О задаче прогнозирования ледовых заторов на реках

И. В. Малыгин

Рассматриваетя задача прогнозирования образования ледовых заторов на северных реках. В случае, когда места заторообразования известны, необходимо прогнозирование мощности явления. Для решения применен комбинаторно-логический подход теории распознавания образов.

Ключевые слова: распознавание образов, тесты.

В работе предлагается решение задачи прогнозирования образования ледовых заторов на реках на примере участка реки Северная Двина между г. Котлас и г. Великий Устюг. Под прогнозом понимается классификация прогнозируемого периода к одному из выделенных сценариев развития ледовой обстановки.

Введем некоторые обозначения. Пусть M — множество наблюдений объекта. Множество M может быть разбито на непересекающиеся подмножества — классы $K_1, \ldots, K_l$. Целиком само разбиение неизвестно, однако в каждом классе есть подмножество элементов, о которых полностью известны их принадлежность и описание (характеристики). Совокупность таких подмножеств всех классов образует *обучающую выборку*: $T_1, \ldots, T_l$, $T_i \subset K_i$, $i = 1, \ldots, l$. Элементы обучающей выборки называются *эталонами*. Для элементов множества M не из обучающей выборки принадлежность к классу не известна. Для распознавания (классификации) предъявляется элемент множества M не входящий в обучающую выборку. Требуется классифицировать этот элемент, то есть отнести его к одному из существующих классов, представленных обучающей выборкой.

Для решения задачи распознавания выбран комбинаторно-логический подход. В этой задаче элементом множества M является известный набор гидрологических и метеорологических данных по отдельным речным постам бассейна Северной Двины за каждый год,

входящий в обучающую выборку. Каждый элемент множества M характеризуется набором признаков, каждый признак принимает либо числовое значение, либо вектор числовых значений. Классы являются результатом экспертной классификации объектов наблюдения (лет) по критерию мощности заторов, обучающую выборку можно рассматривать как классифицированные данные за прошедшие годы. Для решения поставленной задачи предлагается следующий алгоритм.

1. Исходные данные

Для исследования ледовой обстановки в качестве признаков заторобразования экспертами установлен ряд гидрологических и метеорологических показателей. Общий список этих признаков представлен в таблице 1.

В районе наблюдения использованы данные шести речных постов: Каликино, Великий Устюг, Медведки, Котлас, Абрамково, Подосиновец. Период наблюдения составил 20 лет: с 1991 по 2010 гг.

База данных для проведения исследования основана на реляционной модели данных. Для доступа к данным в основной таблице используется составной ключ из трех полей: год наблюдения, название речного поста, название признака. Допускается неполное заполнение таблицы. При фиксированном номере года (1991) в таблице 2 представлен пример заполнения базы данных.

При проведении опытной эксплуатации системы были приняты два возможных сценария (класса) ледохода на исследуемом участке:

1. слабый ледовый затор (в данный класс попадают и ситуации, когда затор произошел выше по течению реки);

2. сильный ледовый затор.

Экспертная классификация периода наблюдения к выделенным сценариям проводилась по фактической ситуации на исследуемом участке.

Указанные сценарии ледохода определяют классы K_1, K_2 периода наблюдения (таблица 3).

№	**Название признака**	**Характеристика признака**	**Единицы измерения**
1	Предледоставный уровень воды	Гидрологический признак	сантиметры
2	Продолжительность осеннего ледохода	Гидрологический признак	сутки
3	Наличие зажоров	Гидрологический признак	есть (1) — нет (0)
4	Особенности температурного режима в период замерзания	Метеорологический признак Дата перехода температуры воздуха через ноль	количество суток с 1 сентября
5	Сумма отрицательных температур воздуха за холодный период	Метеорологический признак	градусы Цельсия
6	Сумма положительных температур воздуха за холодный период	Метеорологический признак	градусы Цельсия
7	Количество дней с положительными температурами воздуха за холодный период	Метеорологический признак	сутки
8	Сумма твердых осадков	Метеорологический признак	миллиметры
9	Особенности температурного режима в период вскрытия	Метеорологический признак Дата перехода температуры воздуха через 0.	количество суток с 1 февраля
10	Толщина льда перед вскрытием	Гидрологический признак	сантиметры
11	Интенсивность роста уровней и расходов воды в период подвижек	Гидрологический признак	сантиметры в сутки

Таблица 1. Список гидрологических и метеорологических показателей.

