Прогнозирование экстремальных событий на основе анализа многомерных разнотипных временных рядов*

Г. С. Лбов, М. К. Герасимов
Институт математики СО РАН, Новосибирск, Россия
e-mail: max_post@ngs.ru

Разработан метод построения логико-вероятностных моделей для прогнозирования экстремальных событий, например, чрезвычайной ситуации природного или техногенного характера. Для прогнозирования экстремальных событий применяются многомерные разнотипные временные ряды, т. е. изменяющиеся во времени как количественные, так и качественные характеристики. В рассматриваемом методе используется взвешенное расстояние (мера близости) в пространстве разнотипных переменных. Для нахождения оптимальных весов предложен метод адаптивного поиска приближенного значения глобального экстремума функции на симплексе.

Ключевые слова: экстремальные события, взвешенное расстояние, многоэкстремальная оптимизация.

Введение

Существующие методы прогнозирования экстремальных событий ориентированы на анализ числовых временных рядов (см., например, [1, 2]). Однако, с нашей точки зрения, любое экстремальное событие связано с определенным сочетанием набора значений разнотипных переменных, измеренных в разные моменты времени. Поэтому при анализе и прогнозировании реальных экстремальных событий и явлений необходимо учитывать следующие особенности:

— методы прогнозирования должны использовать комплексное описание, содержащее как можно более полную информацию обо всех факторах, потенциально влияющих на возникновение экстремальных событий;

— реальные данные представляют собой многомерные временные ряды, которые могут одновременно включать числовые, бинарные и символьные последовательности;

— задачу следует решать в условиях, когда имеющиеся ряды имеют относительно малую длину;

— экстремальные ситуации по своей сути являются редкими событиями, поэтому количество соответствующих им прецедентов в эмпирической информации мало по отношению к общему объему выборки.

Как показывают теоретические и экспериментальные исследования, наиболее подходящим подходом для решения прикладных задач, характеризующихся указанными

*Работа выполнена при финансовой поддержке РФФИ (гранты № 09-07-12087-офи-м, 10-01-00113а и 08-07-00136а) и междисциплинарного интеграционного проекта СО РАН № 83.

особенностями, является построение логико-вероятностных моделей [3, 4] изучаемых объектов или явлений. Логико-вероятностные модели строятся в классе логических решающих функций от разнотипных переменных и представляются списком логических закономерностей ("знаний"), отражающих внутренние причинно-следственные связи сложных объектов или явлений и обладающих достаточно высокой прогнозирующей способностью. Такого рода закономерности близки к естественному языку логических суждений, что допускает совместную обработку результатов статистического анализа и знаний экспертов.

В данной работе для прогнозирования экстремальных событий предложен метод логико-вероятностных моделей. В разделе 1 дана формальная постановка задачи, в разделе 2 описывается метод предсказания возникновения экстремальных ситуаций, в разделе 3 рассматривается метод оценки количественного показателя экстремальных событий, использующий взвешенное расстояние в пространстве разнотипных переменных. Для нахождения оптимальных коэффициентов-весов предложен метод адаптивного поиска приближенного значения глобального экстремума функции на симплексе (разделы 4 и 5). В разделе 6 приведен пример решения прикладной задачи прогнозирования экстремальных гидрологических событий.

1. Постановка задачи

Рассматривается некоторый процесс, потенциально приводящий к экстремальным событиям. Пусть имеется $\mathfrak{N}$ различных реализаций наблюдаемого процесса $a^1, \ldots, a^i, \ldots, a^{\mathfrak{N}}$ (каждая реализация a^i представляет собой многомерный разнотипный временной ряд). Предполагается, что в последовательные моменты времени $t_1, \ldots, t_\nu, \ldots, t_T$ проведены измерения по набору переменных $X = \{X_1, \ldots, X_j, \ldots, X_n\}$. Данный набор может одновременно содержать количественные и качественные переменные. Имеется целевая количественная переменная Y, превышение определенного порога которой в будущий момент времени t_{T+1} означает экстремальную ситуацию. Таким образом, для анализа имеется обучающая выборка

$$V = \left(x_1^i(t_\nu), \ldots, x_j^i(t_\nu), \ldots, x_n^i(t_\nu), y^i\right), \quad i = \overline{1, \mathfrak{N}}, \quad \nu = \overline{1, T},$$

где $x_j^i(t_\nu)$ — значение переменной X_j для реализации a^i в момент времени t_ν, $y^i = Y(a^i)$ — значение целевой переменной Y для реализации a^i в момент времени t_{T+1}. Обозначим $x^i(t_\nu) = \left(x_1^i(t_\nu), \ldots, x_j^i(t_\nu), \ldots, x_n^i(t_\nu)\right)$. Предполагается, что многомерный случайный процесс таков, что в предысториях $\left(x^i(t_1), \ldots, x^i(t_\nu), \ldots, x^i(t_T)\right)$ существуют закономерности, отражающие возникновение экстремальных событий.

