Известия Гомельского государственного университета
имени Ф. Скорины, №3(18), 2003

УДК 519.9:618.32

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

П.Н.Стрибук

Введение

Актуальность этой темы вызвана возросшим в последнее десятилетие интересом к фактору активного самодеятельного проявления исследуемого объекта, в частности, – к субъективному фактору социальной системы. 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 корреляционно-регрессионного анализа.

Общая схема исследования

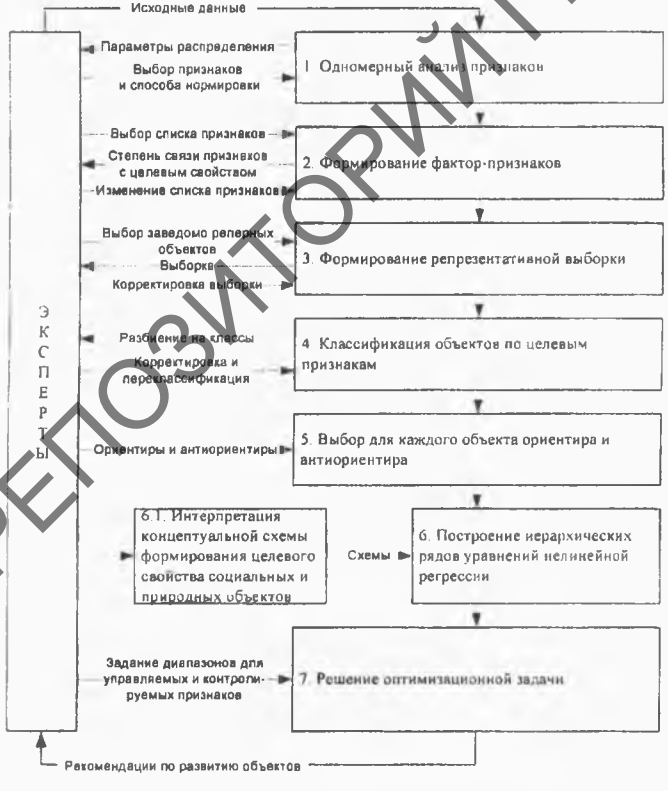

Рисунок 1 Общая схема исследования

Схема исследования причинно-следственных связей в социальных и природных системах реализуется экспертом поэтапно с помощью программно-технологического обеспечения МОНАДА [2] (см. рис. 1).

1. Одномерный анализ признаков с целью корректной нормировки данных на основе оценки распределения (описание дано в [1]).
2. Формирование фактор-признаков, осуществляемое в диалоге с экспертом.
3. Формирование экспертом репрезентативной выборки детально известных (реперных) объектов (описано в [5]).
4. С помощью модуля классификации статистического пакета МОНАДА объекты, представленные в виде точек многомерного пространства 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 предоставляет 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 факторпризнакам, 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).

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 реального перспективного направления развития и предотвращения ошибок, допущенных в управлении объектом-антиориентиром.
6. По выборке реперных объектов 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 (описание дано в [1]).
7. На основе иерархической системы уравнений и заданных экспертом диапазонов допустимого изменения признаков для каждого исследуемого объекта решается оптимизационная задача, показывающая, в каком направлении необходимо развиваться, чтобы максимально приблизиться к объекту-ориентиру и удалиться от объекта-антиориентира.

Рассмотрим подробно ранее не опубликованный алгоритм классификации, который используется в пакете МОНАДА.

Классификация

В модуле классификации используется следующий алгоритм:

1. Рассматривается исходное множество M объектов выборки, задается K – дискретный размер локальной выборочной области (K-области).

1.1. Упорядочение всех объектов обучающей выборки с помощью относительной шкалы «периферия-центр», построенной по K-матрице ближайших соседей в пространстве признаков. K-матрица ближайших соседей состоит из элементов $\{num_{ij}\}$, $i = \overline{1, N}$, $j = \overline{1, K}$, K<<N, где num_{ij} – исходный номер j-го ближайшего соседа для i-го объекта. Шкала «периферия-центр» характеризует степень «центристости» i-ой точки пространства по отношению к окружающим ее точкам и рассчитывается по формуле:

$$C_i = \sum_{l=1}^{N} W(\mathrm{rank}(i, l)), \qquad (1)$$

где rank(i, l) – ранг i-го объекта по отношению к l-му, показывающий, который он по удаленности сосед,

$$W(r) = \begin{cases} K - r + 1, & r \le K \\ 0, & r > K \end{cases}. \qquad (2)$$

