МЕТОДЫ ПОВЫШЕНИЯ И ОЦЕНКИ КАЧЕСТВА ОБУЧАЮЩЕЙ ВЫБОРКИ ДЛЯ ЗАДАЧ НЕЙРОСЕТЕВОГО ПРОГНОЗИРОВАНИЯ ВРЕМЕННЫХ РЯДОВ

Хмелевой С.В.
Донецкий национальный технический университет, г. Донецк
кафедра автоматизированных систем управления
E-mail: hmelevoy_sergey@ukr.net

Abstract
Khmilovyy S.V. Methods of increase and estimation of sample's quality for neural forecasting task. *In issue the methods of increasing sample quality for neural forecasting task are submitted. Parameters of a data quality estimation – repeatability and discrepancy is introduced. Effectiveness of ARMA and ARIMA models is compared. The method of increase of sample quality - reduction the data to a uniform kind is investigated.*

Введение

Предварительная подготовка данных является неотъемлемой частью любого процесса прогнозирования. Поэтому актуальность данной задачи всегда высока. В данной работе показана эффективность различных представлений данных временных рядов (AR, MA, ARMA, ARIMA) и их применимость к различным типам задач.

Введены критерии оценки качества обучающей выборки (ОВ) — повторяемость и противоречивость. Понятие противоречивости модифицировано для более общей оценки качества ОВ. Для того, чтобы модель могла строить достоверный прогноз, необходимо стремиться к тому, чтобы количество обучающих наборов в различных классах было соизмеримо, т.е. ОВ должна быть равномерной и выполнялся некоторый баланс. В статье исследовано влияние преобразования данных к равномерному виду на точность прогнозирования.

Эффективность различных представлений данных временных рядов показана в [1], [2], [3]. Используемые в работе понятия повторяемости и противоречивости приведены в [4], равномерности выборки — в [5].

Целью статьи является описание и модификация некоторых аспектов предварительной подготовки данных применительно к задачам нейросетевого прогнозирования временных рядов.

Эффективность различных представлений данных временных рядов

Чаще всего для прогнозирования временных рядов используются модели данных, называемые ARMA (autoregressive moving average), и ARIMA (autoregressive integrated moving average).

Авторегрессия (AR) содержит элементы, которые последовательно зависят друг от друга. Такую зависимость можно выразить следующим уравнением:

$$X_t = \xi + \phi_1 \cdot X_{t-1} + \phi_2 \cdot X_{t-2} + \ldots + \varepsilon, \quad (1)$$

где ξ — константа (свободный член),

ϕ_1, ϕ_2 — параметры авторегрессии.

Скользящее среднее (MA) представляет случай, где каждый элемент ряда подвержен суммарному воздействию предыдущих ошибок. В общем виде это можно записать следующим образом:

$$X_t = \mu + \varepsilon_t - \theta_1 \cdot \varepsilon_{t-1} - \theta_2 \cdot \varepsilon_{t-2} - \ldots, \quad (2)$$

где μ — константа,

θ_1, θ_2 — параметры скользящего среднего,

ε — ошибка.

Общая ARMA модель представляет из себя сумму AR и MA моделей, которая объединяет формулы (1) и (2).

$$X_t = \xi + \phi_1 \cdot X_{t-1} + \phi_2 \cdot X_{t-2} + \ldots + \varepsilon_t - \theta_1 \cdot \varepsilon_{t-1} - \theta_2 \cdot \varepsilon_{t-2} - \ldots \quad (3)$$

ARIMA модель отличается от ARMA модели тем, что в качестве ряда используются разность между элементами обычного ряда, используемого в ARMA модели (производится дифференцирование ряда). Переход к ARIMA модели необходим в случае, когда данные имеют нестационарность (например, если данные имеют циклический характер). Хотя иногда для того, чтобы убрать нестационарность, недостаточно и двойного дифференцирования.

Нейронная сеть (НС), базирующаяся на AR модели порядка p, использует в качестве входов кроме текущего значения временного ряда еще p входов, на которые поступают значения лагов входного ряда. НС, базирующаяся на MA модели порядка q, использует в качестве входов еще и q входов, на которые поступают значения ошибки прогнозирования ряда на текущем, прошлом шаге (до q-го шага прогнозирования). НС, базирующаяся на ARMA модели, имеет p входов AR модели и q входов MA модели. НС, базирующаяся на ARIMA модели, имеет ту же структуру, что и ARMA-НС, но в качестве входных значений используются проинтегрированные значения временного ряда (ВР).