Будем считать, что реализации a^i могут быть упорядочены по значению целевой переменной Y и отнесены к одному из K классов, $K \geq 2$ (например, Y — уровень воды в реке, классы — "наводнение", "половодье", "норма", "маловодье"). Обозначим через γ^i номер класса для реализации a^i, через $I_k = \{i | \gamma^i = k\}$ — множество номеров реализаций соответствующего класса, $k = \overline{1, K}$. Будем для определенности полагать, что экстремальные события относятся к первому классу, а события, наиболее близкие к ним по значению целевой переменной Y, — ко второму.

Пусть для новой реализации a^* измерены наблюдения по набору X в моменты времени $t_1, \ldots, t_\nu, \ldots, t_T$. Для реализации a^* необходимо оценить вероятность возникновения экстремальной ситуации в момент времени t_{T+1} (первая задача) и, если вероятность высока, оценить значение Y (вторая задача). Предполагается, что распределение

$\mathsf{P}\big(x(t_1), \ldots, x(t_\nu), \ldots, x(t_T), y\big)$ неизвестно и, следовательно, необходимо оценить его на основе анализа обучающей выборки.

Особенность рассматриваемых задач прогнозирования, кроме разнотипности переменных, состоит в том, что количество наблюдений экстремальных событий в эмпирической информации по отношению к общему объему выборки мало, а возможные потери от ошибочного предсказания отсутствия экстремального события достаточно велики.

Введем следующие обозначения: D_j и D_y — области определения переменных X_j и Y соответственно, $D = \prod\limits_{j=1}^{n} D_j$ — пространство разнотипных переменных.

Назовем множество $E \subseteq D$ *прямоугольной областью*, если $E = \prod\limits_{j=1}^{n} E_j$, где $E_j = [\alpha_j, \beta_j]$ — некоторый интервал, если X_j — количественная переменная, E_j — некоторый список имен, если X_j — качественная переменная, т. е. переменная с неупорядоченным набором значений.

2. Прогнозирование экстремальных событий

Особенностью первой поставленной задачи является то, что для построения решающей функции распознавания необходимо использовать трехмерную таблицу “объект—свойство—время”, тогда как в стандартных методах распознавания применяются двумерные таблицы “объект—свойство”. Поэтому в данной работе используется способ построения решающей функции распознавания в два этапа, причем на каждом этапе решение строится по двумерной таблице “объект—свойство” [4].

Этап 1. Для каждого момента времени t_ν, $\nu = \overline{1, T}$, выполним следующие действия:

1) из таблицы V выделим двумерную таблицу

$$V_\nu = \left(x_1^i(t_\nu), \ldots, x_j^i(t_\nu), \ldots, x_n^i(t_\nu), \gamma^i\right), \quad i = \overline{1, \mathfrak{N}};$$

2) с помощью алгоритма распознавания образов (например LRP [4]) получим логическую решающую функцию f_ν. Этой функции соответствует пара $\langle \alpha_\nu, r_\nu(\alpha_\nu) \rangle$, где $\alpha_\nu = \left\{E_\nu^1, \ldots, E_\nu^\chi, \ldots, E_\nu^{M_\nu}\right\}$ — разбиение пространства D на прямоугольные области, $r_\nu(\alpha_\nu)$ — набор решений для разбиения α_ν: $r_\nu(\alpha_\nu) = \left\{r_\nu^1, \ldots, r_\nu^\chi, \ldots, r_\nu^{M_\nu}\right\}$, r_ν^χ — номер класса для множества E_ν^χ .

Этап 2. После выполнения этапа 1 построены T решающих правил f_ν, $\nu = \overline{1, T}$, и соответствующие разбиения пространства D. Введем для каждого t_ν, $\nu = \overline{1, T}$, новую номинальную переменную Z_ν с областью определения $D_{Z_\nu} = \{1, \ldots, \chi, \ldots, M_\nu\}$. Будем считать, что переменная Z_ν приняла значение $z_\nu^i := Z_\nu(a^i) = \chi$, если точка $x^i(t_\nu) \in E_\nu^\chi$. Это преобразование позволяет многомерную реализацию $(x^i(t_1), \ldots, x^i(t_\nu), \ldots, x^i(t_T))$ представить в виде символьной последовательности $(z_1^i, \ldots, z_\nu^i, \ldots, z_T^i)$.

