Анализ совокупности разнотипных временных рядов с использованием логических решающих функций*

В. Б. Бериков[1], И. А. Пестунов[2], М. К. Герасимов[1]
[1]*Институт математики им. С.Л. Соболева СО РАН, Новосибирск, Россия*
[2]*Институт вычислительных технологий СО РАН, Новосибирск, Россия*
e-mail: berikov@math.nsc.ru, pestunov@ict.nsc.ru, max_post@ngs.ru

Разработаны алгоритмы анализа совокупности многомерных разнотипных временных рядов, основанные на логических решающих функциях (деревьях решений). Предложен алгоритм построения коллективного решения по набору различных деревьев. Проведено исследование эффективности метода на прикладной задаче анализа влияния природных факторов на заболеваемость клещевым энцефалитом.

Ключевые слова: многомерные разнотипные временные ряды, логические решающие функции, коллективное решение, клещевой энцефалит.

Введение

В настоящее время большую актуальность имеют задачи анализа и прогнозирования динамически меняющихся характеристик множества объектов, известные в литературе как задачи обработки панельных данных [1, 2] или задачи анализа совокупности (ансамбля) многомерных временных рядов [3]. Для решения такого рода задач существуют методы, основанные на использовании различных видов моделей панельных данных и сведении исходной задачи к задаче оценивания параметров регрессионной модели. Модели позволяют агрегировать данные по всем объектам и получать надежные решения, даже при сравнительно небольшом объёме наблюдений по каждому из объектов.

Для обоснования полученных решений применяются различные предположения о вероятностном распределении исследуемых характеристик (как правило, о его нормальности). Выбор оптимальной сложности регрессионной модели, обеспечивающей разумный компромисс между ее сложностью и точностью на обучающей выборке ограниченного объёма, является пока нерешенной проблемой.

В представленной работе решается задача построения моделей временных рядов при условии, что длина рядов является малой (сравнимой с числом используемых характеристик). В этом случае проверка гипотез о согласии постулируемого распределения с имеющимися данными является затруднительной. Кроме того, предполагается, что интервалы наблюдений за различными объектами могут не полностью совпадать, что осложняет применение трендовых моделей. Наконец, считается, что объекты могут описываться как количественными, так и качественными характеристиками. В этом случае применение традиционных подходов [1] (введение dummy-переменных, рассмотрение

*Работа выполнена при финансовой поддержке РФФИ (гранты № 11-07-00346а, 10-01-00113а и 11-07-12083-офи-м-2011).

пробит-, логит-моделей и т. п.), особенно при большом числе значений качественных переменных, становится затруднительным.

Для решения рассматриваемой задачи предлагается использовать класс логических решающих функций (ЛРФ, решающих деревьев) распознавания динамических объектов по предыстории [4]. Применение данного класса функций позволяет получить хорошо интерпретируемую логическую модель, характеризующую общие закономерности поведения анализируемых объектов, даёт возможность рассматривать случай разнотипных переменных, когда среди исходного набора имеются как количественные, так и качественные характеристики, и сочетает другие особенности, присущие классу ЛРФ: отсутствие предположений о параметрической модели распределения, автоматическое выделение наиболее информативных переменных, прогнозирование целевого показателя по подмножеству наиболее информативных для каждого рассматриваемого случая характеристик, что повышает надёжность решения.

Для построения решения применяется коллективный (ансамблевый) подход, основанный на том, что для повышения надёжности прогнозирования можно привлекать не одну, а несколько решающих функций. Эти решающие функции могут быть получены различными методами (или одним методом, но с использованием различных параметров), по разным выборкам наблюдений, а также по различным подсистемам переменных. Такое разностороннее рассмотрение задачи, как правило, приводит к улучшению качества прогнозирования и к бóльшему пониманию закономерностей исследуемого явления.

Работа имеет следующую структуру. В первом разделе даётся формальная постановка задачи, во втором описывается алгоритм построения ЛРФ для прогнозирования характеристик динамических объектов, в третьем приводится алгоритм построения коллективного решения на основе различных вариантов ЛРФ, в четвертом предложен алгоритм построения ЛРФ, использующий кусочно-линейную регрессию, в пятом проводится исследование разработанных алгоритмов на примере прикладной задачи анализа влияния природных факторов на заболеваемость клещевым энцефалитом в нескольких эндемичных территориях России.