Преобразование данных для увеличения различимости классов

Равномерность данных в ОВ (равномерное распределение по классам) в природе встречается достаточно редко, чаще всего используется нормальное распределение объектов по классам. В процессе нормализации данные приводятся к определенному допустимому диапазону (чаще всего [0..1] или [–1..1]). Следствием нормализации является то, что при сильной неравномерности закона распределения допустимый диапазон используется не полностью. В нем присутствуют как слабо заполненные участки, так и участки скученности значений исходной величины. Слабо заполненные участки приводят к тому, что в процессе обучения НС плохо «запоминает» эти значения. А участки скученности, где на относительно небольших отрезках располагается значительное количество значений исходной величины, оказываются слабо различимыми, что приводит к снижению качества обучения.

При решении этой проблемы можно идти в двух направлениях:

1. повышение чувствительности НС за счет изменения параметров функции активации;
2. повышение равномерности распределения исходной величины.

Реализация первого подхода [6] влечет за собой вмешательство в синтез НС (изменение функции ошибки сети, алгоритма обучения и т.д.), что сложнее в реализации.

Преимущество второго подхода [5] заключается прежде всего в его простоте, т.к. такое преобразование исходных данных фактически является перекодировкой, повышающей информативность. Значения ОВ необходимо наиболее равномерно перераспределить по интервалу значений исходных данных: в областях скученности данных — «растянуть», в «пустых» местах — «сжать». Преобразование значений исходной величины x_i выполняется в соответствии с плотностью их распределения p(x) по диапазону (рис.1).

Значение преобразованной величины x_i' (для подачи на вход НС) вычисляются на основании значения исходной величины x_i по формуле (4).

$$x_i' = S(x_i), \quad (4)$$

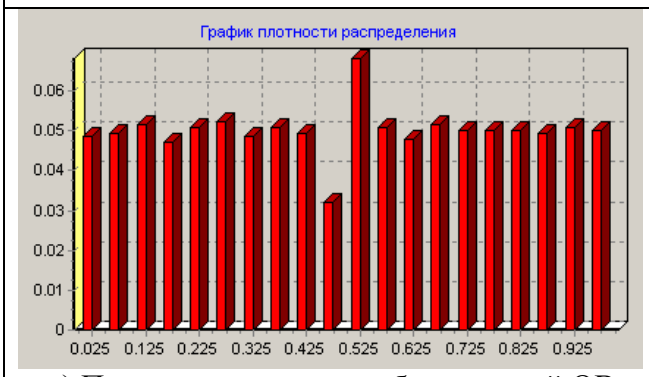
Физически величине x_i' соответствует площадь $S(x_i)$ фигуры, ограниченная значением $x_1=0$ и $x_2=x_i$, т.е. с учетом всех предыдущих значений x.

$S(x_i) = P(X < x_i)$, где P — интегральная вероятность значений исходной величины x (плотность распределения исходной величины x). Т.е. фактически S(x) представляет функцию дифференциальной вероятности значений исходной величины x (функцию распределения исходной величины x).

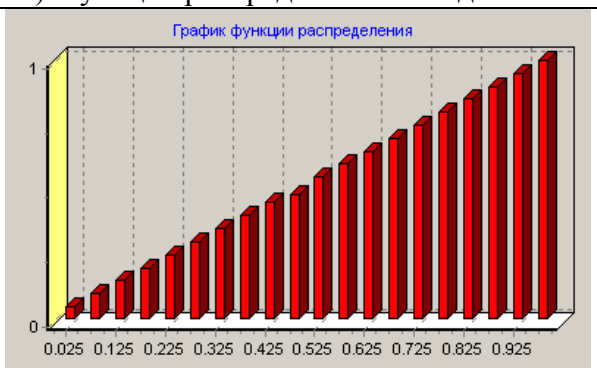
Для восстановления исходной неравномерности распределения по диапазону над выходными величинами НС производится обратная обработка. Вычисляется значение x_i, которому соответствует значение P_i.

