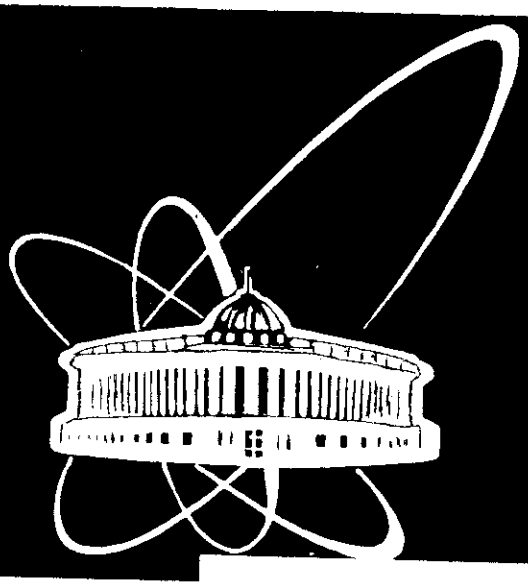

ОБЪЕДИНЕННЫЙ
ИНСТИТУТ
ЯДЕРНЫХ
ИССЛЕДОВАНИЙ

Дубна

P10-99-105

В.Н.Самойлов

МЕТОДЫ АНАЛИЗА ДАННЫХ ФИЗИЧЕСКИХ ЭКСПЕРИМЕНТОВ И СЛОЖНЫХ ПРОЦЕССОВ

Направлено в журнал «Известия РАН. Теория и системы управления»

1999

1. Введение

В настоящее время не известен механизм установления нормируемости и оценки состояния эволюции технологических процессов, несмотря на наличие ряда традиционных методов создания разного рода нормативов [1, 2]. Знание этого механизма необходимо для построения модели измерения в реальном времени и соответствующей информационной модели, определяющих эффективность моделирования слабоструктуризованных процессов [3]. Умение выделить из набора данных существенные параметры создает необходимые предпосылки для более строгого решения задачи с применением методов оптимизации [4]. Независимо от вида и масштаба рассматриваемого объекта, для описания дерева целевых функций и сетей, отображающих взаимосвязи, всегда актуальна задача минимизации состава исходных данных. С такой же проблемой сталкивается пользователь как при постановке задачи и формировании цели, так и при оценке полученного результата. Для решения этой задачи естественно использовать метод анализа соответствий, построенный на основе метода главных компонент и выбора χ^2 - метрики [5-7]. Метод обеспечивает одновременное представление на факторной плоскости как наблюдений, так и переменных вместе с существующими между ними связями, позволяет установить отношение эквивалентности на множестве объектов и построить подходящую модель для синтеза состава информации методом группирования переменных.

Группирование переменных представляет собой разбиение переменных на классы эквивалентности, как правило, не пересекающиеся. В целом методы группирования можно разделить на геометрические и алгебраические. В алгебраических методах рассматриваются отношения лишь между парами исходных элементов [8-10]. В геометрических методах исходные данные, представленные в виде совокупности точек в векторном пространстве большой размерности, требуется спроецировать в пространство более низкой размерности с помощью методов факторного

анализа [4,11] либо выявить на основе анализа центров динамических классов наиболее однородные разбиения и их представления в виде графов [12]. Существует большое число методов группирования, которые различаются между собой исходными данными, способами оценки сходства, процедурами группирования и видом представления результатов. Однако в любом случае присутствуют этапы измерения сходства и группирования.

Известны методы простого, полного и среднего связывания переменных, восходящего иерархического группирования (основанные на введении ультраметрики) и т. д. Методы группирования, основанные на обучающих выборках, не могут использоваться для анализа информационной структуры, так как наличие обучающей выборки подразумевает знание информационной структуры. Кроме того, большинство методов группирования имеют сходные недостатки, например, наличие цепного эффекта и необходимость установления некоторых порогов произвольным образом и т. д. Таким образом, разработка методов группирования свободных от подобных недостатков является актуальной задачей.

В настоящей работе предложен механизм нормируемости на основе выявления зон устойчивого функционирования технологий. Сформулирован метод анализа соответствий для модифицированной модели измерений с унифицированным представлением качественных переменных и без огрубления результатов измерений количественных переменных. Построен итерационный метод динамического группирования переменных, использующий понятие «эталонных» точек, и реализующий его конструктивный алгоритм, связанный с анализом соответствующим образом построенных графов.

Структура работы следующая: в разделе 2 рассмотрены составляющие жизненного цикла и вопросы нормируемости функциональных характеристик эволюционных технологических процессов; в разделе 3.1 обсуждается анализ соответствия данных информационной модели; в разделах 3.2 и 3.3 предложены унифицированная и

модифицированная модели измерений, критерий поиска факторных осей для анализа соответствий; в разделах 4.1 и 4.2 рассмотрены метод динамического группирования переменных и итерационный метод группирования графов.

В заключении обсуждаются перспективы применения предложенного подхода для моделирования сложных процессов.

2. Жизненный цикл и нормируемость функциональных характеристик сложных процессов

Как известно, все эволюционные процессы проходят следующие стадии развития [1, 2]:

- становление (развитие);
- функционирование (стабилизация);
- затухание.

Эти этапы объединены условным понятием «жизненный цикл процесса». С позиции информационного отображения три указанных этапа существенно различаются по следующим признакам:

- источники информации;
- возможности обработки информации;
- оценка состояния объекта.

На рис. 1 представлены примеры типичных жизненных циклов (ЖЦ) процессов. Так как кривые ЖЦ строятся, как правило, из реально решенных задач, носящих локальный характер, наиболее типичным является процесс, представленный графиком (а), где участок "1-2" отражает не развитие процесса, а его совершенствование и корректировку до получения устойчивого состояния. Однако значительный промежуток времени $t_1 - T_2$ весьма часто создает ложное представление о роли и месте этого процесса по отношению к общей тенденции развития. В результате теряется острота внимания к необходимости упреждения своевременного выявления точки "2", т.е. точки перехода процесса в режим затухания.

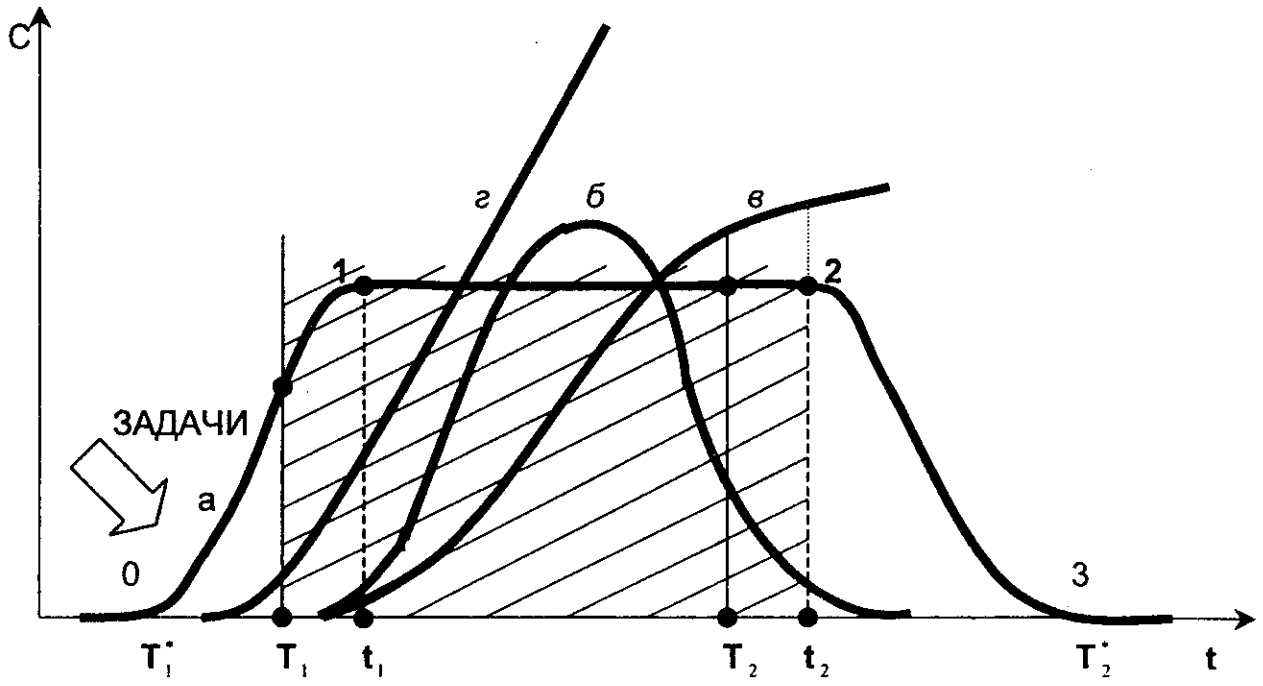

Рис.1. Типы жизненных циклов.

