Электронный научно-технический журнал Октябрь 2005
http://www.bru.mogilev.by

ПРОЕКТИРОВАНИЕ РЕГРЕССИОННОЙ МОДЕЛИ ОЦЕНКИ ЗНАНИЙ[1]

Н.В. Якубова, **Н.М. Щербо**

Объектом исследования в данной работе являются экспертные системы как способ автоматизировать и упростить решение некоторых трудноразрешимых проблем и вопросов в оценке знаний. Целью является автоматизация прогноза оценки путем определения и ранжирования факторов, влияющих на оценку за курс. Результатом является страница сайта, на которой преподаватели могут получить данные о том, какие факторы и в какой степени влияют на выставление оценки, а студенты смогут получить прогноз оценки на экзамене на основе обработанных данных.

Ключевые слова: регрессионная модель, оценка, знания

1. Введение

Система разработана для автоматизации прогноза оценки путем определения и ранжирования факторов, влияющих на оценку за курс. Эта цель реализуется созданием страницы сайта, на которой преподаватели могут получить данные о том, какие факторы в какой степени влияют на выставление оценки, а студенты смогут получить прогноз оценки на экзамене на основе обработанных данных статистики. А также подтвердить или опровергнуть гипотезу, что оценка зависит не только от знаний и умений студента, но и от личностных особенностей студента и преподавателя.

Данная работа имеет три режима. Администратору предоставляется возможность выбрать режим работы системы, в зависимости от которого будут ограничены права пользователей системы.

В двух режимах происходит заполнение базы фактов (администратором либо пользователем), затем на основе обработанных данных статистики получаются коэффициенты, и строится регрессионная модель. Уравнение проверяется на адекватность и получается значение возможной ошибки.

В третьем режиме студент выбирает 5-ти или 10-ти бальную систему оценки, затем вводит информацию о себе, а для выявления личностных качеств может пройти тестирование. Далее он получает прогнозируемую оценку и значение возможной ошибки при данных условиях. [1], [2]

2. Математическая модель

Чтобы построить многофакторную регрессионную модель результативного признака оценка, предварительно необходимо отобрать факторные признаки в модель.

Чтобы выяснить, существует ли линейная связь между исследуемыми величинами, находится коэффициент корреляции:

[1] Работа выполнена по заказу кафедры «Автоматизированные системы управления» и деканата электротехнического факультета Белорусско-Российского университета

$$R_{xy} = \frac{\sum_{j=1}^{N} x_j y_j - N\overline{XY}}{(N-1)\sqrt{S_x^2 \cdot S_y^2}}, \quad (1)$$

где средние выборочные значения исследуемых величин:

$$\overline{X} = \frac{1}{N}\sum_{j=1}^{N} x_j, \qquad \overline{Y} = \frac{1}{N}\sum_{j=1}^{N} y_j; \quad (2)$$

S_x^2, S_y^2 - выборочные дисперсии исследуемых величин:

$$S_x^2 = \frac{1}{N-1}\sum_{j=1}^{N}(x_j - \overline{X})^2, \quad S_y^2 = \frac{1}{N-1}\sum_{j=1}^{N}(y_j - \overline{Y})^2. \quad (3)$$

Если коэффициент корреляции близок к 1, то можно считать, что между исследуемыми величинами имеется линейная связь, причем с увеличением X увеличивается Y. Если коэффициент корреляции близок к -1, то линейная связь существует, но с ростом X уменьшается Y. Если коэффициент корреляции близок к нулю, то величины X и Y не связаны друг с другом, или связь между ними нелинейная.

Чтобы выяснить, можно ли считать коэффициент корреляции значимым (т.е. близким к 1 или к –1), определяется следующий критерий:

$$T = \left|R_{xy}\right|\sqrt{\frac{N-2}{1-R_{xy}^2}}. \quad (4)$$

Этот критерий сравнивается с величиной, определяемой по таблицам распределения Стьюдента и обозначаемой как $T_{табл}$ или $T_{\alpha/2;s}$. Для определения $T_{табл}$ назначается величина α (обычно - из диапазона от 0,05 до 0,1), называемая уровнем значимости. Находится также параметр распределения Стьюдента, называемый числом степеней свободы (s). В задаче, связанной с проверкой значимости коэффициента корреляции, s=N-2.