2. Процедура сравнения однородных признаков за разные годы

Сравнение числовых значений однородных признаков за разные годы производится следующим образом: если различие в числовых

1991 год Пост	Номер признака					
	1 (м)	2 (сутки)	3 (да/нет)	4 (сутки)	5 (C°)	6 (C°)
Каликино	337	13	0	59	−1449	3,9
Вел. Устюг	121	10	0			
Медведки	135	5	0			
Котлас	214	22	0			
Абрамково	112	23	0			
Подосиновец	94	1	0			

Пост	Номер признака				
	7 (сутки)	8 (мм)	9 (сутки)	10 (см)	11 (см/сутки)
Каликино	6	192,44	34	41	76
Вел. Устюг				57	76
Медведки				67	42
Котлас				50	80
Абрамково				64	59
Подосиновец				48	61

Таблица 2. Пример исходных данных за 1991 г.

Год	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
№ класса	1	1	2	2	2	1	2	1	2	2

Год	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
№ класса	1	1	2	1	1	2	2	1	1	1

Таблица 3. Экспертная классификация периода наблюдения.

значениях находится в определенном допуске, то полагается, что два сравниваемых года по этому признаку одинаковые, в противном случае — различные. Для определения различия векторных признаков в алгоритме вычисляется сумма изменений значений по всем N постам:

$$H_k(\Gamma_n, \Gamma_m) = \frac{1}{N} \sum_{j=1}^{N} |p_{knj} - p_{kmj}|,$$

где $H_k(\Gamma_n, \Gamma_m)$ — величина различия n-го и m-го годов по k-му признаку, p_{knj} и p_{kmj} — числовые значения k-го признака на j-м посте в

n-й и m-й годы; далее производится сравнение с пороговым значением δ_k.

3. Формирование таблицы сравнения классов, построение множества тестов

Напомним процесс формирования таблицы сравнения [1, 4]. Для каждой пары лет из разных классов с использованием процедуры сравнения однородных признаков определяется обобщенный вектор различия этой пары лет по всем признакам. Вектор формируется следующим образом: если в результате работы процедуры сравнения установлено различие в паре лет по i-му признаку, то в качестве координаты с номером i обобщенного вектора принимается 1, в противном случае 0:

$$V(\Gamma_n, \Gamma_m) = (v_1, \ldots, v_{11}),$$
$$v_i = 1, \text{ если } H_i(\Gamma_n, \Gamma_m) > \delta_i,$$
$$v_i = 0, \text{ если } H_i(\Gamma_n, \Gamma_m) < \delta_i,$$
$$i = 1, \ldots, 11.$$

Этот вектор не является нулевым, так как года берутся из разных классов. Равенство нулю этого вектора означало бы, что выбранная процедура сравнения является грубой моделью реальной ситуации. Совокупность всех таких векторов составляет таблицу сравнения классов. Исследуемый период разбит на два класса, следовательно, в таблице сравнения содержится 99 булевских векторов размерности 11. Далее по таблице сравнения классов строится множество тестов.

4. Классификация текущего года

По входным данным X текущего года, исходным данным периода наблюдения, принятой процедуре сравнения признаков и построенному множеству тестов производится голосование за принадлежность текущего года одному из классов. Голосование представлено алгоритмом В. Б. Кудрявцева [4].

Фиксируется тест t и год Γ_n из периода наблюдения:

1. $g(X, \Gamma_n, t) = 1$, если для каждого i-го признака такого, что i-я координата t равна 1, i-й признак распознаваемого года X «совпал» с i-м признаком года Γ_n;

2. $g(X, \Gamma_n, t) = 0$, если хотя бы для одного i-го признака такого, что i-я координата t, равной 1, i-й признак распознаваемого года X «не совпал» с i-м признаком года Γ_n.

Производится суммирование голосов по всем тестам и всем годам из класса K_j, $j = 1, 2$:

$$G_j = \frac{1}{|K_j|} \sum_t \sum_{\Gamma_n \in K_j} g(X, \Gamma_n, t),$$

где

$$g(X, \Gamma_n, t) = \prod_{i=1}^{s} (1 - t_i \theta(H_i(X, \Gamma_n) - \delta_i)),$$

$$\theta(x) = \begin{cases} 0, & x < 0, \\ 1, & x \geqslant 0 \end{cases} \quad \text{— функция Хэвисайда,}$$

s — число признаков, $t = (t_1, \ldots, t_s)$ — тест, $X = (X_1, \ldots, X_s)$ — распознаваемый элемент [1].