Учитывая преобразование $x \to z$, трехмерному массиву данных V ставим в соответствие двумерную таблицу

$$V_{(Z)} = \left(z_1^i, \ldots, z_\nu^i, \ldots, z_T^i, \gamma^i\right), \quad i = \overline{1, \mathfrak{N}}.$$

С помощью алгоритма распознавания образов построим логическую решающую функцию $f_{(Z)}$. Этой функции соответствует пара $\big\langle \alpha_{(Z)}, r_{(Z)}(\alpha_{(Z)}) \big\rangle$, где

$\alpha_{(Z)} = \left\{E^1_{(Z)}, \ldots,\ E^m_{(Z)}, \ldots, E^M_{(Z)}\right\}$ — разбиение пространства $D_{(Z)} = \prod_{\nu=1}^{T} D_{Z_\nu}$, $r_{(Z)}(\alpha_{(Z)}) = \left\{r^1_{(Z)}, \ldots,\ r^m_{(Z)}, \ldots, r^M_{(Z)}\right\}$, $r^m_{(Z)}$ — номер класса для множества $E^m_{(Z)}$. Каждому множеству $E^m_{(Z)}$ из $\alpha_{(Z)}$ ставятся в соответствие относительные частоты $\bar{P}^m_1, \ldots, \bar{P}^m_k, \ldots, \bar{P}^m_K$ для всех K образов.

Таким образом, реализации a^* по наблюдениям $(x^*(t_1), \ldots, x^*(t_\nu), \ldots, x^*(t_T))$ можно поставить в соответствие множество $E^m_{(Z)}$ и, следовательно, наиболее вероятный номер класса $r^m_{(Z)}$ (номер с максимальной относительной частотой), а также оценки вероятностей принадлежности в момент времени t_{T+1} к тому или иному классу.

3. Оценка количественной характеристики экстремальных событий

Прямое оценивание количественного показателя экстремального события затруднительно ввиду малого числа таких событий в обучающей выборке. Поэтому для решения второй поставленной задачи предлагается следующая идея. Все предыстории $\big(x^i(t_1), \ldots, x^i(t_\nu), \ldots, x^i(t_T)\big)$ (т. е. реализации a^i) можно упорядочить по потерям и соответственно упорядочить подобласти, к которым принадлежат рассматриваемые реализации. Сравнивая степени близости оцениваемой предыстории к областям истинности найденных закономерностей, можно оценить значение целевой переменной Y. Для этого необходимо ввести понятие расстояния (меры близости) между множествами в многомерном пространстве разнотипных переменных, причем можно использовать меры, предложенные в работах [5 – 7]. В данной работе применяется способ задания расстояния между прямоугольными областями в многомерном разнотипном пространстве [6].

Пусть $E^1 \subseteq D$ и $E^2 \subseteq D$ — прямоугольные области, $E^s = \prod_{j=1}^{n} E^s_j$, $E^s_j \subseteq D_j$, $s = \overline{1,2}$. Для этих множеств введем расстояние

$$\rho(E^1, E^2) = \sum_{j=1}^{n} \lambda_j \rho_j(E^1_j, E^2_j)$$

или, по аналогии с евклидовым расстоянием,

$$\rho(E^1, E^2) = \sqrt{\sum_{j=1}^{n} \lambda_j (\rho_j(E^1_j, E^2_j))^2},$$

где $0 \leq \lambda_j \leq 1$, $\sum_{j=1}^{n} \lambda_j = 1$. Величины $\rho_j(E^1_j, E^2_j)$ задаются в зависимости от типа переменной:

$$\rho_j(E^1_j, E^2_j) = \frac{\mu(E^1_j \Delta E^2_j)}{\mu(D_j)}, \text{ если } X_j \text{ — качественная переменная,}$$

$$\rho_j(E^1_j, E^2_j) = \frac{r^{12}_j + \theta \cdot \mu(E^1_j \Delta E^2_j)}{\mu(D_j)}, \text{ если } X_j \text{ — количественная переменная,}$$

где $\mu(E)$ — мера множества E, $r^{12}_j = \left|\dfrac{\alpha^1_j + \beta^1_j}{2} - \dfrac{\alpha^2_j + \beta^2_j}{2}\right|$, если $E^s_j = [\alpha^s_j, \beta^s_j]$, $s = \overline{1,2}$.

