УДК 519.2

Г. С. Лбов, Г. Л. Полякова

МЕТОД ПРОГНОЗИРОВАНИЯ В КЛАССЕ ЛОГИЧЕСКИХ РЕШАЮЩИХ ФУНКЦИЙ*

Рассматривается метод прогнозирования в классе логических решающих функций. Выбор наилучшего разбиения в пространстве переменных осуществляется на основе перебора, использующего идеологию «метода ветвей и границ». Эффективность предложенного метода иллюстрируется результатами решения прикладных задач.

Ключевые слова: анализ эмпирической информации, логические закономерности, прогнозирование.

В статье задача прогнозирования целевой переменной [1–3] решается в классе логических решающих функций. Особенностями указанной задачи являются небольшое число временных отсчетов, необходимость в «наглядности» получаемых результатов. Рассматривается метод прогнозирования в классе логических решающих функций. При этом выбор наилучшего разбиения в многомерном пространстве переменных осуществляется на основе перебора, использующего идеологию «метода ветвей и границ». Предлагаемый метод осуществляет полный перебор различных разбиений многомерного пространства переменных на заданное число прямоугольных областей. Значительное сокращение перебора всевозможных вариантов осуществляется путем выбора только тех вариантов, которые обеспечивают заданную статистическую надежность прогнозирования целевой переменной для любой прямоугольной области.

Отметим, что решение подобного рода задач связано с некоторыми особенностями:

1) отсутствие априорной информации о распределениях в пространстве переменных (характеристик);

2) в трудноформализуемых областях исследователь вынужден включать большое число потенциально полезных переменных (характеристик) из-за сложности изучаемого явления;

3) малое число наблюдений (объектов), как правило, сравнимое с числом переменных;

4) исследователя при изучении сложных объектов нередко интересует не только решение, дающее хорошее качество прогноза, но и сама форма представления такого решения для получения информации о внутренних причинно-следственных связях между характеристиками изучаемых объектов.

При решении статистических задач выявления закономерностей с указанными выше особенностями возникает ряд проблем.

Во-первых, задачи приходится решать в условиях отсутствия априорной информации о виде функций распределения. Любое предположение (например, о нормальности распределения, линейной регрессионной зависимости, независимости переменных, марковости процесса) ставит вопрос о соответствии выбранного предположения истинным ограничениям. Как вводить предположения и какие? В этом состоит первая проблема.

Во-вторых, в условиях малого числа наблюдений и высокой размерности пространства переменных возникает проблема статистической устойчивости получаемых решений. Из теоретических исследований вытекает, что более сложные функциональные зависимости используются для построения решений, больше переменных и меньше число наблюдений (объем выборки), тем больше вероятность получения решения, сильно отличающегося от оптимального. Суть проблемы устойчивости статистических решений заключается в следующем. С одной стороны, сильное ограничение на класс решений ставит вопрос об адекватности наших предположений истинному распределению: чем больше такое несоответствие, тем хуже решение. С другой стороны, чем более сложный класс функций используется при малом объеме выборки, тем выше вероятность получить недостоверное решение. Так, например, может оказаться, что линейная функция, заданная на всем множестве переменных, будет менее информативной по сравнению с линейной функцией, заданной на подмножестве этих переменных. Таким образом, при построении решения необходимо стремиться к максимальной сложности используемого класса решений (для ослабления ограничений на истинное распределение), но при этом сложность класса решений не должна превышать некоторого порога, задаваемого объемом выборки. При малом объеме выборки класс решений должен иметь малую меру сложности. При увеличении объема выборки этот класс должен позволять постепенно увеличивать свою сложность, вплоть до получения оптимального решения при произвольном распределении. Класс решений, обладающий таким свойством, будем называть универсальным классом. Вопрос о соотношении сложности используемого класса решений и объема выборки – наиболее важный и трудный в общих теоретических исследованиях, связанных с построением решений на основе ограниченной эмпирической информации.

