УДК 339.54

С. В. Вышемирская, Ф. Б. Рогальский, И. Е. Шупик

ИНФОРМАЦИОННАЯ ТЕХНОЛОГИЯ РАЗБИЕНИЯ МНОЖЕСТВ НА ГРУППЫ СРЕДСТВАМИ ПРОГРАММНОГО КОМПЛЕКСА SPSS

Введение. Социальные и экономические объекты, как правило, характеризуются достаточно большим числом параметров, образующих многомерные векторы. Особое значение в экономических и социальных исследованиях приобретают задачи изучения взаимосвязей между компонентами этих векторов, причем эти взаимосвязи необходимо выявлять на основании ограниченного числа многомерных наблюдений [1]. В данной работе такие задачи рассматриваются применительно к компьютеризации процессов корректировки физического состояния студентов. Дело в том, что состояние здоровья молодежи, в том числе и студентов многих технических вузов, в последнее время имеет тенденцию к ухудшению, существенно увеличилось количество студентов, которые по состоянию здоровья относятся к специальной медицинской группе. Поэтому управление физическим состоянием студентов в процессе физического воспитания путем применения оздоровительных методик представляет несомненный интерес[7–8]. Рассмотрение индивида как элемента кибернетической системы требует разбиения исходного множества студентов на отдельные элементы для индивидуального оценивания состояния индивида, контроля физической нагрузки и рекомендаций по корректировке его физического состояния.

Использование методов снижения размерности многомерного пространства, позволяющих без существенной потери информации перейти от первоначальной системы большого числа наблюдаемых взаимосвязанных факторов к системе существенно меньшего числа скрытых (ненаблюдаемых) факторов, определяющих вариацию первоначальных признаков, помогает обрабатывать большие массивы разнородной информации. К таким методам относятся методы компонентного и факторного анализа. Методы многомерной классификации предназначены для разделения совокупностей объектов (характеризующиеся большим числом признаков) на классы, в каждый из которых должны входить объекты, в определенном смысле однородные или близкие.

Постановка задачи. Рассмотрим построение модели поддержки принятия решения для разбиения по группам студентов, имеющих разные нозологические формы заболеваний. В ходе анкетирования 220 студентов специальной группы было выявлено 14 нозологических форм, причем у большинства студентов встречается несколько нозологий, которые могут быть как следствием основного заболевания, так и совсем от него не зависеть. В результате задача разбиения студентов на группы усложняется тем, что количество таких групп может быть очень велико. Заниматься с каждым студентом индивидуально преподаватель физического воспитания не имеет возможности, поэтому необходимо наиболее рационально сгруппировать студентов, не упуская из внимания тот факт, что для разных нозологий предполагается использовать различные комплексы упражнений. Более того, разные нозологические формы предполагают взаимоисключающие комплексы физической нагрузки. Для формирования групп наиболее рационально объединяющих студентов, которым рекомендуется выполнять одинаковые комплексы упражнений использован кластерный анализ. Однако в ходе исследований возникла необходимость уменьшения массива исходной информации, поэтому была разработана математическая модель факторного анализа, с помощью которой удается уменьшить размерность указанного массива информации.

Изложение основного материала. Предположим, что $X = (X_1, X_2, \ldots, X_k)^T$ — исходный k-мерный случайный вектор. Каноническая модель факторного анализа для центрированного вектора $\overset{\circ}{X} = X - MX$ имеет следующий вид (1):

$$\overset{\circ}{X} = AF + \overset{\circ}{\varepsilon} \qquad (1)$$

где F = (F_1, F_2, …, F_m) — центрированный и нормированный случайный вектор некоррелированных m общих факторов для всех исходных случайных величин X_i ($m < k$), $A = \left(a_{ij}\right) \in R^{k \times m}$ – (неслучайная) матрица нагрузок случайных величин X_i на факторы F_j, $\overset{\circ}{\varepsilon} = \left(\overset{\circ}{\varepsilon}_1, \overset{\circ}{\varepsilon}_2, \ldots, \overset{\circ}{\varepsilon}_k\right)^T$ – нормально распределенный центрированный вектор ε специфических факторов ε_1, ε_2, …, ε_k, некоррелированных как между собой, так и с общими факторами [2].