Практически всегда возникает стандартная ситуация, когда возвращение назад к выявлению первоначального этапа ЖЦ "0-1" находится вне поля внимания и чаще всего забыто, а появление точки 2 после длительного периода $t_1 - T_2$ стабильного функционирования процесса часто бывает настолько неожиданным, что приводит к появлению различного рода нештатных ситуаций. Здесь следует особо отметить, что недостаточное внимание к процессам, подобным (а), приводит к привлечению дополнительных средств, и формирует принципиально неверное отношение к задаче, выдвигая на первый план догматические решения, характерные только для участка "1-2". Процесс чаще всего рассматривается локально и автономно, а режим "1-2" становится тормозом по отношению к рассмотрению каждого ЖЦ процесса как незамкнутой, локальной системы со своими измерениями, оценками и технологиями принятия решений.

Из вышесказанного вытекает, что областью измерения функциональности, как основного свойства информационной технологии [3], являются:

1. Участки кривых ЖЦ "0-1" – для получения ответов на вопрос "Что

делать?".

2. Участки «1-2" - для вопросов и ответов "Что делать?" и "Как делать?".

3. Участки «1-3» - для процессов с длительным временем протекания на участках "1-2", как наиболее характерных процессов совершенствования и корректировки и как результат проявления постановки автономных, локальных задач.Эти участки кривой жизненного цикла процесса должны быть выявлены при рассмотрении и информационном отображении структуры и содержания процессов высоких технологий [3]. Так, на первом этапе, когда процесс только "выкристаллизовывается", основным источником для информационного описания являются задачи, которые отражают объективные потребности в данный момент времени *t*. По мере накопления опыта в решении задач, характерных для данного процесса, источником информации для его отображения, кроме локальных задач, является накопленный опыт [3]. Этот опыт трансформируется в нормативы, позволяющие выделить и оценить стабильные состояния процесса. Под нормативами понимается комплекс повторяющихся операций, отдельные данные, задачи и комплексы задач, которые должны обеспечивать функционирование процесса.

Рассмотрим жизненный цикл технологического процесса в виде зависимости его некоторой относительной характеристики C от времени функционирования t (рис. 1). Период жизненного цикла ($T_{жц}$) определяется длиной интервала $[T_1^*, T_2^*]$, где T_1^* - время, соответствующее началу жизненного цикла, T_2^* – время, соответствующее концу жизненного цикла. Следовательно, период жизненного цикла есть

$$T_{жц} = T_2^* - T_1^* .$$

Задача информационного описания объекта характеризуется своим временным интервалом $[T_1, T_2]$, где T_1 - время, соответствующее началу решения задачи информационного описания, T_2 - время, соответствующее

концу решения задачи информационного описания. Время решения задачи $T_{рз}$ информационного описания, определяемое как

$$T_{рз} = T_2 - T_1$$

должно удовлетворять следующему ограничению:

$$T_{рз} \le T_{\text{жц}},$$

которое следует из неравенств:

$$T_2 \le T_2^* \ , \ T_1 \ge T_1^* \ .$$

Отметим, что для различных объектов можно схематично составлять различные графики $C(t)$, поэтому для их сравнительной оценки зададим три типа графиков (см. рис. 1) за счет выделения трех основных интервалов на условной кривой, описывающей жизненный цикл объекта.

Данные интервалы, соответствующие выделенным выше этапам жизненного цикла, в дальнейшем будем именовать стадиями жизненного цикла объекта, а именно:

- стадией "становления" - интервал (0,1) , соответствующий этапу становления;
- стадией "стабильного функционирования" - интервал (1,2);
- стадией "затухания" - интервал (2, 3).

Тогда временная ось разбивается на три интервала:

$$I_1 = (T_1^*, t_1),\ I_2 = (t_1, t_2),\ I_3 = (t_2, T_2^*) \ .$$

Точки 0, 1, 2 и 3 на кривой $C(t)$ назовем характерными точками жизненного цикла, а соответствующие точки временной оси - переходными точками, которые являются условно нормативными.

Строго говоря, изображение процесса схематично, это - лишь его условная геометрическая интерпретация. На самом деле в теории систем отображение процесса является достаточно сложной проблемой [14]. Следует отметить, что основная цель подобной геометрической интерпретации состоит в получении значений переходных точек на временной оси. Далее необходим анализ свойств стадий жизненного цикла

по трем признакам, приведенным выше, с позиции информационного отображения.

В случае, когда переходные точки удается определить на основе функциональной зависимости, то свойства функции $C(t)$ предположительно должны быть следующими:

- для временного интервала I_1 характерно монотонное возрастание функции $C(t)$;

- на интервале I_2 функция монотонно возрастает и затем монотонно убывает, при этом выполняется условие:

$$\Delta\varepsilon_1 \geq A \cdot \Delta\varepsilon_2,\ \Delta\varepsilon_3 \geq A \cdot \Delta\varepsilon_2$$

$$\Delta\varepsilon_i = \max_{t \in I_i} \left| \varepsilon_i(t) - r_i(t) \right|,\ i = 1,2,3$$

где A – положительная константа, $A \geq \left| \dot{I}_2 \right|$, $\varepsilon_i(t)$ и $r_i(t)$ нормативное и частное отклонения аппроксимации характеристик функции, т.е., величина относительного изменения $\Delta\varepsilon_i$, функции $C(t)$ на интервалах I_1 и I_3 по модулю отличается от величины относительного изменения $\Delta\varepsilon_2$ функции $C(t)$ на интервале I_2 на константу A;

– на интервале I_3 функция монотонно убывает.

По наличию и характеру поведения кривых на стадиях жизненного цикла можно выделить три типа кривых, задающих условно характер поведения жизненных циклов, что и изображено схематично на рис. 1 с указанием характерных точек 0, 1, 2 и 3 на графике (а).

Наибольший интерес представляют стадии "становления" и "стабильного функционирования", т.е. интервалы кривой (0, 1) и (1, 2), поэтому в дальнейшем основное внимание будет уделено стадиям жизненного цикла, соответствующим первым двум интервалам времени (I_1 и I_2).

Задачу информационного описания применительно к различным интервалам будем определять, исходя из двух противоположных концепций, суть которых состоит в следующем:

- для интервала (0, 1) характерно исследование возможных отображений задач, т.е. понимание сути исходных задач;

- для интервала (1, 2) (интервала поддержания функционирования) заданное информационное отображение формируется от информационного описания к установлению закономерностей в классификации задач.

Общим для стадии жизненного цикла является обязательное представление сравнимых характеристик, с помощью которых либо решается задача информационного описания, исходя из первой концепции (на интервале I_1), либо она определена второй концепцией (на временном интервале I_2). Отметим важный аспект, заключающийся в том, что работая на уровне первого приближения, исследуется решение "причина-следствие", не раскрывая сути их связи через функциональные законы, поэтому возникает необходимость определения возможных областей образования структур и времени их существования.

Системный анализ позволяет находить элементы системы и их взаимоотношения. Однако возможности различных системных представлений одного и того же объекта открывают путь не только к всестороннему анализу, но и к внесению произвола в интерпретации объекта. Вследствие такой интерпретации результатов исследования нередко объект, как некоторая объективная целостность, исчезает из рассмотрения, и остается лишь предмет исследования, определяемый условиями данной задачи. Эта целостность, утрачиваемая в рамках данной специальной области науки, может быть восстановлена на основе перехода от системного анализа к изучению структуры, поскольку структурный подход позволяет выработать принципы отбора необходимых решений среди многообразия системных рассмотрений. Очевидно, что ни один из этих подходов сам по себе не имеет преимущества перед другими:

"структура немыслима вне системы, как и система в своей основе всегда структурна" [15].

