УДК 519.2

Ярмоленко Н.В.,
кандидат технических наук доцент кафедры информатики и математики Санкт-Петербургский военный институт войск национальной гвардии Российской Федерации,
Россия, г. Санкт-Петербург

ПРИМЕНЕНИЕ КЛАСТЕРНОГО АНАЛИЗА В ПРОЦЕССЕ ОБРАБОТКИ И ИНТЕРПРЕТАЦИИ СТАТИСТИЧЕСКИХ ДАННЫХ ПСИХОЛОГИЧЕСКИХ ИССЛЕДОВАНИЙ

Аннотация: *В статье поясняется использование методов математической статистики для обработки данных психологических исследований. Формулируются задачи кластерного анализа, анализируются свойства кластера и понятие расстояния и меру близости между объектами.*

Ключевые слова: *математико-статистические модели и методы; кластер; кластерный анализ; выборочная совокупность; матрица.*

Annotation: *The article explains how to use the methods of mathematical statistics for data processing in psychological research. The problems of cluster analysis are formulated, the cluster properties and the concept of distance and proximity between objects are analyzed.*

Keywords: *mathematical and statistical models and methods; cluster; cluster analysis; sample; matrix.*

На современном этапе изучение психологических явлений проводят с применением квалифицированно организованного неоднородного выборочного наблюдения с целью получить вероятностные оценки параметров исследуемого объекта или процесса в условиях значительной экономии времени, экономических и финансовых ресурсов. Автоматизация наиболее трудоемких операций обработки и анализа массивов психодиагностических данных

выполняется с целью:

− диагностики индивидуальных особенностей (личностных качеств), личностно-профессионального становление, мотивационный профиль и психологического сопровождения личного состава;

− выявления лиц, склонных к срывам социально-психологической адаптации и девиантным формам поведения;

− психологического сопровождения боевого дежурства (боевой службы), караульной и внутренней (дежурной, вахтенной) служб;

− оценки актуального психологического состояния, как отдельных военнослужащих, так и воинских коллективов (подразделений, смен, караулов и т.п.);

− изучения социально-психологических процессов в группах и воинских коллективах (оценка психологической совместимости,

− профилактика конфликтов, комплектование малых групп для выполнения особо ответственных заданий);

− психодиагностического обеспечения кадровых решений; оценки функциональной готовности к выполнению профессиональных обязанностей, оценки влияния факторов деятельности на работоспособность и др.

Последние достижения в развитии информационных технологий позволяют пользователю с минимальными навыками работы на современных IBM-совместимых персональных компьютерах успешно применять их на практике для различных вычислительных процедур прикладного статистического анализа, в том числе и для психодиагностического обследования. Как известно, особенно актуальной проблема трудоемкости становится при многомерном анализе данных. Поэтому эффективная, грамотная, всесторонняя статистическая обработка данных даже небольшого объема практически невозможна без использования вычислительной техники.

Функциональные возможности современного компьютера позволяют обрабатывать большие объемы информации, а информационные технологии сделали доступными для широкого пользователя самые современные методы

статистического анализа. Разработанные статистические пакеты прикладных программ представляют современную форму программного обеспечения для овладения основами статистических методов.

Благодаря уменьшению объема изучаемой совокупности, используется объективная возможность углубить программу исследования, т.е. получить более детальную информацию об объекте изучения. Военный психолог стоит перед лицом множества индивидуальных наблюдений, характеризующих их параметров и выявление структуры взаимосвязей между ними. И поэтому применение методов многомерного статистического анализа необходимо для принятия обоснованного решения, оценки надежности и точности выводов[1].

Для разделения неоднородной совокупности наблюдаемых объектов на однородные группы, классифицировать эти группы наблюдения выполняет кластерный анализ, разделяя их таким образом, что в каждой группе они похожи друг на друга больше, чем объекты из других классов. Такие группы называют кластерами. Итак, задача, которая стоит перед кластерным анализом, состоит в разбиении множества объектов на определенное количество кластеров таким образом, чтобы каждый объект выборочной совокупности принадлежал одному кластеру (подмножеству объектов) с определенными свойствами.

Обычной формой представления исходных данных в задачах кластерного анализа служит матрица

$$X = \begin{pmatrix} x_{11} & \dots & x_{1j} & \dots & x_{1k} \\ \dots & \dots & \dots & \dots & \dots \\ x_{i1} & \dots & x_{ij} & \dots & x_{ik} \\ \dots & \dots & \dots & \dots & \dots \\ x_{n1} & \dots & x_{nj} & \dots & x_{nk} \end{pmatrix},$$

каждая строка которой представляет результаты измерений k рассматриваемых признаков у одного из обследованных объектов. В конкретных ситуациях может представлять интерес как группировка объектов, так и группировка признаков. В тех случаях, когда разница между двумя этими задачами не существенна, например, при описании некоторых алгоритмов, мы будем пользоваться только термином «объект», включая в это понятие и термин «признак» [2].