Пусть $\Sigma_X = M\left(\overset{\circ}{X}\,\overset{\circ}{X}^T\right)$ – ковариационная матрица вектора X, а $\Sigma_\varepsilon = M\left(\overset{\circ}{\varepsilon}\,\overset{\circ}{\varepsilon}^T\right)$ – ковариационная матрица (диагональная) вектора ε с диагональными элементами, равными $M\overset{\circ}{\varepsilon}^2 = D\overset{\circ}{\varepsilon} = D\overset{\circ}{\varepsilon}_i = v_i$.

Построим систему уравнений для нахождения матриц A и Σ_ε. С учетом условий на векторах F и $\overset{\circ}{\varepsilon}$ получим [3]:

$$\Sigma_X = M\left[\left(AF+\overset{\circ}{\varepsilon}\right)\left(AF+\overset{\circ}{\varepsilon}\right)^T\right] = M\left(AFF^TA^T\right)+M\left(AF+\overset{\circ}{\varepsilon}^T\right)+$$

$$+M\left(\overset{\circ}{\varepsilon}F^TA^T\right)+M\left(\overset{\circ}{\varepsilon}\overset{\circ}{\varepsilon}^T\right) = M\left(AIA^T\right)+AM\left(F\overset{\circ}{\varepsilon}^T\right)+A^TM\left(\overset{\circ}{\varepsilon}F^T\right)+\Sigma_\varepsilon = AA^T+\Sigma_\varepsilon.$$

Тогда

$$\Sigma_X = AA^T+\Sigma_\varepsilon.$$

или

$$\begin{cases}\operatorname{cov}\left(X_i,X_p\right)=\sum\limits_{j=1}^{m}a_{ij}a_{pj},\ i=1,2,\ldots,k,\ p=1,2,\ldots,k,\ i\neq p\\ \operatorname{cov}\left(X_i,X_i\right)=DX_i=\sum\limits_{j=1}^{m}a_{ij}^2+v_i,\ i=1,2,\ldots,k.\end{cases} \quad (2)$$

Таким образом, ковариации исходных случайных величин полностью воспроизводятся матрицей нагрузок, а для воспроизведения их дисперсий помимо нагрузок нужны дисперсии υ_i специфических факторов. И далее, так как

$$M\left(\overset{\circ}{X}F^T\right) = M\left[\left(AF+\overset{\circ}{\varepsilon}\right)F^T\right] = M\left(AFF^T+\overset{\circ}{\varepsilon}F^T\right) = AM\left(FF^T\right)+M\left(\overset{\circ}{\varepsilon}F^T\right) = A,$$

то ковариации $\operatorname{cov}\left(X_i,F_j\right)=a_{ij}$.

В системе (2) k^2 уравнений, а число неизвестных (a_{ij} и υ_i) равно $mk + k < < k(k + 1)$. Если допустить, что k, m и матрица Σ_X таковы, что решение этой системы существует, то это решение не единственно [3].

Действительно, пусть V – ортогональная матрица размером $m\times m$. Проведем тождественные преобразования модели (2):

$$\overset{\circ}{X} = AF+\overset{\circ}{\varepsilon} = AIF+\overset{\circ}{\varepsilon} = A\left(VV^T\right)F+\overset{\circ}{\varepsilon} = \left(AV\right)\left(V^TF\right)+\overset{\circ}{\varepsilon} \quad (3)$$

В преобразованной модели вектор общих факторов — это вектор $\widetilde{F}=V^TF$, а матрица нагрузок $\widetilde{A}=AV$.

Следовательно, если решение системы (2) существует, то оно не единственно [4].

