Известия Гомельского государственного университета
имени Ф.Скорины, №6(15), 2002

*УДК 519.9:618.32*

# Выявление причинно-следственных связей в компьютерном моделировании социальных и природных систем

П.Н.Стрибук, А.Н.Осипенко, Н.Б.Осипенко

## Введение

Актуальность этой темы вызвана возросшим в последнее десятилетие интересом к фактору активного или самодеятельного проявления системы, в частности, — к субъективному фактору в социальной сфере. Как показывает практика, отсутствие соответствующего аппарата экспертной интерпретации и моделирования процессов порождения субъектом своих действий не позволяет выходить ни на эффективный прогноз его поведения, ни на разработку способов коррекции его жизнедеятельности. Одной из первых во всем многообразии возникающих здесь задач является задача разработки аналитико-статистических моделей поведения социальных и природных систем. Специфика конкретной системы состоит в том, что для нее, как правило, трудно подобрать готовый алгоритм корректной обработки из-за отсутствия необходимого числа доступных информативных признаков. В свою очередь, чтобы обеспечить автоматизацию синтеза эффективной модели, необходимо подготовить информационную среду для снабжения процедур конструирования модели разнообразными априорными знаниями со стороны эксперта. То есть, для автоматизации исследования активной системы первостепенное значение приобретает этап концептуального моделирования целевого функционирования путем выделения дерева объясняющих факторов и построения сети причинно-следственных связей с помощью прежде всего корреляционно-регрессионного анализа. Для разрешения этой проблемы авторами разработано программно-технологическое обеспечение МОНАДА [2].

## Общая схема исследования

Схема исследования причинно-следственных связей в социальных и природных системах в среде пакета МОНАДА реализуется экспертом поэтапно.

1.Одномерный анализ признаков с целью корректной нормировки данных на основе оценки распределения.

2.Формирование фактор-признаков.

3.Формирование репрезентативной выборки детально известных (реперных) объектов.

4.С помощью модуля классификации пакета МОНАДА объекты, представленные в виде точек многомерного пространства фактор-признаков, разбиваются на классы таким образом, чтобы в один класс попали по возможности наиболее похожие по этим признакам, а в разные классы — значимо различающиеся между собой объекты. При этом модуль классификации реализован в режиме интерактивного взаимодействия, предоставляя эксперту возможность самостоятельно удалять из класса отдельные объекты (в первую очередь речь идет о реперных объектах) в случае обнаружения в одном классе значимо различных с его точки зрения объектов. Далее осуществляется переклассификация с участием эксперта до тех пор, пока не будет обеспечена однородность классов как по фактор-признакам, так и по мнению эксперта о сходстве и различии реперных объектов по дополнительным качественным аспектам. Отметим, что в результате классификации может образоваться значительное число однообъектных классов (уникальные объекты). Для биологических и социальных систем это явление не случайно. Во многих случаях систематизации таких объектов упорядоченный по убыва-

нию ряд численностей классов подчиняется так называемому
многообъектных классов и столько же или больше однообъект

5.По результатам классификации и визуализации распредел
пространстве (путем проекции его на плоскость) эксперт по ка
жайший объект-ориентир и ближайший объект-антиориентир
ального перспективного направления развития и предотвращ
управлении объектом-антиориентиром. Желательно, чтобы в ч
тиров попадали более известные реперные объекты (в ряде с
тальное обследование новых объектов).

6.По выборке реперных объектов с использованием фактор
признаков качественного характера строятся несколько иерарх
линейной регрессии для описания сети причинно-следственны
нению эффективности управления как по отдельным элемента
в целом по объекту. Основное назначение этих уравнений сост
оценить вклад того или иного управляемого параметра в целе
фективности.

Остановимся кратко на основных компонентах пакета МО

**Одномерный анализ и классификация**

С целью обеспечения адекватности одномерного анали
природной системе в предлагаемом методе реализованы возм
«ядерной» аппроксимации функции распределения признаков
гипотез), взвешивания конкретных значений (что, например, по
вать данные, полученные из различных источников), обработки
тых выборок, расщепления смеси, подбора преобразования,
подвыборки для последующего регрессионного анализа, визуал
ретического распределений.

Одномерный анализ проходит по следующим этапам:

1.Нахождение эмпирических статистических характеристи
ваться при дальнейшей работе с данными в том случае, если
сующееся теоретическое распределение и рассчитать по нему те

2.«Ядерная» аппроксимация функции распределения. Выбо
ции вызван тем, что на практике очень часто приходится работ
которым «ядерная» аппроксимация дает лучшую оценку параме
ды (гистограмма и полигон частот) [1]. К тому же на «сверхмал
ют. Особенно это актуально при построении регрессионных мо
формация имеется лишь по небольшому числу объектов.