Структуру или структурное образование объекта определим на основе понятия отношения [12]. Пусть имеется система не обязательно различных множеств,

$$A_1, A_2, \ldots, A_n \ .$$

Рассмотрим их декартово произведение

$$D = A_1 \otimes A_2 \otimes \ldots \otimes A_n .$$

Предположим, что кортеж $(a_1, a_2, \ldots, a_n)$ - элемент множества D:

$$a_i \in A_i, i = 1,2,\ldots,n$$

удовлетворяет некоторому условию $(a_1, a_2, \ldots, a_n) \in F$, тогда говорят, что элементы находятся в отношении F : $F(a_1, a_2, \ldots, a_n)$.

Пусть заданы конечные множества $E_1, E_2, \ldots, E_N$ и элементы $e_j \in E_j$, $j = 1,2,\ldots N$ находятся в отношении F, т.е. задано $F\ (e_1, e_2, \ldots, e_N)$, и данное отношение допускает геометрическую интерпретацию в виде графа $G[E, F(E)]$. Данный граф будем называть структурой, порожденной отношением на декартовом произведении множеств $E = E_1 \otimes E_2 \otimes \ldots \otimes E_N$. Введённые понятия естественно использовать для построения информационных моделей слабоструктуризованных и структуризованных процессов.

Так как выделены участки кривых (0, 1) и (1, 2) как составные части одного целого, то представляется целесообразным построение структур по единым методологическим принципам с обязательным выполнением условия классификации переменных по составу и содержанию характеристик (параметров, показателей и признаков). Данную процедуру назовем параметрическим наполнением. Определение понятия структуры сводится к утверждению, что это способ организации – взаимосвязи отдельных частей и элементов. Структура - относительно устойчивое единство элементов, их отношений и целостности объекта, инвариантная

характеристика системы. Последнее свойство характерно для интервала (1, 2), поэтому возникает необходимость выделения в полной структуре:

- постоянной части;
- условно-постоянной;
- переменной части

относительно фиксированного интервала времени $\Delta t = I_2$.

Рассмотрим понятие "функциональность" применительно к точкам 0, 1, 2, 3 (рис 1). Очевидно, что характер зависимостей, отражающих функциональность для участков "0-1", "1-2" и "2-3" носит различный характер. Если для участка "2-3", т.е. для процессов функционирования и затухания, характер функциональных зависимостей позволяет обеспечить только режим поддержки заданных функциональных характеристик, то на участке "0-1" первостепенным является выявление уровня стабилизации, который конкретизируется на участке "1-2". Таким образом, для участков кривых "2-3", т.е. для участка функциональной поддержки, свойства функциональности проявляются только в вопросах и ответах «Как делать», здесь главными являются процессы регулирования по нормативным и частным отклонениям. Ответы на вопрос «Что именно регулировать?» должны быть получены на ранних стадиях процесса, т.е. на участке кривой "0-1", и уточнение вопросов и ответов "Что делать" на участке "1-2". Если участок "0-1" служит для построения возможных траекторий развития и течения процессов (см. графики в, г), то участок "1-2" – для возможности постановки вопросов и получения ответов "Что" и "Как" в ретроспективе и анализа состояния в данный момент времени и в перспективе (см. графики а, б). Если перспектива формируется как "затухание", то роль участка "1-2" сводится к коррекции параметров и собственно процесса регулирования (см. график а). Если мы имеем случай, когда неизвестны границы участков "1-2" и "2-3", выяснение "Что" и "Как" продолжается до тех пор, пока не будет определена зона перехода процесса в затухающий режим.

3. МОДИФИЦИРОВАННЫЙ МЕТОД АНАЛИЗА СООТВЕТСТВИЙ

3.1. Информационная модель

Метод анализа соответствий [5-7] применяется к исходным данным, представленным в виде таблиц сопряженности или наблюдений, и обеспечивает возможность одновременного представления на факторной плоскости (ФП) как переменных, так и наблюдений. Как отмечалось ранее, проблема выбора переменных встречается при решении многих задач [4,12]. При этом принципы построения процедур выбора переменных определяются априорными знаниями о решаемой задаче и характером представления исходной информации. В корреляционном анализе обработка происходит без преобразования исходных переменных, а в анализе главной компоненты (АГК) - с преобразованием. Причем в информационной модели (ИМ) допускаются самые слабые шкалы при минимальном числе априорных предпосылок, т.е., по существу, в ИМ определяются лишь отношения эквивалентности.

Анализ соответствий (АС) позволяет одновременно представлять наблюдения и переменные вместе с существующими между ними связями и дает возможность наглядно представить отношения эквивалентности на множестве объектов E и получить содержательные выводы как о полноте информационной модели, так и о виде модели обобщенных связей.

Непосредственная оценка полноты информационной модели в соответствии с ее формальным определением оказывается чрезвычайно громоздкой задачей из-за ее комбинаторного характера. По взаимному расположению точек-наблюдений на факторной плоскости можно получить приближенную оценку полноты, причем ее можно уточнить в результате построения факторных плоскостей для контролируемых и для выходных переменных, то есть для набора C и Y. Анализ взаимного расположения точек-переменных на факторной плоскости позволяет выявить связи между переменными.

Анализ соответствий принципиально не отличается от анализа главной компоненты и может рассматриваться как частный случай разложения Карунена-Лоэва, в котором используется χ^2 - метрика:

$$d^2(i,k)=\sum_{j=1}^{N}\frac{1}{c_j}\left(\frac{y_{ij}}{r_i}-\frac{y_{kj}}{r_k}\right)^2,$$

где i и k - два объекта (или признака), y_{ij}, c_j и r_i - оценки совместной и маргинальных вероятностей по отношению к объектам (или признакам) [13]. Метрика такого типа не является чем-то исключительным, хотя в ней «прямой» угол оказывается отличным от прямого угла с точки зрения обычной евклидовой метрики.

Рассмотрим следующий пример, который дает формальное обоснование возможности совместного рассмотрения наблюдений и переменных на ФП, например, в плоскости R^2, некоторой точки $a=(a_1,a_2)$ и произвольной точки $z=(x_n,y_n)$ на прямой, заданной осью $\vec{u}$. Мера близости между a и z в соответствии с χ^2 -метрикой определяется следующим образом (см. рис. 1):

$$\rho^2(\vec{a},\vec{z})=\alpha_1(x_n-a_1)^2+\alpha_2(y_n-a_2)^2,$$

где α_i – вес i -ой координаты, $i=1,2$. Для того, чтобы построить проекцию радиус-вектора $\vec{OA}=\vec{a}$ точки a на $\vec{u}$, на прямой $u=(u_1,u_2)$ следует найти подбором подходящего значения параметра k точку $x=(x_n=ku_1,y_n=ku_2)$, наименее удаленную от точки a,

$$\rho^2(\vec{a},\vec{z})=\alpha_1(ku_1-a_1)^2+\alpha_2(ku_2-a_2)^2.$$

Из условия равенства нулю производной этого выражения по параметру k, т.е.

$$\left.\frac{d\rho^2}{dk}\right|_{k=k_{\chi^2}}=0,\ \alpha_1\left(k_{\chi^2}u_1-a_1\right)u_1+\alpha_2\left(k_{\chi^2}u_2-a_2\right)u_2=0,$$

следует требуемое выражение для параметра k в R^2 и, аналогично, в R^n:

$$k_{\chi^2} = \frac{\alpha_1 a_1 u_1 + \alpha_2 a_2 u_2}{\alpha_1 u_1^2 + \alpha_2 u_2^2} \qquad \text{и} \qquad k_{\chi^2} = \frac{\sum_{i=1}^{n} \alpha_i a_i u_i}{\sum_{i=1}^{n} \alpha_i u_i^2}.$$