После проведения факторного анализа удается уменьшить размерность входного массива информации. Далее можно воспользоваться кластерным анализом. Методы кластерного анализа позволяют разбить изучаемую совокупность объектов на группы однородных в некотором смысле объектов, называемых кластерами или классами. Наибольшее распространение получили два подхода к задаче классификации: эвристический, реализующий некоторую схему разделения объектов на классы, исходя из интуитивных соображений, и экстремальный, реализующий схему разделения на основе заданного критерия оптимальности. Наиболее трудным в задаче классификации является определение меры однородности объектов.

В настоящее время основное значение имеют эвристические критерии, которые, по мере их использования в анализе данных, постоянно модифицируются и уточняются, в том числе на основе аппроксимационных или статистических соображений. В отличие от данных о сходстве, удачный критерий для таблиц объект–признак был найден достаточно рано: сумма расстояний от объектов до центров соответствующих кластеров, как показано на рис. 1.

Рис. 1. Суммарные расстояния между точками, представляющими объекты и центрами соответствующих кластеров, представленными звездами – суть критерия метода *к*-средних

Переходя к математической формулировке, рассмотрим предобработанную матрицу объект – признак $Y=(y_{iv})$, где столбцы $v=1, \ldots, V$ соответствуют признакам, а строки $i \in I$ – объектам. Кластерная структура метода κ-средних задается разбиением S множества объектов на K непересекающихся кластеров, $S=\{S_1, S_2, \ldots, S_K\}$, представляемых таким образом через кластеры S_k, и центроиды $c_k=(c_{k1}, c_{k2}, \ldots, c_{kV})$, $k = 1, 2, \ldots, K$. Тогда минимизируемым критерием метода является сумма расстояний $d(y_i, c_k)$ от объектов y_i до соответствующих центроидов c_k (4):

$$W(S,c)=\sum_{k=1}^{K}\sum_{i\in S_k} d(y_i, c_k) \tag{4}$$

В случае, когда $d(y_i, c_k)$ представляет собой квадрат Евклидова расстояния, данный критерий может быть выражен как аппроксимационный критерий наименьших квадратов в простой модели, такой, что используется в дисперсионном анализе [5]. Согласно этой модели, каждый объект, представленный строкой $y=(y_1, y_2, \ldots, y_{iv})$ матрицы Y, равен, с точностью до небольших погрешностей, центроиду соответствующего кластера c_k: $y_{iv}=c_{kv}+e_{iv}$ для всех $i \in S_k$, и всех $v=1, 2, \ldots, V$.

Сумма квадратов погрешностей этой модели равна (5):

$$L^2 = \sum_{i\in I}\sum_{v\in V} e_{iv}^2 = \sum_{k=1}^{K}\sum_{i\in S_k}\sum_{v\in V}\left(y_{iv} - c_{kv}\right)^2 , \tag{5}$$

что, в свою очередь, совпадает с выражением для критерия $W(S, c)$ при расстоянии d, равном квадрату Евклидового расстояния [2].

Метод k-средних применим не только в данных предположениях «кластеры – сферы одного и того же радиуса». Из того, что утверждение M влечет утверждение H вовсе не следует, что H влечет M. Ситуация здесь сходна с той, что возникает при применении линейной регрессии: линейная регрессия действительно работает для нормальных распределений, но может применяться и просто для того, чтобы аппроксимировать данные, не обязательно Гауссовые, прямой линией.

Для решения задачи разбиения студентов специальной медицинской группы на подгруппы, в которых студенты могли бы заниматься физическими упражнениями, разрешенными и рекомендуемыми для определенных нозологий, в работе рассмотрена возможность использования кластерного анализа.

Для реализации математической модели, предложенной выше, нами был выбран программный комплекс SPSS Statistics (аббревиатура англ. «Statistical Package for the Social Sciences» – «статистический пакет для социальных наук») – компьютерная программа для статистической обработки данных, один из лидеров рынка в области статистических продуктов, предназначенных для проведения прикладных исследований в социальных науках. В рамках выбранного нами программного инструментария реализованы такие функции:

- ввод и хранение данных;
- возможность использования переменных разных типов;
- частотность признаков, таблицы, графики, таблицы сопряжённости, диаграммы;
- первичная описательная статистика;
- маркетинговые исследования;
- анализ данных маркетинговых исследований.