По указанным выше причинам, а именно при отсутствии информации о виде распределения, линейной зависимости, наличия малого количества наблюдений (объектов) относительно числа переменных (характеристик), традиционные статистические методы оказываются малоподходящими при решении указанных задач.

* Работа выполнена при финансовой поддержке РФФИ (проект № 10-01-00113-а), проект № 09-07-12087-офи_м. Интеграционного гранта СО РАН (№ 83).

Как показывают теоретические и экспериментальные исследования [1–3], построение логико-вероятностных моделей изучаемых явлений на основе такой информации оказывается достаточно перспективным направлением. Модели представляют собой список логических закономерностей, обладающих достаточно высокой прогнозирующей способностью. Кроме того, в отличие от классических методов статистического анализа результаты в рамках указанной модели представляются на языке, близком к естественному языку логических суждений, что облегчает интерпретацию результатов.

Постулируется, что изучаемое явление характеризуется лишь небольшим числом закономерностей (несколько десятков). Для решения поставленной задачи был использован как алгоритм ТЕМР [3], так и его модификация. Оба алгоритма обнаруживают все закономерности с заданной прогнозирующей способностью. Ниже приводится описание алгоритма ТЕМР и его модификации.

Алгоритм обнаружения логических закономерностей. Как указывалось выше, при решении каждой из 18 задач выявления закономерностей между природными факторами и заболеваемостью клещевым энцефалитом имеем таблицу данных $T = \{x^i, y^i\}$, где $x^i = (x_1^i, ..., x_{12}^i)$, $i = 1, ..., 18$. Так как объем выборки ($N = 20$) является весьма малым, а размерность пространства переменных относительно велика (от 12 до 48), статистически надежные закономерности могут быть получены лишь при огрублении статистического материала.

В данном случае диапазон значений каждой переменной разбивается на ряд интервалов. Интервалы переменной *Y* назовем образами. Из исходной таблицы *T* создаем таблицу $v = \{z^i, u^i\}$, где *z* и *u* – соответствующие интервалы переменных *x* и *y*.

Для каждого сочетания переменных $z_1, ..., z_{12}$ перебираются все наборы интервалов рассматриваемого подмножества переменных (логические высказывания, в данном случае конъюнкции). Пример конъюнкции интервалов: $S = (z_5 = 2) \wedge (z_7 = 1) \wedge (z_{11} = 3)$. Естественно, ее можно записать через границы интервалов указанных переменных, что и делается при описании результатов изучения влияния природных факторов на заболеваемость клещевым энцефалитом.

Логические закономерности определяются отдельно для каждого образа-интервала *Y*. Для этого образ с номером s, $s = 1, ..., k$, назовем первым образом, а объединение всех остальных – вторым образом. Обозначим конъюнкцию через $S(a, E)$, где a – имя объекта (номер переменной); E – некоторая область многомерного пространства переменных. Если набор значений переменных для рассматриваемого объекта принадлежит области E, то будем говорить, что $S(a, E)$ для этого объекта принимает значение «истина», иначе – «ложно». Для любой конъюнкции $S(a, E)$ можно определить по таблице данных v число объектов первого образа $N(1, S)$ и число объектов второго образа $N(2, S)$, на которых указанная конъюнкция истинна.

Конъюнкцию $S(a, E)$ назовем логической закономерностью S^*, характеризующего первый образ, если выполняются неравенства

$$\frac{N(1, S)}{N(1)} \geq \delta, \quad \frac{N(2, S)}{N(2)} \leq \beta,$$

где δ и β – некоторые параметры, $0 \leq \beta < \delta \leq 1$. Чем больше δ и меньше β, тем сильнее логическая закономерность. Множество всех закономерностей обозначим через S^*. Конъюнкцию $S(a, E)$ называем потенциальной логической закономерностью для первого образа (обозначим ее через S'), если выполняются неравенства

$$N(1, S)N(1) \geq \delta, \quad N(2, S)N(2) > \beta.$$

Множество потенциальных закономерностей обозначим через S'. Очевидно, что из $S' \in S'$ можно, вообще говоря, получить закономерность S^* последовательным присоединением предикатов, т. е. $S' \wedge J(a, E_j) \wedge ....$ Если для некоторой конъюнкции $S(a, E)$ выполняется неравенство $N(1, S) / N(1) < \delta$, то конъюнкция S по определению не является закономерностью и присоединение к ней какого-либо предиката не даст закономерности (множество таких конъюнкций обозначим через *S*). Таким образом, любая конъюнкция $S(a, E)$ может быть трех типов: S^*, S' и S.

