УДК 338.27.015 (075.8)

ПРИМЕНЕНИЕ СИСТЕМНО-АНАЛИТИЧЕСКОГО МЕТОДА «ДЕРЕВО РЕШЕНИЙ» С МЕТОДОМ ВЫБОРА АТРИБУТА CART ДЛЯ ПОСТРОЕНИЯ СИСТЕМЫ КРАТКОСРОЧНОГО ПРОГНОЗИРОВАНИЯ

И.М. Сунагатов, В.И. Батищев

Самарский государственный технический университет
443100, г. Самара, ул. Молодогвардейская, 244

Рассмотрен и применен аналитический метод «Дерево решений: CART». Рассмотрены варианты получения прогнозных значений энергопотребления. Проанализированы ошибки прогнозирования.

***Ключевые слова**: классификационно-регрессионные деревья решений, CART, методы прогнозирования, энергопотребление, математическое моделирование.*

Прогнозирование режимных параметров и технико-экономических показателей является одной из важных задач как при планировании, так и при ведении текущих режимов энергообъединения. Составляя планы по различным показателям на предстоящие сутки, неделю, месяц, квартал, год, службы и отделы ЭО решают задачу планирования энергобаланса – соотношения между потребностью в электроэнергии (мощности) и средствами ее удовлетворения. Одним из важнейших показателей при планировании является уровень ожидаемого электропотребления в целом по объединению, группам и отдельным потребителям. В этом смысле величина прогноза электропотребления является опорным показателем для планирования балансов электроэнергии и мощности.

Для построения прогноза поведения системы необходимо понимание ее функционирования и представление ее структуры. Для построения структуры воспользуемся системно-аналитическим методом «дерево решений» с алгоритмом выбора атрибута CART [1, 2, 3].

В алгоритме CART каждый узел дерева решений согласно методу дихотомии имеет два листа. На каждом шаге построения дерева правило, формируемое в узле, делит заданное множество примеров (обучающую выборку) на две части – часть, в которой правило выполняется и часть, в которой правило не выполняется [4]. Для выбора оптимального правила используется функция оценки качества разбиения

$$Gini(T) = 1 - \sum_{i=1}^{n} p_i^2 \to \min , \qquad (1)$$

где p_i^2 – вероятность появления класса i в наборе данных T.

Входными данными являются временные ряды значений потребления электроэнергии и соответствующих им значений окружающей температуры. Для анализа использованы реально зарегистрированные временные ряды Z_t, содержащие по 8329 значений, снятых с интервалом 1 час. Ряды квантованы на суточные выборки Z_{tj}^{M}, где $j = 1,2,...,n,\ M = 24.$

На основе визуального анализа значений годичного интервала ряда потребления

Ильдар Маратович Сунагатов, аспирант.

Виталий Иванович Батищев (д.т.н., проф.), заведующий кафедрой «Информационные технологии».

выдвигаются гипотезы о подобии суточных выборок. Критерием для проверки гипотез примем линейный коэффициент корреляции Пирсона. В случае подтверждения гипотезы наличия подобия прогнозные значения анализируемых факторов F_t^M определяются путем экстраполяции выборки по предыдущим ее значениям или значениям аппроксимирующей выборки.

Значениями аппроксимирующей выборки $\bar{Z}_t^M$ являются средние значения:

$$\bar{Z}_t^M = \left\{ \frac{1}{n} \sum_{j=1}^{n} B_{tj} \right\}, j = 1,2,...,24,$$

где $B_{tj} = Z_t^M - \min(Z_t^M)$.

Экстраполируя данные, получим прогнозные значения факторов:

$$F_t^M = P_{ct} + \bar{Z}_t^M + e, \quad (2)$$

где P_{ct} – текущие значения энергопотребления.

Среднее значение ошибки экстраполирования аппроксимирующей выборки M также будет являться входными данными для построения дерева решения.

Рассмотрим различные варианты характера изменения значений температуры и энергопотребления на отдельных участках временных рядов (табл. 1).

Таблица 1

Характер изменения значений температуры и энергопотребления на отдельных участках временных рядов

№	Температура	Потребление	Участок	Ошибка, %
1	Снижение	Увеличение	$[Z_t^M; Z_{t+150M}^M]$	5
2	Увеличение	Снижение	$[Z_t^M; Z_{t+150M}^M]$	5
3	Снижение	Увеличение	$[Z_{t+150M}^M; Z_{t+450M}^M]$	4
4	Увеличение	Снижение	$[Z_{t+150M}^M; Z_{t+450M}^M]$	4
5	Снижение	Увеличение	$[Z_{t+450M}^M; Z_{t+366M}^M]$	5
6	Увеличение	Снижение	$[Z_{t+450M}^M; Z_{t+366M}^M]$	5
7	Снижение	Снижение	$[Z_{t+180M}^M; Z_{t+210M}^M]$	5
8	Увеличение	Увеличение	$[Z_{t+180M}^M; Z_{t+210M}^M]$	5
9	Отсутствие изменения	Отсутствие изменения	$[Z_t^M; Z_{t+366M}^M]$	0,5

С учетом выявленной в результате статистического анализа данных регрессионной зависимости значений энергопотребления от температуры окружающей среды можно ввести поправку в прогнозный уровень, относительно которого будет производиться экстраполяция выборки на сутки вперед:

$$P_{ft} = P_{ct} + K_1(T_{ct} - T_{t+1}), \quad (3)$$

где T_c – текущие значения температуры;

T_{t+1} – прогнозные значения температуры;

K_1 – коэффициент линейной регрессии (в рассматриваемом примере $K_1 = 16,7$).

