Многоуровневая классификация при обнаружении движения цен*

А. А. Кузьмин
senatormipt@gmail.com

В данной работе рассматривается один из возможных методов прогнозирования, основанный на модели логистической регрессии. Предлагается способ разметки пучка временных рядов и построения матрицы объект — признак. Алгоритм проверяется на синтетических пучках временных рядов вида зашумленных синусов и периодических трапеций. Как вариант практического применения, алгоритм тестируется на данных о потреблении электроэнергии.

Ключевые слова: *логистическая регрессия, разметка временных рядов, потребление электроэнергии.*

Multi-level classification upon detection of price movement*

A. A. Kuzmin
Moscow Institute of Physics and Technology

This research describes one of the possible methods of forecasting, which is based on logistic regression model. A method of marking the time series beam and building a matrix of attributes and objects is proposed. Algorithm is tested on synthetic time series beams, which have the form of noisy sine and periodic trapezium. As the variant of practical application, algorithm is tested on energy consumption data.

Keywords: *logistic regression, time series marking, electricity consumption.*

Введение

Временным рядом называют последовательность, упорядоченную по времени. Это могут быть значения биржевых индексов, цены на определенный товар, характеристики пациента, среднесуточная температура в городе и т.д.[4]. Пучком временных рядов называется набор зависимых временных рядов, например метеорологические данные: температура, давление, скорость и направление ветра [3], [4]. Таким образом, имея все эти данные за некоторый прошедший период, мы гораздо точнее сможем предсказать значение температуры, нежели если бы мы имели лишь данные о температуре. Предполагая зависимость всех временных рядов в исследуемом пучке, будем предсказывать значение одного ряда, основываясь на предыстории всего пучка.

Мы будем рассматривать лишь пучки временных рядов одинаковой длины, представляющие собой дискретную последовательность с одинаковым шагом по времени. В данной работе ставится задача прогнозирования разметки: возрастет или уменьшится его значение в следующий момент. Для этого проводится разметка ряда и строится матрица объект — признак, как, например, в [3],[4], [5]. В качестве модели будем рассматривать логистическую регрессию (см. [1],[6]).

Научный руководитель В. В. Стрижов

Постановка задачи

Пусть у нас имеется N временных рядов длиной T с шагом по времени t. Для удобства пронормируем шаг, т.е. возьмем $t = 1$ и аналогично пересчитаем T. Будем называть каждый из рядов — $a_i, i \in \{1, 2 \dots N\}$, значение ряда a_i в момент времени t будем обозначать $a_{i,t}$. Таким образом, имеем матрицу с неизвестным столбцом $a_{i,T+1}$, где каждая строка — временной ряд. Пусть, для определенности, нам надо предсказать поведение ряда a_1 в момент $T + 1$.

$$\mathbf{A} = \begin{pmatrix} a_{1,1} & a_{1,2} & \dots & a_{1,T-1} & a_{1,T} & a_{1,T+1} \\ a_{2,1} & a_{2,2} & \dots & a_{2,T-1} & a_{2,T} & a_{2,T+1} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ a_{N,1} & a_{N,2} & \dots & a_{N,T-1} & a_{N,T} & a_{N,T+1} \end{pmatrix}$$

Определим множество элементов разметки ряда как: $\mathcal{M} = \{+1, 0, -1\}$. Разметим интересующий нас ряд a_1, причем разметку будем проводить следующим образом: вместо элемента $a_{1,t}$ будем ставить $+1$, если следующий элемент этого ряда $a_{1,t+1} > a_{1,t}$, 0 — если $a_{1,t+1} = a_{1,j}$ и -1, если $a_{1,t+1} < a_{1,j}$. Получим ряд, состоящий из -1, 0 и 1:

$$\mathbf{A} = \begin{pmatrix} 0 & -1 & \dots & 1 & -1 & 0 \\ a_{2,1} & a_{2,2} & \dots & a_{2,T-1} & a_{2,T} & a_{2,T+1} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\ a_{N,1} & a_{N,2} & \dots & a_{N,T-1} & a_{N,T} & a_{N,T+1} \end{pmatrix}.$$

