Обнаружение статистических закономерностей при решении задачи прогнозирования температуры приземного воздуха

Копылов А. Н.[1], Синегубов С. В.[2]

[1]*Копылов Алексей Николаевич / Kopylov Alexey Nikolaevich — кандидат технических наук, доцент, старший преподаватель;*

[2]*Синегубов Сергей Владимирович / Sinegubov Sergey Vladimirovich — кандидат технических наук, доцент, кафедра высшей математики,*

Воронежский институт МВД России, г. Воронеж

Аннотация: *рассмотрен один из подходов к прогнозированию температуры приземного слоя воздуха на примере г. Воронежа.*

Ключевые слова: *обнаружение закономерностей, прогнозирование, температура воздуха, уровень значимости, таблица сопряженности.*

Несмотря на то, что оправдываемость прогнозов метеорологических величин и явлений погоды за последние десятилетия возросла, проблема повышения точности прогнозов актуальна и на сегодняшний день [1]. В общем случае при решении задачи прогнозирования метеорологических величин (например, температуры воздуха) не всегда требуется знать их значения в заданный момент времени. В ряде случаев достаточно знать, будет ли температура воздуха выше либо ниже по сравнению с текущей.

Рассмотрим задачу среднесрочного прогнозирования погоды на примере Воронежа на основе архива данных за период с 01.02.2011 по 31.10.2013. При этом в качестве исходных данных возьмем только температуру воздуха по состоянию на 13:00 каждого из $n = 1002$ дней рассматриваемого промежутка. Разобьем исходную последовательность на две: обучающую (первые $n_1 = 750$ отсчетов) и тестовую (остальные $n_2 = 252$ отсчета). Задачу прогнозирования будем решать как для исходного временного ряда, так и для ряда, полученного из исходного путем вычитания сезонной компоненты и среднего значения. Таким образом, если y_i $(i = \overline{1,n})$ – отсчеты исходного ряда (температуры воздуха), то значения второго ряда могут быть рассчитаны в соответствии с формулой:

$$\tilde{y}_i = y_i - a\sin(\omega i) - b\cos(\omega i) - c, \qquad (1)$$

где $\omega = 2\pi/365.25$ – угловая частота сезонной компоненты, неизвестные a, b и c можно рассчитать в соответствии с методом наименьших квадратов [2, 3], исходя из минимизации $\sum_i \tilde{y}_i^2$ $(i = \overline{1,n_1})$, либо то же самое, исходя из решения системы линейных алгебраических уравнений:

$$\begin{cases} a\sum_i \sin^2(\omega i) + b\sum_i \sin(\omega i)\cos(\omega i) + c\sum_i \sin(\omega i) = \sum_i y_i \sin(\omega i), \\ a\sum_i \cos(\omega i)\sin(\omega i) + b\sum_i \cos^2(\omega i) + c\sum_i \cos(\omega i) = \sum_i y_i \cos(\omega i), \\ a\sum_i \sin(\omega i) + b\sum_i \cos(\omega i) + c\sum_i 1 = \sum_i y_i . \end{cases} \qquad (2)$$

Прогнозировать изменение температуры воздуха будем исходя из обнаруженных на обучающей последовательности статистических закономерностей (СЗ). При этом под статистической закономерностью (по аналогии с вероятностной закономерностью в [4]) будем подразумевать правило $A_1\&A_2\&\ldots\&A_k \to A_0$ (где $A_0, A_1, \ldots, A_k$ — некоторые атомарные формулы), удовлетворяющее следующим условиям:

1) оценка условной вероятности $\hat{p}\left(A_0 \mid A_1\&A_2\&\ldots\&A_k\right) \neq 0$,

2) оценка условной вероятности $\hat{p}\left(A_0 \mid A_1\&A_2\&\ldots\&A_k\right)$ правила строго больше оценок условных вероятностей каждого из его подправил.

В качестве атомарных формул для исходного временного ряда, в частности, можно взять следующие:

1) $y_{i-m1} < y_{i-m2}$,
2) $y_{i-m1} - y_{i-m1-1} < y_{i-m2} - y_{i-m2-1}$,
3) $(y_{i-m1} + y_{i-m1-1})/2 < (y_{i-m2} + y_{i-m2-1})/2$,
4) $\min(y_{i-m1},\ y_{i-m1-1}) < \min(y_{i-m2},\ y_{i-m2-1})$,
5) $\max(y_{i-m1},\ y_{i-m1-1}) < \max(y_{i-m2},\ y_{i-m2-1})$, (3)

где $m1$ и $m2$ — некоторые натуральные числа либо ноль.