С учетом поправки прогнозные значения будут рассчитываться по формуле

$$F_t^M = P_{ft} + \bar{Z}_t^M + e. \quad (4)$$

В табл. 2 приведены результаты анализа ошибок прогнозирования с учетом введенных выше поправок, отнесенные к некоторым характерным интервалам.

Таблица 2

Значения ошибки на выборках с определенным сдвигом

№	Сдвиг	Ошибка, %	Соответствие
1	t_{-7M}	5	Праздничный – выходной (воскресенье)
2	t_{-6M}	4	Суббота при 6-дневной рабочей неделе
3	t_{-5M}	3	Предпраздничный (пятница)
4	t_{-1M}	3	Послепраздничный (понедельник)
5	$t_{-2M}, t_{-3M}, t_{-4M}$	2	Вторник, среда, четверг при 5-дневной рабочей неделе

Дальнейшее уточнение прогноза может быть обеспечено рациональным выбором участков опорной выборки исходных данных. В табл. 3 показано изменение ошибок прогнозирования энергопотребления в рабочие дни, обусловленных ошибками аппроксимации суточного графика при переходе от участка $[Z_t^M; Z_{t+150M}^M]$ к $[Z_{t+150M}^M; Z_{t+450M}^M]$ и от $[Z_{t+150M}^M; Z_{t+450M}^M]$ к $[Z_{t+450M}^M; Z_{t+366M}^M]$.

Таблица 3

Изменение ошибки аппроксимации в зависимости от выбора участка исходной выборки

№	Сдвиг	Ошибка, %	Ошибка аппроксимации, %
1	$[Z_t^M; Z_{t+150M}^M] \to [Z_{t+150M}^M; Z_{t+180M}^M]$	$-4\% \to +2\%$	$-2\% \to +0\%$
2	$[Z_{t+150M}^M; Z_{t+180M}^M] \to [Z_{t+180M}^M; Z_{t+210M}^M]$	$+2\% \to +3\%$	$+0\% \to +1\%$
3	$[Z_{t+180M}^M; Z_{t+210M}^M] \to [Z_{t+210M}^M; Z_{t+450M}^M]$	$+3\% \to +2\%$	$+1\% \to +0\%$
4	$[Z_{t+210M}^M; Z_{t+450M}^M] \to [Z_{t+450M}^M; Z_{t+366M}^M]$	$+2\% \to -4\%$	$+0\% \to -2\%$

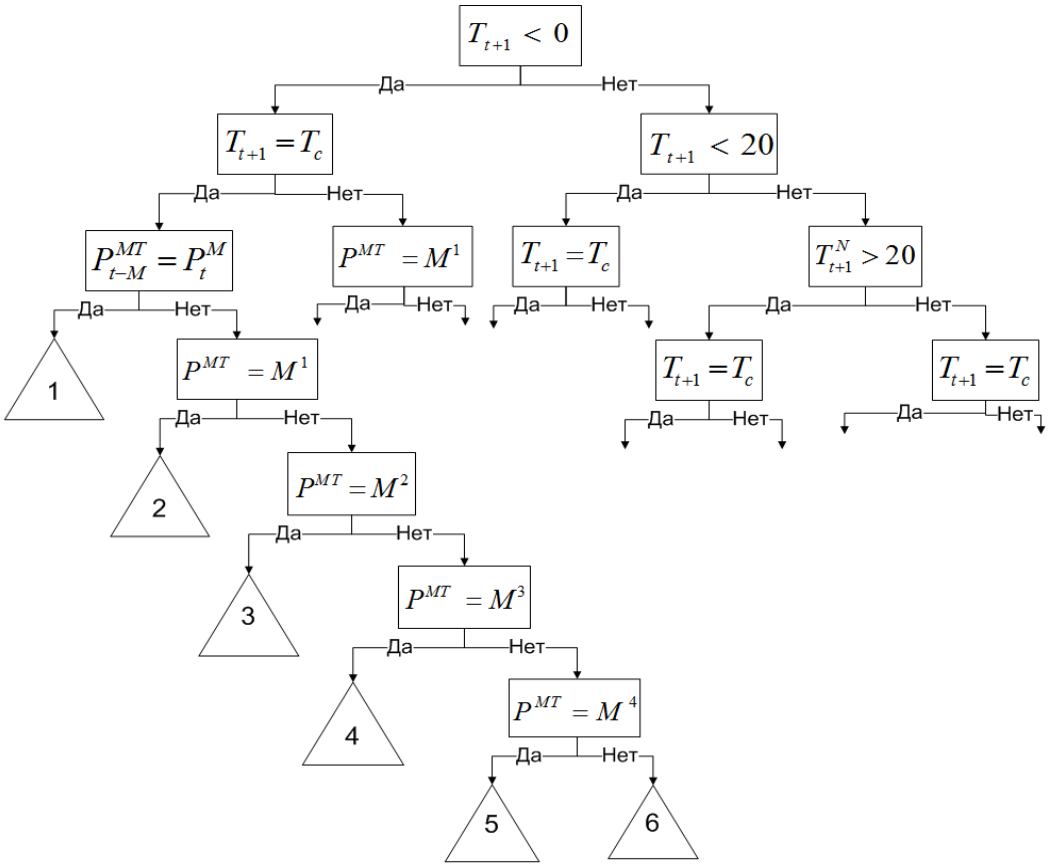
В результате учета температурного сезона ошибка прогнозирования снижается до ~ 2 %.

