УДК 519.237.8

АЛГОРИТМИЗАЦИЯ ПРОЦЕДУРЫ КЛАСТЕРНОГО АНАЛИЗА ДЛЯ ПРОГНОЗИРОВАНИЯ УСПЕВАЕМОСТИ СТУДЕНТОВ

В.А. Шевченко, к.т.н., доцент, А.И. Кудин, к.т.н., доцент, ХНАДУ

Аннотация. *Рассмотрены существующие методы кластерного анализа. Обоснована модификация метода k-средних Мак-Кина для организации прогнозирования успеваемости студентов. Разработан алгоритм процедуры прогнозирования успеваемости студентов на основе модифицированного метода кластерного анализа k-средних Мак-Кина, позволяющий автоматизировать процедуру прогнозирования.*

Ключевые слова: *кластерный анализ, центр кластера, алгоритм, прогнозирование, успеваемость, блок-схема.*

АЛГОРИТМІЗАЦІЯ ПРОЦЕДУРИ КЛАСТЕРНОГО АНАЛІЗУ ДЛЯ ПРОГНОЗУВАННЯ УСПІШНОСТІ СТУДЕНТІВ

В.А. Шевченко, к.т.н., доцент, А.І. Кудін, к.т.н., доцент, ХНАДУ

Анотація. *Розглянуто існуючі методи кластерного аналізу. Обґрунтована модифкація методу k-середніх Мак-Кіна для організації прогнозування успішності студентів. Розроблено алгоритм процедури прогнозування успішності студентів на основі модифікованого методу кластерного аналізу k-середніх Мак-Кіна, що дозволяє автоматизувати процедуру прогнозування.*

Ключові слова: *кластерний аналіз, центр кластера, алгоритм, прогнозування, успішність, блок-схема.*

ALGORITHMIZATION OF THE CLUSTER ANALYSIS PROCEDURE FOR FORECASTING OF STUDENTS' PROGRESS

V. Shevchenco, Candidate of Technical Science, Associate Professor, KhNAHU,
A. Kudin, Candidate of Technical Science, Associate Professor, KhNAHU

Abstract: *Existing methods of the cluster analysis are considered. Justified is the modification of the k-averages of method McKean for the organization of forecasting of students' academic performance. An algorithm for the procedure for predicting students' progress based on the modified method of cluster analysis of k-averages McKean has been developed, which makes it possible to automate the prediction procedure.*

Keywords: *cluster analysis, cluster center, algorithm, forecasting, academic achievement, block diagram.*

Введение

В настоящее время существует огромное количество алгоритмов кластер-анализа. Наиболее распространенную группу методов кластеризации составляют методы, основывающиеся на иерархической агломеративной процедуре. Смысл иерархический агломеративной процедуры заключается в следующем [1]. Перед началом кластеризации все объекты считаются отдельными кластерами, т.е. имеется p = n кластеров, каждый из которых включает по одному элементу. На первом шаге алгоритма определяются два наиболее близких или сходных объекта, которые объединяются в один кластер, общее количество

которых сокращается на 1. Итеративный процесс повторяется, пока на последнем (p-1)-м шаге все классы не объединятся. На каждом последующем шаге агломеративной процедуры требуется пересчет лишь одной строки и одного столбца матрицы Δ, т.е. рассчитываются расстояния от образованного кластера до каждого из оставшихся кластеров.

Итерационные процедуры пытаются найти наилучшее разбиение, ориентируясь на заданный критерий оптимизации [1]. В начале последовательных итераций в качестве центра выбирается один из элементов $x_l \in X$ и формируется кластер S из элементов, удаленных от него не далее чем на заданное расстояние $\tau > 0$, определяющее радиус кластера:

$$\begin{aligned} &\text{если } d_{ix_l} \le \tau, \text{то } x_i \in S, \\ &\text{если } d_{ix_l} > \tau, \text{то } x_i \notin S, \end{aligned} \qquad (1)$$

где d_{ix_l} - расстояние от некоторого элемента x_i до центра кластера x_l.

Далее процедура повторяется для остальных элементов. После выполнения очередного шага выясняется, достигнуто ли желательное разбиение. Существуют различные условия определения критерия остановки процедуры: 1) получено определенное заранее количество кластеров; 2) все кластеры содержат более определенного числа элементов; 3) кластеры обладают требуемым соотношением внутренней однородности и разнородности между собой.