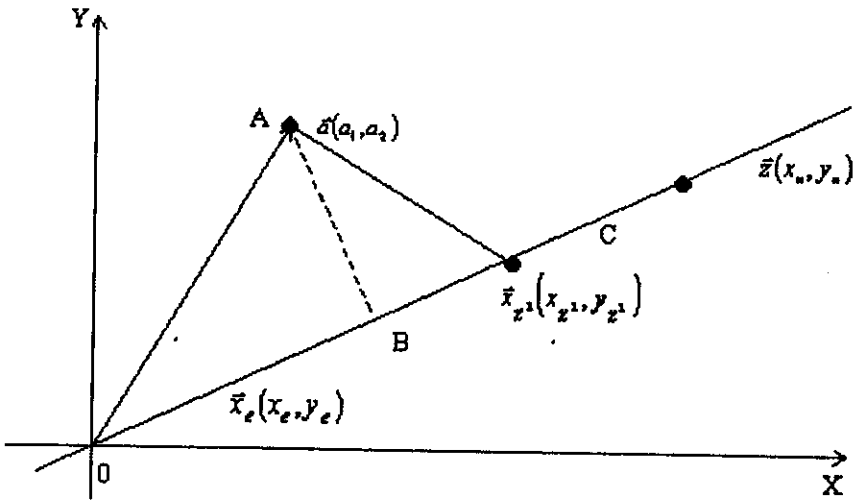

Рис.1 Проецирование в соответствии с χ^2 - метрикой.

Таким образом, координаты проекции радиус-вектора $\vec{OA}$ точки a на $\vec{u}$ при использовании χ^2 - метрики равны:

$$x_{\chi^2} = k_{\chi^2} u_1 \qquad \text{и} \qquad y_{\chi^2} = k_{\chi^2} u_2.$$

Заметим, что только в случае однородных весов $\alpha_1 = \alpha_2$ эти координаты совпадают с координатами проекции при использовании обычной евклидовой метрики:

$$\left\|\vec{OA}\right\|_e^2 = \left\|\vec{OB}\right\|_e^2 + \left\|\vec{AB}\right\|_e^2 \quad , \quad \left\|\vec{OA}\right\|_{\chi^2}^2 = \left\|\vec{OC}\right\|_{\chi^2}^2 + \left\|\vec{AC}\right\|_{\chi^2}^2.$$

Действительно,

$$\left\|\vec{OC}\right\|^2_{\chi^2} + \left\|\vec{AC}\right\|^2_{\chi^2} = k_{\chi^2}\left(\alpha_1 u_1^2 + \alpha_2 u_2^2\right) + \alpha_1\left(a_1 - k_{\chi^2} u_1\right)^2 + \alpha_2\left(a_2 - k_{\chi^2} u_2\right)^2 =$$

$$= \alpha_1 a_1^2 + \alpha_2 a_2^2 = \left\|\vec{OA}\right\|^2_{\chi^2}$$

Как указывалось выше, АС предназначен для обработки данных, представленных в виде таблиц сопряженности (ТС). Отмеченная возможность его применения непосредственно (без предварительного перехода к ТС) к таблицам наблюдения типа «объект-признак» базируется на предварительном преобразовании исходных данных и соответствующей модели измерений. Поэтому целесообразно перечислить основные преобразования, связанные с наиболее распространенными типами модификации данных, с которыми приходится иметь дело при решении практических задач. Ими же, в частности, определяются те преобразования или возможности по формированию массивов, которые необходимо предусмотреть в программах АС, предназначенных для использования на ЭВМ.

С точки зрения применимости АС таблицы сопряженности характеризуются тремя следующими основными свойствами: однородностью данных, симметрией в трактовке переменных и наблюдений, а также исчерпывающим представлением областей возможных значений. Однородность данных означает, что все величины, представленные в соответствующей матрице K, имеют одинаковый характер, а сумма элементов в строке или столбце строго положительна. Симметрия в математической трактовке наборов I и J приводит к их единообразной обработке, причем сумма элементов в строке или столбце равна единице. По существу, этой симметрией и объясняется возможность одновременного представления на факторной плоскости наблюдений и переменных. Исчерпываемость представления областей возможных значений означает, что градациями $i \in I$ и $j \in J$ исчерпываются все

возможные значения анализируемых переменных. Это требование совпадает с требованием к ИМ при классификации переменных.

Таким образом, для применимости АС к таблицам наблюдений у последних следует сформировать аналогичные свойства. Для получения этих свойств у различных типов моделей измерений могут потребоваться преобразования следующих типов: переход к унифицированному представлению и переход к представлению типа таблицы сопряженности. Во втором случае данные сначала преобразуются в однородные.

3.2 Унифицированная модель измерений

Для дальнейшего изложения большое значение будет иметь преобразование модели измерений, связанное с переходом к унифицированному представлению данных, поэтому рассмотрим его более подробно. Унификация проводится в два этапа, что можно представить в виде соответствующей последовательности операций по модификации исходной модели измерений (МИ):

$$МИ = \langle E, V\{K_\nu, Y_\nu\}_{\nu \in V} \rangle .$$

На первом этапе множество возможных значений K_ν каждой количественной переменной ν разбивается на m_ν непересекающихся подмножеств и представляется в виде $K_\nu = \bigcup_{p=1}^{m_\nu} K_p^{(\nu)}$, причем попадание измерения в $K_p^{(\nu)}$ означает запись результата измерения по ν в виде градации $K_p^{(\nu)}$, т.е. вводится новая функция $y'_\nu : \bigwedge_{\eta \in E} y'_\nu(\eta) = K_p^{(\nu)}$, если $y_\nu(l) \in K_p^{(\nu)}$, а K_ν заменяется на $K_{\nu 1} = \{K_1^{(\nu)}, \ldots, K_{m_\nu}^{(\nu)}\}$. В случае неколичественных переменных соответствующие K_ν и Y_ν остаются без изменений.

На втором этапе каждая переменная v заменяется совокупностью из m_v величин $Z_\gamma = \{z_1^{(\gamma)},...,z_{m_v}^{(\gamma)}\}$, причем для каждой из них определяется такая функция измерений $\beta_{z_p}^{(v)}$, что $\beta_{z_p}^{(v)}(l)=1$ при $y_v(l)\in K_p^{(v)}$ и $\beta_{z_p}^{(v)}(l)=0$, в противном случае $(v\in V;\ p=1,...,m_v)$. В результате в качестве множества возможных значений каждой переменной $z_p^{(v)}$ используется множество $\beta_{z_p}^{(v)}=\beta=\{0,1\}$. Если теперь пронумеровать наблюдения $l(i=1,...,n)$, то каждой переменной $z_p^{(v)}$ соответствует столбец измерений $b_{rp}=\left(\beta_{z_p}^{(v)}(l_i),\ i=1,...,n\right)$, а переменной v- матрица $B_v=\left(b_{r1},...,b_{rp},...,b_{rm_v}\right)=\left[b_{ip}\right]_{i=1,...,n}^{p=1,...,m_v}$. Таким образом, модель измерений приводится к унифицированному виду УМИ:

$$УМИ=\left\langle E, Z=\bigcup_{v\in V} Z_v, \{\beta_z=\beta, \beta_z\}_{z\in Z}\right\rangle.$$

Если теперь пронумеровать переменные $z_p^{(v)}\left(j=1,...,m;\ m=\sum_{v\in V} m_v\right)$, то УМИ можно сопоставить матрицу наблюдений типа «объект-признак» вида $B=\left[b_{ij}=\beta_j(l_i)\right]_{i=1,...,n}^{j=1,...,m}$, а после перенумерации переменных $v(l=1,...,t)$ представить B в блочном виде $[B_1...B_l...B_t]$.

В связи с рассмотрением УМИ, отличительной чертой АС является свойство так называемой распределительной эквивалентности χ^2 – метрики. Его можно сформулировать следующим образом: если два столбца матрицы M с одинаковыми "профилями", $\bigwedge_{i\in I} m_{ij_1}=m_{ij_2}$, заменить на один столбец с весом, равным сумме весов столбцов j_1 и j_2, то результат проведения АС останется неизменным. Это свойство позволяет избежать субъективности при выборе тех или иных градаций для унифицированного представления качественных и количественных

переменных. Вместе с тем переход от МИ к УМИ, в результате огрубления результатов количественных измерений при замене y_γ на y'_γ, приводит к тому, что УМИ оказывается неэквивалентна МИ $(R_V \neq R_Z)$, то есть, происходит потеря информации за счет перехода к более грубой шкале. Это общий недостаток всех методов, направленных на обработку разнотипных переменных. Именно поэтому весьма актуальной на сегодняшний день является проблема разработки алгоритмов, позволяющих обрабатывать разнородную информацию с сохранением соответствующей природы каждой переменной.