Как видим, более центристой является та точка, которая чаще всего встречается среди ближайших соседей других точек. Для каждого C_i – считаем C_i' количество повторений в ряде $\{C_j\}$. Сортируем $\{C_j'\}$ по возрастанию центристости $\{C'_{(i)}\}$. По этому ря-

ду формируется ряд критерия скачка:

$$R_i^C = \frac{C'_{\{i\}}+C'_{\{i+1\}}+C'_{\{i+2\}}+C'_{\{i+3\}}}{C'_{\{i-4\}}+C'_{\{i-3\}}+C'_{\{i-2\}}+C'_{\{i-1\}}}. \quad (3)$$

1.2. Формируется ряд по убыванию расстояния до первого ближайшего соседа $\left\{Rast^1_{\{j\}}\right\}$. По этому ряду формируется ряд критерия скачка:

$$R_j^{Rast} = \frac{Rast^1_{\{j-4\}} + Rast^1_{\{j-3\}} + Rast^1_{\{j-2\}} + Rast^1_{\{j-1\}}}{Rast^1_{\{j\}} + Rast^1_{\{j+1\}} + Rast^1_{\{j+2\}} + Rast^1_{\{j+3\}}}. \quad (4)$$

1.3. В ряде (3) находятся первый и второй максимум, отсекая соответственно множества объектов M_1^C и M_2^C.

1.4. Ряд (4) сортируем по возрастанию. Q_{90} и Q_{95} – 0,90-квантиль и 0,95-квантиль соответственно. Начиная с середины с «окном» в 10 точек, находим первые позиции, для которых значения как минимум пяти точек окна превышают Q_{90} и Q_{95}. Отсекаем множества объектов M_2^{Rast} и M_1^{Rast}, соответственно.

1.5. Формируем два множества объектов $M_1 = M_1^C \cup M_1^{Rast}$ и $M_2 = \left(M_2^C \cup M_2^{Rast}\right) \setminus M_1$. Множество M_1 объявляется *множеством уникальных объектов* и исключается из дальнейшего процесса классификации. Множество M_2 объявляется *множеством точек разряженной зоны* и не участвует во втором этапе классификации.

Примечание: Данный этап (1.2–1.5), как правило, выполняется лишь при достаточно большом количестве объектов.

2. Рассматривается множество объектов $M' = M \setminus \left(M_1 \cup M_2\right)$. Начиная с самой центристой точки $O_j \in M'$, строим k-области:

2.1. Для $k = \overline{2, K}$:

2.1.1. $k^0(O_j)$ – множество k ближайших соседей точки O_j.

2.1.2. $$k^t(O_j) = k^{t-1}(O_j) \cup \left(\bigcup_{m \in k^{t-1}(O_j)} \left(k^0(m), \text{при усл.} \left\| k^{t-1}(O_j) \cap k^0(m) \right\| \geq [k/2]+1 \right) \right), \quad (5)$$

где $\|X\|$ – мощность множества X.

2.1.3. $k^\infty(O_j) = k^t(O_j)$, если $k^t(O_j) = k^{t-1}(O_j)$.

2.1.4. Из пары $((k\text{-}1)^\infty(O_j), k^\infty(O_j))$ выбираем $k^\infty(O_j)$, если

$$\left(\left\| k^\infty(O_j) \right\| < \left\| (k-1)^\infty(O_j) \right\| + k \right) \vee \left(Asim(k^\infty(O_j)) / Asim((k-1)^\infty(O_j)) < \delta \right),$$

где δ – порог чувствительности по асимметрии,

$$Asim(k^\infty(O_j)) = \frac{\left\| \left\{ O \in k^\infty(O_j) : c(O, \overline{O}) > M_c(k^\infty(O_j)) + \Theta \cdot y_c(k^\infty(O_j)) \right\} \right\|}{\left\| k^\infty(O_j) \right\|}, \quad (6)$$

где $M_\rho(k^\infty(O_j))$ – среднее расстояние точек k-области до ее центра $\overline{O}$,
$\sigma_\rho(k^\infty(O_j))$ – среднеквадратическое отклонение расстояний точек k-области от ее центра $\overline{O}$,
Θ – порог просеивания точек класса по асимметрии,
и $(k\text{-}1)^\infty(O_j)$ в противном случае.

2.2. Выполнение процедуры построения k-областей 2.1 осуществляется согласно принципу решета Эратосфена [3] (в качестве очередного эталона выбирается непокрытый построенными k-областями объект с наибольшей центристостью).