Назовем конъюнкцию $S(a, E) = J(a, E_{j_1}) \wedge ... \wedge J(a, E_{j_m})$ конъюнкцией длины m, где E_j область значений переменной Z_j.

Алгоритм обнаружения логических закономерностей состоит в последовательном выполнении следующих шагов.

Шаг 1. Рассматриваются всевозможные конъюнкции длины 1, т. е. конъюнкции вида $S(a, E) = J(a, E_j)$, $E_j \in W_j$, $j = 1, ..., n$. Если $S(a, E) \in S^*$, то она включается в список закономерностей и соответствующее подмножество E_j исключается из дальнейшего перебора; если $S(a, E) \in S'$, то соответствующее подмножество E_j оставляется для дальнейшего перебора; если $S(a, E) \in S$, то соответствующее подмножество E_j исключается из дальнейшего перебора. Обозначим через W_j^1 множество подмножеств E_j, оставленных для дальнейшего перебора после выполнения шага 1 алгоритма.

Шаг 2. Рассматриваются всевозможные конъюнкции длины 2, т. е. конъюнкции вида $S(a, E) = J(a, E_j) \wedge J(a, E_l)$, $j \neq l$; $E_j \in W_j^1$, $E_l \in W_l^1$. Если $S \in S^*$ или $S \in S'$, то соответствующие подмно-

жества E_j и E_l исключаются из дальнейшего перебора. Если $S \in S^*$, то соответствующая конъюнкция включается в список закономерностей.

Шаг 3. Рассматриваются всевозможные конъюнкции длины 3, т. е. конъюнкции вида $S(a,E) = J(a,E_j) \wedge J(a,E_l) \wedge J(a,E_m), \quad j \neq l \neq m, \; j \neq m; \; E_j \in W_j^2, \; E_l \in W_l^2, \; E_m \in W_m^2$, где W_j^2 содержит подмножества E_j, оставленные для перебора после шага 2 алгоритма. Далее аналогично рассматриваются конъюнкции длины 4, 5 и т. д.

Предполагается, что реальные таблицы данных таковы, что число элементов множества W_j^m резко уменьшается с увеличением числа шагов m, а также, что уже при небольшом числе шагов происходит останов программы. Из схемы алгоритма следует, что с его помощью обнаруживаются все логические закономерности, характеризующие s-й образ. Последовательно применяя данный алгоритм для каждого из k образов, получаем k списков логических закономерностей: $\{S_1^1, ..., S_{d_1}^1\}, ..., \{S_1^k, ..., S_{d_k}^k\}$. Сделаем сквозную нумерацию полученных закономерностей: $S_1, ..., S_d$, где $d = \sum_{s=1}^{k} d_s$. Заметим, что при малых значениях β ($\beta \simeq 0$) и больших значениях δ ($\delta \simeq 1$) логических закономерностей может не быть; при фиксированном малом β (например, $\beta = 0{,}05$) при уменьшении δ ($\delta \to \beta$) число закономерностей может резко возрастать, а их качество падать. Поэтому при фиксированном значении β необходимо выбрать такую величину δ, чтобы число закономерностей было небольшим (например, десять закономерностей на каждый образ). Этой цели можно достигнуть последовательно уменьшая δ с некоторым шагом $\Delta\delta$. По критерию $F = \delta - \beta$ можно упорядочить полученные закономерности по их вероятностной прогнозирующей способности.

Особенность модифицированного алгоритма ТЕМР состояла в переборе всевозможных конъюнкций сразу заданной длины, указанной в параметрах алгоритма, и нового критерия качества дерева решений.