Решающее правило F_1 алгоритма заключается в отнесении текущего распознаваемого года к тому классу, за который подано больше голосов:

$F_1(X) = j_0$, где $G_{j_0} = \max\limits_j G_j$.

В случае совпадения числа голосов за разные классы применяется алгоритм голосования Ю. И. Журавлева [2].

Как и выше, фиксируется тест t и эталон Γ_n из обучающей выборки. Считается количество единичных координат теста таких, что значение признака распознаваемого элемента «не совпало», то есть они отличаются в смысле процедуры сравнения с признаками эталона. Таким образом, определяется общее число «не совпадений» $v(X, \Gamma_n, t)$ эталона Γ_n и распознаваемого элемента X по фиксированному тесту t. По смыслу $v(X, \Gamma_n, t)$ выражает число голосов «против» принадлежности распознаваемого вектора X к классу, которому принадлежит эталон Γ_n. Далее, производится суммирование значений v по всем тестам и всем элементам из класса K_j, $j = 1, 2$:

$$V_j = \frac{1}{|K_j|} \sum_t \sum_{\Gamma_n \in K_j} v(X, \Gamma_n, t), \text{ где}$$

$$v(X, \Gamma_n, t) = \sum_{i=1}^{s} t_i \theta(H_i(X, \Gamma_n) - \delta_i).$$

Решающее правило F_2 алгоритма заключается в отнесении текущего распознаваемого элемента к тому классу, за который подано меньше голосов «против»:

$$F_2(X) = j_0, \text{ где } V_{j_0} = \min_j V_j.$$

Таким образом, общее решающее правило алгоритма имеет вид:

$$F(z) = \begin{cases} 1, & z = -1, \\ 2, & z = 1, \end{cases}$$

где $z = \operatorname{sgn}(G_1 - G_2) - (1 - \operatorname{sgn}^2(G_1 - G_2)) \operatorname{sgn}(V_1 - V_2)$. Видно, что в случае различного числа голосов $F = F_1$, иначе $F = F_2$.

5. Результат и выводы

В таблице 4 представлены результаты машинной обработки данных за период наблюдения по описанному выше алгоритму. Из периода наблюдения последовательно удалялся каждый год, оставшиеся годы рассматривались в качестве обучающей выборки, удаленный год подавался на вход прогнозному алгоритму в качестве распознаваемого, производилось распознавание. Как видно из таблицы 4 число правильно классифицированных лет составило 85%.

Возможность использования теоретико-вероятностного подхода при решении задач прогнозирования предполагает наличие статистически значимых обучающих выборок. В исследуемой проблеме эти выборки не удовлетворяют такому требованию, информативно-полные однородные данные возможно структурировать за небольшой период наблюдения (15–20 лет). Для малообъемных выборок разработан комбинаторно-логический подход. В его рамках тестовые процедуры голосования решают задачи распознавания для параметров различной природы, структуры, описания, происхождения [1, 3].

Год	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000
№ класса	1	1	2	2	2	1	2	1	2	2
Результат прогноз-я	1	1	**1**	2	2	1	2	1	2	2

Год	2001	2002	2003	2004	2005	2006	2007	2008	2009	2010
№ класса	1	1	2	1	1	2	2	1	1	1
Результат прогноз-я	1	1	2	1	1	**1**	2	**2**	1	1

Таблица 4. Сравнение результатов прогнозирования с фактическими результатами.

Использование прогноза заторообразования позволяет учитывать этот фактор при составлении прогноза наводнений, что в свою очередь может сократить ущерб для экономической и хозяйственной деятельности на исследуемом участке.

Автор выражает благодарность научному руководителю профессору И. К. Лурье, а также академику В. Б. Кудрявцеву и профессору С. В. Алешину за внимание к работе и ценные обсуждения.

Список литературы

[1] Алешин С. В. Распознавание динамических образов. Часть 1. — М.: МГУ, 1996.

[2] Дмитриев А. Н., Журавлев Ю. И., Кренделев Ф. П. О математических принципах классификации предметов и явлений // Дискретный анализ. — Новосибирск: ИМ СО АН СССР, 1966. — Вып. 7. — С. 3–15.

[3] Кудрявцев В. Б., Андреев А. Е. Теория тестового распознавания // Интеллектуальные системы. — 2006. Т. 10, вып. 1–4. — С. 95–140.

[4] Константинов Р. М., Королева З. Е., Кудрявцев В. Б. Комбинаторно-логический подход к задачам прогноза рудоносности // Проблемы кибернетики. — М.: Наука, 1976. — Вып. 31. — С. 5–38.