Как в случае с ТС, так и в случае с таблицами наблюдений (ТН), использование алгоритма АС для обработки разнотипных данных приводит к потере информации. Применение АС к ТС или к УМИ приводит, прежде всего, к потере информации для количественных переменных в результате замены последних номинальными величинами.

Можно показать, что справедливо следующее утверждение: для унифицированных данных поиск факторных осей в анализе соответствий на основе определения направлений, наименее искажающих взаимные расстояния между точками, эквивалентен определению таких векторов, которые составляют минимальные углы с подпространствами, порождаемыми градациями заданных переменных. Это утверждение объясняет попытку представления количественных переменных в виде одновременных подпространств пространства R^n без изменения R_V.

3.3 Модифицированная модель измерений

Пусть задана модель измерений МИ:

$$МИ = \langle E, V\{K_v, Y_v\}_{v \in V} \rangle,$$

в которой определен набор разнотипных переменных $v = F \cup Q$, причем в F входят качественные переменные, а в Q – количественные переменные $(q \in Q)$. Для перечисления элементов наборов F и Q

обозначим мощности этих множеств $t_Q = \overline{\overline{Q}}$ и $t_F = \overline{\overline{F}}$ (где символ $\overline{\overline{Q}}$ обозначает мощность – число элементов – множества) $t = t_Q + t_F$. Представим теперь все количественные переменные в стандартизованном виде, то есть выполним над столбцами $\vec{K}_q$, соответствующими каждой переменной $q \in Q$, преобразование вида $\vec{K}'_q = \frac{\vec{K}_q - \vec{m}_q}{B_q}$, где $\vec{m}_q$ - n-мерный вектор, все компоненты которого равны m_q, а m_q и $B_q = \sqrt{\frac{1}{n}\left(\vec{K} - \vec{m}_q\right)^T\left(\vec{K}_q - \vec{m}_q\right)}$ - математическое ожидание и среднеквадратичное отклонение переменной q. Другими словами, $\vec{K}'_q = \frac{1}{B_q} P\vec{K}_q$, где $P = \{P_{ij}$ - идемпотентная матрица размерности $n \times n$. В результате, в данной МИ соответствующим образом изменяется Y_q и K_q. Представим теперь данные по всем неколичественным переменным в унифицированном виде, что, как было показано выше, приводит к образованию множеств Z, $B_z = \{0,1\}$ и $\beta_z(l)$ $(z \in Z)$, причем $\overline{\overline{Z}} = m_Z = \sum_{l'=1}^{t_F} m_{l'}$.

Введем далее вместо каждой количественной переменной $q_{l''}$ две переменные $x_1^{(l'')}$ и $x_2^{(l'')}$ $\left(l'' = 1,...,t_Q\right)$. Сопоставим им столбцы измерений, определяющие новые функции измерений $\gamma_{x_1}^{(l'')}$ $\left(\gamma_{x_2}^{(l'')}\right)$ и соответствующие области изменения $H_{x_p}^{(l'')}(p = 1,2)$, следующим образом:

$$\vec{h}_{x_1}^{(l'')} = a\vec{K}_{q_1} + \vec{b}_1 \quad \text{и} \quad \vec{h}_{x_2}^{(l'')} = -a\vec{K}_{q_1} + \vec{b}_2,$$

где $\vec{b}_1$ и $\vec{b}_2$ - n-мерные векторы из констант $b_1 = b$ и $b_2 = (1-b)$, соответственно, причем $0 < b < 1$, а $a = \sqrt{b(1-b)} = \sqrt{b_1 b_2}$. Обозначим множество всех переменных типа $x_1^{(l'')}$ через X_1, а типа $x_2^{(l'')}$ - через X_2. Пусть далее $X = X_1 \cup X_2$, причем $\overline{\overline{X}} = m_X = 2t_Q$.

Модель измерений с таким представлением количественных переменных и с унифицированным представлением качественных переменных называется модифицированной моделью измерений (ММИ):

$$ММИ = \langle E, Y = Z \cup X_1 \cup X_2, \{\beta_z = \{0,1\}, \beta_z\}_{z \in Z}, \{H_x, \gamma_x\}_{x \in X} \rangle,$$

причем $\overline{\overline{Y}} = m = m_Z + m_X$. Соответствующая схема преобразования «размножения» переменных при переходе от МИ к ММИ показана на рис.2.

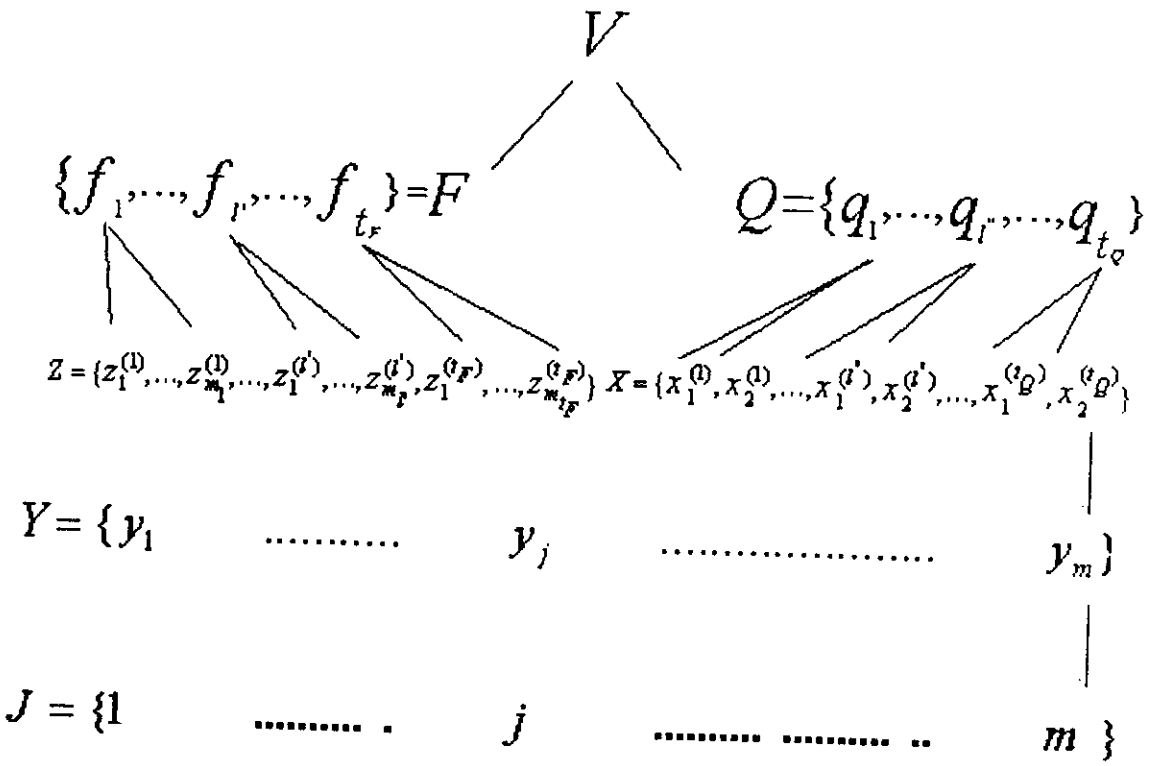

Рис. 2. Схема преобразования переменных в ММИ

В ММИ каждому $j \in J$ соответствует область возможных значений m. Тогда для ММИ совокупностью функций β_z и γ_x вместе с набором E (или множеством индексов $j \in J$) определяется матрица $M = [B, H]$. Здесь B - ранее определенный блок матрицы, соответствующий качественным переменным, H - матрица составленная из всех пар вектор-столбцов $\vec{h}_1^{(l'')} = \vec{h}_{x_1}^{(l'')}$ и $\vec{h}_2^{(l'')} = \vec{h}_{x_2}^{(l'')}\left(l'' = 1,...,t_Q\right)$, их i-е компоненты равны $h_{1i}^{(l'')} = \gamma_{x_1}^{(l'')}(i)$ и $h_{2i}^{(l'')} = \gamma_{x_2}^{(l'')}(i)$. Таким образом, матрицу $M = [B, H]$ можно представить в виде

$$M = \left[\vec{b}_1 ... \vec{b}_{m_z} \vec{h}_1^{(1)} \vec{h}_2^{(1)} ... \vec{h}_1^{(t_Q)} \vec{h}_2^{(t_Q)}\right] .$$