Для нормировки от 0 до 1 и выполнения свойств расстояния необходимым и достаточным является условие $0 \le \theta \le 1/2$ [6].

Пусть $\bar{y}$ — порог, превышение которого целевой переменной означает экстремальную ситуацию. Обозначим

$$\widehat{y} = \max_{i=\overline{1,\mathfrak{N}}} y^i, \quad \Delta y^i = \frac{y^i - \bar{y}}{\widehat{y} - \bar{y}}, \quad i = \overline{1,\mathfrak{N}}.$$

Для определенности будем полагать, что в построенном решении $f_{(Z)}$ экстремальным событиям соответствуют множества $E^m_{(Z)}$, т. е. $r^m_{(Z)} = 1$, $m \in \{1,\ldots,M\}$. Рассмотрим набор переменных $Z(m) \subseteq \{Z_1,\ldots,Z_\nu,\ldots,Z_T\}$, информативных для описания множества $E^m_{(Z)}$, т. е. переменная $Z_\nu \in Z(m)$, если $(E^m_{(Z)})_\nu \ne D_{Z_\nu}$. Каждой переменной из набора $Z(m)$ соответствует момент времени t_ν, влияющий на возникновение экстремальных событий. Для переменной $Z_\nu \in Z(m)$ определим в соответствующей решающей функции f_ν множества второго класса, т. е. множества E^χ_ν такие, что $r^\chi_\nu = 2$.

Найдем взвешенное расстояние между объектами первого (т. е. экстремальными событиями) и второго классов. Будем использовать следующую гипотезу: существует такой набор значений коэффициентов λ_j (весов), при котором мера различия экстремальных событий от неэкстремальных по набору переменных X соответствует мере различия по целевой переменной Y:

$$\rho\left(\{x^i(t_\nu)\}, E^\chi_\nu\right) \approx \Delta y^i \quad \text{для } i \in I_1.$$

Таким образом, определив коэффициенты λ_j, можно по степени близости к наблюдавшимся ранее экстремальным событиям оценить количественную характеристику прогнозируемой экстремальной ситуации.

Для формализации гипотезы введем функционал

$$F(\lambda_1,\ldots,\lambda_j,\ldots,\lambda_n) = \sum_{\nu|Z_\nu \in Z(m)} \sum_{\chi|r^\chi_\nu = 2} F^\chi_\nu(\lambda_1,\ldots,\lambda_j,\ldots,\lambda_n),$$

$$\text{где} \quad F^\chi_\nu(\lambda_1,\ldots,\lambda_j,\ldots,\lambda_n) = \sum_{i \in I_1} \left(\rho(\{x^i(t_\nu)\}, E^\chi_\nu) - \Delta y^i\right)^2.$$

Заметим, что область значений коэффициентов $\lambda_1,\ldots,\lambda_n$ (обозначим ее через Λ) представляет собой многомерный симплекс

$$\Lambda = \Big\{\lambda = (\lambda_1,\ldots,\lambda_n)\colon \sum_{j=1}^n \lambda_j = 1, \quad \lambda_j \ge 0, \;\; j = 1,\ldots,n\Big\} \subset \mathbb{R}^n.$$

При других способах задания расстояния (меры близости) между объектами могут быть получены другие многомерные области значений коэффициентов.

Для нахождения оптимальных коэффициентов-весов λ_j, $j = \overline{1,n}$, минимизирующих функционал F, далее предложен метод адаптивного поиска приближенного значения глобального экстремума функции на симплексе.

Процедура оценки количественной характеристики экстремального события осуществляется следующим образом.

Рассмотрим новую реализацию a^*. Если по построенной логической решающей функции $f_{(Z)}$ следует, что будет экстремальное событие и $(z_1^*, \ldots, z_\nu^*, \ldots, z_T^*) \in E_{(Z)}^m$, то по каждому из указанных множеств E_ν^χ можно сделать частную оценку

$$y_\nu^\chi \approx \bar{y} + \rho\left(\{x^i(t_\nu)\}, E_\nu^\chi\right) \cdot (\widehat{y} - \bar{y}).$$

