И. А. ПРИХОДЬКО, Е. С. ФИЛАТОВА, Д. А. ШИЛЬНИКОВА
Санкт-Петербургский государственный электротехнический университет «ЛЭТИ»
им. В.И. Ульянова (Ленина), Санкт-Петербург

ПРОГНОЗИРОВАНИЕ ВРЕМЕННЫХ РЯДОВ НА ОСНОВЕ МЕТОДА РЕЛЕВАНТНЫХ ВЕКТОРОВ

Выполнено исследование сравнительной эффективности методов релевантных и опорных векторов для решения задачи регрессии на примере прогноза электропотребления. Для формирования прогностической модели выполнен анализ исследуемого временного ряда: рассмотрен закон распределения, проверена коррелированность данных обучающей выборки. Показана необходимость предварительной обработки обучающей выборки для повышения точности прогноза. В результате исследования подтверждена сравнительная эффективность метода релевантных векторов для решения задачи прогнозирования временных рядов.

Введение. Задача прогнозирования временных рядов решена на основе методов опорных и релевантных векторов (support vector machine, SVM, relevance vector machine, RVM).

Сравнительным преимуществом методов является свойство разреженности модели. В методе SVM прогностическая модель формируется на основе опорных векторов, в методе RVM – на основе релевантных векторов.

По сравнению с широко используемыми методами на основе нейросетевых алгоритмов, преимущество метода опорных векторов в том, что параметры регрессионной модели определяются на основе решения задачи квадратичного программирования. Ограничением в применении является неустойчивость по отношению к шуму в исходных данных. [1].

Метод RVM, используя байесовский подход, позволяет получить для задач регрессии полностью вероятностную модель.

Устойчивость к шуму в исходных данных – преимущество метода релевантных векторов [1].

Предлагаемый доклад посвящен исследованию сравнительной эффективности методов SVM и RVM для предсказания предполагаемого электропотребления при наличии шума в исходных данных.

Основная часть.

Прогностическая модель строится на основе уравнения линейной регрессии.

Задачей построения уравнения линейной регрессии является оценка неизвестной вещественной функции

$$y = f(x) + \varepsilon, \tag{1}$$

где для модели на основе метода опорных векторов $f(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle + w_0$, $x \in R^n$ – вектор переменных временного ряда, вектор $\boldsymbol{w} = (w, ..., w_n) \in R^n$ и смещение $w_0 \in R$ – весовые коэффициенты, ε – допустимая ошибка аппроксимации.

Для модели на основе метода релевантных векторов в (1) $f(x) = \sum_{i=1}^{m} w_i \phi_i(x)$, $\phi_i(x)$ – базисная функция, $\varepsilon \sim N(0, \sigma)$ – остатки модели, распределённые по нормальному закону с математическим ожиданием равным нулю и среднеквадратичным отклонением σ.

В применении к решаемой задаче в (1) $y = P_{t+1}$ – прогнозируемое электропотребление за t +1 день недели.

При использовании метода опорных векторов задача нахождения параметров сводится к задаче квадратичной оптимизации и формулируется в виде минимизации функционала [1]:

$$\min_{w,w_0,\xi,\xi^*}\left[\frac{1}{2}w^T w + C\sum_{i=1}^{p}\left(\xi_i + \xi_i^*\right)\right],$$

при ограничениях

$$y_i - w^T x_i - w_0 \le \varepsilon + \xi_i,$$
$$w^T x_i + w_0 - y_i \le \varepsilon + \xi_i^*, \xi_i \ge 0, \xi_i^* \ge 0, i = \overline{1, p},$$

где параметр C – положительная константа, задающая штраф на ошибку; $\xi_i > 0$ – набор дополнительных переменных, характеризующих величину ошибки на объектах x_i, $i = 1, \ldots, p$. Штрафное слагаемое в функционале $\frac{1}{2}\|w\|^2$ вводится согласно принципу регуляризации и означает, что среди всех векторов w, минимизирующих функционал (2), наиболее предпочтительны векторы с минимальной нормой. Второе слагаемое функционала штрафует любые отклонения $f(x)$ от y большие, чем ε для всех обучающих данных.