Метод решения задачи прогнозирования. Рассмотрим задачу прогнозирования [3–6] для каждой переменной Y_l, l = 1, …, m, отдельно, т. е. определим решающую функцию F набором функций $\{\overline{f}_1, ..., \overline{f}_l, ..., \overline{f}_m\}$, где $\overline{f}_l$ – отображение из B в D_{Y_l}. Построение функции $\overline{f}_l$ осуществляется в два этапа [4–6].

Этап I. На основе временного ряда q для фиксированного номера d, $d \in \{1, …, R\}$, и фиксированного номера l организуется обучающая выборка в виде таблицы данных $v_d = \{x^{\mu-d}, y_l^{\mu}\}$, $\mu = R + 1, …, N$. По таблице данных v_d строим логическую решающую функцию $<\alpha_d, r(\alpha_d)>$ с разбиением $\alpha_d = \{E_d^1, E_d^t, ..., E_d^{M_d}\}$ множества $D = \prod_{j=1}^{n} D_{X_j}$ на M_d подмножеств (с помощью алгоритма обнаружения логических закономерностей). Введем новую переменную Z_d с множеством значений $D_{Z_d} = \{1, ..., t, ..., M_d\}$ следующим образом: если предыстория $(x_1^{\mu-d}, ..., x_n^{\mu-d}) \in E_d^t$, то $z_d = t$. Указанные выше вычисления проводятся для всех номеров предысторий (d = 1, …, R). Результатом первого этапа является набор разбиений (α, …, α_d, …, α_R) и соответствующий ему набор переменных (Z_1, …, Z_d, …, Z_R). Это позволяет для любого момента времени t_μ, $\mu = R + 1, …, N$, n-мерную предысторию длины R, представленную набором ($x^{\mu-R}$, …, $x^{\mu-d}$, …, $x^{\mu-1}$), $x^{\mu-d} \in D$, задать в виде одномерной последовательности $z^{\mu} = (z_R^{\mu}, ..., z_d^{\mu}, ..., z_1^{\mu})$, $z_d^{\mu} \in D_{Z_d}$.

Этап II. Временному n-мерному ряду $q = \{x_j^v\}$, j = 1, …, n, v = 1, …, N ставится в соответствие таблица данных $v = \{z^{\mu}, y_l^{\mu}\}$, $\mu = R + 1, …, N$. По таблице v строится логическая решающая функция $\overline{f}_l$, которая позволяет прогнозировать значение переменной Y_l по предыстории длины R, l = 1, …, n. Заметим, что задача прогноза для любого набора $\{Y_{i_1}, ..., Y_{i_m}\} \subseteq \{X_1, ..., X_n\}$, m = 1, …, n решается аналогично.

Из-за сложности статистической задачи (многомерность, нестационарность, малое число наблюдений) возникла необходимость в «огрублении» статистической информации (использовалось небольшое число интервалов значений переменных, включая и показатель заболеваемости).

Совместный анализ климатических и астрофизических факторов в природных очагах клещевого энцефалита проводился на основе следующей статистической информации: температура воздуха (ТВ), относительная влажность воздуха (ОВВ) в приземном слое; количество осадков; данные солнечной радиации: прямой радиации на горизонтальную поверхность, прямой радиации на перпендикулярную поверхность, отраженной радиации; заболеваемость людей клещевым энцефалитом. Исследование влияния взаимозависимых природных факторов на заболеваемость людей клещевым энцефалитом заключалось в совместной статистической обработке 12 временных рядов. Каждый временной ряд представлял собой набор значений среднемесячных показателей перечисленных 6 природных факторов и годовых значений заболеваемости людей в течение 20 лет (с 1991 по 2010 гг.). Таким образом, каждому природному фактору соответствовала таблица из 12 столбцов (переменные X_1, …, X_{12}) и 20 строк (наблюдения или объекты). Для каждой строки указывается количество заболевших за год (переменная Y). Требовалось найти статистические закономерности (взаимосвязи) между всеми X_1, …, X_{12} и Y. В силу того, что природных фак-

торов – 6, а рассматриваемых регионов – 3, возникает 18 задач выявления закономерностей, решение которых и представлено в данной работе.