Если выполняется условие $T > T_{\alpha/2;s}$, то коэффициент корреляции можно считать значимым. Это означает, что с вероятностью, равной 1-α, можно считать, что между исследуемыми величинами имеется линейная связь. Если $T < T_{\alpha/2;s}$, то коэффициент корреляции не является значимым. В этом случае можно считать, что линейной связи между исследуемыми величинами нет.

Коэффициенты модели A_i (i=1,…,M) показывают, на сколько в среднем изменится выходная переменная Y при увеличении входной переменной X_i на единицу (при неизменных значениях остальных входных переменных).

Коэффициент A_0 представляет собой приближенную оценку выходной переменной Y в случае, когда все входные переменные равны нулю.

Также проанализируем влияние изменения переменной X на изменение Y не в абсолютных величинах, а в процентах. Для этого используется величина, называемая коэффициентом эластичности:

$$E = A_1 \cdot \frac{\overline{X}}{\overline{Y}}. \quad (5)$$

Коэффициент эластичности показывает, на сколько процентов в среднем изменяется переменная Y при увеличении X на один процент.

Линейная модель связи выходной переменной Y с входными переменными X_1, $X_2,\ldots,X_M$ имеет следующий вид: $Y = A_0+A_1\cdot X_1+ A_2\cdot X_2+\ldots+ A_M\cdot X_M$. Значения коэффициентов A_0, $A_1,\ldots$, A_M находятся по методу наименьших квадратов. Можно доказать, что сумма квадратов ошибки будет минимальной, если коэффициенты A_0, $A_1,\ldots$, A_M определяются путем решения следующей системы из M+1 уравнения:

$$\begin{aligned}
&A_0\cdot N + A_1\cdot\sum_{j=1}^{N}x_{1j} + A_2\cdot\sum_{j=1}^{N}x_{2j} + \ldots + A_M\cdot\sum_{j=1}^{N}x_{Mj} = \sum_{j=1}^{N}y_j\\
&A_0\cdot\sum_{j=1}^{N}x_{1j} + A_1\cdot\sum_{j=1}^{N}x_{1j}^2 + A_2\cdot\sum_{j=1}^{N}x_{1j}\cdot x_{2j} + \ldots + A_M\cdot\sum_{j=1}^{N}x_{1j}\cdot x_{Mj} = \sum_{j=1}^{N}x_{1j}\cdot y_j\\
&A_0\cdot\sum_{j=1}^{N}x_{2j} + A_1\cdot\sum_{j=1}^{N}x_{2j}\cdot x_{1j} + A_2\cdot\sum_{j=1}^{N}x_{2j}^2 + \ldots + A_M\cdot\sum_{j=1}^{N}x_{2j}\cdot x_{Mj} = \sum_{j=1}^{N}x_{2j}\cdot y_j\\
&\ldots\\
&A_0\cdot\sum_{j=1}^{N}x_{Mj} + A_1\cdot\sum_{j=1}^{N}x_{Mj}\cdot x_{1j} + A_2\cdot\sum_{j=1}^{N}x_{Mj}\cdot x_{2j} + \ldots + A_M\cdot\sum_{j=1}^{N}x_{Mj}^2 = \sum_{j=1}^{N}x_{Mj}\cdot y_j,
\end{aligned} \qquad (6)$$

где x_{ij}, i=1,…,M, j=1,…,N – значения входных переменных, известные из статистических данных (таким образом, для каждой входной переменной должно быть известно N значений);

y_j, j=1,…,N – значения выходной переменной, также известные из статистических данных (при этом каждое значение выходной переменной y_j соответствует набору значений входных переменных $x_{1j}, x_{2j},\ldots,x_{Mj}$).

Что касается нашей системы, рассмотрим, какие факторы в большей мере влияют на оценку.

Построенная модель должна быть проверена на адекватность, т.е. на соответствие исходным данным. Модель является адекватной (достаточно точной), если фактические величины y_j (j=1,...,N), известные из статистических данных, близки к модельным значениям $\hat{y}_j$, определяемым путем подстановки известных значений x_j (j=1,…,N) в построенную модель.