1. Постановка задачи

Предположим, что изучается некоторый набор объектов $A = \{a_1, ..., a_N\}$. Для описания свойств объектов используется общий набор характеристик (переменных) $X_1(t)$, ..., $X_n(t)$, зависящих от времени. Пусть D_j — область определения характеристики X_j, $j = 1, ..., n$. Характеристики могут быть как количественными, так и качественными. Для количественной характеристики X_j область ее определения является интервалом числовой оси: $D_j \subset R$. Для качественной характеристики $D_j = \{u_{j,1}, ..., u_{j,k_j}\}$ есть множество некоторых значений (имен). Качественные характеристики будем разделять на порядковые (на множестве значений характеристики задан некоторый порядок), номинальные (множество значений характеристики неупорядоченно) и булевы. Обозначим $D = D_1 \times ... \times D_n$.

Пусть переменные измеряются в последовательные моменты времени $t_1, ..., t_m, ..., t_M$. Для определенности будем предполагать, что измерения проводятся через равные временные интервалы. Для каждого объекта a_i существуют начальный и конечный моменты измерений t_{m_i}, $t_{M_i} \in \{t_1, ..., t_M\}$. Таким образом, имеем набор из N многомерных разнотипных временных рядов размерности n и различной длины $T_i = M_i - m_i + 1$,

$i = 1, ..., N$. Значение характеристики X_j для объекта a_i в момент времени t_m обозначим как $x_{i,j}(t_m)$, $m = m_i, ..., M_i$, $i = 1, ..., N$, $j = 1, ..., n$. Вектор $(x_{i,1}(t_m), ..., x_{i,n}(t_m))$ обозначим через $x_i(t_m)$.

Пусть дополнительно определена общая для всех объектов прогнозируемая случайная характеристика Y. В зависимости от типа Y будем рассматривать различные типы задач прогнозирования:

1) Y — номинальная характеристика: $Y \in D_Y = \{\omega_1, ..., \omega_K\}$, где K — число классов (образов). Задачу данного типа по аналогии с обычной задачей распознавания образов назовём задачей распознавания по набору многомерных временных рядов;

2) Y — количественная характеристика. В этом случае имеем задачу прогнозирования количественной характеристики по совокупности многомерных временных рядов (т. е. задачу регрессионного анализа).

Пусть задана некоторая величина L — лаг, или "глубина предыстории", причём $1 \le L << \min T_i$. Рассмотрим произвольный момент времени t_m и набор предыдущих моментов времени $t_{m-1}, t_{m-2}, \ldots, t_{m-L}$. Для произвольного объекта a_i, $i = 1, ..., N$, и момента времени t_m, $m = m_i + L, ..., M_i$, сформируем таблицу предысторий $v_{i,m} = \begin{pmatrix} x_i(t_{m-1}) \\ \vdots \\ x_i(t_{m-L}) \end{pmatrix}$, состоящую из наблюдаемых значений характеристик за L предшествующих моментов времени (размерность таблицы — L строк и n столбцов).

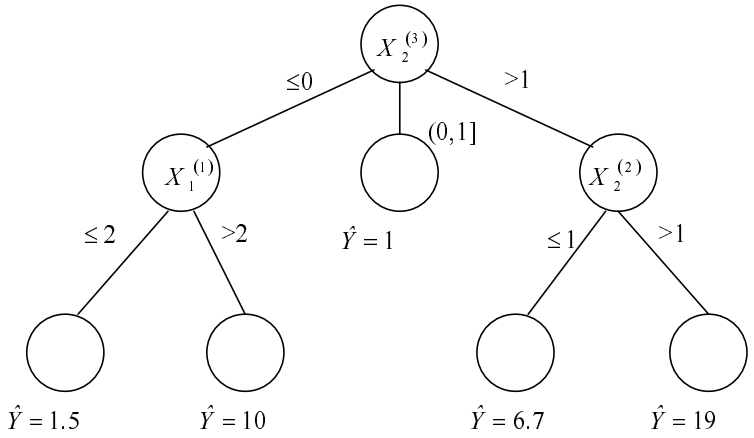
Обозначим через $Y_{i,m}$ случайную величину — возможное значение переменной Y для объекта a_i в момент времени t_m. Предполагается, что $Y_{i,m}$ зависит от поведения объекта в прошлом.

