 133–140. DOI: 10.18698/2542-1468-2024-3-133-140

References

- [1] Zade L.A. *Ponyatie lingvisticheskoy peremennoy i ego primeneniye k prinyatiyu priblizitel'nykh resheniy* [The concept of a linguistic variable and its application to approximate decision making]. Moscow: Mir, 1976, 165 p.
- [2] Voskoglou M.G. Fuzzy Logic as a Tool for Assessing Students' Knowledge and Skills. *Education Science*, 2013, v. 3(2), pp. 208–221.
- [3] Lin C.T., Lee C.S. Neural Network based Fuzzy Logic Control and Decision System. *IEEE Transactions on Comput*, 1991, v. 40, no. 12, pp. 1320–1336.
- [4] Mendes R. R., Voznika F. D., Freitas, A. A., Nievola J. C. Discovering fuzzy classification rules with genetic programming and co-evolution. *Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery*, 2001, pp. 314–325.
- [5] Poleshchuk O., Komarov E. The determination of students' fuzzy rating points and qualification levels. *1st International Fuzzy Systems Symposium — FUZZYSS'2009*, 1–2 October, 2009, Ankara, Turkey, 2009, pp. 218–224.
- [6] Ilahi R., Widiaty I., Gafar A. Abdullah Fuzzy system application in education. *IOP Conference Series Materials Science and Engineering*, 2018, v. 434 (1), p. 012308.
- [7] Darwish A., Poleshchuk O. Fuzzy Models for Educational Data Mining. *J. of Telecommunications*, 2012, v. 15, no. 2, pp. 8–22.
- [8] Ruspini E.H. A new approach to clustering. *Information and Control*, 1969, v. 15, pp. 22–32.
- [9] Tamura S., Higuchi S., Tanaka K. Pattern classification based on fuzzy relations. *IEEE Transactions on Systems Man and Cybernetics SMC1*, 1971, no.1, pp. 61–66. DOI:10.1109/TSMC.1971.5408605
- [10] Zadeh L.A. Similarity relations and fuzzy orderings. *Information Sciences*, 1971, v. 3, pp. 177–200.
- [11] Hwang C.L., Lin N.J. *Group decision making under multiple criteria*. Berlin: Springer, 1987. 400 p.
- [12] Dunn J.C. A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *J. of Cybernetics*, 1973, v. 3, pp. 32–57.
- [13] Ruspini E.H. *Recent developments in fuzzy clustering*. Fuzzy Set and Possibility Theory. N.Y.: Pergamon Press, 1982, pp. 133–146.
- [14] Ryjov A.P. The Concept of a Full Orthogonal Semantic Scope and the Measuring of Semantic Uncertainty. *Fifth International Conference Information Processing and Management of Uncertainty in Knowledge-Based Systems*, Iran, 4–7 December, 1994, pp. 33–34.
- [15] Bezdek J.C. Selected applications in classifier design. *Pattern recognition with fuzzy objective function algorithms*, 1981, v. 2, pp. 203–239.
- [16] Dubois D., Prade H. Ranking Fuzzy Numbers in Setting of Possibility Theory. *Information Science*, 1983, v. 30, pp. 183–224.
- [17] Borisov A.N., Krumberg O.A., Fedorov I.P. *Decision making on the basis of fuzzy models: Examples of use*. Riga: Zinatne, 1990. 184 p.
- [18] Ryjov A. Fuzzy Linguistic Scales: Definition, Properties and Applications. In: Reznik L., Kreinovich V. (eds) *Soft Computing in Measurement and Information Acquisition*. Studies in Fuzziness and Soft Computing, 2003, v. 127.
- [19] Poleshchuk O.M. *Klasternyy analiz ekspertnoy informatsii na osnove Z-chisel* [Cluster analysis of expert information based on Z-numbers]. *Lesnoy vestnik / Forestry Bulletin*, 2022, vol. 26, no. 1, pp. 143–148. DOI: 10.18698/2542-1468-2022-1-143-148
- [20] Sabic D.A., Pedrycz W. Evaluation on fuzzy linear regression models. *Fuzzy Sets and Systems*, 1991, v. 39, pp. 51–63.
- [21] Tanaka H., Ishibuchi H., Yoshikawa S. Exponential possibility regression analysis. *Fuzzy Sets and Systems*, 1995, v. 69, pp. 305–318.