Определим весовые коэффициенты частных решений: $\omega_\nu^\chi = 1 - F_\nu^\chi / |I_1|$. После получения всех частных решений можно сделать общий прогноз:

$$y^* = Y(a^*) \approx \frac{\sum\limits_{\nu | Z_\nu \in Z(m)} \sum\limits_{\chi | r_\nu^\chi = 2} \omega_\nu^\chi y_\nu^\chi}{\sum\limits_{\nu | Z_\nu \in Z(m)} \sum\limits_{\chi | r_\nu^\chi = 2} \omega_\nu^\chi}.$$

4. Метод многоэкстремальной оптимизации на симплексе

Если не вводить ограничений на класс функций, то в соответствии с теоремой об отсутствии бесплатного завтрака (No Free Lunch Theorem) различные алгоритмы поиска глобального экстремума функции имеют одинаковую среднюю эффективность [8, 9]. Поскольку свойства функции $F(\lambda)$ в прикладных задачах неизвестны, то при решении указанной оптимизационной задачи можно применять только слабые ограничения на класс оптимизируемых функций. В данной работе используются достаточно общие свойства класса $\{F(\lambda)\}$ в виде некоторой "разумной" гипотезы, приведенной ниже.

Для нахождения оптимальных коэффициентов-весов будем использовать модификацию метода поиска приближенного значения глобального экстремума функции (ADAPT [3, 4, 10]). Отметим, что метод ADAPT является развитием метода СПА [11]. Модификация метода ADAPT, предлагаемая в работе, заключается в том, что область поиска представляется в виде симплекса.

Рассмотрим функцию $F(\lambda)$, $\lambda \in \Lambda$. Обозначим $\lambda^* = \arg\min_{\lambda \in \Lambda} F(\lambda)$. Пару $\langle \lambda^i, F(\lambda^i) \rangle$ будем называть *испытанием*. Под алгоритмом поиска приближенного значения глобального минимума функции понимается некоторая процедура последовательного планирования N испытаний с целью нахождения минимально возможного значения функции. Общее число испытаний N разбивается на R групп: $N = N_1 + \ldots + N_R$.

Обозначим через $N^\tau = \sum\limits_{i=1}^{\tau} N_i$ — число испытаний, проведенных за τ шагов поиска (под шагом поиска понимается проведение N_i испытаний), $\tau = 1, ..., R$. Первая группа испытаний планируется таким образом, чтобы для любого $E \subset \Lambda$ число испытаний, соответствующее множеству E, было примерно равно $\dfrac{N_1 V(E)}{V(\Lambda)}$, где $V(E)$ — объем области E. Другими словами, точки $\lambda^1, \ldots, \lambda^{N_1}$ планируем таким образом, чтобы они были максимально равномерно распределены по множеству Λ.

Для вычисления объемов в предлагаемом методе используется метод Монте-Карло. Как показано в работе [12], этот метод эффективен для вычисления многомерных интегралов, в частности объемов. Кроме того, метод Монте-Карло позволяет определить объемы и более сложных, чем симплекс, областей.

Пусть проведено N^τ испытаний ($\tau = 1, \ldots, R-1$) и получена соответствующая таблица $v^\tau = \{\lambda^i, F(\lambda^i)\}$, $i = 1, ..., N^\tau$. С помощью варианта алгоритма LRP для аппроксимации функций (см., например, [4]) построим наилучшую регрессионную функцию $\bar{f}^\tau$

в классе логических решающих функций. Функции $\bar{f}^{\tau}$ соответствует некоторое разбиение множества Λ: $\alpha_\tau = \{E_\tau^1, \ldots, E_\tau^{M_\tau}\}$. Множества E_τ^t имеют вид

$$E_\tau^t = \{\lambda \in \Lambda\colon c_j < \lambda_j \le d_j,\, j = 1, \ldots, n\}\,.$$

Будем использовать следующую гипотезу: вероятность достижения глобального минимума функции в области E зависит от числа проведенных испытаний, объема множества E и результатов проведенных ранее в точках множества E испытаний. Предполагается, что чем больше объем множества E (а, значит, и его неисследованность) и меньше по сравнению с другими областями полученные при испытаниях значения функции, тем больше вероятность достижения глобального минимума в этом множестве.

Введем вспомогательную функцию

$$\kappa_b(z) = \frac{b+1}{(1+bz)^2}, \quad 0 \le z \le 1, \quad 0 \le b \le \infty.$$

Выбор данной функции определяется следующими ее свойствами:

1) при $b = 0$ получаем $\kappa_b(z) = 1$;

2) при $b > 0$ имеем монотонно убывающую функцию по z;

3) чем больше значение параметра b, тем больше скорость убывания функции по z;

4) $\int\limits_0^1 \kappa_b(z)dz = 1$ при любом b.