Обычно, при использовании метода SVM, решается двойственная задача, уравнение регрессии (1) выражается через двойственные переменные

$$f\left(x_i\right) = \sum_{i=1}^{p}\left(\alpha_i - \alpha_i^*\right)K\left(x_i, x\right) + w_0,$$

где α_i, α_i^* – множители Лагранжа, $K(x_i, x)$ – ядерная функция. Если $\alpha_i \ne 0, \alpha_i^* \ne 0$, x_i – опорный вектор.

Уравнение (1) для модели на основе метода RVM может быть переписана в виде $y = \Phi \cdot w + \varepsilon$, где Ф – матрица $n \times m$, i-й столбец которой образован значениями базисной функции $\phi_i(x)$ во всех точках тренировочного набора, а ε – вектор остатков.

Весовые коэффициенты рассчитываются по формулам (2), (3) [2].

Формула расчета весов для метода наименьших квадратов:

$$w = (\Phi^T \Phi)^{-1}\Phi^T y. \quad (2)$$

Формула расчета весов для максимума апостериорной вероятности:

$$w = (\Phi^T \Phi + \sigma^2 \cdot \psi)^{-1}\Phi^T y, \quad (3)$$

где ψ – матрица дисперсии.

Формируется разреженная модель, в которой весам присваивается априорное гауссовское распределение с нулевым математическим ожиданием и различными значениями дисперсии для каждого веса. Разреженность достигается с помощью апостериорной вероятности множества весов, которая стремится к нулю [3].

В рассматриваемом примере обучающая выборка представляет собой почасовые данные электропотребления за рабочие дни декабря 2020 года и января 2021 года[1]. Для выделения основной зависимости, данные были разделены по дням недели (с понедельника по пятницу) с использованием агрегирования средним.

[1] http://www.atsenergo.ru

Исследование выполнено в среде Anaconda с применением языка программирования Python.

На основе визуального анализа гистограммы и огибающей частот (характеризующей плотность вероятности) значений электропотребления в течение суток, получено, что закон распределения энергопотребления не соответствует гауссовскому.

Для построения регрессионной модели, опираясь на расчеты коэффициента корреляции Спирмена и соответствующего значения уровня значимости – *p*-value, получено, что предикторы и предиктанты коррелированны.

Для определения вида регрессии построена зависимость предикторов по месяцам. На основе гистограммы получено, что зависимость предикторов имеет линейный вид, поэтому для решения задачи прогнозирования использована модель линейной регрессии.

Предварительная обработка данных обучающей выборки.

Предикторы задаются как агрегированные средние по дню недели данные электропотребления. Предиктанты представляют собой агрегированные средние данные электропотребления с лагом в один день: прогноз на вторник осуществляется на основе данных за понедельник, прогноз на понедельник – основе данных за пятницу.

Для получения прогнозных значений формируется ансамбль регрессионных моделей, каждой модели соответствует пара предиктор – предиктант по соответствующему дню недели, для каждой модели ансамбля рассчитываются метрики с применением предикторов и предиктантов тестового набора.

В работе рассмотрен ансамбль моделей, в качестве примера работы модели на всех рисунках приведен прогноз электропотребления на пятницу.

Для выявления необходимости нормализации данных при обучении RVM моделей выполнен эксперимент, в котором произведено сравнение качества прогноза моделей, обученных с применением нормализованных и ненормализованных данных.

Стационарность остатков модели подтверждена с применением теста Дики–Фулера.