При решении задачи алгоритмом ТЕМП [4; 5] значения показателя заболеваемости были разделены на 2 диапазона: 1 диапазон соответствовал низкому уровню заболеваемости от 2 до 17 человек на 100 тысяч населения, 2 диапазон – высокому уровню заболеваемости от 17 до 62 заболевших на 100 тысяч населения. Это дало возможность обнаружить логические закономерности, как правило, включая конъюнкции значений не более трех переменных для прогноза уровня заболеваемости. При решении данной задачи модифицированным алгоритмом ТЕМП [6] количество образов (диапазонов значений целевой переменной) задавалось равным k, где $k > 2$.

Приведем некоторые из полученных закономерностей.

Высокий уровень заболеваемости КЭ возникает при совместном (одновременном) выполнении условий: «ТВ в январе – от –20 до –11°С» и «ТВ в августе – от 15 до 18,7 °С» и «ТВ в октябре – от 2,1 до 5,6 °С».

Высокий уровень заболеваемости КЭ определяется совместным выполнением условий: «ОВВ в феврале – от 72,5 до 83,5 %» и «ОВВ в апреле – от 58 до 73 %» и «ОВВ в декабре – от 79,8 до 85 %».

Полученная нами оценка заболеваемости КЭ в НСО в 2009 г. имела значение 7 человек на 100 тысяч населения, реальная заболеваемость составила 6,5 человек на 100 тысяч населения.

Предложен метод прогнозирования целевой переменной на основе обнаружения логических закономерностей на обучающей выборке, который включает перебор различных разбиений многомерного пространства переменных на заданное число прямоугольных областей. Значительное сокращение перебора всевозможных вариантов осуществляется путем выбора только тех вариантов, которые обеспечивают заданную статистическую надежность прогнозирования целевой переменной для любой прямоугольной области.

Эффективность предложенного метода иллюстрируется результатами решения прикладной задачи, состоящей в исследовании влияния природных факторов на заболеваемость клещевым энцефалитом населения Новосибирской, Иркутской областей и Республики Горный Алтай.

Библиографические ссылки

1. Лбов Г. С. Метод анализа многомерных разнотипных временных рядов в классе логических решающих функций // Доклад РАН. 1994. Т. 339. № 6.

2. Лбов Г. С., Методы обработки разнотипных экспериментальных данных. Новосибирск : Наука, 1981.

3. Лбов Г. С., Старцева Н. Г. Логические решающие функции и вопросы статистической устойчивости решений. Новосибирск : Изд-во Ин-та математики, 1999.

4. Исследование влияния природных факторов на заболеваемость клещевым энцефалитом / Г. С. Лбов, Г. Л. Полякова, В. Н. Бахвалова [и др.] // Вестник Новосиб. гос. ун-та. 2010. № 3. С. 31–37. (Серия: Биология. Клиника Медицина).

5. Лбов Г. С., Полякова Г. Л., Пестунов И. А. Метод прогнозирования на основе анализа коротких временных рядов // AIS-IT'10 : тр. Конгресса по интеллектуальным системам и информац. технол. М. : Физматлит, 2010. Т. 1. С. 264–271.

6. Лбов Г. С., Полякова Г. Л. Решение задачи прогнозирования в классе логических решающих функций // Математическая биология и биоинформатика : докл. III Междунар. науч. конф. (10–15 окт. 2010, Пущино). М. С. 226–227.

G. S. Lbov, G. L. Polyakova

FORECASTING METHOD IN CLASS CLASS OF LOGICAL DECISION FUNCTIONS

In this paper we suggest a method using the class of logical decision functions. A search of optimal division of variable space is based on sort out similar to branch and bound method. Results of statistical analysis are represented in the form of logical rules reflected cause-effect relations of object under investigation. Effectiveness of the suggested method is shown by solving applied problems in the sphere of ecology.

Keywords: analyses of empirical information, logical rules, forecasting