Оставшиеся объекты принадлежат другим кластерам с иными свойствами. К этим свойствами относят плотность, дисперсия, размеры, форма и отделимость. Плотность − это свойство, которое позволяет определить кластер как скопление точек в пространстве данных, относительно плотное по сравнению с другими областями пространства, содержащими либо мало точек, либо не содержащими их вовсе. Дисперсия характеризует степень рассеяния точек в пространстве относительно центра кластера, т.е. насколько близко друг к другу расположены точки кластера. Свойство кластера − размеры - тесно связано с дисперсией; если кластер можно идентифицировать, то можно измерить и его «радиус». Это свойство полезно лишь в том случае, если рассматриваемые кластеры являются гиперсферами (т.е. имеют круглую форму) в многомерном пространстве, описываемом признаками. Форма − это расположение точек в пространстве. Если кластеры имеют удлиненную форму, то вместо размера можно вычислить его «связность» - относительную меру расстояния между точками. Отделимость характеризует степень перекрытия кластеров и насколько далеко друг от друга они расположены в пространстве [3].

В кластерном анализе различают понятие расстояния и меру близости между объектами. Меры близости отличаются от расстояний тем, что они тем больше, чем более похожи объекты. Каждый элемент совокупности – это отдельный кластер и между кластерами по какой-то из описанных выше формул найдено расстояние в виде матрицы D, симметричной относительно главной диагонали:

$$D = \begin{pmatrix} d_{11} & d_{12} & \dots & d_{1n} \\ d_{21} & d_{22} & \dots & d_{2n} \\ \dots & \dots & \dots & \dots \\ d_{n1} & d_{n2} & \dots & d_{nn} \end{pmatrix},$$

где $d_{ij} = d(Y_i, Y_j)$.

Из всех расстояний выбираем наименьшее и объединяем соответствующие этому расстоянию элементы в одну группу (кластер). В результате два столбца (и две строки) матрицы преобразуются в один. Пересчитываем на основе одного из принципов расстояния между вновь образованным кластером и остальными

элементами. Получаем новую матрицу расстояний D_1, в которой будет на одну строку и один столбец меньше по сравнению с D. Далее в матрице D_1 находим минимальное расстояние и аналогично объединяем соответствующие кластеры и т. д. Процесс может быть закончен на любой стадии в зависимости оттого, какое количество групп требуется в задаче.

Для того чтобы выяснить, входят ли два конкретных элемента в одну группу или нет, необходимо определить некую меру близости между ними. Иначе говоря, необходимо ввести в рассмотрение некую формулу, которая определяет расстояние между двумя различными объектами. В качестве этой формулы достаточно часто используются:

1. Евклидово расстояние:

$$d_{ij} = d(Y_i, Y_j) = \sqrt{\sum_{k=1}^{m} (x_{ki} - x_{kj})^2}.$$

2. Квадрат Евклидова расстояния:

$$d_{ij} = d(Y_i, Y_j) = \sum_{k=1}^{m} (x_{ki} - x_{kj})^2.$$

3. «Взвешенное» евклидово расстояние:

$$d_{ij} = d(Y_i, Y_j) = \sqrt{\sum_{k=1}^{m} w_k (x_{ki} - x_{kj})^2}.$$

4. Хеммингово расстояние: $d_{ij} = d(Y_i, Y_j) = \sum_{k=1}^{m} |x_{ki} - x_{kj}|$.

5. Расстояние Чебышева: $d_{ij} = d(Y_i, Y_j) = \max_{1 \le k \le n} |x_{ki} - x_{kj}|$.

Из всех расстояний выбираем наименьшее и объединяем соответствующие этому расстоянию элементы в одну группу (кластер) [4]. В результате два столбца (и две строки) матрицы преобразуются в один. Пересчитываем на основе одного из принципов расстояния между вновь образованным кластером и остальными элементами. Получаем новую матрицу расстояний D_1, в которой будет на одну строку и один столбец меньше по сравнению с D. Далее в матрице

D_1 находим минимальное расстояние и аналогично объединяем соответствующие кластеры и т. д. Процесс может быть закончен на любой стадии в зависимости оттого, какое количество групп требуется в задаче.