Из условия поставленной задачи следует, что у нас есть массив данных, состоящий из 13 независимых переменных (утверждений), в различных аспектах описывающих текущее состояние здоровья 220 студентов специальной медицинской группы. Основной задачей проводимого факторного анализа является группировка схожих по смыслу утверждений в макрокатегории с целью сократить число переменных и оптимизировать структуру данных.

В качестве исходной информации была использована таблица Microsoft Excel со всеми необходимыми данными. После запуска программы на выполнение мы получили следующие результаты.

Как уже было отмечено выше, число групп (компонентов) факторной модели определяется при помощи расчета «характеристических чисел» (*Eigenvalues*). Эти показатели характеризуют полноту отображения исходной информации в построенной факторной модели [6].

Значения этих показателей содержатся в таблице *«Total Variance Explained»*, которая выводится на экран компьютера среди прочих результатов факторного анализа (табл. 1).

В первом столбце табл. 1 (*Component*) указывается число компонентов различных вариантов факторной модели. В четвертом столбце этой таблицы (*Cumulative*, %) показан процент информации, сохраненной в процессе группировки исходного массива переменных с помощью факторной модели. Например, если число факторов в факторной модели равно числу переменных исходного массива, т.е. группировка переменных не производится, исходная информация будет сохранена на 100%.

Во втором столбце таблицы *(Total)* указываются значения «характеристических чисел» *(Eigenvalues)*. В рассматриваемом примере было задано условие: значение «характеристических чисел» должно быть больше единицы (*Eigenvalues over* 1). Максимальное значение компонентов факторной модели, в которой

данный показатель превышает единицу, составляет 7. Это означает, что оптимальное число групп (факторов) в факторной модели составляет 7.

Таблица 1

Определение числа компонентов факторной модели Total Variance Explained

Total Variance Explained									
Compo-nent	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	1,463	11,255	11,255	1,463	11,255	11,255	1,415	10,887	10,887
2	1,406	10,815	22,070	1,406	10,815	22,070	1,289	9,913	20,799
3	1,245	9,574	31,645	1,245	9,574	31,645	1,282	9,860	30,659
4	1,235	9,503	41,147	1,235	9,503	41,147	1,278	9,832	40,490
5	1,141	8,773	49,921	1,141	8,773	49,921	1,226	9,430	49,921
6	1,060	8,157	58,077						
7	1,039	7,995	66,072						
8	,942	7,247	73,319						
9	,922	7,090	80,410						
10	,876	6,740	87,150						
11	,745	5,729	92,879						
12	,651	5,011	97,890						
13	,274	2,110	100,000						
Extraction Method: Principal Component Analysis.									

Как видно из данных, представленных в табл. 1, факторная модель, состоящая из семи факторов, сохраняет лишь 66,072% исходной информации. Как отмечалось ранее, при группировке исходного массива переменных потеря информации неизбежна, в то же время при построении факторной модели следует стремиться к минимизации потерь информации.

В ходе формирования задания на проведение факторного анализа также был построен график *«Screen plot»*, с помощью которого можно также определить оптимальное число групп (рис. 2),.