При анализе соответствий ММИ заменяется на анализируемую АММИ, тогда поиск факторных осей сводится к минимизации выражения

$$W = \frac{1}{t}\left(\sum_{l'=1}^{t_F} \cos^2 \varphi_{l'} + \sum_{l''=1}^{t_Q}\left(\cos^2 \xi_{l''}^{(1)} + \cos^2 \xi_{l''}^{(2)}\right)\right),$$

где $\varphi_{l'}$ - угол между факторной осью $\vec{u}$ и подпространством градаций переменной $f_{l'}$, $\xi_{l''}^{(p)}$ - угол между $\vec{u}$ и переменной $x_p^{(l'')}$ $(p = 1,2)$, получающейся в результате «размножения» переменной $q_{l''}$.

За счет выбора различных значений коэффициента b можно получать различные преобразования столбцов количественных переменных. Например, в качестве пары коэффициентов можно использовать:

$$\left(a = \frac{\sqrt{3}}{4},\ b = \frac{1}{4}\right);\ \left(a = \frac{2}{5},\ b = \frac{1}{5}\right) \text{ и т.п.}$$

При этом для каждой пары преобразований $a\vec{k}' + \vec{b}$ и $-a\vec{k}' + \left(\vec{1} - \vec{b}\right)$ условиям утверждения будет удовлетворять и пара $a\vec{k}' + \left(\vec{1} - \vec{b}\right)$ и

$-a\vec{k}+\vec{b}$. Это же преобразование можно сделать зависящим не от одного параметра b (так как $a=\sqrt{b(1-b)}$), а от двух коэффициентов a_1 и b_1:

$$\vec{h}_1=\frac{\sqrt{a_1(b_1-a_1)}\ \vec{k}+\vec{a_1}}{b_1} \quad \text{и} \quad \vec{h}_2=\frac{-\sqrt{a_1(b_1-a_1)}\ \vec{k}+(\vec{b_1}-\vec{a_1})}{b_1}, \quad \text{где}$$

$$0<a_1/b_1<1.$$

В случае любой переменной q для порождаемых ею переменных имеет место равенство $R_{x_1}=R_{x_2}=R_q$, то есть преобразования ТН, связанные с переходом к ММИ, приводят к эквивалентной МИ, так как $R_Y=R_{Y'}$. Это преобразование заключается в удвоении количественных и в переходе к унифицированному представлению качественных переменных. Заметим, что в результате преобразования исходный столбец $\vec{k}$ трансформируется некоторым образом при проведении АС. Приведенное утверждение позволяет добавить к числу характерных преобразований исходной ТН и, следовательно, МИ и ИМ, еще одно, связанное с переходом к ММИ, что позволяет обрабатывать разнотипные переменные без потерь с точки зрения сохранения отношения эквивалентности. Предложенный метод автор применял при обработке данных таблиц сопряжённости и наблюдений [3].

4. СИНТЕЗ СОСТАВА ИНФОРМАЦИИ МЕТОДОМ ГРУППИРОВАНИЯ ПЕРЕМЕННЫХ

4.1 Метод динамического группирования

На основе исходных данных (в виде таблицы измерений, частот или качественных оценок) требуется сгруппировать переменные в однородные классы (с точки зрения их сходства, сходства их профилей или совпадения качественных характеристик), причем отличные переменные должны при этом разделяться. Эти группы могут соответствовать либо разбиению множества переменных, либо вложенной иерархии.

Разбиением множества E называется такое семейство $\rho(E)$ классов, что

$$\begin{matrix}\forall A \in \rho(E) \\ \forall B \in \rho(E)\end{matrix} \Rightarrow A \cap B \neq \varnothing \qquad \text{или} \qquad A \cup B = A = B,$$

а в объединении получается все исходное множество $\bigcup_i A_i = E$.

Иерархией подмножеств из E называют такое семейство $\varphi(E)$, что:

1. любое одноэлементное множество принадлежит (E);

2. $E \in \varphi(E)$;

3. и $\forall A \in \varphi(E)$ и $\forall B \in \varphi(E) \Rightarrow (A \bigcap B \neq \varnothing)$ *или* $\Big((A \supset B) \ \vee \ (B \supset A)\Big)$.

Иерархия представляет собой множество вложенных разбиений.

Метод динамического группирования построен на использовании понятия «эталонных» точек и является итерационным. Кратко суть этого метода заключается в следующем:

- на первом шаге задаем начальные центры группирования;
- на основании некоторого правила выделяем группы, формируемые вокруг каждого центра группирования;
- для каждой такой отдельной группы по некоторому правилу определяем новые центры группирования. При этом для определения новых центров группирования не обязательно использование правила, совпадающего с правилом выделения групп;
- проверяется устойчивость итерационного процесса, т.е. проверяется возможность достижения такого шага, когда разбиение на следующем шаге совпадает с разбиением на предыдущем шаге. Если режим процесса группирования не является устойчивым, то следует переход ко второму пункту. Иначе процесс группирования считается законченным.

При реализации такого метода группирования удается избавиться от ряда недостатков, присущих большинству из известных методов группирования, а именно:

- возможность получения единичных групп (в данном случае под единичной группой понимается такая группа, которая содержит лишь центр группирования);
- отсутствие цепного эффекта, то есть отсутствие тенденций включения в одну группу двух элементов разбиваемого множества, связанных цепочкой «близких» элементов;
- возможность получения разбиения исходного множества как на непересекающиеся классы, так и на группы, пересечение которых не является пустым;
- отсутствие необходимости задания произвольных порогов для окончания процесса группирования.

Рассмотрим реализацию метода группирования, общая схема которого была приведена выше. Введем следующие определения.

E – множество объектов, которое необходимо разбить на группы. По определению оно должно быть конечным.

S_K – множество разбиений E на n частей $(n \le K)$.

K – число групп. В данном случае под разбиением множества E понимается совокупность подмножеств E_i, объединение которых $\bigcup_i^n E_i$ есть все множество E, а попарное пересечение не обязательно должно быть пустым.

L_K - множество последовательностей из K элементов, принадлежащих L, причем ни в одной такой последовательности нет хотя бы двух одинаковых элементов, т. е.

$$L_K = \left\{ L = (a_1, ..., a_n) \mid a_i \in E \wedge a_i \neq a_j \forall i \neq j (i, j = 1 \div K) \right\}$$

Другими словами, L есть множество множеств эталонных элементов для K групп.

$V_K = L_K \otimes S_K$ – декартово произведение L_K и S_K, т.е. элемент $v \in V_K$ представляет собой упорядоченную пару элементов $L \in L_K$ и $S \in S_K$. Такая пара определяется множеством из K эталонных элементов и некоторым разбиением множества E, причем, в общем случае, эталонные элементы и разбиение могут и не соответствовать друг другу. В данном случае под соответствием понимается формирование рассматриваемого разбиения из данного множества эталонных элементов.

Теперь можно формально определить критерий разбиения множества E на группы:

$$W : V_K \to R^+ .$$

Здесь каждому элементу из V_K (множеству эталонных элементов из E и разбиению S) ставится в соответствие по определенному правилу некоторое неотрицательное число, т. е. оценка "качества" разбиения есть число. После формального определения критерия можно ввести понятия локального и глобального оптимума (минимума). Локальным оптимумом на подмножестве $C \in V_K$ будем называть такой элемент $v^* \in C$, для которого выполняется условие

$$W(v^*) = \min_{v \in c} W(v).$$

Если подмножество C совпадает с множеством V_K, то есть $C = V_K$, то в этом случае v^* будет являться глобальным оптимумом.