В психологических исследованиях объектами для кластеризации могут быть как испытуемые, так и признаки (шкалы психодиагностических методик, показатели социально-демографических характеристик и т.д.), а также корреляции, результаты социометрии и т.д.

Особенности мотивационного профиля личности военнослужащего исследованы математической процедурой кластерного анализа для решения вопроса о группировке испытуемых. Программная реализация алгоритмов кластерного анализа широко представлена в различных инструментах Data Mining, которые позволяют решать задачи достаточно большой размерности. Например, агломеративные методы реализованы в пакете SPSS, дивизимные методы – в пакете Statgraf.

Для разбиения объектов на кластеры использованы иерархические англомеративные методы кластерного анализа. К первому кластеру отнесены испытуемые с высокими значениями как активности, так и общения, но невысокими показателями состояния устремлений и житейского состояния. Во второй кластер отправили испытуемых с высоким показателем мотивационного профиля устремления как житейского, так и учебного, а так же социального статуса. В третьем кластере отнесены испытуемые, обладающими значительными значениями житейского (идеального) состояния, но с невысокими показателями общественной полезности. Четвертый кластер характеризуется высокими показателями житейского (идеального) состояния, однако показатель учебное идеальное имеет низкую величину. В пятый кластер отправили испытуемых с показателями профессионального (реального) состояния высокими, но показатель учебного (идеального) – низкого уровня.

Анализируя полученные итоги исследования, необходимо обратить внимание на тот факт, что в мотивационном профиле испытуемых превалирует аппарат социального поддержания жизненного процесса и потребления, однако,

значительно выше идеальная повседневная и идеальная рабочая мотивация. Таким образом, высокий показатель идеальной мотивации говорит о наличии развивающихся мотивов, значительном уровне побуждения и устремления. Наиболее показательные признаки, которые характеризуют исследуемую выборку испытуемых, представлены показателями: поддержание жизнеобеспечения и идеальное положение устремлений. Если вести классификацию мотивационного профиля, то при этом выделяют группу акмеориентированных испытуемых (кластер 1), выбравших образовательную направленность, где наблюдается поступательное прогрессивное личностно-профессиональное развитие, и группу, имеющую затруднения в собственном становлении (кластер 5).

Оценка мотивационного профиля испытуемых позволяет констатировать, что испытуемые, обладающие выраженной духовной и гражданской направленностью, ориентированы на развитие, творческий подход к решению возникающих задач, удовлетворение личных мотивов в контексте социальных интересов. Их личностно-профессиональное становление предполагает непрерывный процесс саморазвития. Интегральный характер психических свойств, связь психологи со всеми науками не только о человеке, но и о мире живого, комплексный характер психологических проблем требует совместных усилий специалистов разных областей знания в исследовании природы психических свойств и функций. Эти уровни и аспекты психического анализа представляют собой совокупность требований при построении математических моделей психических свойств и форм деятельности [5].

Таким образом, кластерный метод позволяет разделить неоднородную совокупность наблюдаемых объектов психологического исследования на однородные группы, элементы которых характеризуются многими признаками, или близкие по определяющим критериям группы (кластеры), пригодные для дальнейшего исследования. В случае необходимости определенным образом систематизировать большие объемы информации к пригодным для дальнейшей обработки кластерный анализ оказывается весьма полезным и эффективным.

ИСПОЛЬЗОВАННЫЕ ИСТОЧНИКИ

1. Айвазян С.А. Многомерный статистический анализ и эконометрика. Труды VIII Международная школы-семинары "Многомерный статистический анализ и эконометрика", Цахкадзор, Республика Армения [Текст]: труды / Центральный экономико-математический ин-т РАН [и др.]. − Москва: ЦЭМИ РАН, 2012. − 216 с.

2. Айвазян С.А., Мхитарян В.С. Прикладная статистка и основы эконометрики. Учебник. - М: ЮНИТИ, 1998. – 1022 с.

3. Ким Дж.-О., Мьюллер Ч.У. и др. Факторный, дискриминантный и кластерный анализ. Под ред. И.С. Енюкова. – М.: Финансы и статистика, 1989. – 215 с.

4. Мандель И.Д. Кластерный анализ. — М.: Финансы и Статистика, 1988. – 176 с.

5. Ярмоленко Н.В. Особенности применения методов математической психологии в профессиональной деятельности военных психологов// Сборник научных статей VIII Межвузовской научно-практической конференции с международным участием «Направления и перспективы развития образования в военных институтах войск национальной гвардии российской федерации». − 2017.− С. 346-349.