2.3. Получаем набор классов Cl_i, $i = \overline{1, N}$.

3. Формируется матрица долей межклассовых пересечений:

$$\Pi(i, j) = \frac{\|Cl_i \cap Cl_j\|}{\|Cl_i\|}, \quad i, j = \overline{1, N}. \tag{7}$$

(Отметим, что матрица (7) в общем случае не симметрическая.)

4. Классы Cl_i и Cl_j объединяются, если $\Pi(i, j)>\gamma$ или $\Pi(j, i)>\gamma$ и при этом выполняется критерий эффективности объединения: $\sigma_\rho(Cl_i \cup Cl_j) \le \alpha \cdot \max(\sigma_\rho(Cl_i), \sigma_\rho(Cl_j))$, то есть дисперсия расстояний значительно не увеличится. Здесь γ – порог пересечения, α – коэффициент допустимого увеличения дисперсии.

5. Все классы просеиваются, формируя множество удаленных точек:

$$M_{del} = M_2 \cup \left(\bigcup_{i=1}^{N} \left\{ O \in Cl_i : c(O, \overline{O}) > M_c(Cl_i) + \Theta \cdot y_c(Cl_i) \right\} \right). \tag{8}$$

6. Точки множества M_{del} последовательно распределяются по классам, если при добавлении точки внутриклассовая дисперсия значительно не увеличится. Остается множество точек M'_{del}.

7. На множестве M'_{del} пытаемся построить классы:

   7.1. Начинаем с самой центристой точки $O_j \in M'_{del}$ и далее действуем согласно принципу решета Эратосфена.

   7.2. Строим класс $1^\infty(O_j)$ (см. 2.1.1.–2.1.3.).

   7.3. Для $k = \overline{2, K}$ строим классы $k^\infty(O_j)$ пока выполняется условие

$$\frac{\|k^\infty(O_j) \cap (M \setminus M'_{del})\|}{\|k^\infty(O_j)\|} < \frac{1}{2}. \tag{9}$$

   7.4. В качестве нового класса рассматриваем $k^\infty(O_j) \setminus (M \setminus M'_{del})$.

8. Проводим процедуру уменьшения пересечений получившихся классов:

   8.1. Последовательно рассматриваются объекты, принадлежащие более чем одному классу.

   8.2. Для каждого из них находится максимальное и минимальное расстояния до содержащих его классов – ρ_{min} и ρ_{max}.

   8.3. Если $\rho_{max} \ge \psi \cdot \rho_{min}$, то объект исключается из наиболее удаленного класса.

   8.4. Повторяем 8.1.–8.3. пока удается уменьшить пересечения классов.

Примечание: В качестве метрики используется

$$\rho(X, Y) = \max_i(|X_i - Y_i|) \cdot \sum_i |X_i - Y_i|. \tag{10}$$

Апробация предложенного выше метода осуществлялась при исследовании связей факторов, объясняющих эффективность работы сельскохозяйственных предприятий [4].

Abstract

The problem of discovering an adequate structure of interdependences of social system's activity factors isconsidered. The method of classification is described.

Литература

1. Стрибук П. Н., Осипенко А. Н., Осипенко Н. Б. Выявление причинно-следственных связей в компьютерном моделировании социальных и природных систем // Известия Гомельского государственного университета имени Ф. Скорины. – 2002. – №6(15). – С.105–109.

2. Осипенко А. Н. Метод и средства автоматизации моделирования активных систем: Автореф. дис... канд. техн. наук. – Гомель: ГГУ, 1997.

3. Айвазян С. А., Бухштабер В. М., Енюков И. С. Мешалкин Л. Д. Прикладная статистика: Классификация и снижение размерности. М.: Финансы и статистика, 1989. – 312 с.

4. Осипенко А. Н., Стрибук П. Н. Систематизация пострадавших в результате чернобыльской катастрофы сельскохозяйственных предприятий и выбор направлений их экономического развития.// Известия Академии аграрных наук РБ. – 2001. – №2. – С. 15–24.

5. Стрибук П. Н. Алгоритм планирования пассивного регрессионного эксперимента // Новые математические методы и компьютерные технологии в проектировании, производстве и научных исследованиях: Материалы VI Республиканской научной конференции студентов и аспирантов. – Гомель: ГГУ, 2001. – С. 58–59

Гомельский государственный
университет им. Ф.Скорины

Поступило 14.04.03