Для временного ряда, полученного в соответствии с (1), можно записать аналогичные формулы. Кроме того, в более общем случае можно рассмотреть и другие атомарные формулы.

Чтобы проверить на обучающем множестве, является ли некоторое правило $A_1\&A_2\&\ldots\&A_k \to A_0$ статистической закономерностью, необходимо проверить выполнимость условий 1 и 2. При этом прежде, чем переходить к оценке условной вероятности $\hat{p}\left(A_0 \mid A_1\&A_2\&\ldots\&A_k\right)$, необходимо проверить, являются ли признаки A_0 и $A_1\&A_2\&\ldots\&A_k$ зависимыми или нет при заданном уровне ошибки первого рода α (α положим равным 0.05) [5-6]. Так как в рассматриваемой задаче метод отбора данных является перекрестным [5], то достаточно построить таблицу сопряженности 2x2 (табл. 1).

Таблица 1. Таблица сопряженности

	A_0	$\overline{A_0}$

$A_1\&A_2\&\ldots\&A_k$	n_{11}	n_{12}
$\overline{A_1\&A_2\&\ldots\&A_k}$	n_{21}	n_{22}

и далее воспользоваться либо критерием Пирсона, либо точным критерием независимости Фишера. Если окажется, что при заданном α гипотезу о независимости признаков A_0 и $A_1\&A_2\&\ldots\&A_k$ отвергаем, то условную вероятность $p(A_0 \mid A_1\&A_2\&\ldots\&A_k)$ можно оценить следующим образом:

$$\hat{p}(A_0 \mid A_1\&A_2\&\ldots\&A_k) = \frac{n_{11}}{n_{11}+n_{12}}. \qquad (4)$$

Статистические закономерности в рассматриваемой задаче будем искать в соответствии с алгоритмом, изложенным в [4], однако при этом наложим дополнительные ограничения на частоту истинности посылки $A_1\&A_2\&\ldots\&A_k$ — данная посылка должна быть истинной не менее, чем в $\gamma = 10\%$ случаев от объема обучающей выборки. Данное ограничение наложено для того, чтобы можно было объективно сравнить оценки условных вероятностей на обучающей и тестовой выборках. В общем случае выбор данного порога произволен, однако чем ниже порог, тем больше на тестовой выборке будет правил $A_1\&A_2\&\ldots\&A_k \to A_0$, для которых при заданном α не будет основания отвергнуть гипотезу о независимости признаков A_0 и $A_1\&A_2\&\ldots\&A_k$.

Для уменьшения объема расчетов допустимыми атомарными формулами A_l будем считать только те, для которых оценка условной вероятности $\hat{p}(A_0 \mid A_l)$ превышает некоторый порог (например, 0.55). Кроме того, дополнительно ограничим максимальное число атомарных формул в посылке $A_1\&A_2\&\ldots\&A_k$ (k положим равным 4). Наложим ограничения на значения i, связанные с тем, что исходные данные могут быть неполными, а также на то, что при построении таблицы сопряженности (табл. 1) значения временного ряда, участвующие в атомарных формулах (3), должны быть определены. Так, например, в исходном временном ряде рассматриваемой задачи отсутствуют 4 значения — 3 в обучающей последовательности и 1 в тестовой. В более общем случае, при большом объеме исходных данных можно воспользоваться алгоритмом, приведенным в [7].

Среди обнаруженных закономерностей будем отбирать только те, для которых оценка условной вероятности на обучающей выборке превысит 0.85. Однако при этом наложим дополнительные ограничения на добавление новой статистической закономерности в архив — новое правило $A_1\&A_2\&\ldots\&A_k \to A_0$ будем добавлять в архив статистических закономерностей, если:

1) правило $A_1\&A_2\&\ldots\&A_k \to A_0$ является уточнением правила $A_1\&A_2\&\ldots\&A_{k-1} \to A_0$ и при этом справедливо неравенство

$\hat{p}(A_0|A_1\&A_2\&\ldots\&A_k) - \hat{p}(A_0|A_1\&A_2\&\ldots\&A_{k-1}) > 0.01$;

2) множество значений $i \in \mathbb{N}_{\le n_1}$, при которых верна посылка $A_1\&A_2\&\ldots\&A_k$ (либо $\overline{A_1\&A_2\&\ldots\&A_k}$), отличается от аналогичного множества для уже добавленных в архив закономерностей на некоторую величину β.