Теперь перейдем к формализации правил формирования групп из эталонных элементов и получения новых эталонных элементов из групп. Для этого введем следующее отображение

$$R : E \otimes N_K \otimes S_K \to R^+ ,$$

которое будем использовать для определения меры близости между элементами из E и подмножеством, входящим в S_K. В этом отображении N_K - множество натуральных чисел от 1 до K.

Определим правило получения групп из эталонных элементов следующим образом:

$$f: L \to S_K, \text{ т. е. } f(L) = S, \text{ где } l \in L_K, \quad S \in S_K.$$

При этом отображении разбиение S_K представляет собой множество групп (подмножеств) элементов E. Отображение f каждому элементу L из L_K, т. е. множеству из K эталонных элементов, ставит в соответствие элемент S из S_K, т. е. некоторое разбиение множества E, причем это отображение должно быть однозначным. Необходимо отметить, что оно не обязательно является отображением на все множество S_K.

Правило получения новых эталонных элементов из групп определим следующим образом:

$$g: S_K \to L_K, \text{ т. е. } g(S) = L, \text{ где } S \in S_K, \; L \in L_K.$$

Отображение g каждому элементу S из S_K ставит в соответствие элемент L из L_K, причем это отображение также должно быть однозначным. Отображение g позволяет получить множество «эталонных» точек из некоторого разбиения множества E на группы.

Метод итеративной группировки заключается в попеременном применении функций f и g, начиная с некоторого $L^{(0)} \in L_K$, $L^{(0)}$ выбирается либо случайным образом, либо по некоторому правилу.

Итеративная процедура группирования работает следующим образом. Выбирается некоторым образом $L^{(0)}$ - начальное (исходное) множество эталонных элементов. Используя отображение f, получаем из $L^{(0)}$ разбиение $S^{(0)}$. При этом $S^{(0)} = f(L^{(0)})$. Затем из разбиения $S^{(0)}$ получаем $L^{(1)}$, т. е. $L^{(1)} = g(S^{(0)})$ и т. д. Таким образом, последовательное применение отображений f и g позволяет получить последовательность пар: $\left(L^{(0)}, S^{(0)}\right)$, $\left(\left(L^{(1)}, S^{(1)}\right), \ldots, \left(L^{(n)}, S^{(n)}\right), \ldots\right)$. В этой последовательности

$S^{(i)} = f\left(L^{(i)}\right)$; $L^{(i+1)} = g\left(S^{(i)}\right)$. Получаемая последовательность может иметь один из двух возможных видов:

- начиная с некоторого номера n_0, все члены последовательности будут одинаковые, то есть существует такое n_0, при котором для любого $m > n_0$ имеем $L^{(m)} = L^{(n_0)}$, $S^{(m)} = S^{(n_0)}$. Это означает, что, начиная с члена последовательности $(L^{(n_0)}, S^{(n_0)})$, итеративная процедура группирования не будет давать нового разбиения множества E на группы. Следовательно, на n_0- шаге работа итеративной процедуры группирования прекращается, то есть в этом случае мы имеем сходящийся итеративный процесс;

-не существует такого n_0, начиная с которого все члены последовательности будут одинаковые, то есть для любого n_O существует такое $m > n_0$, при котором $L^{(m)} \neq L^{(n_0)}$ и $S^{(m)} \neq S^{(n_0)}$. Это означает, что наш итеративный процесс не сходится.

4.2 Итеративный метод группирования графов

Введение отображений f и g определяет структуру множества V_K. Можно построить граф, в котором вершинами будут элементы из V_K, то есть пары (L,S). Две вершины (L,S) и (L_1,S_1) соединяются дугой, если выполняются условия:

$$L_1 = g(S);\ S_1 = f(L_1).$$

Видно, что определенный таким образом граф является ориентированным.

На языке графов итеративный метод группирования можно описать следующим образом. Выбираем начальную вершину $(L^{(0)}, S^{(0)})$. Затем переходим к другой вершине $(L^{(1)}, S^{(1)})$, которая связана дугой с

начальной и т. д. Таким образом, переходим от одной вершины $(L^{(i)}, S^{(i)})$ графа к другой $(L^{(i+1)}, S^{(i+1)})$, с которой предыдущая вершина связана дугой. В силу однозначности отображений f и g из каждой вершины может выходить не более одной дуги. Может оказаться, что из некоторой вершины нет дуги к другой вершине графа. Это означает, что данная вершина имеет петлю, а сам итеративный процесс является сходящимся. Возможен и второй вариант, когда мы никогда не попадаем на вершину с петлей. Это означает, что мы движемся в графе по циклу (причем простому, в силу однозначности f и g). Но структуру графа определяет выбор отображений f и g, поэтому необходимо выбирать их так, чтобы они приводили к сходящимся итеративным процессам. Прежде чем перейти к рассмотрению условий, при которых итеративный процесс будет сходящимся, покажем, что достижение локального и глобального оптимума имеет место лишь для первого случая. Однако практика подтверждает его высокую эффективность и при использовании во втором случае.

Предположим, что отображение f и g выбраны таким образом, что итеративный процесс является сходящимся, а переход от одной вершины к другой улучшает критерий разбиения. В этом случае граф представляет собой или одно дерево, или совокупность несвязанных деревьев, а сама итеративная процедура заключается в следующем. Начиная с некоторой начальной вершины графа, последовательно, переходя от вершины к вершине, приближаемся к корню дерева, к которому относится начальная вершина. Если граф представляет собой одно дерево, то такая процедура приведет к вершине, являющейся корнем данного дерева, т. е. достигаем глобального оптимума.

В том случае, когда граф представляет собой совокупность несвязанных деревьев, достижение корня некоторого дерева не означает того, что достигнут глобальный оптимум. Тогда можно говорить лишь о достижении локального оптимума. Для того, чтобы найти глобальный оптимум, надо применять итеративную процедуру для нахождения

локального оптимума на одном дереве; затем в качестве начальной вершины выбрать вершину другого дерева и опять применить итеративную процедуру. Проделав это с каждым деревом и получив значения локального оптимума для каждого дерева графа, можно получить глобальный оптимум.

Рассмотрим требования к выбору отображений f и g, которые позволяют получать сходящийся итеративный процесс. Критерий W определим следующим образом:

$$W : V_K \to R^+ : \; W(v) = W\left[(L,S)\right] = \sum_{i=1}^{K} R(a_i, i, S).$$

Введем последовательности $\{v_n\}$ и $\{w_n\}$ следующим образом. Пусть h будет отображением $V_K \to V_K$ таким, что

$$V_K \; э \; v = (L,S) \Rightarrow h(v) = (g(S), f[g(S)]) \in V_K .$$

Следовательно, последовательность $\{v_n\}$ определяется выбором $v_O = (L^{(0)}, \; S^{(0)})$ и заданием правила перехода от v_n к v_{n+1}, то есть в нашем случае $v_{n+1} = h(v_n)$. Последовательность $\{w_n\}$ определяется из последовательности $\{v_n\}$ следующим образом: $w_n = W(v_n)$. Каждый член последовательности $\{w_n\}$ является неотрицательным числом. Итерационная процедура будет сходящейся, если последовательность $\{w_n\}$ сходится. Отсюда следует, что необходимо сделать так, чтобы последовательность $\{w_n\}$ была убывающей, т.е., чтобы для любого n выполнялось неравенство

$$w_{n+1} \le w_n .$$

Рассмотрим теперь конкретное задание отображений f и g для решения поставленной задачи, т.е. для анализа информационной структуры. В качестве меры «близости» пары переменных предлагается

использовать парную меру взаимосвязи - парный коэффициент корреляции [4]. В этом случае отображение принимает следующий вид:

$$S_i\{x \in E \mid D(x,a_i) \geq M\},$$

где $(S_1, S_2, .., S_K)$ - разбиение исходного множества E на группы $S_i(1,2,...,K)$; $D(x,a_i)$ - мера «близости» (подобия) пары переменных x и a_i; a_i - переменная, являющаяся центром i-ой группы; M - некоторое фиксированное число. Заданное таким образом отображение f позволяет получать группы, пересечение которых не обязательно должно быть пустым.