- [22] Chang Y.-H.O. Hybrid fuzzy least-squares regression analysis and its reliability measures. *Fuzzy Sets and Systems*, 2001, v. 119, pp. 225–246.
- [23] Domrachev V.G., Poleshuk O.M. A regression model for fuzzy initial data. *Automation and Remote Control*, 2003, v. 64, no. 11, pp. 1715–1724.
- [24] Yager R. R., Filev D. P. On the issue of defuzzification and selection based on a fuzzy set. *Fuzzy Sets Syst.*, 1993, v. 55, pp. 255–272.
- [25] Altunin A.E., Semukhin M.V. *Modeli i algoritmy prinyatiya resheniy v nechetkikh usloviyakh* [Models and algorithms for decision making in fuzzy conditions]. Tyumen: Izd-vo Tyumenskogo gos. un-ta [Tyumen State Publishing House Univ.], 2002, 268 p.
- [26] Liu F., Mendel J.M. Encoding words into interval Type-2 fuzzy sets using an interval approach. *IEEE Trans. Fuzzy Systems*, 2008, v. 16, no. 6, pp. 187–201.
- [27] Runkler T.A., Katz C. Fuzzy clustering by particle swarm optimization. *Proceedings of the IEEE International Conference on Fuzzy Systems*, Oslo, 1-4 November, 2006. Berlin, 2006, pp. 34-41.
- [28] Phyo O., Chaw E. Comparative Study of Fuzzy PSO (FPSO) Clustering Algorithm and Fuzzy C-Means (FCM) Clustering Algorithm. *National J. of Parallel and Soft Computing*, 2019, v. 1, no. 1, pp. 62–67.
- [29] Zadeh L.A. Fuzzy logic and approximate reasoning. *Synthese*, 1975, v. 80, pp. 407–428.
- [30] Aliev R., Guirimov B.: Z-number clustering based on general Type-II fuzzy sets. *Advances in Intelligent Systems and Computing*, 2018, v. 896, pp. 270-278.
- [31] Jamal M., Khalif K., Mohamad S. The implementation of Z-numbers in fuzzy clustering algorithm for wellness of chronic kidney disease patients. *J. of Physics: Conference Series*, 2018, v. 1366, pp. 201-209.
- [32] Zadeh L.A. A note on Z-numbers. *Inf. Sci.*, 2011, v. 14(181), pp. 2923–2932. DOI: 10.1016/j.ins.2011.02.022
- [33] Komarov E.G., Poleshchuk O.M., Poyarkov N.G. *Izuchenie vzaimosvyazey mezhdu kachestvennymi priznakami pri nechetkoy iskhodnoy informatsii. Obozrenie prikladnoy i promyshlennoy matematiki* [Studying the relationships between qualitative features with fuzzy initial information]. [Review of Applied and Industrial Mathematics], 2005, v. 12, iss. 4, pp. 992–993.
- [34] Darwish A., Poleshchuk O., Komarov E. A new fuzzy linear regression model for a special case of interval type-2 fuzzy sets. *Applied Mathematics & Information Sciences*, 2016, v. 10, no. 3, pp. 1209–1214.
- [35] Poleshchuk O.M. Formalization, Prediction and Recognition of Expert Evaluations of Telemetric Data of Artificial Satellites Based on Type-II Fuzzy Sets. *Machine Learning and Data Mining in Aerospace Technology. Studies in Computational Intelligence*, 2020, v. 836, pp. 39–64.

Authors' information

Poleshchuk Ol'ga Mitrofanovna✉ — Dr. Sci. (Tech.), Professor, Head of Higher Mathematics and Physics Department of the BMSTU (Mytishchi branch), poleshchuk@mgul.ac.ru

Komarov Evgeniy Gennad'evich — Dr. Sci. (Tech.), Professor, Head of the Department of Information and Measuring Systems and Instrumentation Technologies of the BMSTU (Mytishchi branch), komarov@mgul.ac.ru

Poyarkov Nikolay Gennad'evich — Cand. Sci. (Tech.), Associate Professor, Dean of the Space Faculty of the BMSTU (Mytishchi branch), poyarkov@mgul.ac.ru

Received 13.11.2023.

Approved after review 22.03.2024.

Accepted for publication 27.03.2024.

Вклад авторов: все авторы в равной доле участвовали в написании статьи
Авторы заявляют об отсутствии конфликта интересов
Authors' Contribution: All authors contributed equally to the writing of the article
The authors declare that there is no conflict of interest