Построение дерева регрессии схоже с процедурой построения дерева классификации. Сначала строится дерево максимального размера, затем уменьшается размер дерева с помощью алгоритма «отсечения» (pruning) до приемлемого размера.

Дерево решений дает возможность работать с многомерными задачами и задачами, в которых существует зависимость выходной переменной от переменных категориального типа. Главный принцип построения – разбиение всего пространства на участки, в которых выходная переменная считается постоянной. При этом следует учитывать, что существует сильная зависимость между объемом обучающей выборки и результирующей ошибкой ответа дерева.

Процесс построения дерева происходит последовательно. На первом шаге мы получаем регрессионную оценку как константу по всему пространству примеров,

которая учитывается как среднее значение выходной переменной в обучающей выборке.



Р и с. 1. Фрагмент дерева решений

Для построения дерева решений имеющиеся данные сведены в табл. 4, где указана степень влияния атрибутов на ошибку.

Результатом получим дерево с максимальной глубиной в 9 узлов, состоящее из категориальных и регрессионных переменных (атрибутов). Первым выбирается атрибут с наибольшей степенью влияния, т. е. тот, который максимально сузит диапазон значений, – прогнозная температура (регрессия). Локализуя температурный сезон, выполняем разбиение ряда температуры сравнением со вторым атрибутом (категориальным) «Температура 1-4».

Третьим атрибутом выбирается сезон недельного потребления, характер которого зависит от второго атрибута.

Визуально фрагмент дерева решений представлен на рисунке.

Таблица 4

Перечень атрибутов

№	Наблюдение	Степень влияния, %	Обозначение	Вид атрибута
1	Температура 1. Зима	1	T^1	Категориальный
2	Температура 2. Осень, весна		T^2	Категориальный
3	Температура 3. Лето		T^3	Категориальный

№	Наблюдение	Степень влияния, %	Обозначение	Вид атрибута
4	Температура 4. Повышение ночной летней		T^4	Категориальный
5	Потребление. Функция зависимости потребления от прогнозной температуры	2	P_f	Регрессионный
6	Потребление 1. Рабочие дни при 5-дневной неделе	1	M^1	Категориальный
7	Потребление 2. Рабочие дни при 6-дневной неделе		M^2	Категориальный
8	Потребление 3. Праздничные и выходные дни		M^3	Категориальный
9	Потребление 4. Предпраздничные дни		M^4	Категориальный
10	Потребление 5. Послепраздничные дни		M^5	Категориальный

Решением будет служить среднее значение дневного потребления, по которому будет экстраполирован график, выбранный по второму и третьему атрибуту. Дерево позволяет сделать прогноз с точностью до 2 %, а также визуализирует основные зависимости.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Ананий В. Левитин.* Алгоритмы: введение в разработку и анализ = Introduction to The Design and Analysis of Aigorithms. – М.: Вильямс, 2006. – С. 409-417.
2. Breiman Leo, Friedman J.H., Olshen R.A. & Stone C.J. (1984). Classification and regression trees. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software.
3. *Hyafil, Laurent; Rivest, RL* (1976). Constructing Optimal Binary Decision Trees is NP-complete. Information Processing Letters 5 (1): 15-17.
4. *Айвазян С.А., Мхитарян В.С.* Прикладная статистика и основы эконометрики. – М.: Юнити, 1998.

Статья поступила в редакцию 7 июля 2013 г.

USE SYSTEM-ANALYTICAL METHOD «DECISION TREES» TO THE PROCESS OF SELECTING THE ATTRIBUTES «CART» FOR BUILDING SYSTEM OF SHORT-TERM FORECAST

I.M. Sunagatov, V.I. Batishchev

Samara State Technical University
244, Molodogvardeyskaya st., Samara, 443100

The analytical method «A tree of decisions: CART» is considered and applied. Options of obtaining expected values of power consumption are considered. Errors of forecasting are analyzed.

Keywords: *classification and regression decision trees, CART, forecasting methods, energy, mathematical modeling.*

Ilidar M. Sunagatov, Postgraduate Student.
Vitaly I. Batishchev (Dr. Sci. (Techn.)), Professor.