Глубиной лагирования Δ назовем отступ по времени. Задавая ее, мы предполагаем, что на значение в момент времени t интересующего нас ряда влияют в различной степени значения всего пучка на временном отрезке $[t - \Delta, t - 1]$. Их мы будем использовать в качестве признаков. Поставим в соответствие каждому значению исследуемого ряда $a_{1,t}$, где $t \in \{\Delta + 1, \dots T\}$ матрицу:

$$\mathbf{A}_t = \begin{pmatrix} a_{1,t-\Delta} & a_{1,t-\Delta+1} & \dots & a_{1,t-2} & a_{1,t-1} \\ a_{2,t-\Delta} & a_{2,t-\Delta+1} & \dots & a_{2,t-2} & a_{2,t-1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ a_{N,t-\Delta} & a_{N,t-\Delta+1} & \dots & a_{N,t-2} & a_{N,t-1} \end{pmatrix}.$$

Для получения строки признаков x_t векторизуем ее:

$$\mathbf{x}_t = \begin{pmatrix} a_{1,t-\Delta} & a_{2,t-\Delta} & \dots & a_{N,t-\Delta} & a_{1,t-\Delta+1} & \dots & a_{N,t-1} \end{pmatrix}.$$

Значение правильного ответа y_t для этого набора признаков — $a_{1,t}$. Теперь имеем $T - \Delta$ обучающих наборов типа объект — признак или матрицу размером $\Delta * (T - \Delta)$

$$\mathbf{X} = \begin{pmatrix} \mathbf{x}_1 & \mathbf{x}_2 & \dots & \mathbf{x}_{T-\Delta} \end{pmatrix}^T$$

и столбец ответов для нее $\mathbf{y} = \begin{pmatrix} y_1 & y_2 & \dots & y_{T-\Delta} \end{pmatrix}^T$. Введем также вектор весов как $\mathbf{w} = \begin{pmatrix} w_1 & w_2 & \dots & w_{\Delta * N} \end{pmatrix}$. Для нахождения параметров будем использовать логистическую регрессию, согласно которой

$$P(y|\mathbf{x}) = \sigma(\langle \mathbf{w}, \mathbf{x} \rangle y),$$

где $\sigma(z)$ — сигмоидная функция:

$$\sigma(z) = \frac{1}{1+\exp(-z)}.$$

Значение $z = \langle \mathbf{w}, \mathbf{x}_i \rangle y$ будем называть отступом и обозначать $M_i(\mathbf{w})$. Критерием качества модели является значение логарифма правдоподобия:

$$L(\mathbf{w}, \mathbf{X}) = \sum_{i=1}^{T-\Delta} \log\left(\sigma\left(\langle \mathbf{w}, \mathbf{x}_i \rangle y_i\right)\right) + const(\mathbf{w})$$

и задача сводится к его максимизации и нахождению неизвестного вектора параметров $\mathbf{w}$.

$$L\left(\mathbf{w}, \mathbf{X}\right) \to \max_{\mathbf{w}}.$$

Задача максимизации $L(\mathbf{w}, \mathbf{X})$ эквивалентна минимизации эмпирического риска

$$\mathbf{Q}(\mathbf{w}, \mathbf{X}) = \sum_{i=1}^{T-\Delta} \log\left(1 + \exp\left(-\langle \mathbf{w}, \mathbf{x} \rangle y_i\right)\right) \to \min_{\mathbf{w}}$$

Описание алгоритма

Для удобства перечислим вспомогательные обозначения, которые будут введены и использованы в дальнейшем:

P — число строк в матрице $\mathbf{X}$.

L — число признаков объекта, или иначе, количество столбцов матрицы $\mathbf{X}$.

m — число обектов, выбираемых из $\mathbf{X}$ для обучения.

λ — параметр шага градиентного спуска.

s — число раз восстановления регрессии и получения векторов $\mathbf{w}$ при фиксированном m, но различных разбиениях $\mathbf{X}$ и $\mathbf{y}$ на обучающие и конрольные выборки.

N — число временных рядов.

i_{max} — максимальное число шагов градиентного спуска.

δ — критерий остановки градиентного спуска

Для определенности будем прогнозировать тенденцию первого ряда. Делаем разметку, описанную в постановке задачи с одним исключением, если $a_{1,t+1} = a_{1,t}$ будем ставить вместо a_i не 0, а -1. Так как оптимальную глубину лагирования Δ экспертно задать затруднительно, проведем вычисления для различных Δ и возьмем наилучшую, сравнивая полученные результаты. Нам не интересны абсолютные значения остальных временных рядов и мы не знаем ничего о том, какой ряд влияет на интересующий нас сильнее, а какой слабее. Все их значения будут входить линейно в σ, поэтому разумно нормировать остальные ряды. Пусть a_i^{max} — максимальное значения ряда a_i. Тогда новые значения $a_{i,t}^n$ нормированных рядов будут:

$$a_{i,t}^n = \frac{a_{i,t}}{a_i^{max}}.$$

Теперь имеем пучок рядов, где первый ряд размеченный, остальные нормированны. Зафиксировав Δ, составляем матрицу признаков $\mathbf{X}$ и столбец ответов к ней $\mathbf{y}$:

$$\mathbf{X}_t = \begin{pmatrix} x_{1,1} & x_{1,2} & \dots & x_{1,L-1} & x_{1,L} \\ x_{2,1} & x_{2,2} & \dots & x_{2,L-1} & x_{2,L} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{P,1} & x_{P,2} & \dots & x_{P,L-1} & x_{T,L} \end{pmatrix}, \quad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_P \end{pmatrix}$$

где $P = T - \Delta$ — число получившихся строк признаков, а $L = N * \Delta$ — число признаков объекта y_i. Теперь разобьем матрицу $\mathbf{X}$ и столбец $\mathbf{y}$ на 2 части: обучающую и контрольную. Случайным образом выберем m строк из матрицы $\mathbf{X}$ и соответствующие этим строкам значения из столбца $\mathbf{y}$. Составим из них обучающую выборку, т.е матрицу и столбец $\mathbf{X}^{ed}, y^{ed}$. Из оставшихся составим $\mathbf{X}^{ch}, y^{ch}$ — контрольную выборку. Число m тоже является параметром и интерестно посмотреть, при каком процентном соотношении между m и общим числом объектов P мы будем получать оптимальный результат. Далее будет описан метод восстановления регрессионной модели методом градиентного спуска [2].

В качестве начального приближения зададим вектор весов модели как нулевой вектор:

$$\mathbf{w} = \begin{pmatrix} w_1 & w_2 & \dots & w_L \end{pmatrix}^T, \quad w_i = 0, \quad i \in \{1, 2, \dots, L\}$$

Тогда для шага k формула имеет вид:

$$\mathbf{w^{(k+1)}} = \mathbf{w^{(k)}} - \lambda \nabla \mathbf{Q}(\mathbf{w^{(k)}}, \mathbf{X}),$$

где λ есть параметр, влияющий на величину градиентного шага, а $\mathbf{Q}(\mathbf{w}, \mathbf{X})$ — эмпирический риск. В данной работе λ это достаточно малая константа, но для оптимизации скорости сходимости ее можно выбирать, например, по правилу Армихо [2]. Посчитаем производную сигмоидной функции:

$$\sigma'(z) = \frac{d}{dz} \frac{1}{1 + \exp(-z)} = \frac{1}{1 + \exp(-z)} \left(\frac{\exp(-z)}{1 + \exp(-z)} \right) = \sigma(z)\sigma(-z).$$

С учетом этого вектор градиента функционала записывается как:

$$\nabla \mathbf{Q}(\mathbf{w}, \mathbf{X}) = -\sum_{i=1}^{m} y_i \mathbf{x_i} \sigma\left(M_i(\mathbf{w})\right),$$

и градиентный шаг будет иметь вид:

$$\mathbf{w^{(k+1)}} = \mathbf{w^{(k)}} - \lambda \sum_{i=1}^{m} y_i \mathbf{x_i} \sigma\left(M_i(\mathbf{w}^{(k)})\right).$$

В качестве критерия остановки градиентного спуска будем использовать минимальное значение разности эмпирического риска на k-ом и $k+1$-ом шаге — δ. Если изменение значения эмпирического риска на протяжении нескольких итераций изменялось меньше чем на δ, то будем считать, что мы нашли оптимальный вектор весов $\mathbf{w}$. На случай, если значение эмпирического риска будет колебаться возле значения с амплитудой большей, чем

заданная δ, введем максимально допустимое количество итераций i_{max}. Для предотвращения переобучения будем проверять значение эмпирического риска, если на протяжении нескольких итераций оно будет возрастать, то останавливаемся.

Полученный вектор весов проверяем на $\mathbf{X}^{ch}$ и $\mathbf{y}^{ch}$. В качестве способа оценки качества полученного вектора весов $\mathbf{w}$ можно использовать, например, площадь под ROC кривой, построенной на контрольной выборке или количество ошибок, допущенных на контрольной выборке. В данной работе используется последний вариант. Так как мы разбиваем матрицу $\mathbf{X}$ и столбец $\mathbf{y}$ случайным образом, проделаем эту процедуру s раз и выберем наилучший вектор весов $\mathbf{w}$ по описанному выше критерию.