Эти свойства функции $\kappa_b(z)$ позволяют формализовать используемую гипотезу. Заметим, что в качестве функции $\kappa_b(z)$ можно выбрать любую функцию, удовлетворяющую указанным свойствам.

Зададим функцию

$$\kappa^u = \int\limits_0^u \kappa_b(z)dz = \frac{(b+1)u}{1+bu}, \quad 0 \le u \le 1.$$

Определим вероятности p_τ^t проведения испытания в множествах E_τ^t, $t = 1, \ldots, M_\tau$. Обозначим через $F_{\min}^t$ минимальное значение функции F при проведенных испытаниях в множестве E_τ^t. Пусть $F_{\min}^{t_1} \le \ldots \le F_{\min}^{t_{M_\tau}}$. По оси z последовательно будем откладывать величины z^i, равные относительным объемам соответствующих множеств:

$$z^i = \frac{V(E_\tau^{t_i})}{V(\Lambda)}, \quad i = 1, \ldots, M_\tau,$$

т. е. вначале откладывается относительный объем наилучшего множества $E_\tau^{t_1}$, которому соответствует $F_{\min}^{t_1}$, затем относительный объем второго по порядку множества $E_\tau^{t_2}$, которому соответствует $F_{\min}^{t_2}$, и т. д. Таким образом, отрезок $[0, 1]$ на оси z разбивается на M_τ отрезков $[z_{i-1}, z_i]$, $i = 1, \ldots, M_\tau$ (будем полагать, что $z_0 = 0$). Обозначим $u_i = \sum\limits_{j=0}^{i} z_j$, $i = 0, \ldots, M_\tau$. Вероятности p_τ^t определяются по формуле

$$p_\tau^{t_i} = \kappa^{u_i} - \kappa^{u_{i-1}} = \int\limits_0^{u_i} \kappa_b(z)dz - \int\limits_0^{u_{i-1}} \kappa_b(z)dz = \frac{(b+1)(u_i - u_{i-1})}{(1+bu_i)(1+bu_{i-1})}, \quad i = 1, \ldots, M_\tau.$$

Таким образом, чем больше объем множества E^t_τ и меньше номер i, определяющий номер этого множества в порядке $E^{t_1}_\tau, \ldots, E^{t_{M_\tau}}_\tau$, тем больше вероятность p^t_τ (согласно указанной выше гипотезе). Зададим линейную зависимость параметра b от числа проведенных испытаний N^τ, т. е. на $(\tau+1)$-м шаге поиска будем использовать величину $b^\tau = \dfrac{N^\tau}{N} b_{\max}$. Величина $b_{\max}$ определяется из следующих соображений. После проведения всех испытаний ($N^R = N$) в соответствии с гипотезой можно указать область $E(\gamma)$ такую, что вероятность нахождения точки λ^* равна $p(\gamma) \simeq 1$. При этом область $E(\gamma)$ достаточно мала, т. е. $\gamma = \dfrac{V(E(\gamma))}{V(\Lambda)} \simeq 0$. Например, $p(\gamma) = 0.95$, а $\gamma = 0.05$. Величина $b_{\max}$ определяется из соотношения

$$p(\gamma) = \int\limits_0^\gamma \kappa_b(z)dz = \frac{(b_{\max}+1)\gamma}{1 + b_{\max}\gamma}.$$

5. Исследование эффективности метода оптимизации

Предложенный алгоритм поиска приближенного значения глобального минимума функции на симплексе был проверен на ряде тестовых функций:

1) $f_1(x) = \cos 2\pi x_1 + \cos 3\pi x_2 + \cos 6\pi x_3$;
2) $f_2(x) = 100(x_1 - x_2)^2 + 100 x_3 x_4$;
3) $f_3(x) = (x_1 - 0.6)^2 + (x_2 - 0.3)^2 - \cos(18x_1 - 10.8) - \cos(18x_2 - 5.4) - \cos(20x_3 - 2)$;
4) $f_4(x) = (x_1 - 0.3)^2 + (x_2 - 0.4)^2 + (x_3 - 0.1)^2 + (x_4 - 0.2)^2$;
5) $f_5(x) = -100 x_1 x_2 x_3$.