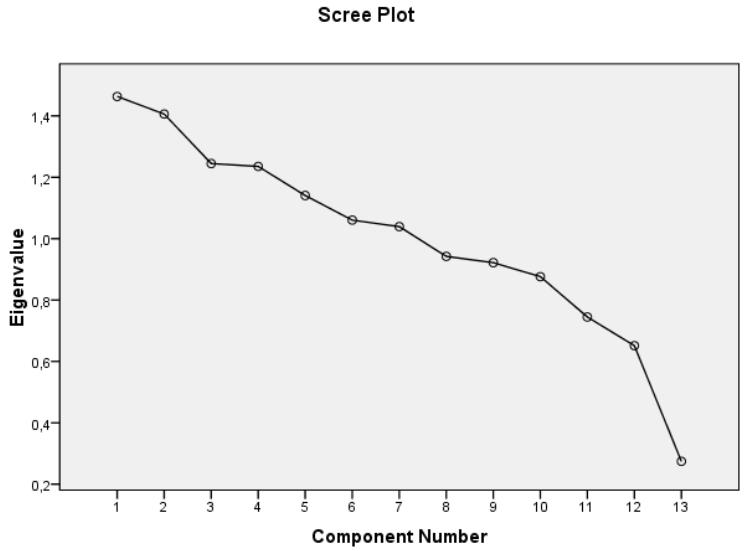

Рис. 2. Графическое определение компонентов факторной модели

На рис. 2 представлен график, отображающий зависимость между «характеристическими числами» (*Eigenvalues)* и числом компонентов факторной модели *(Component Number).* При изменении количества факторов с 7 до 13 данный график представляет собой практически линейную функцию, а

при уменьшении числа факторов с 8 до 7 происходит «перелом» графика. Это означает, что оптимальное число компонентов факторной модели (факторов) равно 7.

Таким образом, результаты графического метода определения числа факторов подтвердили результаты расчетного метода (табл. 1). В результате применения обоих методов оптимальное число компонентов факторной модели составило 7.

Следующим шагом в представлении результатов факторного анализа является ротированная матрица компонентов (табл. 2).

Таблица 2

Ротированная матрица компонентов факторной модели

Rotated Component Matrix[a]					
	Component				
	1	2	3	4	5
x5	-,738				
x3	,737				
x10	,499				
x11		,706			
x8		,573			
x6		,503			
x4			,793		
x12			,623		
x2				-,654	
x1				,654	
x13				-,507	
x7					-,716
x9					,562
Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.					
a. Rotation converged in 5 iterations.					

Метод извлечения: анализ главных компонентов. Метод ротации: «*Varimax*» с нормализацией Кайзера. Ротация получена за 5 итераций.

В табл. 2 представлены коэффициенты корреляции, характеризующие связи между переменными исходного массива данных и компонентами построенной факторной модели (факторами). Согласно общему правилу проведения факторного анализа в одну группу (под одним фактором) собираются переменные исходного массива, имеющие наиболее тесную связь (самое большое значение коэффициента корреляции) с данным компонентом факторной модели. Здесь отмечены максимальные значения коэффициентов корреляции, свидетельствующие о наиболее тесной взаимосвязи переменных исходного массива с компонентами факторной модели. Как отмечалось ранее, при построении факторной модели неизбежна частичная потеря информации. Потеря информации особо ощутима в случае, если отдельные переменные исходного массива данных имеют высокие значения коэффициентов корреляции сразу с несколькими факторами.

Проведя с помощью пакета SPSS факторный анализ имеющихся данных, мы сократили количество рассматриваемых факторов с 13 до 7. Теперь посредством кластерного анализа по методу k-средних можно провести разделение всех студентов на 5 групп. В качестве исходной информации используем результат факторного анализа.

Среди данных, выдаваемых *SPSS* в качестве результатов кластерного анализа, в первую очередь на экран выводится таблица, содержащая результаты сравнения объектов исследования. Первоочередность представления этих данных в качестве результатов обусловливается агломеративным алгоритмом иерархического кластерного анализа (рис. 3). В нашем примере в качестве показателя, характеризующего степень сходства (различия) объектов исследования, был выбран квадрат евклидова расстояния (*Squared Euclidean Distance)*. Чем меньше этот показатель, тем больше сходство сравниваемой пары объектов исследования.

Для определения очередности последующего объединения объектов исследования в кластеры необходимо заново определить квадрат евклидова расстояния между вновь созданным кластером и прочими кластерами.