3.Проверка гипотез о нормальности или логнормальности р
нормальности или логнормальности упрощает дальнейшую раб
дятся теоретические параметры, нет необходимости подбирать
ния к нормальному виду, при выделении подвыборки признак
ным распределением, не приводит к сильному разрежению выб
нию информативных объектов).

4.Проверка на бимодальность и расщепление смеси. Бимода
кает в случае, когда не была обеспечена однородность выборки
объектов из разных классов (например, измерения биохимичес
ров, в котором наряду со здоровыми коровами имеются больные

5.Подбор преобразования выборки. Практически все методы
раются на предположение о нормальности распределения входя
личин. Когда мы имеем несимметричную эмпирическую фун
смысл преобразовать статистические данные так, чтобы получ

нию ряд численностей классов подчиняется так называемому распределению Ципфа (10-15 многообъектных классов и столько же или больше однообъектных) [5].

5.По результатам классификации и визуализации распределения объектов в многомерном пространстве (путем проекции его на плоскость) эксперт по каждому объекту намечает ближайший объект-ориентир и ближайший объект-антиориентир с целью выбора наиболее рационального перспективного направления развития и предотвращения ошибок, допущенных в управлении объектом-антиориентиром. Желательно, чтобы в число ориентиров и антиориентиров попадали более известные реперные объекты (в ряде случаев придется провести детальное обследование новых объектов).

6.По выборке реперных объектов с использованием фактор-признаков и дополнительных признаков качественного характера строятся несколько иерархических рядов уравнений нелинейной регрессии для описания сети причинно-следственных связей, приводящих к изменению эффективности управления как по отдельным элементам (например, отраслям), так и в целом по объекту. Основное назначение этих уравнений состоит в том, что они позволяют оценить вклад того или иного управляемого параметра в целевой показатель изменения эффективности.

Остановимся кратко на основных компонентах пакета МОНАДА.

## Одномерный анализ и классификация данных

С целью обеспечения адекватности одномерного анализа данных о социальной или природной системе в предлагаемом методе реализованы возможности: учета объема (при «ядерной» аппроксимации функции распределения признаковых значений и при проверке гипотез), взвешивания конкретных значений (что, например, позволяет корректно обрабатывать данные, полученные из различных источников), обработки данных с пропусками и сжатых выборок, расщепления смеси, подбора преобразования, выделения репрезентативной подвыборки для последующего регрессионного анализа, визуализации эмпирического и теоретического распределений.

Одномерный анализ проходит по следующим этапам:

1.Нахождение эмпирических статистических характеристик, которые будут использоваться при дальнейшей работе с данными в том случае, если не удастся подобрать согласующееся теоретическое распределение и рассчитать по нему теоретические параметры.

2.«Ядерная» аппроксимация функции распределения. Выбор данного вида аппроксимации вызван тем, что на практике очень часто приходится работать с малыми выборками, по которым «ядерная» аппроксимация дает лучшую оценку параметров, чем стандартные методы (гистограмма и полигон частот) [1]. К тому же на «сверхмалых» выборках они не работают. Особенно это актуально при построении регрессионных моделей, так как детальная информация имеется лишь по небольшому числу объектов.

3.Проверка гипотез о нормальности или логнормальности распределения. Обнаружение нормальности или логнормальности упрощает дальнейшую работу с выборкой (легко находятся теоретические параметры, нет необходимости подбирать преобразования для приведения к нормальному виду, при выделении подвыборки признак, согласующийся с нормальным распределением, не приводит к сильному разрежению выборки и, как следствие, удалению информативных объектов).

4.Проверка на бимодальность и расщепление смеси. Бимодальность, как правило, возникает в случае, когда не была обеспечена однородность выборки и информация бралась для объектов из разных классов (например, измерения биохимического состава крови стада коров, в котором наряду со здоровыми коровами имеются больные).

5.Подбор преобразования выборки. Практически все методы многомерного анализа опираются на предположение о нормальности распределения входящих в модель случайных величин. Когда мы имеем несимметричную эмпирическую функцию распределения, имеет смысл преобразовать статистические данные так, чтобы получить вид функции плотности

наиболее приближенный к симметричному, с тем, чтобы после проверки гипотезы о согласии подобрать теоретические параметры распределения и использовать данный признак в дальнейшем анализе.

6. Оценка теоретических параметров распределения и формирование отчета.

7. Нормирование данных на основе полученной информации о выборке.

