УДК 519.237.8

ПРОГНОЗИРОВАНИЕ УСПЕВАЕМОСТИ СТУДЕНТОВ НА ОСНОВЕ МЕТОДОВ КЛАСТЕРНОГО АНАЛИЗА

В.А. Шевченко, доц., к.т.н.,
Харьковский национальный автомобильно-дорожный университет

Аннотация. *Предложена методика прогнозирования успеваемости студентов на основе методов кластерного анализа. Приведены результаты проведенного эксперимента, подтверждающие эффективность разработанной методики прогнозирования успеваемости.*

Ключевые слова: *прогнозирование, успеваемость, кластерный анализ, матрица исходных данных, матрица расстояний.*

ПРОГНОЗУВАННЯ УСПІШНОСТІ СТУДЕНТІВ НА ОСНОВІ МЕТОДІВ КЛАСТЕРНОГО АНАЛІЗУ

В.О. Шевченко, доц., к.т.н.,
Харківський національний автомобільно-дорожній університет

Анотація. *Запропоновано методику прогнозування успішності студентів на основі методів кластерного аналізу. Наведено результати проведеного експерименту, що підтверджують ефективність розробленої методики прогнозування успішності.*

Ключові слова: *прогнозування, успішність, кластерний аналіз, матриця вихідних даних, матриця відстаней.*

PROGNOSTICATION OF STUDENTS PROGRESS ON THE BASIS OF CLUSTER ANALYSIS METHODS

V. Shevchenko, Assoc. Prof., Ph. D. (Eng.),
Kharkiv National Automobile and Highway University

Abstract. *The method of prognostication of students progress on the basis of methods of cluster analysis has been offered. The results of the experiment confirming the efficiency of the developed method of prognostication have been given.*

Key words: *prognostication, progress, cluster analysis, matrix of initial data, matrix of distances.*

Введение

В настоящее время существуют сотни методов прогнозирования. Виды математических методов прогнозирования: корреляционный анализ, регрессионный анализ, кластерный анализ, факторный анализ и др.

Анализ публикаций

Рассмотрением сущности прогнозирования в области обучения занимались Б.С. Гершунский [1], В.И. Загвязинский [2], А.Ф. Присяжная [3], Р.В. Майер [4] и др.

В ходе анализа публикаций был сделан вывод, что для достоверного прогнозирования успеваемости студентов больше всего подходят методы кластерного анализа, поскольку кластерный анализ позволяет производить разбиение объектов не по одному параметру, а по целому набору признаков. Кроме того, кластерный анализ позволяет рассматривать множество исходных данных практически произвольной природы.

Цель и постановка задачи

По результатам проведенного анализа в области педагогического прогнозирования были поставлены следующие цели:
1. Разработать процедуру прогнозирования успеваемости студентов на основе методов кластерного анализа.

2. Для проверки эффективности разработанной процедуры прогнозирования провести эксперимент с целью сравнения реальной и прогнозируемой успеваемости студентов.

Выбор метода кластерного анализа для прогнозирования успеваемости студентов

Для решения поставленной задачи – разработки процедуры прогнозирования успеваемости студентов из множества алгоритмов кластеризации наиболее подходящим, на наш взгляд, является алгоритм *k*-средних Мак-Кина, в котором сам пользователь должен задать искомое число конечных кластеров, обозначаемое *k*. Принцип классификации заключается в следующем:
– выбираются или назначаются *k* наблюдений, которые будут первичными центрами кластеров;
– остальные наблюдения приписываются к ближайшим заданным кластерным центрам;
– текущие координаты первичных кластерных центров заменяются на кластерные средние;
– предыдущие два шага повторяются до тех пор, пока изменения координат кластерных центров не станут минимальными.

Однако алгоритм Мак-Кина предполагает, что кластерные центры выбираются из существующего набора данных для кластеризации. Для решения поставленной задачи такой подход не приемлем, так как могут быть группы студентов с различной успеваемостью; например, группы, где нет двоечников, или, наоборот, нет отличников, или много троечников. Если выбирать кластерные центры из данных каждой студенческой группы, то для каждой группы распределение студентов на кластеры в зависимости от их успеваемости будет различным, и может случиться, что студент с хорошей успеваемостью попадет в кластер плохой успеваемости и наоборот. Необходимо определить такие кластерные центры, значения которых не зависят от набора классифицируемых данных и обеспечивают распределение студентов на кластеры в соответствии с существующими параметрами успеваемости: до 60 баллов – плохо, от 60 до 75 баллов – удовлетворительно, от 75 до 90 баллов – хорошо, свыше 90 баллов – отлично.