Второе условие введено, в том числе, для исключения идентичных либо почти идентичных правил. Так, например, атомарная формула 1) в (3) при $k = m + 2$ будет идентична атомарной формуле 3) при $k = m + 1$.

В качестве атомарной формулы A_0 рассмотрим условия $y_{i+7} < y_i$ ($y_{i+7} > y_i$), т. е. температура воздуха в 13:00 через семь дней будет меньше (больше) температуры воздуха по состоянию на 13:00 рассматриваемого дня. Аналогичную формулу A_0 возьмем и для ряда, заданного соотношением (1). Положим в атомарных формулах (3) $m1 = \overline{0,8}, m2 = \overline{m1 + 1,9}$.

Результаты эксперимента приведены в табл. 2.

Таблица 2. Результаты эксперимента

	A_0	$y_{i+7} < y_i$	$y_{i+7} > y_i$	$\tilde{y}_{i+7} < \tilde{y}_i$	$\tilde{y}_{i+7} > \tilde{y}_i$
$\gamma = 10\%$	Количество обнаруженных закономерностей	0	0	995	0
	Среднее арифметическое оценок условных вероятностей отобранных СЗ на тестовой выборке	0	0	0.84	0
$\gamma = 7\%$	Количество обнаруженных закономерностей	0	10	6421	89
	Среднее арифметическое оценок условных вероятностей отобранных СЗ на тестовой выборке	0	0.65	0.85	0.73

Таким образом, несмотря на достаточную простоту математической модели, видно, что при $\gamma = 10\%$ точность среднесрочного прогнозирования уменьшения температуры воздуха (с поправкой на изменение, связанное с сезонной компонентой) достаточно высока. При снижении γ до 7% число статистических

закономерностей увеличилось. Также появились закономерности, отвечающие за увеличение температуры, однако точность обнаруженных закономерностей на тестовой выборке оказалась ниже.

К недостаткам данной модели следует отнести тот факт, что прогнозировать изменение температуры можно лишь в том случае, если на текущий момент посылка $A_1 \& A_2 \& \ldots \& A_k$ оказалась истинной хотя бы у одной из обнаруженных статистических закономерностей. Очевидно, что для повышения точности прогнозирования и увеличения количества обнаруженных статистических закономерностей желательно использовать дополнительную информацию о состоянии окружающей среды: атмосферном давлении, влажности и т. п.

Литература

1. *Васильев А. А., Вильфанд Р. М.* Прогноз погоды: монография. — М.: Гидрометеорологический науч.-исслед. центр РФ. — 2008. — 60 с.
2. *Копылов А. Н.* Основы вычислительной математики: учебное пособие. — Воронеж: ВИ МВД России, 2012. — 183 с.
3. *Родин В. А., Синегубов С. В.* Применение метода наименьших квадратов для выравнивания экспериментальных данных, характеризующих поток информации интенсивного режима работы ПЦО // Вестник Воронежского института МВД России. — 1999. — № 2 (4). — С. 152–155.
4. *Демин А. В., Витяев Е. Е.* Разработка универсальной системы извлечения знаний «Discovery» и ее применения // Вестник НГУ. Серия: Информационные технологии. — 2009. — Т. 7. — Вып. 1. — С. 73-83.
5. *Флейс Дж.* Статистические методы для изучения таблиц долей и пропорций. Пер. с англ. Под ред. и с предисл. Ю. Н. Благовещенского. — М.: Финансы и статистика, 1989. — 319 с.
6. *Думачев В. Н.* Теория вероятностей и математическая статистика: учебник. — Воронеж: Воронежский ин-т МВД России, 2006. — 199 с.
7. *Копылов А. Н.* Алгоритм поиска статистических закономерностей при решении задач двухклассовой классификации // Вестник Воронежского института МВД России. — 2015. — № 2. — С. 233-238.