Для выявления групп взаимосвязанных переменных целесообразно отображение g задавать с использованием некоторого коэффициента множественной взаимосвязи $R(a_i,i,S)$. В качестве такого коэффициента можно использовать множественный коэффициент корреляции. Его преимущество состоит в том, что можно получить его распределение (предельное) и оценки доверительного интервала, т.е. вероятностную меру достоверности полученного результата. Основным недостатком использования множественного коэффициента корреляции является корректность его использования лишь в случае линейных зависимостей между переменными. В том случае, когда зависимости между переменными являются нелинейными, предлагается использовать непараметрический коэффициент корреляции.

Если под $R(a_i,i,S)$ иметь в виду коэффициент множественной взаимосвязи, то отображение можно задать следующим образом:

$$L = \left\{ a_i \in E,\ i = 1,2,\ldots,K \mid R(a_i,i,S) = \min_{x \in S_i} R(x,i,S) \right\}$$

где a_i - новые «эталонные» точки, т.е. новые центры группирования; $R(x,i,S)$ - значение взаимосвязи переменной x со всеми другими переменными, входящими в i-ую группу разбиения S.

Исходными данными для данного алгоритма является матрица парных взаимосвязей $M = \{m_{ij}\}(i, j = 1,2,\ldots,N)$ и совокупность эталонных точек $(a_1^{(0)}, a_2^{(0)}, \ldots, a_K^{(0)})$. Элемент m_{ij} матрицы M представляет собой оценку меры i-ой и j-ой переменной. Причем, если известно, что i-ая и j-ая переменные находятся в функциональной связи, то данный элемент матрицы M равен 1. В противном случае присваивается значение, равное модулю оценки меры взаимосвязи, которая выбирается в зависимости от характера и вида каждой из этих переменных. За «эталонные» точки, т.е. за начальные центры группирования, выбираются критериальные переменные.

Схема работы алгоритма итеративного группирования взаимосвязанных переменных заключается в следующем:

а) берется начальный центр группирования и в группу, центром которой является эта переменная, относят все те переменные, значение меры взаимосвязи которых с центром группирования превышают некоторый фиксированный порог;

б) последовательно рассматриваются все переменные, включенные в данную группу, и подсчитывается значение коэффициента множественной взаимосвязи. При этом рассматриваемая переменная группы принимается за зависимую, а все остальные переменные группы считаются влияющими на эту выделенную переменную. Таким образом, подсчитывается столько коэффициентов множественной взаимосвязи, сколько переменных входят в группу;

в) за новый центр группирования выбирается такая переменная группы, для которой соответствующее значение коэффициента множественной взаимосвязи максимально;

г) новый центр группирования сравнивается с предыдущим центром формируемой группы. Если они совпадают, то формирование группы по данному начальному центру заканчивается. В противном случае переходим к пункту б);

д) проверяется, все ли начальные центры группирования рассмотрены. Если нет, то выбирается очередной начальный центр

группирования и переходят к пункту а). В противном случае процедура группирования считается законченной;

е) если в качестве коэффициента множественной взаимосвязи использовался не множественный коэффициент корреляции, то для каждой группы относительно ее центра группирования подсчитывается множественный коэффициент корреляции.

На каждом шаге итерационного процесса на печать выводятся следующие данные:

- начальный центр группирования для данной группы;
- список переменных, входящих в полученную группу на данном шаге итерации;
- значение коэффициента множественной взаимосвязи для нового центра группы;
- идентификатор переменной, выбранной в качестве нового центра группы на данном шаге итераций.

5. Заключение

Модифицированный метод анализа соответствий применяется при исследовании процессов, которые характеризуются данными, представлениями в различных шкалах измерения (качественными и количественными характеристиками), и позволяет совместно обрабатывать качественно-количественную информацию без потерь для описании структур, взаимосвязей. формирования сравнимых целей и оценок результатов. Метод группирования переменных свободен от известных недостатков и полезен при формировании целей и результатов, сокращении объемов информации. Предложенная методика позволяет не только повысить качество моделирования, но и более строго структурировать процесс, выделяя области переходных границ в жизненном цикле процесса высокой технологии.

Автор выражает искреннюю благодарность И.В.Пузынину и коллективу Научного центра прикладных исследований по разработке методов анализа данных физических экспериментов и сложных технологических процессов.

Литература

1. Негойце К. Применение теории систем к проблемам управления. Пер. с англ. – М Мир, 1981, 179с.
2. Растригин Л.А., Марков В.А. Кибернетические модели познания. Вопросы методологии. – Рига.: Зинатне, 1976, 264с.
3. Самойлов В. Н. ОИЯИ, Р10-99-104, Дубна, 1999.
4. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. –М.: Наука, 1982, 472с.
5. Глушков В.М. Макроэкономические модели и принципы построения ОГАС. –М.: Статистика. 1975, 312с.
6. Налимов В.В., Чернова Н.А. Статистические методы планирования экстремальных экспериментов. –М.: Наука, 1965, 284с.
7. Нейлор Т. и др. Машинные имитационные эксперименты с моделями экономических систем. –М.: Мир, 1975, с. 500.
8. Кини Р. Л., Райфа Х. Принятие решений. –М.: Советское радио, 1972, 550с.
9. Крамер Г. Математические методы статистики. –М.: Мир, 1975.
10. Бодин Н.А. Оценка параметров распределения по группированным выборкам // Тр. МИАН, Наука, 1976, т. X I, с. 110-154.
11. Аркадьев А.Г., Браверман Э.М. Обучение машины классификации объектов.–М.: Наука, 1971, 192с.
12. Липский В. Комбинаторика для программистов. –М. : Мир, 1988, 213с.
13. Лоэв М. Теория вероятностей. –М.; ИЛЛ, 1962, с.719.
14. Нечеткие множества в моделях управления и искуственного интелекта. Под ред. Д.А. Поспелова. –М.; Наука, 1986, 311с.
15. Вилкас Э.И., Майминас Е.Э. Решения: теория, информация, моделирвание. –М.: Радио и связь, 1981, 285с.

Рукопись поступила в издательский отдел
14 апреля 1999 года.

Самойлов В.Н.
Р10-99-105

Методы анализа данных физических экспериментов и сложных процессов

Проведен анализ составляющих жизненного цикла эволюции технологического процесса. Предложено решение проблемы нормируемости — установление границ действия нормативов, соответствующих характеру протекания процесса. Разработан модифицированный метод анализа соответствий для исследования сложных многофакторных процессов, которые характеризуются данными, представленными в различных шкалах измерения, качественными и количественными характеристиками. Предложен итерационный метод динамического группирования переменных для формирования целей и результатов, позволяющий сократить объем информации и не только повысить качество моделирования, но и структурировать процесс.

Работа выполнена в Лаборатории вычислительной техники и автоматизации и в Научном центре прикладных исследований ОИЯИ.

Препринт Объединенного института ядерных исследований. Дубна, 1999

Samoilov V.N.
P10-99-105

Data Analysis Methods of Physical Experiments and Complex Processes

Analysis of components of life cycle of evolution of the technological process is carried out. The solving of the normalization problem — establishing of boundary of normative actions corresponding to a character of the processes run is proposed. A modified method of the corresponding analysis for study of the complex multi-fragment processes, characterized by data base presented in difference measure scale, quality and quantity characteristics. An iteration method of the dynamical grouping of the variables for formation aims and results is suggested, reducing the information volume and to increase not only the modeling quality and providing process structuring.

The investigation has been performed at the Laboratory of Computing Techniques and Automation and at the Scientific Center of Nuclear Research, JINR.

Preprint of the Joint Institute for Nuclear Research. Dubna, 1999

Редактор М.И.Зарубина. Макет Н.А.Киселевой

Подписано в печать 29.04.99.
Формат 60 × 90/16. Офсетная печать. Уч.-изд. листов 3,15
Тираж 345. Заказ 51352. Цена 3 р. 78 к.

Издательский отдел Объединенного института ядерных исследований
Дубна Московской области