Кроме того, по алгоритму Мак-Кина после добавления какого-либо данного в кластер необходимо произвести пересчет центра кластера. В этом случае значение кластерного центра будет изменяться, что также приведет к искажению результатов кластеризации.

Следовательно, метод *k*-средних Мак-Кина целесообразно применить для решения поставленной задачи после некоторой модификации.

Модификация метода *k*-средних Мак-Кина

Модифицируем алгоритм Мак-Кина исходя из следующих допущений:

1. При решении поставленной задачи необходимо задать такие кластерные центры, которые представляют собой усредненные значения каждого параметра для каждого класса.

2. Заданные центры должны оставаться неизменными на протяжении всей процедуры кластеризации.

Постановка задачи кластеризации

Известно множество объектов X, представляющих собой данные по успеваемости n студентов, состоящие из m признаков: $X = \{X_1, X_2, ..., X_m\}$. Множество объектов X описывается множеством векторов измерений X_j, $j = \overline{1,m}$. Требуется разбить выборку X на четыре типологические группы, характеризующие успеваемость студентов: «отлично», «хорошо», «удовлетворительно» и «плохо». Следовательно, задаем число кластеров $k = 4$.

Процедура кластеризации

1. Зададим матрицу исходных данных в соответствии с формулой (1), где x_{ij} – j-й параметр i-го объекта, m – количество парамет-

ров; n – количество студентов (объектов кластеризации)

$$X = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1j} & \dots & x_{1m} \\ x_{21} & x_{22} & \dots & x_{2j} & \dots & x_{2m} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{i1} & x_{i2} & \dots & x_{ij} & \dots & x_{im} \\ \dots & \dots & \dots & \dots & \dots & \dots \\ x_{n1} & x_{n2} & \dots & x_{nj} & \dots & x_{nm} \end{bmatrix}. \quad (1)$$

2. Назначим первичные центры кластеров. Для этого для каждого кластера определим эталонные значения параметров как усредненные данные по каждой типологической группе студентов, полученные при моделировании процесса формирования компетенций у студентов [5]. Эталонные значения будут использованы в качестве центров будущих кластеров, вокруг которых группируются наиболее близкие объекты по значениям выбранных параметров. Эталонные значения параметров кластеризации приведены в табл. 1.

Таблица 1 Эталоны для прогнозирования

Типолог. группы	Начальн. знания	Знания по теме	Кол-во пропусков
Класс 5	85	95	0
Класс 4	75	85	0
Класс 3	60	70	0
Класс 2	40	40	2

Вокруг эталонов собираются объекты, близкие по своим параметрам. В качестве объектов кластеризации в данной задаче выступают студенты, а в качестве параметров – факторы, значения которых можно оценить в начальный момент изучения дисциплины:
– уровень начальных знаний студентов;
– уровень компетенций, сформированных у студентов по первой теме дисциплины;
– количество пропусков занятий студентами на момент составления прогноза.

3. Поскольку выбранные признаки имеют разные единицы измерения, исходные данные нормируем вместе с добавленными к ним эталонами по формуле (2)

$$z_{ij} = \frac{x_{ij} - \overline{x}_j}{\sigma_j}, \quad i = \overline{1, n+4},\ j = \overline{1, m}, \quad (2)$$

где z_{ij} – нормированное значение j-го параметра i-го объекта; x_{ij} – исходное значение j-го параметра i-го объекта; $\overline{x}_j$ – среднее значение j-го параметра по всем объектам; σ_j – среднеквадратическое отклонение x_{ij}.

4. Для нормированных данных строим матрицу расстояний D (3)

$$D = \begin{bmatrix} 0 & d_{1,2} & \dots & d_{1,n} & d_{1,n+4} \\ d_{2,1} & 0 & \dots & d_{2,n} & d_{2,n+4} \\ \dots & \dots & \dots & \dots & \dots \\ d_{n,1} & d_{n,2} & \dots & 0 & d_{n,n+4} \\ d_{n+4,1} & d_{n+4,2} & \dots & d_{n+4,n} & 0 \end{bmatrix}. \quad (3)$$

Расстояния между объектами вычисляем по Евклидовой метрике (4)

$$d_{ij} = \frac{1}{m}\sqrt{\sum_{k=1}^{m}\left(z_{ik} - z_{jk}\right)^2}, \quad (4)$$

где d_{ij} – расстояние между i-м и j-м объектами; m – количество признаков кластеризации; z_{ik} – нормированное значение i-го объекта по k-му признаку; z_{jk} – нормированное значение j-го объекта по k-му признаку.