Будем полагать, что при заданных предыдущих значениях $X_1(t), \ldots, X_n(t)$ условное распределение $Y_{i,m}$ для объекта a_i зависит только от значений этих рядов в L предыдущих моментов времени (от таблицы предысторий $v_{i,m}$). Кроме того, предположим, что данная зависимость одна и та же для всех возможных значений m и для всех объектов. Это означает, что статистические свойства рядов, определяющие зависимость, неизменны во времени. В дальнейшем, имея в виду указанное свойство, будем пропускать индекс, соответствующий моменту времени t_m.

Обозначим характеристики с учётом предыстории как $X_j^{(l)}$, что означает: характеристика X_j в l-й предыдущий момент времени относительно текущего момента. Пусть $X^{(l)} = (X_1^{(l)}, \ldots, X_n^{(l)})$. Обозначим через $\tilde{X} = (X^{(1)}, \ldots, X^{(L)})$ предысторию длины L относительно текущего момента времени. Пусть $x_j^{(l)}$ — значение переменной X_j в l-й предыдущий момент. Обозначим через $v = (x_j^{(1)}, ..., x_j^{(L)})$, $j = 1, ..., n$, таблицу предысторий относительно текущего момента.

Пусть Y — качественная характеристика. Рассмотрим решающую функцию f, которая произвольной таблице предысторий $v \in D^L$ для объекта a_i сопоставляет решение — прогнозируемый класс $\hat{\omega} = f(v, i)$. Для этой решающей функции можно определить критерий качества как риск неправильного распознавания (вероятность ошибки) $Q_f = P\left(Y_i \ne f(v, i)\right)$.

Пусть Y — количественная характеристика, f — некоторая решающая функция, по которой строится решение — прогноз для объекта a_i: $\hat{Y} = f(v, i)$. Критерий качества (функцию риска) в данном случае можно определить как математическое ожидание квадрата ошибки $Q_f = E\left(Y_i - f(v, i)\right)^2$.



Пример дерева решений

Для вычисления критерия качества решающей функции необходимо знать вероятностное распределение характеристик. Однако это распределение, как правило, неизвестно. Поэтому будем использовать эмпирический критерий — найденную по наблюдениям частотную оценку вероятности ошибки или выборочную дисперсию.

Пусть задан некоторый класс Ф решающих функций. Требуется выбрать из указанного класса функцию, оптимальную по заданному критерию, т. е. построить общую для всех объектов модель зависимости характеристики Y от таблицы предысторий. Модель должна позволять получать прогнозируемое значение $\hat{Y}$ характеристики Y в будущий момент времени по значениям остальных характеристик за L прошлых моментов. Примером аналогичной модели, используемой при эконометрическом анализе панельных данных, является линейная модель с распределённым лагом и фиксированными индивидуальными эффектами [1].

Заметим, что для каждого объекта в принципе можно строить индивидуальную модель зависимости. Однако при этом не учитываются общие закономерности, которые, как предполагается, характерны для всех изучаемых объектов.

В качестве класса решающих функций будем рассматривать класс логических решающих функций от разнотипных переменных (класс деревьев решений, для задачи регрессионного анализа называемых также деревьями регрессии; см., например, [3]). В этом случае функцию f можно представить в виде некоторого дерева решений (см. рисунок), имеющего следующую особенность: в его вершинах проверяются высказывания относительно некоторых переменных X_j в определенный l-й отсчет времени назад (относительно текущего момента). Цепочка проверяемых высказываний ведет из корня дерева (на рисунке корню соответствует верхняя вершина) в терминальную вершину, которой приписано прогнозируемое значение Y: для качественной целевой характеристики Y — прогнозируемый класс $\omega \in \{\omega_1, ..., \omega_K\}$, для количественной характеристики Y — значение $Y = c = \text{const}$. Будем называть указанный класс классом деревьев решений для прогнозирования по предыстории.