Можно показать, что глобальный минимум функции $f_1(x)$ на симплексе достигается при $x_1 = \pi/2$, $x_2 = \pi/3$, $x_3 = \pi/6$; $f_2(x)$ — при $x_1 = x_2$, $x_3 \vee x_4 = 0$; $f_3(x)$ — при $x_1 = 0.6$, $x_2 = 0.3$, $x_3 = 0.1$; $f_4(x)$ — при $x_1 = 0.3$, $x_2 = 0.4$, $x_3 = 0.1$, $x_4 = 0.2$; $f_5(x)$ — при $x_1 = x_2 = x_3 = 1/3$.

Для каждой функции проведено многократное моделирование (100 запусков алгоритма). В табл. 1 приводятся некоторые результаты тестирования. Обозначим для каждой функции число испытаний через N, истинное значение глобального минимума через $F_{\min}$, найденное среднее приближенное значение глобального минимума через $F^*_{\min}$. Кроме того для каждой функции приводится величина σ^*, вычисленная по формуле

$$\sigma^* = \sqrt{\frac{\sum_{i=1}^{100}(F_{\min} - F^i_{\min})^2}{100}},$$

Т а б л и ц а 1. Результаты исследования метода оптимизации на тестовых примерах

f	$F_{\min}$	$F^*_{\min}$	σ^*	$F^0_{\min}$	σ^0	N
$f_1(x)$	-3	-2.94	0.12	-2.87	0.17	150
$f_2(x)$	0	0.34	0.51	0.53	0.73	40
$f_3(x)$	-3	-2.89	0.15	-2.79	0.25	200
$f_4(x)$	0	0.015	0.020	0.018	0.022	60
$f_5(x)$	-3.70	-3.63	0.10	-3.60	0.13	60

Т а б л и ц а 2. Результаты исследования метода оптимизации на функции $f_6(x) = -\exp(-2\rho)\cos(400\rho)$

$F^*_{\min}$	σ^*	$F^0_{\min}$	σ^0	N
-0.80	0.21	-0.80	0.21	200
-0.86	0.17	-0.83	0.18	300
-0.91	0.11	-0.85	0.16	400

где $F^i_{\min}$ — приближенное значение глобального минимума, найденное при i-м запуске алгоритма. Для сравнения также приведены значения $F^0_{\min}$ и σ^0 для случайного поиска без адаптации при том же числе испытаний.

Рассмотрим функцию $f_6(x) = -\exp(-2\rho)\cos(400\rho)$, где

$$\rho = \sqrt{\sum_{j=1}^{3}(x_j - 1/3)^2}.$$

Данная функция характеризуется большим числом локальных экстремумов, "резкими" перепадами положительных и отрицательных значений и малым объемом отрицательной подобласти (около 0.000035 от объема симплекса), в которой достигается глобальный минимум $F_{\min} = -1$. Результаты приведены в табл. 2.

Для сравнения тестовые функции были исследованы средствами MATLAB при случайных значениях начальной точки. Результаты показали, что на "простых" функциях f_2, f_4 и f_5 был найден глобальный минимум. В то же время на более сложных функциях выявилось преимущество предложенного метода оптимизации. Так, для функции f_1 при 100 запусках 14 раз был найден глобальный минимум $F_{\min} = -3$ и 42 раза найдено значение $F_{\min} = -1.76$. Для функции f_6 при 100 запусках наилучшим результатом были значения -0.80 (три раза) и -0.71 (четыре раза).

6. Пример решения прикладной задачи

Методы, разработанные в рамках данного подхода, были применены для прогноза экстремальных ситуаций на реках Сибири [13, 14]. В частности, оценивалось возникновение в марте экстремальной по маловодью ситуации в контрольной точке "Барнаул". Были обработаны среднемесячные данные замеров стока р. Обь за период с 1937 по 1990 г. Использовались следующие дополнительные характеристики: температура воздуха и количество выпавших осадков за сентябрь и октябрь предыдущего года (станции "Онгудай" и "Волчиха"). Для контроля рассматривались данные с 1991 по 2000 г. За этот период сток р. Обь в марте колебался от 191 до 590 м^3/с и наблюдалось одно маловодье (1998 г.). Разработанными методами было верно предсказано маловодье 1998 года и сделан неверный прогноз маловодья на 1999 год. В остальные годы правильно предсказано отсутствие маловодья.

Приведем некоторые из полученных закономерностей.

1. При стоке в сентябре больше 1150 м^3/с, сентябрьских осадках в Онгудае меньше 164 мм и стоке в октябре меньше 994 м^3/с в марте будет маловодье.

2. При стоке в сентябре больше 1150 м^3/с, сентябрьских осадках в Онгудае меньше 164 мм и стоке в октябре больше 994 м^3/с в марте маловодья не будет.