Проверка на синтетических данных

Проверим метод логистической регрессии на синтетических данных, например на зашумленных синусах и периодических зашумленных трапециях. На входе имеются $N = 7$ схожих временных рядов вида произведения синусов (см. рис.1), или трапеций (см. рис.2) одинаковой длины $T = 100$ и с одинаковым временным шагом $t = 1$.

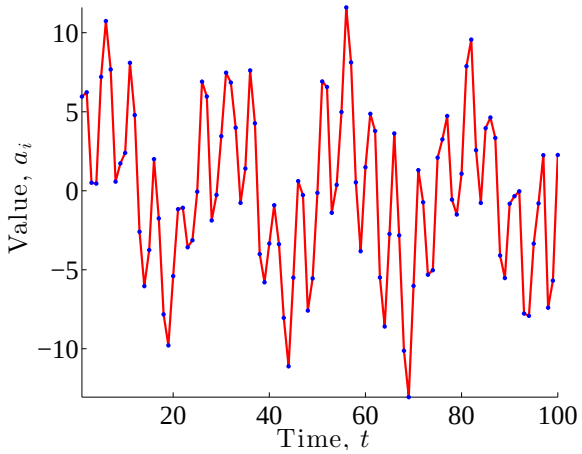

Рис. 1. Пример ряда — синус.

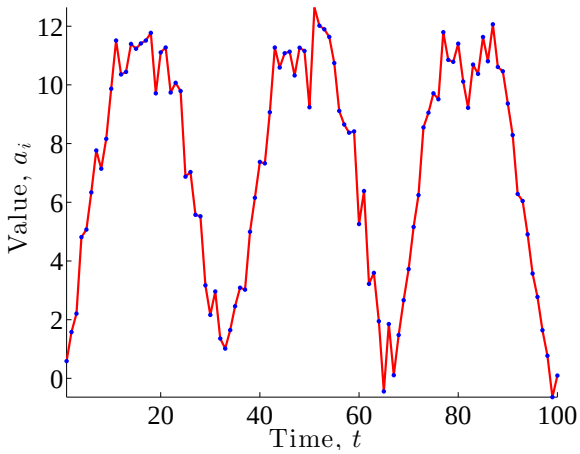

Рис. 2. Пример ряда — трапеция.

Восстановим регрессию, задав следующие параметры: $i_{max} = 1000$, $\lambda = 0.005$, $m = 70$, $\delta = 0.001$, $s = 20$. Проверим эффективность алгоритма при разной глубине лагирования Δ.

В левом столбце приведены графики для синусов, в правом для трапеций. На рис.3 и рис.4 иллюстрируется сходимость метода. По оси ординат откладывается значение эмпирического риска, а по оси абсцис — номер шага градиентного спуска. Как видно из графиков, метод градиентного спуска сходится монотонно.

Далее приведены таблицы с процентными соотношениями числа ошибок к общему числу объектов и значениями AUC для ROC кривых для различных Δ (см. таб. 1 и 2). Так же приводятся ROC кривые для разного значения параметра Δ (см. рис.5 – 12). Синим цветом отображаюстя ROC кривые, построенные по контрольной выборке, красным — по обучающей. Зеленым цветом проведена кривая, соответствующая случайному предсказанию (исходы $+1$ и -1 полагаются равновероятными).

Как можно видеть из результатов, при увеличении параметра Δ качество модели заметно улучшается, затем при некотором значении Δ_{opt} мы имеем минимальный процент ошибок, максимальное значение AUC для контрольной выборки и достаточно высокое

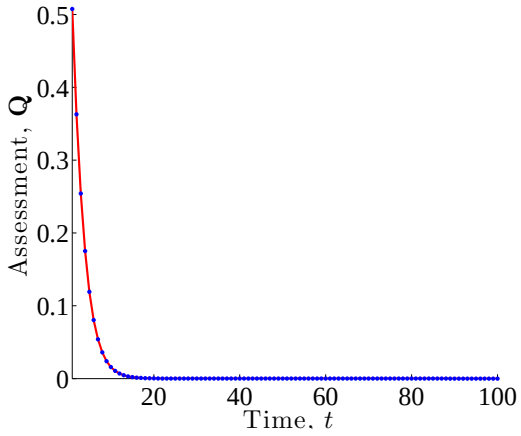

Рис. 3. Зависимость **Q** от номера градиентного шага, синусы.