Чтобы выполнить проверку модели на адекватность, требуется найти модельные значения $\hat{y}_j$ (j=1,…,N), а также следующие вспомогательные величины:

$$Q_r = \sum_{j=1}^{N}(\hat{y}_j - \overline{Y})^2, \quad Q_e = \sum_{j=1}^{N}(y_j - \hat{y}_j)^2. \qquad (7)$$

Величина Q_r называется суммой квадратов, обусловленной моделью, а величина Q_e - остаточной суммой квадратов, или суммой квадратов ошибки (эта величина уже упоминалась выше при описании метода построения модели).

Для проверки модели на адекватность находится следующий критерий:

$$F = \frac{Q_r / k}{Q_e /(N - k - 1)}, \qquad (8)$$

где k - количество коэффициентов модели, не считая A_0 (для модели с одной входной переменной k=1).

Этот критерий сравнивается с величиной, определяемой по таблицам распределения Фишера и обозначаемой как $F_{табл}$ или $F_{\alpha,s1,s2}$. Для определения $F_{\alpha,s1,s2}$ назначается величина α (обычно - из диапазона от 0,05 до 0,1), называемая уровнем значимости. Находятся также параметры распределения Фишера, называемые числами степе-

ней свободы (s_1, s_2). В задачах, связанных с проверкой адекватности линейных моделей, $s_1 = k$, $s_2 = N-k-1$.

Если выполняется условие $F > F_{\alpha,s1,s2}$, то построенная линейная модель является адекватной, т.е. она достаточно точно описывает связь между исследуемыми величинами.

Для оценки точности модели применяется также величина, называемая коэффициентом детерминации:

$$R^2 = \frac{Q_r}{Q_r + Q_e}. \quad (9)$$

Эта величина показывает, какая часть разброса значений выходной переменной Y (т.е. различий между величинами $y_1, y_2, \ldots, y_N$) объясняется разбросом значений входной переменной X (то есть различиями между величинами $x_1, x_2, \ldots, x_N$) [3], [4].

3. Заключение

Научная работа тему «Проектирование регрессионной модели оценки знаний» обеспечила достижение всех поставленных целей.

Для реализации используются Интернет-сервер Apache с установленным языком PHP и базой MySQL [5, 6, 7].

Разработанная экспертная система может быть использована в качестве прогнозной системы на сайте учебного заведения. Кроме того, обеспечение дружественного интерфейса позволяет любому пользователю овладеть ей и использовать в своих целях.

Литература

1. *Якубова Н.В., Крутолевич С.К.* Факторы, влияющие на результат экзаменов. // Межрегиональная науч.-техн. конф. студентов и аспирантов «Информационные технологии, энергетика и экономика»: Материалы докладов в 4-х т. – Смоленск: филиал ГОУ ВПО МЭИ(ТУ), 2005. Т. 1 – С. 92-94.
2. *Якубова Н.В., Крутолевич С.К.* Факторы, влияющие на оценку. // Материалы, оборудование и ресурсосберегающие технологии: материалы междунар. науч.-техн. конф., Могилев, 21-22 апр. 2005 г. В 2 ч. Ч. 1. – Могилев: ГУ ВПО «Бел.-Рос. ун-т», 2005. – С. 418-419.
3. Методы анализа и принятия решений в слабоструктурированных задачах: Учеб. пособие для вузов /*С.С.Смородинский, Н.В.Батин* и др.; Под ред. *С.С.Смородинского*. – М.: 2002. – 120 с.
4. *Щербо Н.М.* // Конспект лекций по курсу «Методы и средства поддержки принятия решений» для специальности Т.10.01 «Автоматизированные системы обработки информации». – Могилев Бел.-Рос. ун-т, 2005.
5. Учебник PHP 4.0 [Электрон. ресурс], 2004. Режим доступа: http://www.compdoc.ru/internet
6. MySQL [Электрон. ресурс], Учебник. Режим доступа: http://www.ponteley.al.ru
7. PHP & MySQL [Электрон. ресурс], Учебник. Режим доступа: http://www.ponteley.al.ru

Якубова Наталья Викторовна
Студентка электротехнического факультета
Белорусско-Российский университет, г. Могилев
Тел.: +375(222) 44-29-12
E-mail: radost_moia@tut.by
Щербо Наталья Михайловна
Старший преподаватель кафедры АСУ
Белорусско-Российский университет, г. Могилев
Тел.: +375(22) 25-63-57