2. Алгоритм построения логической решающей функции

При прогнозировании по предыстории качество любого заданного дерева решений f можно оценить следующим образом. Обозначим через $y_{i,m}$ и $\hat{y}_{i,m}$ соответственно истин-

ное и прогнозируемое значения Y для объекта a_i и момента времени t_m. Тогда зададим критерий качества как эмпирический риск

$$\hat{Q}_f = \frac{1}{W}\sum_{i=1}^{N}\sum_{m=m_i+L}^{M}\delta(y_{i,m},\hat{y}_{i,m}),$$

где для задачи распознавания динамического объекта

$$\delta(y_{i,m},\hat{y}_{i,m}) = \begin{cases} 0, & \text{если } \ y_{i,m}=\hat{y}_{i,m}, \\ 1, & \text{иначе,} \end{cases}$$

для задачи прогнозирования количественной характеристики

$$\delta(y_{i,m},\hat{y}_{i,m}) = \left(y_{i,m}-\hat{y}_{i,m}\right)^2,$$

$W = \sum\limits_{i=1}^{N} T_i - N\,L$ — количество суммируемых величин, $m = m_i + L, ..., M_i$, $i = 1, ..., N$. Величина $\hat{Q}_f$ есть частота несовпадений прогноза и реального значения Y(для задачи распознавания) либо средняя квадратическая погрешность прогноза (для задачи регрессионного анализа). Лучшему варианту дерева соответствует меньшее значение критерия.

Сформируем множество всех наблюдаемых таблиц предысторий: $V = \{v_{i,m}\}$, $m = m_i + L, ..., M_i,\ i = 1, ..., N$. Исходной информацией для построения дерева решений для прогнозирования по предыстории является набор таблиц $V = \{v_{i,m}\}$ с указанными для каждой таблицы значениями прогнозируемой характеристики $y_{i,m}$, $m = m_i + L, ..., M_i$, $i = 1, ..., N$. Будем говорить, что указанные таблицы используются для "обучения" прогнозированию.

Таким образом, возникает задача обработки четырёхмерной таблицы данных (объект — переменная — предыстория — время). Однако имеющиеся алгоритмы распознавания или регрессионного анализа с применением деревьев решений используют в качестве входной информации двумерные таблицы. Ниже описывается способ представления данных, при котором в зависимости от типа прогнозируемой характеристики Y исходная задача сводится к задаче распознавания образов или регрессионного анализа.

Для унификации записи введём дополнительную качественную переменную X_{ind}, кодирующую каждый из объектов и необходимую для учёта решающей функцией индивидуальных особенностей объектов. Эта переменная не изменяется с течением времени.

Рассмотрим следующий алгоритм построения ЛРФ для прогнозирования по предыстории.

1. Каждую предысторию $v_{i,m}$, где $i = 1, ..., N,\ m = m_i + L, ..., M_i$, "вытянем" в строку $s_{i,m}$, дополненную значениями индивидуальной и целевой переменных:

$$s_{i,m} = \left(x_{i,1}(t_{m-1}), ..., x_{i,n}(t_{m-1}), ..., x_{i,1}(t_{m-L}), ..., x_{i,n}(t_{m-L}), x_{i,ind}, y_{i,m}\right)$$

(длина строки равна $n\,L + 2$).

2. Из полученных строк сформируем таблицу данных размерности $W \times (n \cdot L + 2)$:

$$B = \begin{pmatrix} s_{1,m_1+L} \\ \vdots \\ s_{1,M_1} \\ \vdots \\ s_{N,m_N+L} \\ \vdots \\ s_{N,M_N} \end{pmatrix}.$$

3. Для двумерной таблицы B с помощью произвольного алгоритма построения дерева решений сформируем ЛРФ.