На первом условии основывается наиболее популярный алгоритм – метод k-средних Мак-Кина, в котором сам пользователь должен задать искомое число конечных кластеров, обозначаемое k. Принцип классификации заключается в следующем:

1) выбираются или назначаются k наблюдений, которые будут первичными центрами кластеров;
2) остальные наблюдения приписываются к ближайшим заданным кластерным центрам;
3) текущие координаты первичных кластерных центров заменяются на кластерные средние;
4) предыдущие два шага повторяются до тех пор, пока изменения координат кластерных центров не станут минимальными.

Для решения поставленной задачи – распределения потока студентов на типологические группы осуществляется по выделенным ранее шести признакам: уровень начальных знаний; средний балл знаний, приобретенных на занятиях; количество пропусков занятий; средний балл знаний после забывания изученных ранее тем; средний балл знаний после самостоятельной работы; моделируемый зачетный балл. Из рассмотренных выше алгоритмов кластеризации наиболее подходит метод k-средних Мак-Кина.

Однако алгоритм Мак-Кина предполагает, что кластерные центры выбираются из существующего набора данных для кластеризации. Для решения поставленной задачи такой подход неприемлем, 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 кластеризции.

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

Анализ публикаций

В [2] Шевченко В.А. приводит модификацию метода k-средних Мак-Кина для решения задачи прогнозирования успеваемости студентов.