8. Выделение подвыборки. На практике нередко возникает задача планирования пассивного регрессионного эксперимента путем выделения по данным массового обследования из исходного множества объектов такого его подмножества, которое удовлетворяет требованиям корректности регрессионных построений (обеспечение по каждому признаку, по крайней мере, одномодальности и симметричности распределения).

При исследовании социальных и природных объектов традиционные алгоритмы машинной классификации разбивают множество объектов на небольшое количество классов, структура которых, как правило, получается неоднородной. Для решения этой проблемы в модуле классификации разработана модификация алгоритма гибкой классификации [2].

## Построение регрессионных уравнений

Для разрешения проблемы реалистичности и адекватности модели социальной или природной системы предлагается вместо одного-двух (чаще всего линейных) уравнений связи с большим числом переменных перейти к построению иерархической сети уравнений взаимосвязи факторов. В работе при построении регрессионных моделей предлагается использовать проведенную пользователем интерпретацию концептуальной схемы формирования целевого свойства социальных и природных объектов [2]. Все признаки в результате такой интерпретации разбиты на группы в соответствии с тем, какой компонент целевого функционирования они описывают. В данной схеме присутствуют следующие компоненты целевого функционирования:

- *Средства* – фрагменты среды, пассивной и активной подсистем функционирования, ресурсы которых используются непосредственно в процессе преобразования предмета функционирования;
- *Фон* – фрагменты среды, активной и пассивной подсистем функционирования, косвенным образом определяющие преобразование предмета функционирования;
- *Носитель целевого функционирования* – организм или цельная природная система, на базе которого происходит развитие «плода»;
- *Носитель узла связывания* – рабочее место для осуществления преобразований целевого функционирования;
- *Субъект* – активная составляющая целевого функционирования;
- *Инструмент* – управляемый сознанием фрагмент активной подсистемы функционирования, посредством которого осуществляется перевод части ресурсов (структурных, энергетических, информационных) от средств и входного предмета к выходному предмету функционирования.

Регрессионная модель, основанная на интерпретации данной концептуальной схемы, схематично может быть представлена следующим образом:

$$ЦС = Средства^{(Субъект\ и\ Инструмент)\cdot Фон\cdot НосительУС} \cdot НосительЦФ^{\gamma}, \qquad (1)$$

при подробном расписывании получим уравнение вида (3).

Другая проблема традиционного подхода заключается в излишней математической «подгонке» коэффициентов модели. Алгоритмы расчета ориентированы на достижение наилучшего результата без учета содержательной природы входящих в модель признаков. Поэтому в конце их работы очень часто получаются абсурдные результаты, обусловленные корреляцией между признаками, неинформативностью некоторых признаков и т. п. На практике приходится выделять условно оптимальный вариант из нескольких удовлетворительно

интерпретируемых и подходящих для дальнейшего анализа (пусть и не оптимальный с математической точки зрения). Для решения этой проблемы предлагается изменить математические процедуры построения регрессионных уравнений путем наложения ограничений на изменение коэффициентов.

Рассмотрим сначала модифицированный алгоритм для построения мультипликативного уравнения регрессии вида

$$y = F(x) = \alpha_0 \cdot x_1^{\alpha_1} \cdot x_2^{\alpha_2} \cdot \ldots \cdot x_n^{\alpha_n}. \quad (2)$$

1.Находятся $\alpha^0_i$ из уравнения регрессии $y = x_i^{\alpha_i^0}$, при этом остаются в рассмотрении лишь те $k$ признаков, для которых $\alpha^0_i$ достаточно велико и совпадает по знаку с соответствующим коэффициентом парной корреляции.

2.По значениям ($\alpha^0_i$) строятся ранги ($r^0_i$), $i=1 \div k$.

3.Задаются пороги $\alpha^H_0$ и $\alpha^B_0$ для $\alpha_0$.

4.Последовательно проходят $t=1 \div t_{max}$ итераций:

- находится шаг ($s_i$), $i=0 \div k$;
- строятся ранги ($r_i$), $i=1 \div k$ по значениям ($\alpha^t_i + s_i$);
- если значение $\alpha_0$ выйдет за пределы $\alpha^H_0$ и $\alpha^B_0$, то $s_0$ уменьшается вдвое;
- сравниваются ранги $r_i$ и $r^0_i$: если отклонение по ним превышает заданный порог (зависит от количества участвующих в построении регрессионного уравнения признаков), то $s_i$ уменьшается вдвое;
- при достижении некоторым $s_i$ заданного порога малости он обнуляется;
- начиная с заданной итерации (например, десятой), проверяем условие вида

$$Err^t \cdot \sum_{i=1}^{k} \left| \frac{\Delta\alpha_i^t}{\alpha_i^{t-1}} \right| \geq \varepsilon \cdot Err^{t-1} \cdot \sum_{i=1}^{k} \left| \frac{\Delta\alpha_i^{t-1}}{\alpha_i^{t-2}} \right|, \quad (3)$$

где $Err^t$ – ошибка модели на $t$-ой итерации, при его выполнении сохраняем состояние $t$-$1$;

- после выполнения условий сходимости или достижения максимального количества итераций выбираем лучший результат среди сохраненных состояний и результата последней итерации.