В настоящей работе для реализации шага 3 применялся рекурсивный R-алгоритм построения дерева решений [4], основанный на методе динамического программирования и позволяющий находить сложные закономерности путём увеличения глубины перебора сочетаний переменных. В этом алгоритме используется модифицированный критерий качества

$$Q_f^{'} = Q_f + \alpha \frac{H}{W},$$

где H — число терминальных вершин дерева, α — параметр регуляризации, призванный снизить эффект "переобучения". Детальное описание алгоритма дано в [4].

Известно, что эмпирический риск может сильно отличаться от "истинного" неизвестного риска. Поэтому для более точного оценивания качества предпочтительно использовать оценку риска, найденную по контрольной выборке наблюдений.

Пусть для каждого объекта a_i, $i = 1, ..., N$, имеются ряды наблюдений $X_1(t), ..., X_n(t)$, $Y(t)$ в моменты времени t_m, $m = M_i + 1, ..., M_i + T_i^c$. Будем считать, что данные ряды используются для контроля качества прогнозирования. Аналогично тому, как это делалось в вышеописанном алгоритме, по соответствующим предысториям длины L сформируем контрольную таблицу данных B^c. Тогда можно сравнить полученные с использованием построенного дерева прогнозируемые по таблице значения $\hat{y}_{i,m}$ с наблюдаемыми значениями $y_{i,m}$ и определить оценку риска по контрольным рядам:

$$Q_f^c = \frac{1}{W_c} \sum_{i=1}^{N} \sum_{m=M_i+L}^{M_i+T_i^c} \delta(y_{i,m}, \hat{y}_{i,m}),$$

где $W_c = \sum\limits_{i=1}^{N} T_i^c - N\,L$.

Заметим, что контрольную таблицу можно формировать не только по контрольным рядам, но и путём случайного отбора некоторых строк таблицы B. Эти строки не используются при построении дерева решений, а участвуют только на этапе контроля. Аналогичным образом для оценивания качества построенных решающих функций можно применять методы скользящего экзамена и кросс-валидации.

3. Построение ансамбля деревьев решений

Для повышения качества алгоритмов распознавания и прогнозирования широко применяется коллективный (ансамблевый) подход (см., например, [5]). При использовании

этого подхода комбинируются результаты, полученные различными алгоритмами или одним алгоритмом, но с разными параметрами настройки. Итоговое коллективное решение принимается с помощью процедуры голосования, в которой вклады различных вариантов решений могут зависеть от степени "компетентности" алгоритмов, оцениваемой по наблюдениям. Этот подход был использован для анализа многомерных разнотипных временных рядов на основе деревьев решений.

Пусть имеется набор деревьев решений для прогнозирования по предыстории, сформированных описанным выше алгоритмом. Рассмотрим произвольную таблицу предысторий v. Каждое дерево даёт свой прогноз для v (при этом также могут учитываться индивидуальные свойства объектов). Общее (коллективное) решение можно определить с помощью методов "голосования" (задача распознавания) или "усреднения" (задача регрессионного анализа). При использовании метода "голосования" наблюдению приписывается тот класс, которому отдаёт предпочтение большинство деревьев из набора. В случае задачи регрессионного анализа прогнозируемое значение получается усреднением прогнозов по всем деревьям регрессии. Кроме простого усреднения или голосования с равными вкладами каждого "голосующего", более целесообразно применять процедуру, в которой учитывается оценка риска, соответствующего каждому дереву: чем меньше риск, тем больший вес имеет соответствующий голос.

Для оценивания качества построенного коллективного решения используются способы, аналогичные описанным в предыдущем разделе.

В настоящее время существуют следующие основные способы формирования ансамблей деревьев решений [4]:

— по разным подсистемам переменных;

— на основе различных подвыборок (алгоритмы бэггинга, бустинга [6] и т. п.).

Поскольку в решаемой задаче предполагается, что количество наблюдений временных рядов не очень велико, а число переменных, напротив, значительно, то был выбран первый принцип. В данной работе использовался алгоритм, основанный на последовательном исключении характеристик. Алгоритм состоит из нескольких этапов.