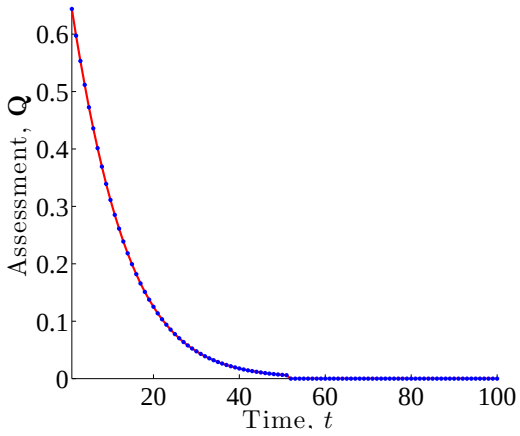

Рис. 4. Зависимость **Q** от номера градиентного шага, трапеции.

Δ	процент ошибок	AUC для обучающей выборки	AUC для контрольной выборки
2	21	0.8253	0.8773
4	16	0.8962	0.9026
6	12	0.9676	0.9697
8	13	0,9588	0.9388

Таблица 1. Результаты для произведения синусов

значение AUC для обучающей. При дальнейшем росте Δ количество ошибок начинает увеличиваться, а значение AUC для контрольной выборки уменьшаться. Тем не менее значение AUC для обучающей выборки продолжает расти. Это объясняется тем, что при увеличении Δ мы увеличиваем количество признаков, и поэтому точнее можем обучиться. Но после экстремального значения Δ_{opt} мы начинаем переобучаться и поэтому на контрольной выборке делаем больше ошибок. Показатели AUC на контрольных выборках при $\Delta \leqslant \Delta_{opt}$ оказываются явно лучше, чем на обучающих. Это связано с тем, что мы выбираем тот вектор весов $\mathbf{w}$, при котором мы делаем наименьшее число ошибок на контрольной выборке.

Δ	процент ошибок	AUC для обучающей выборки	AUC для контрольной выборки
2	39	0.6669	0.7222
4	25	0.7239	0.7306
6	38	0.7108	0.7048
8	42	0,7680	0.5865

Таблица 2. Результаты для трапеций

Стоит отметить, что для синусоидальных данных алгоритм работает на порядок лучше. Сравним результаты (см. таб. 3) при оптимальном значении $\Delta_{opt} = 8$ для синусов и $\Delta_{opt} = 6$ для трапеций. Обучаясь по 70% объектам, мы имеем процент ошибок для трапеций вдвое больший, чем для синусов, а значения AUC явно ниже.

тип ряда	процент ошибок	AUC на обучении	AUC на контроле
синусы	12	0.9676	0.9697
трапеции	25	0.7239	0.7306

Таблица 3. Сравнение результатов

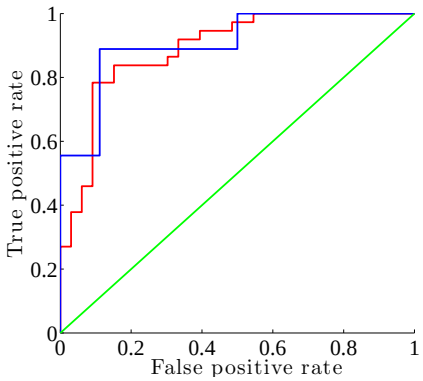

Рис. 5. ROC кривая, синусы, $\Delta = 2$.

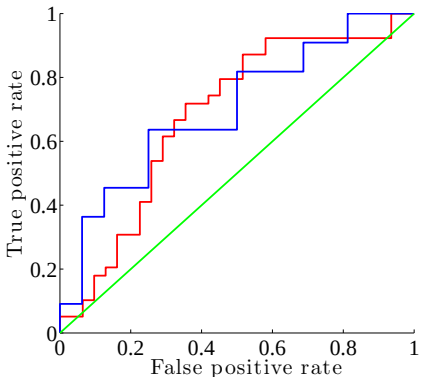

Рис. 6. ROC кривая, трапеции, $\Delta = 2$.

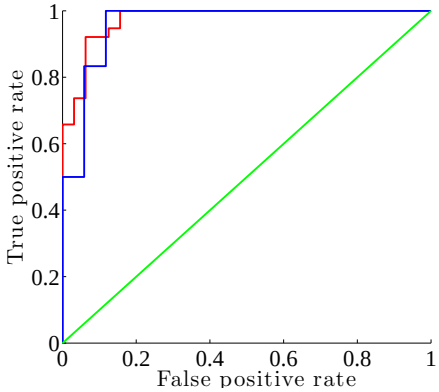

Рис. 7. ROC кривая, синусы, $\Delta = 6$.