Наша оценка стока в марте 1998 г. имела величину 254 м^3/с, реальный сток составил 251 м^3/с.

Для сравнения была построены различные регрессионные модели. Результаты показали преимущество предложенного подхода. Так, для регрессионной модели оценка стока в марте 1998 г. составила 478 м^3/с.

Необходимо подчеркнуть, что имеющаяся информация представляет собой числовые (количественные) временные ряды, тогда как разработанные методы могут быть использованы для решения более общих задач, в которых информация об исследуемых событиях описывается как количественными, так и качественными характеристиками.

Заключение

Экстремальные события, как правило, вызываются уникальными комбинациями сложных причинно-следственных зависимостей, поэтому число соответствующих прецедентов мало по отношению к общему объему выборки, при этом возможные потери от экстремальных событий достаточно велики. Применение логико-вероятностных моделей для построения решающей функции прогнозирования дает возможность оценить не только вероятность появления экстремальных событий, но и само значение заданной целевой количественной характеристики. Для этого использовано взвешенное расстояние (меры близости) в разнотипном пространстве. Предложен метод поиска приближенного значения глобального экстремума функции для нахождения оптимальных коэффициентов-весов, позволяющий прогнозировать даже уникальные события, когда величина целевой функции превышает значения, достигавшиеся ранее. Результаты решения тестовых и прикладных задач показали эффективность данного метода прогнозирования экстремальных событий.

Список литературы

[1] Hallerberg S., Bröcker J., Kantz H. Prediction of extreme events // Lecture Notes Earth Sci. 2008. Vol. 112. P. 35–59.

[2] Denny M., Hunt L., Miller L., Harley C. On the prediction of extreme ecological events // Ecological Monographs. 2009. Vol. 79, No. 3. P. 397–421.

[3] Лбов Г.С. Методы обработки разнотипных экспериментальных данных. Новосибирск: Наука, 1981. 160 с.

[4] Лбов Г.С., Старцева Н.Г. Логические решающие функции и вопросы статистической устойчивости решений. Новосибирск: Институт математики СО РАН, 1999. 212 с.

[5] Загоруйко Н.Г. Прикладные методы анализа данных и знаний. Новосибирск: Институт математики СО РАН, 1999. 268 с.

[6] Лбов Г.С., Герасимов М.К. Введение расстояния между логическими высказываниями в задачах прогнозирования // Искусств. интеллект. 2004. № 2. С. 105–108.

[7] Викентьев А.А. Расстояние на высказываниях экспертов и мера опровержимости (информативности) высказываний с помощью моделей некоторых теорий // Докл. 12-й Всероссийской конф. “Математические методы распознавания образов”. М.: МАКС Пресс, 2005. С. 60–63.

[8] Wolpert D.H., Macready W.G. No Free Lunch Theorems for Search / Techn. Rep. SFI-TR-95-02-010. Sante Fe. 1995.

[9] Wolpert D.H., Macready W.G. No free lunch theorems for optimization // IEEE Trans. Evolut. Comp. 1997. Vol. 1(1). P. 67–82.

[10] Лбов Г.С., Бериков В.Б., Зенкова Н.А. Экспериментальное сравнение алгоритмов адаптивного поиска глобального экстремума функции // Тр. учредит. конф. междунар. ассоциации "Нетрадиционные методы оптимизации". Красноярск, КИКТ, 1992.

[11] Лбов Г.С. Выбор эффективной системы зависимых признаков //Вычисл. системы. 1965. Вып. 19. С. 21–34.

[12] Метод статистических испытаний (метод Монте-Карло) / Н.П. Бусленко, Д.И. Голенко, И.М. Соболь, В.Г. Срагович, Ю.А. Шрейдер. М.: Гос. изд-во физ.-мат. литературы, 1962. 332 с.

[13] Лбов Г.С., Бериков В.Б., Герасимов М.К. Прогнозирование экстремальных гидрологических ситуаций на основе анализа многомерных временных рядов и экспертных знаний // Тр. междунар. науч. конф. "Экстремальные гидрологические события: Теория, моделирование и прогнозирование". М.: ИВП РАН, 2003. С. 26–30.

[14] Лбов Г.С., Герасимов М.К. Прогнозирование экстремальных ситуаций на основе совместного анализа временных рядов и экспертных высказываний // Научный вестник НГТУ. 2007. № 3(28). С. 13–24.

Поступила в редакцию 26 января 2010 г.,
с доработки — 23 апреля 2010 г.