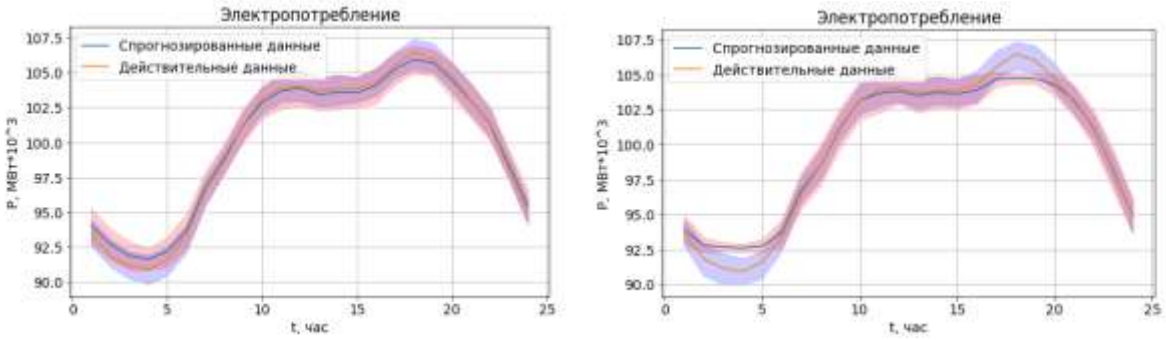
Приведённая средняя ошибка прогнозирования без нормализации данных равна 1.059 %, после нормализации – 0.15 %.

Выполнено сравнение результатов прогнозирования для различных ядерных функций: линейной(linear), радиальной базисной (rbf) и сигмоидной (sig).

Приведенные средние ошибки прогнозирования при использовании RVM модели для функций linear, rbf, sig составили, соответственно, 0,15 %, 0.186 %, 0,158 %. В исследованиях в качестве ядерной функции использована линейная функция. Для SVM модели наименьшая приведённая средняя ошибка прогнозирования составила 0.56 % при использовании радиальной базисной функции, которая применялась в сравнительных экспериментах.

Для эксперимента с добавлением аддитивного шума использовался метод моделирования Монте-Карло. Для данного эксперимента к первоначальным данным P_t прибавлялась случайная выборка чисел с нормальным распределением P_{noise}: $P_t + P_{noise}$, $P_{noise} \sim N(0, 1)$. Полученные зашумленные данные проходили через нормализацию и в дальнейшем строилась регрессия (рисунок).

По полученным спрогнозированным данным высчитывались метрики. Данный эксперимент повторялся 100 раз, каждые новые метрики записывались в массив, затем, из него были высчитаны средние метрики всего массива данных. Этот опыт проводился с целью получить усредненные показания, так как шум – случайная составляющая и может оказаться удачной, в контексте регрессионного анализа или наоборот, неудачной в одной определённой реализации сгенерированного шума. Максимальная амплитуда составляющей аддитивного шума равна 2.65 МВт*10^3.



а) RVM модель б) SVM модель

Рисунок. Прогнозирование электропотребления на данных с шумом

Приведённая средняя ошибка прогнозирования для зашумленных данных модели RVM составляет 0.99 %, для модели SVM – 1.07 %.

Заключение. В результате сравнительного исследования точности прогноза электропотребления подтверждена перспективность использования метода релевантных векторов.

ЛИТЕРАТУРА

1. **Воронцов К.В**. Лекции по методу опорных векторов. 2007. Режим доступа: URL: http://machinelearning.ru/.
2. Tipping M.E. Sparse Bayesian Learning and the Relevance Vector Machine. *Journal of Machine Learning Research*. 2001. T. 1. P. 211–244.
3. Fokoué E., Sun D., Goel P. Fully Bayesian analysis of the relevance vector machine with an extended hierarchical prior structure. *Statistical Methodology*. 2011. vol. 8(1) P. 83–96.

I.A.Prikhodko, E.S.Filatova D.A.Shilnikova (Saint Petersburg Electrotechnical University “LETI”, St. Petersburg)

Forecasting of a temporary row on the basis of the relevance vector machine

A study was made to find out a comparative effectiveness of the relevance vector machine and the support vector machine methods for solving the problem of regression on the example of power consumption forecast. To form a predictive model, an analysis of the time series under investigation was performed: the distribution law is considered, the correlation of the data of the training sample is checked. The necessity of the preprocessing of the training sample to improve forecast accuracy is shown. As a result of the study, the effectiveness of the relevance vector machine method for solving the problem of time series forecasting was confirmed.