Результаты расчета квадратов евклидова расстояния для каждого этапа формирования кластеров не выводятся на экран компьютера. Среди данных, выводимых на экран в качестве результатов кластерного анализа, предоставляются лишь результаты сравнения кластеров на этапе, когда каждый объект исследования рассматривается как кластер. Результаты кластерного анализа нагляднее всего

представляются в виде дендограммы. Дендограмма является графическим изображением таблицы «График агломерации». При построении дендограммы SPSS нормирует значения коэффициента, характеризующего степень гетерогенности формируемых кластеров, по шкале от нуля до 25. Дендограмма иллюстрирует увеличение разнородности кластеров по мере их укрупнения. Максимальное значение шкалы дендограммы 25 характеризует максимально возможную степень гетерогенности кластеров, когда все объекты исследования объединены в один кластер. Если объекты исследования разделить на два кластера, то данные кластеры будут значительно более разнородны. Степень их разнородности по шкале дендограммы понизится примерно до 7. После кластерного анализа можно проводить дополнительные исследования, в ходе которых оцениваются особенности выделенных кластеров [6].

Рис. 3. Агломеративный алгоритм иерархического кластерного анализа

Выводы. В условиях большого числа разных (бинарных, вещественных, вероятностных) критериев для описания состояния возникают трудности при разбиении объектов на группы по определённым признакам. В работе рассмотрена возможность использования инструментария пакета статистической обработки информации SPSS Statistica для решения поставленной задачи. В качестве математического аппарата выбран кластерный анализ, которому предшествует факторный анализ. В результате обработки накопленных данных удалось получить результаты, которые можно использовать в работе преподавателя физического воспитания для разбиения студентов специальной медицинской группы на подгруппы по нозологиям.

ЛІТЕРАТУРА:

1. Калинина В. Н., Соловьев В. И. Введение в многомерный статистический анализ: Учебное пособие / ГУУ. – М., 2003. – 92 с.
2. Орлова И.В. Экономико-математические методы и модели. Выполнение расчетов в среде EXCEL / Практикум: Учебное пособие для вузов. – М.: ЗАО «Финстатинформ», 2000. – 136 с.
3. Дубров А. М., Мхитарян В. С., Трошин Л. И. Многомерные статистические методы: Учебник. – М.: Финансы и статистика, 2000. – 352 с.
4. Колемаев В. А., Староверов О. В., Турундаевский В. Б. Теория вероятностей и математическая статистика: Учеб. пособие. – М.: Высшая школа, 1991. – 400 с.
5. Миркин Б. Г. Методы кластер-анализа для поддержки принятия решений: обзор : препринт WP7/2011/03 [Текст] / Б. Г. Миркин; Национальный исследовательский университет «Высшая школа экономики». – М.: Изд. дом Национального исследовательского университета «Высшая школа экономики», 2011. – 88 с.
6. Бююль А., Цефель П. SPSS: искусство обработки информации. Анализ статистических данных и восстановление скрытых закономерностей. – СПб.: «ДиаСофтЮП», 2005. – 608 с.
7. Шупік І.Є. Задачі комп'ютеризації процесів коригування фізичного стану студентів // Вестник Херсонского национального технического университета. – 2004. – № 2(20). – С. 303 – 305.

8. Шупик И.Е., Рогальский Ф.Б. Особенности управления физическим состоянием студентов в процессе физического воспитания // Інтелектуальні системи прийняття рішень та прикладні аспекти інформаційних технологій: Матеріали науково-практичної конференції. Том 4 – Херсон: Видавництво Херсонського морського інституту, 2005. – с. 173 -175.

ВИШЕМИРСЬКА Світлана Вікторівна - к.т.н., доцент кафедри інформатики та комп'ютерних технологій Херсонського національного технічного університету.

Наукові інтереси:

- системи підтримки прийняття рішень, соціотехнічні системи.

РОГАЛЬСЬКИЙ Франц Борисович - к.т.н., професор, завідувач кафедри інформатики і комп'ютерних технологій Херсонського національного технічного університету.

Наукові інтереси:

- інформаційні та керуючі системи; моделювання в технічних та економічних системах.

ШУПІК Ігор Євгенович - завідувач кафедри фізичного виховання Херсонського національного технічного університету.

Наукові інтереси:

- використання інформаційних технологій у фізичній культурі і спорті.