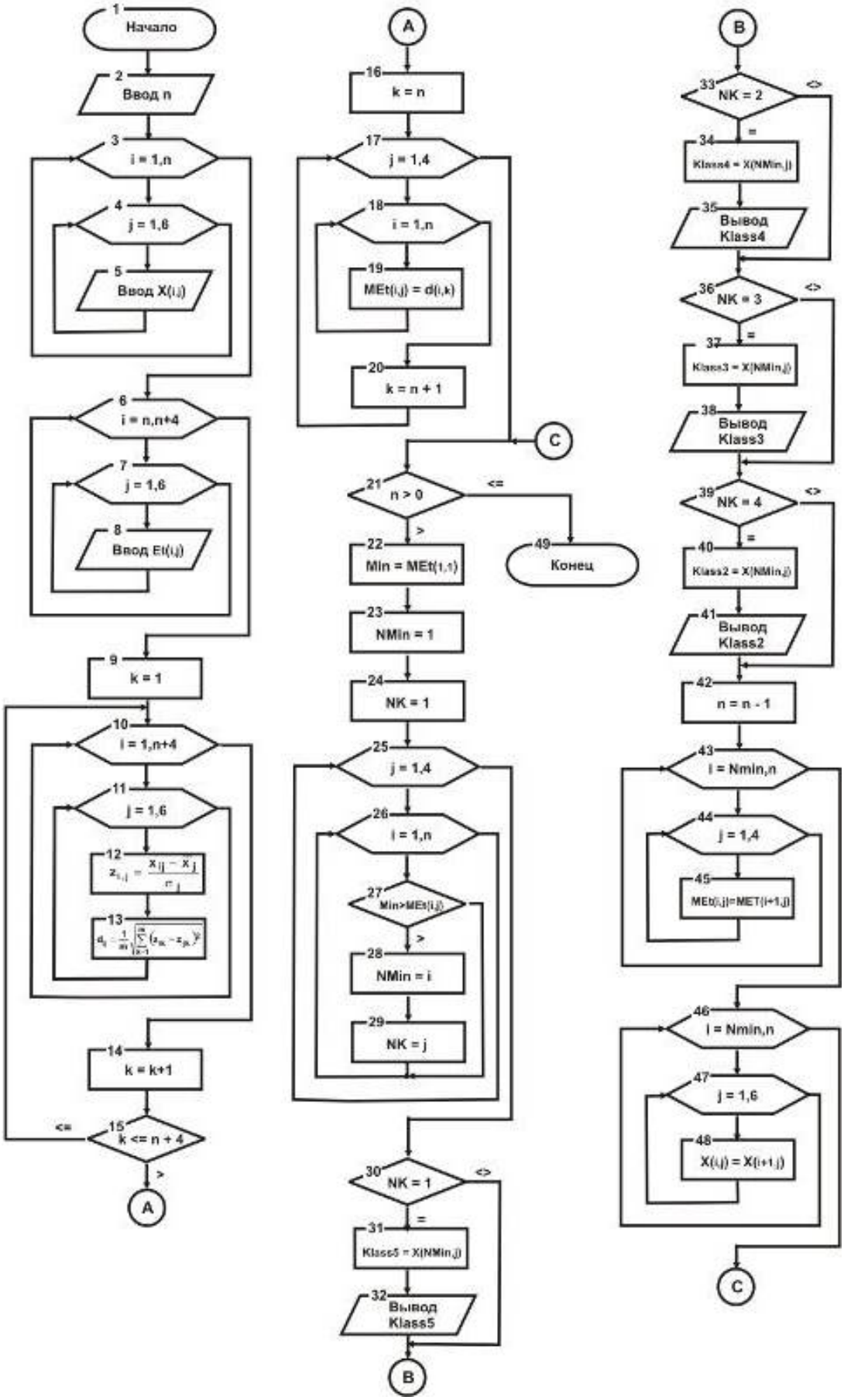

Рис. 1. – Схема алгоритма разработанной процедуры классификации студентов в зависимости от их успеваемости

В [3] представлены результаты апробации модифицированного метода *k*-средних Мак-Кина при прогнозировании успеваемости.

Таким образом, автор подтверждает достоверность результатов прогнозирования успеваемости студентов с помощью модифицированного метода *k*-средних Мак-Кина.

Постановка задачи

Однако для автоматизации процедуры прогнозирования успеваемости студентов с помощью модифицированного метода *k*-средних Мак-Кина необходимо провести формализацию разработанной процедуры, для этого целесообразно разработать алгоритм процедуры прогнозирования.

Алгоритм процедуры прогнозирования успеваемости студентов

Схема алгоритма для разработанной процедуры кластеризации по модифицированному методу *k*-средних Мак-Кина, который описан в [2], приведена на рис. 1.

Описание алгоритма.

Блок 1 – начало алгоритма.

Блок 2 – ввод количества студентов *n*, подлежащих распределению на группы.

Блоки 3-5 – ввод матрицы исходных данных $X_{i,j}$.

Блоки 6-8 – ввод в матрицу исходных данных эталонов кластеризации $Et_{i,j}$.

Блоки 9-15 – нормирование исходных данных (блок 16) и вычисление матрицы расстояний *D* (блок 17).

Блоки 16-20 – определение матрицы эталонных расстояний $MEt_{i,j}$.

Блок 21 – проверка условия наличия нераспределенных объектов.

Блоки 22-29 – определение минимального эталонного расстояния, номера объекта и кластерного эталона, которые находятся на этом минимальном расстоянии.

Блоки 30-41 – распределение объекта с минимальным расстоянием в кластер, соответствующий номеру кластерного эталона.

Блок 42 – сокращение на единицу количества нераспределенных объектов.

Блоки 43-45 – удаление из матрицы эталонных расстояний минимального эталонного расстояния.

Блоки 46-48 – удаление из матрицы исходных данных объекта с минимальным эталонным расстоянием.

Блок 49 – конец алгоритма.

Выводы

Предложенный алгоритм может быть реализован на любом алгоритмическом языке для автоматизации процедуры прогнозирования успеваемости студентов.

Литература

1. Мандель И.Д. Кластерный анализ / И.Д. Мандель – М.: Финансы и статистика, 1988. – 176 с.
2. Шевченко В.А. Прогнозирование успеваемости студентов на основе методов кластерного анализа / В.А. Шевченко / Вестник ХНАДУ: сб. науч. тр. –Харьков: ХНАДУ. – 2015. – Вып. 68. – С.18–21.
3. Шевченко В.А. Прогнозирование успеваемости студентов с помощью методов кластерного анализа/ В.А. Шевченко / Экспертные оценки элементов учебного процесса: прогр. и матер. XV межвуз. науч.-практ. конф., Харьков: НУА. – 2013. С.112–115.

Рецензент: В.М. Колодяжный, профессор, д.ф-м.н., ХНАДУ.

Статья поступила в редакцию 26 мая 2017 г.