5. Из матрицы расстояний выделяем эталонную матрицу расстояний, которая представляет собой матрицу расстояний (5) от каждого объекта до эталонных данных

$$MEt = \begin{bmatrix} d_{1,n+1} & d_{1,n+2} & d_{1,n+3} & d_{1,n+4} \\ d_{2,n+1} & d_{2,n+2} & d_{2,n+3} & d_{2,n+4} \\ \dots & \dots & \dots & \dots \\ d_{i,n+1} & d_{i,n+2} & d_{i,n+3} & d_{i,n+4} \\ \dots & \dots & \dots & \dots \\ d_{n,n+1} & d_{n,n+2} & d_{n,n+3} & d_{n,n+4} \end{bmatrix}, \quad (5)$$

где MEt – эталонная матрица расстояний.

6. В эталонной матрице определим минимальное значение расстояния, номер объекта и кластерного эталона, которые находятся на этом минимальном расстоянии.

7. Выбранный объект припишем к соответствующему кластеру.

8. Из матрицы исходных данных и матрицы эталонных расстояний удалим данные об объекте, который был приписан к кластеру.

Пункты 6 – 8 повторим до тех пор, пока все объекты не будут разнесены по кластерам.

Разработанная процедура прогнозирования успеваемости студентов имеет реализацию в виде макроса на языке VBA.

Описание и результаты эксперимента

Для проверки эффективности применения метода формирования индивидуальных траекторий для самостоятельной работы на основе кластерного анализа для организации индивидуализации самостоятельной работы потока студентов был проведен эксперимент со студентами трех групп (всего 61 студент) дорожно-строительного факультета ХНАДУ, изучающими информатику в осеннем семестре.

В качестве исходных данных использованы три фактора: начальный уровень знаний студентов (оценен в начале первого занятия), знания, полученные студентами на занятии по первой теме (оценены на первой лабораторной работе), и количество пропусков занятий (эксперимент проводился на втором занятии). По этим исходным данным был составлен прогноз успеваемости для каждого студента по дисциплине «Информатика».

По окончании изучения дисциплины прогнозируемые баллы студентов сравнили с баллами, которые студенты получили на зачете по информатике.

Сравнительные данные эксперимента приведены в табл. 2 и на рис. 1.

Таблица 2 Зачетные и прогнозные данные

	Отлично	Хорошо	Удовлетв.	Плохо	Итого
Зачетный балл	3	11	40	7	61
Прогнозный балл	2	13	38	8	61

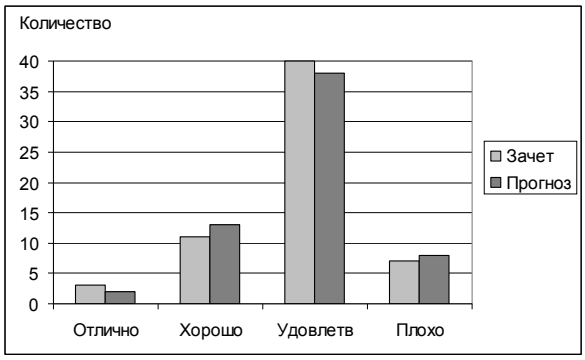

Рис. 1. Сравнительная диаграмма зачетных и прогнозных данных

Вывод

Результаты проведенного эксперимента показали, что прогнозируемая успеваемость студентов отличается от реальной не более, чем на 3,3 %. Следовательно, процедура на основе модифицированного метода *k*-средних Мак-Кина действенна и может использоваться для прогнозирования успеваемости студентов.

Литература

1. Гершунский Б.С. Прогностические методы в педагогике / Б.С. Гершунский. – К.: Вища школа, 1979. – 240 с.
2. Загвязинский В.И. Педагогическое предвидение / В.И. Загвязинский. – М.: Знание, 1987. – 77 с.
3. Присяжная А.Ф. Прогнозирование как функция педагога (от будущего учителя до профессионала): монография / А.Ф. Присяжная. – Челябинск: Образование, 2006. – 306 с.
4. Майер Р.В. Классификация учебных фактов методом кластерного анализа / Р.В. Майер // Проблемы учебного физического эксперимента: сб. науч. и метод. работ. – 1998. – Вып. 5. – С. 12–19.
5. Шевченко В.А. Концепция построения модели приобретения знаний студентами по дисциплине «Информатика» / В.А. Шевченко // Вестник ХНАДУ: сб. науч. тр. – 2012. – Вып. 56. – С. 26–29.

Рецензент: В.В. Бондаренко, профессор, к.пед.н., ХНАДУ.

Статья поступила в редакцию 17 февраля 2015 г.