1. На первом этапе используются все имеющиеся характеристики. Характеристика, которая соответствует корню построенного дерева, является наиболее информативной, поскольку, во-первых, именно от неё зависит дальнейшее движение по дереву, во-вторых, при ветвлении корневой вершины применяется полный набор наблюдений.

2. При построении второго дерева используются все характеристики, кроме отобранной на предыдущем этапе. Это делается с целью получить вариант дерева, наиболее радикально отличающийся от предыдущего. Известно, что для достижения наилучшего качества решающие функции, входящие в ансамбль, должны быть как можно более "разнообразными", т. е. каждому варианту решающего правила должна отвечать своя "область компетенции" — некоторая подобласть пространства переменных, для которой ошибка прогнозирования минимальна.

3. На следующих этапах последовательно исключаются характеристики, соответствующие корневым вершинам уже построенных деревьев. Поскольку исключаются наиболее информативные характеристики, качество деревьев, как правило, от этапа к этапу может лишь ухудшаться.

Алгоритм прекращает работу как только будет построено заданное число деревьев либо показатель качества достигнет заданной минимально допустимой величины.

Кроме решения задачи прогнозирования, ансамбль деревьев решений может быть использован для оценивания важности анализируемых факторов. Для этого перемен-

ной, отобранной для ветвления при построении каждого дерева, присваивается вес, который пропорционален уменьшению ошибки прогнозирования для вершины, содержащей эту переменную. После построения ансамбля веса переменных усредняются по всем вариантам решений.

4. Построение кусочно-линейной регрессии

Если целевая характеристика Y — количественная, то дерево решений даёт достаточно грубую кусочно-постоянную модель регрессионной зависимости. Для повышения степени соответствия модели данным в терминальных вершинах дерева можно использовать линейные регрессии Y по набору $\tilde{X} = (X^{(1)}, \ldots, X^{(L)})$ и переменной X_{ind}. В этом случае логическая решающая функция представляет собой кусочно-линейную регрессию.

Пусть имеются количественные переменные X_1, X_2, а также переменная X_{ind}, принимающая значения из множества $\{C, \bar{C}\}$. В этом случае пример логической решающей функции:

если $X_{ind} = C$ и $X_1^{(2)} \le a$, то $Y = b_1 X_1^{(1)} + c_1$;

если $X_{ind} = C$ и $X_1^{(2)} > a$, то $Y = b_2 X_1^{(1)} + b_3 X_1^{(2)} + c_2$;

если $X_{ind} = \bar{C}$, то $Y = b_4 X_1^{(1)} + b_5 X_2^{(1)} + c_3$,

где $a, b_1, b_2, b_3, b_4, b_5, c_1, c_2, c_3$ — некоторые константы.

Для построения логической решающей функции используется следующий алгоритм последовательного ветвления вершин дерева.

В каждой конечной вершине $u^{(h)}$, $h = 1, ..., H$, определены решение — линейная регрессия $f^{(h)}$ и значение величины

$$R^{(h)} = \sum_{i|b_{i,*} \in B^{(h)}} (y_i - f^{(h)}(\tilde{x}_i, x_{i,ind}))^2,$$

где $b_{i,*}$ — строка таблицы данных B под номером i, которая содержит строку $\tilde{x}_i$ предысторий; $B^{(h)}$ — подтаблица предысторий (строк таблицы B), соответствующих вершине $u^{(h)}$. Для каждой вершины $u^{(h)}$ рассматриваются возможные разбиения таблицы $B^{(h)}$ по характеристикам $X_j^{(l)}$, $j = 1, \ldots, n$, $l = 1, \ldots, L$. Для качественных характеристик $X_j^{(l)}$ рассматриваются разбиения столбца под номером $j + (l-1)n$ таблицы $B^{(h)}$ на подмножества, для количественных характеристик $X_j^{(l)}$ — на интервалы. Границы интервалов определяются на основе выборки. Например, их можно задать на равном расстоянии от соседних выборочных значений характеристики $X_j^{(l)}$.