Описанный выше алгоритм построения уравнения простой мультипликативной регрессии можно применить и для построения уравнения регрессии более общего вида:

$$y = F(x) = \alpha_0 \cdot x_1^{\alpha_0^1 + \alpha_1^1 \cdot z_1^1 + \ldots + \alpha_{k_1}^1 \cdot z_{k_1}^1} \cdot x_2^{\alpha_0^2 + \alpha_1^2 \cdot z_1^2 + \ldots + \alpha_{k_2}^2 \cdot z_{k_2}^2} \cdot \ldots \cdot x_n^{\alpha_0^n + \alpha_1^n \cdot z_1^n + \ldots + \alpha_{k_n}^n \cdot z_{k_n}^n}. \quad (4)$$

Для этого преобразуем (4) к (2) следующим образом:

$$y = \alpha_0 \cdot x_1^{\alpha_0^1} \cdot \left( x_1^{\beta_1^1 \cdot z_1^1 + \ldots + \beta_{k_1}^1 \cdot z_{k_1}^1} \right)^{\gamma_1} \cdot x_2^{\alpha_0^2} \cdot \left( x_2^{\beta_1^2 \cdot z_1^2 + \ldots + \beta_{k_2}^2 \cdot z_{k_2}^2} \right)^{\gamma_2} \cdot \ldots \cdot x_n^{\alpha_0^n} \cdot \left( x_n^{\beta_1^n \cdot z_1^n + \ldots + \beta_{k_n}^n \cdot z_{k_n}^n} \right)^{\gamma_n}, \quad (5)$$

где $\beta^i_j$ фиксируются после построения системы регрессионных уравнений вида

$$y = x_i^{\beta_1^i \cdot z_1^i + \ldots + \beta_{k_i}^i \cdot z_{k_i}^i} = \left( x_i^{z_1^i} \right)^{\beta_1^i} \cdot \left( x_i^{z_2^i} \right)^{\beta_2^i} \cdot \ldots \cdot \left( x_i^{z_{k_i}^i} \right)^{\beta_{k_i}^i}, \; i=1 \div n. \quad (6)$$

Данные алгоритмы вошли в разработанное математическое программное обеспечение интеллектуального анализа данных, которое успешно применяется в РНИУП «Институт радиологии» при построении радиоэкологических и социально-экономических моделей.

Апробация предложенного выше метода осуществлялась при исследовании связей факторов, объясняющих эффективность работы сельскохозяйственных предприятий [6]. Проинтерпретировав полученные результаты на конкретном хозяйстве, можно сделать выводы о наличии у него резервов повышения эффективности экономической деятельности, о задействовании наиболее перспективных рычагов управления, предложить наиболее подходящие решения по выходу из кризисной ситуации.

**Abstract**

The problem of discovering an adequate structure of interdependences of social system's activity factors is decided. The method of using conceptual knowledge for making the nonline multidimension regression model is described.

**Литература**

1. Д.В.Гаскаров, В.И.Шаповалов, Малая выборка, М.: Статистика, 1978.
2. А.Н.Осипенко, Метод и средства автоматизации моделирования активных систем: Автореф. дис... канд.техн.наук.: ГГУ, Гомель (1997).
3. С.А.Айвазян, И.С.Енюков, Л.Д.Мешалкин, Прикладная статистика: Основы моделирования и первичная обработка данных, М.: Финансы и статистика, 1983.
4. С.А.Айвазян, В.М.Бухштабер, И.С.Енюков, Л.Д.Мешалкин, Прикладная статистика: Классификация и снижение размерности, М.: Финансы и статистика, 1989.
5. Ю.А.Шрейдер , А.А.Шаров, Системы и модели, М.: Радио и связь, 1982.
6. А.Н.Осипенко, П.Н.Стрибук, Систематизация пострадавших в результате чернобыльской катастрофы сельскохозяйственных предприятий и выбор направлений их экономического развития, Известия Академии аграрных наук РБ, № 2 (2001), С. 15–24.

Поступило 20.05.2002