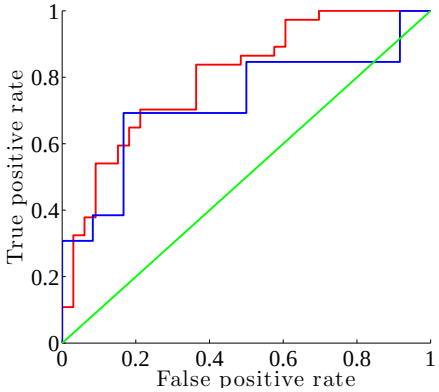

Рис. 8. ROC кривая, трапеции, $\Delta = 4$.

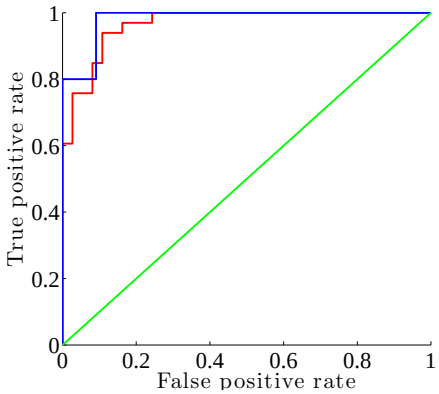

Рис. 9. ROC кривая, синусы, $\Delta = 8$.

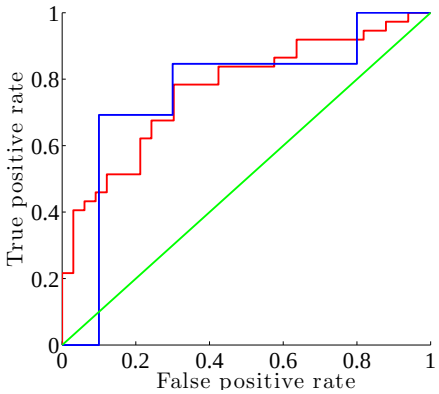

Рис. 10. ROC кривая, трапеции, $\Delta = 6$.

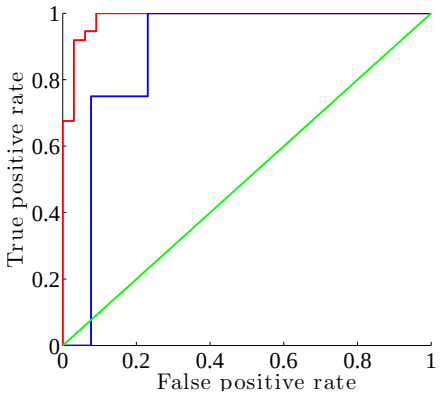

Рис. 11. ROC кривая, синусы, $\Delta = 12$.

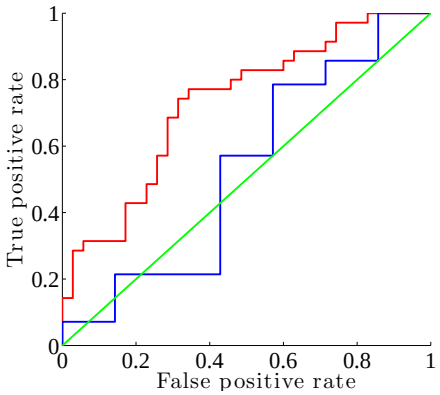

Рис. 12. ROC кривая, трапеции, $\Delta = 8$.

Прогноз потребления электроэнергии

Проверим работоспособность алгоритма на данных о потребленииии электроэнергии с 01.01.08 по 25.04.08. Данные имеют вид, приведенный в таб.4. Количество строк в таблице — 2767. Связь между этими временными рядами явно прослеживается. Например, количество потребляемой энергии явно зависит от температуры, так как при низкой температуре люди начинают использовать электронагреватели, а при высокой — кондиционеры, потребляемая мощность которых довольно высокая. От времени суток зависит количество электроэнергии, тратящееся на освещения, а от дня недели — количество людей на работе и дома, что в свою очередь влияет на количество работающей аппаратуры. Исходя из всего вышеперечисленного, будем считать, что все эти ряды довольно сильно коррелируют между собой. Поэтому целесообразно исследовать их всех вместе, как временной пучок. Каждая строчка в наших обозначениях будет соответствовать интервалу $t = 1$, тогда длина этих рядов будет $T = 2767$. Приведем зависимость значения потребления электроэнергии от времени (см. рис. 14) для небольшого временного интервала.