В зависимости от выбранного разбиения вершине $u^{(h)}$ ставятся в соответствие вершины $u^{(h_q)}$, $q = 2, 3, \ldots$, в которых с помощью стандартного метода наименьших квадратов строятся линейные регрессии $f^{(h_q)}$ (при этом учёт качественных характеристик проводится с помощью методики их сведения к dummy-переменным [1]). Далее можно определить величины

$$R^{(h_q)} = \sum_{i|b_{i,*} \in B^{(h_q)}} (y_i - f^{(h)}(\tilde{x}_i, x_{i,ind}))^2,$$

где $B^{(h_q)}$ — подтаблицы таблицы $B^{(h)}$, соответствующие вершинам $u^{(h_q)}$.

Качество разбиения для вершины $u^{(h_q)}$ определяется величиной $\Delta R^{(h)} = R^{(h)} - \sum\limits_q R^{(h_q)}$: чем больше $\Delta R^{(h)}$, тем выше качество разбиения. С помощью перебора всех конечных вершин дерева выбирается наилучшее ветвление.

Алгоритм прекращает работу при достижении заданных уровня ошибки и числа вершин, невозможности дальнейшего ветвления дерева или отсутствии при ветвлении существенного улучшения качества.

Данный алгоритм реализует пошаговую процедуру поиска наилучшего дерева решений. Отметим, что эта процедура не гарантирует нахождения оптимального дерева решений.

5. Анализ влияния природных факторов на заболеваемость клещевым энцефалитом

Разработанная методика была применена для изучения влияния природных факторов на заболеваемость клещевым энцефалитом (КЭ). Известно, что динамика этой заболеваемости носит циклический характер и сильно отличается для различных эндемичных регионов, что указывает на влияние природных и антропогенных факторов, специфичных для данных территорий [7]. Существуют также общие закономерности динамики для различных регионов, определяемые схожими механизмами функционирования паразитарной системы КЭ. На заболеваемость влияет весьма много факторов (как природных, так и антропогенных, в том числе профилактических, мероприятий). Не по всем из указанных факторов ведется регистрация данных. В то же время наиболее доступными являются метеорологические данные, а также наблюдения за активностью Солнца.

Для выявления общих закономерностей использовались ряды данных, описывающие динамику заболеваемости и метеорологических наблюдений в Новосибирске (1991 – 2010 гг.), Иркутске (1990 – 2010 гг.) и Горно-Алтайске (1995 – 2010 гг.). Таким образом, имеется $N = 3$ объекта (a_1 — “Новосибирск”, a_2 — “Иркутск”, a_3 — “Горно-Алтайск”), описываемые набором следующих переменных:

Y — относительный показатель заболеваемости КЭ (число заболевших на 100 тыс. населения);

$X_1, \ldots, X_{12}$ — среднемесячная температура воздуха (с октября предыдущего года по сентябрь текущего);

$X_{13}, \ldots, X_{24}$ — среднемесячная относительная влажность воздуха в соответствующие месяцы;

$X_{25}, \ldots, X_{36}$ — среднемесячное количество осадков.

Кроме того, использовались данные об активности Солнца: X_{37} — среднегодовой показатель индекса пятнообразования (число Вольфа).

Длина анализируемых рядов $T_1 = 20$, $T_2 = 21$, $T_3 = 16$. Из-за сложности задачи применялось “огрубление” информации: интервал изменения заболеваемости был разделен на три диапазона: до 10 чел./100 тыс. (низкая заболеваемость); от 10 до 20 чел./100 тыс. (средняя заболеваемость); выше 20 чел./100 тыс. (высокая заболеваемость). Таким образом, решается задача распознавания по набору динамических объектов. Значения ряда заболеваемости для каждого из городов за последние два года использовались как контрольные.

В результате применения разработанного алгоритма был построен ансамбль деревьев решений для прогнозирования по предыстории, для которого ошибка на контроле составила 0.167, т. е. одно значение ряда из шести было предсказано неверно. Проведено сравнение с существующим бустинг-алгоритмом построения ансамбля деревьев

решений [6], для которого исходные панельные данные были преобразованы в таблицу в соответствии с алгоритмом, описанным в разделе 2. Для построенного ансамбля ошибка составила 0.33 (т. е. два значения из шести были предсказаны неверно). Таким образом, на контрольных данных предложенный алгоритм оказался более точным.

Список наиболее важных факторов, составленный по результатам работы алгоритма построения ансамбля, содержит следующие переменные: средняя температура августа, среднее количество осадков в апреле, принадлежность к Горно-Алтайску, средняя влажность воздуха в марте, а также средняя температура в ноябре и среднее количество осадков в октябре предыдущего года и числа Вольфа.

Рассмотрены и другие постановки задач. Например, требовалось найти закономерности, определяющие качественное поведение показателя заболеваемости (уменьшение или увеличение по сравнению с предыдущим уровнем) в зависимости от метеорологических факторов. Найдена закономерность: если средняя температура воздуха в апреле была больше 3.8 °C, то в текущем году заболеваемость повышается, в противном случае — уменьшается, справедливая для 78 % годовых показателей заболеваемости. На контрольных данных был неправильно классифицирован один из шести объектов.

Была рассмотрена также задача построения кусочно-линейной регрессии для количественной оценки уровня заболеваемости. Приведём пример полученных закономерностей (через X_{38} обозначена переменная — численность клещей в соответствующем году):

если X_{ind} = “Иркутск” и $X_{37} \leq 93.8$, то $\ln Y = 2.562 + 0.01 \cdot X_{37} - 0.014 \cdot X_{38}$. Все коэффициенты и модель в целом значимы, множественный коэффициент корреляции равен 0.75, стандартная ошибка составляет 0.5;

если X_{ind} = “Иркутск” и $X_{37} > 93.8$, то $\ln Y = 1.225 + 0.008 \cdot X_{37} + 0.017 \cdot X_{38}$. Все коэффициенты и модель в целом значимы, множественный коэффициент корреляции равен 0.95, стандартная ошибка составляет 0.14.

Таким образом, в периоды “высокой” и “низкой” активности Солнца (выше или ниже указанного порога для чисел Вольфа) структура зависимостей между заболеваемостью и анализируемыми факторами меняется. Причина изменений, возможно, состоит в различном механизме влияния характеристик паразитарной системы клещевого энцефалита на заболеваемость в указанные периоды.

Заключение

В работе предложены алгоритмы анализа ансамбля многомерных разнотипных временных рядов, основанные на классе логических решающих функций. В отличие от существующих методов анализа временных рядов разработанные алгоритмы позволяют выделять логические закономерности, описывающие динамику изучаемых объектов, дают возможность анализировать как количественные, так и качественные (порядковые, номинальные, булевы) переменные, не требуют предположений о параметрической модели распределения. На основе предложенных алгоритмов создан алгоритм построения коллективного решения.

При решении прикладной задачи анализа влияния природных факторов на заболеваемость клещевым энцефалитом в трёх эндемичных регионах России разработанный алгоритм показал более высокую точность прогнозирования на контрольных данных, чем существующий бустинг-алгоритм построения коллектива деревьев решений.

Список литературы

[1] Johnston J., DiNardo J. Econometric Methods. 4th Ed. McGraw-Hill, 1997.

[2] Frees E.W. Longitudinal and Panel Data. Cambridge Univ. Press, 2004.

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

[4] Лбов Г.С., Бериков В.Б. Устойчивость решающих функций в задачах распознавания образов и анализа разнотипной информации. Новосибирск: Ин-т математики СО РАН, 2005. 218 с.

[5] Журавлёв Ю.И., Рязанов В.В., Сенько О.В. Распознавание. Математические методы. Программная система. Практические применения. М.: Фазис, 2006.

[6] Freund Y., Schapire R. A decision-theoretic generalization of on-line learning and an application to boosting // J. of Computer and System Sci. 1997. Vol. 55, No. 1. P. 119–139.

[7] Бериков В.Б., Лбов Г.С., Полякова Г.Л. и др. Анализ факторов, влияющих на заболеваемость клещевым энцефалитом, с использованием логико-вероятностных и корреляционно-регрессионных моделей // Эпидемиология и вакцинопрофилактика. 2011. № 6(61). С. 25–34.

Поступила в редакцию 22 марта 2012 г.,
с доработки — 9 июня 2012 г.