потребление электроэнергии в МВт*ч	дата	день недели	час	температура
1366,74115	01.01.08	2	0:00	-11,9
1333,16888	01.01.08	2	1:00	-12,0
1293,20544	01.01.08	2	2:00	-12,0
1302,07739	01.01.08	2	3:00	-12,0
...	...	...	...	...

Таблица 4. Вид данных о потреблении электроэнергии

Как можно заметить из рис.14, данные имеют вид синусоид. Поэтому основываясь на результатах для синтетических синусоид, можно предположить, что результаты для этих данных также будут хорошими.

Мы будем прогнозировать тенденцию потребления электроэнергии (увеличится оно или уменьшится в следующий момент времени) т.е. первый ряд таблицы. Как и для синтетических данных, размечаем его, а затем строим матрицу $\mathbf{X}$ и столбец $\mathbf{y}$. Восстановим регрессию, задав следующие параметры: $i_{max} = 1000$, $\lambda = 0.001$, $m = 70\%$, $\delta = 0.001$, $s = 10$. Величину Δ будем изменять от 1 до 60, с шагом 1 и найдем оптимальное значение для наших данных. Оценивать качество модели при заданном Δ будем по трем парамет-

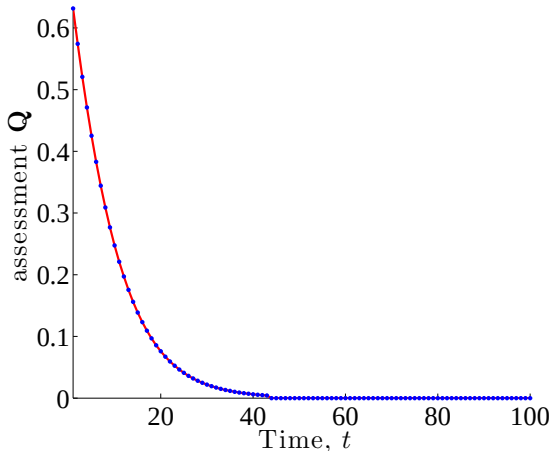

Рис. 13. Зависимость **Q** от номера градиентного шага, синусы.

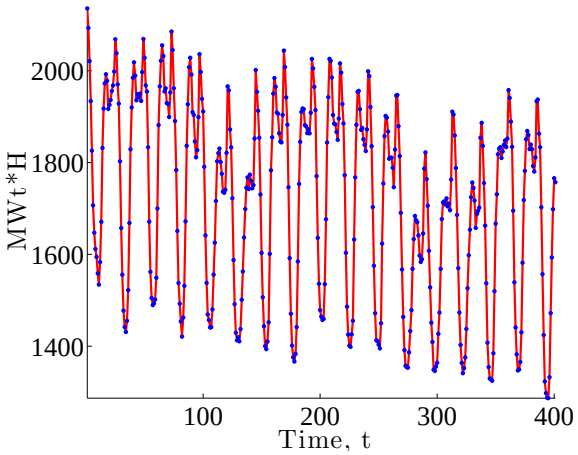

Рис. 14. Зависимость потребления электроэнергии от времени

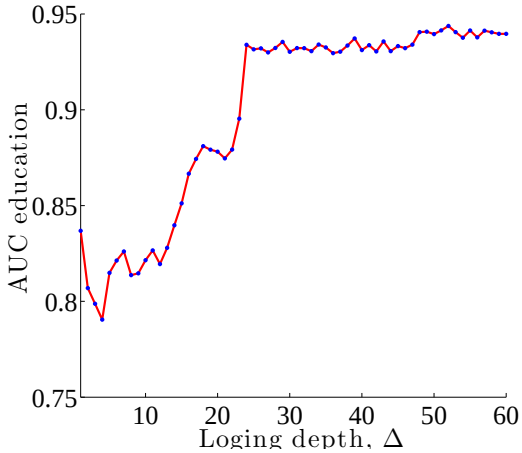

Рис. 15. Зависимость AUC на обучающей выборке от Δ.

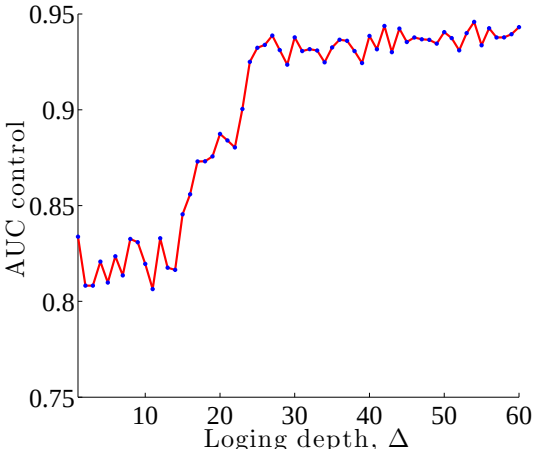

Рис. 16. Зависимость AUC на контрольной выборке от Δ.

рам: AUC для ROC кривой на обучающей выборке, AUC для ROC кривой на контрольной выборке и процент ошибок от общего числа объектов. На рисунках (16, 15, 17) показаны зависимости этих параметров от Δ. Можно заметить, что начиная с $\Delta = 24$ значения выходят на плато и практически не изменяются. Число ошибок медленно уменьшается, но это скорее связано с переобученностью, чем с улучшением качества алгоритма. А так как при увеличении Δ время восстановления и объем обрабатываемых данных сильно увеличивается, разумно взять $\Delta_{opt} = 25$, т.к. дальнейшее увеличение не даст нам значительного выигрыша.

Значение $\Delta_{opt} = 24$ можно оъяснить и с логической точки зрения. Так как в сутках 24 часа (а данные повторяются периодически с периодом в сутки), шаг по времени наших данных есть $t = 1$час, то наиболее полную информацию о таком ряде мы будем получать, зная его предысторию за предшествующий период, т.е 24 часа или 24 шага по времени.

Литература

[1] *К.В.Воронцов.* Лекции по линейным алгоритмам классификации. / К.В.Воронцов.

Δ	процент ошибок	AUC для обучающей выборки	AUC для контрольной выборки
25	14.1	0.9276	0.9398

Таблица 5. Результаты прогноза потребления энергии для Δ_{opt}

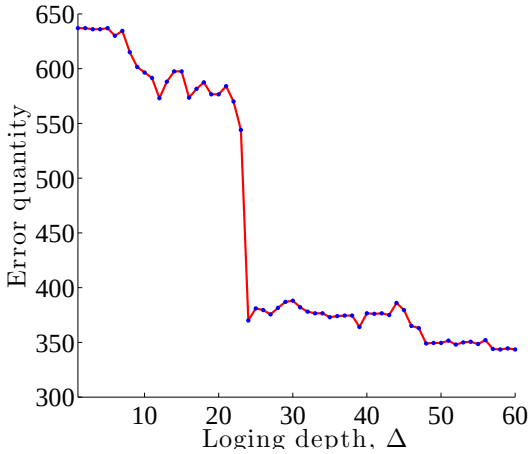

Рис. 17. Зависимость количества ошибок на данных от Δ.

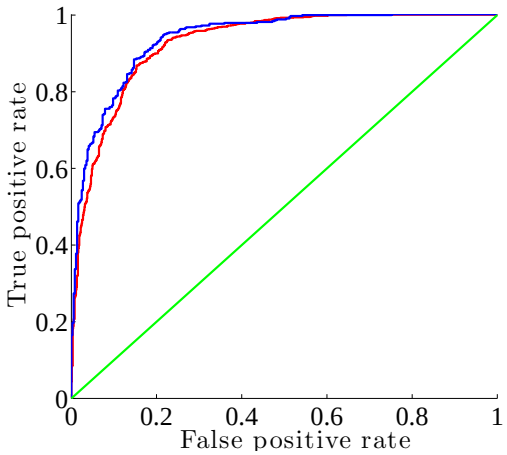

Рис. 18. ROC кривые при оптимальном Δ.

[2] *В.Г.Жадан.* Численные методы решения задач оптимизации / В.Г.Жадан.

[3] *Н.В.Филипенков.* — О задачах анализа пучков временных рядов с изменяющимися закономерностями. — Master's thesis, 2006.

[4] *Н.В.Филипенков.* Об алгоритмах прогнозирования процессов с плавно меняющимися закономерностями: Ph.D. thesis. — 2010.

[5] *Б.А.Романенко.* Событийное моделирование и прогноз финансовых временных рядов / Б.А.Романенко. — 2011.

[6] *Ng, A.* Classification and logistic regression / A. Ng.