а) Плотность распределения исходной ОВ

б) Функция распределения исходной ОВ



в) Плотность распр. преобразованной ОВ



г) Функция распр. преобразованной ОВ

Рисунок 1 — Плотности и функции распределения исходной и преобразованной ОВ

Предварительная оценка качества обучающей выборки

Для оценки качества ОВ необходим критерий, характеризующий сходство образов внутри каждого из классов ОВ, и показатель, характеризующий степень внутренней противоречивости ОВ, характеризующий сходство образов в разных классах. Он должен иметь невысокую трудоемкость расчета. В [4] для этих целей предложено использовать такие критерии оценки качества ОВ, как повторяемость и противоречивость.

Противоречивыми можно считать наборы, описывающие одинаковые ситуации, но принадлежащие к разным классам. Поскольку каждая ситуация описывается не дискретными значениями, а действительными числами, нельзя говорить о точном совпадении ситуаций. Для снижения трудоемкости расчетов можно перейти от описания входного вектора в виде значений ВР к его описанию в виде номеров классов, к которым принадлежат соответствующие значения ВР. Вместо расчета расстояний в таком случае производится покоординатное сравнение векторов.

Для двух наборов: $A=\{a_1,a_2,...,a_n,c_a\}$ и $B=\{b_1,b_2,...,b_n,c_b\}$, где $a_i,b_i(i\in 1;n)$ — описывает распознаваемую ситуацию в терминах временного ряда, c_a,c_b — признак распознаваемого образа. Переходим к векторам $A'=\{a'_1,a'_2,...,a'_n,c'_a\}$ и $B'=\{b'_1,b'_2,...,b'_n,c'_b\}$, где $a'_i,b'_i(i\in 1;n)$ — номер класса, соответствующего значению прогнозируемой величины $a_i,b_i(i\in 1;n)$, c'_a,c'_b — номер класса распознаваемого образа.

Наборы А и В считаются повторяющимися, если $a'_i=b'_i(i\in 1;n)$ и $c'_a\neq c'_b$.

Повторяемость наборов для класса c_i

$$\rho_i = \frac{n_i^{\rho}}{n_i^{c}}, \tag{5}$$

где n_i^{ρ} — число повторяющихся наборов в классе i;

n_i^{c} — общее число наборов в классе i.

Повторяемость ОВ

$$\rho_l = \frac{1}{n_c}\sum_{i=1}^{n_c}\rho_i, \tag{6}$$

где n_c — общее число классов в ОВ.

Наборы A и B считаются противоречивыми, если $a_i' = b_i'(i \in 1;n)$ и $c_a' \neq c_b'$.

Противоречивость наборов A и B

$$\delta_{ab} = \left|\frac{c_a - c_b}{n_c - 1}\right|, \tag{7}$$

где δ_{ab} — противоречивость наборов A и B;

n_c — общее число классов в ОВ.

Противоречивость ОВ

$$\delta_l = \frac{1}{n_l}\sum_{i=1}^{n_l}\sum_{j=1}^{n_l}\delta_{ij}, \tag{8}$$

где n_l — число наборов в ОВ.

Ввод понятий повторяемости и противоречивости дает мощное средство, которое позволяет определить качество ОВ, а следовательно и успешность обучения НС до его проведения. Качество выборки и успешность обучения определяются значениями параметров повторяемости и противоречивости и их сочетанием.

Расчет повторяемости и противоречивости по данным формулам производить несколько затруднительно, поскольку обычно значения ВР не сгруппированы по классам. Поэтому повторяемость ОВ предлагается считать не по классам, а по элементам ОВ.

Повторяемость ОВ. В этом случае формула (6) немного изменяется и принимает вид:

$$\rho_l = \frac{1}{n}\sum_{i=1}^{n}\rho_i, \tag{9}$$

где n — общее число элементов в ОВ.

Данное понятие противоречивости не учитывает того, что возможна различная противоречивость внутри одного класса. Функция противоречивости должна зависеть от количества точек в классе и от того, как взаимно распределены элементы в классе. Возможны следующие распределения для максимальной противоречивости данных в классе:

- Значения признаков образов делятся на 2 одинаковые группы и принимают взаимно противоположные граничные возможные значения.
- Значения признаков образов равномерно «размазываются» в диапазоне возможный значений.

И в том и в другом случае противоречивость данных в классе будет максимальной.

Минимальной противоречивость данных в классе будет в том случае, когда все признаки образов принимают одинаковое значение.

Допустим, что количество элементов в классе i = n_i, максимальный номер класса=1 и минимальный номер класса =0.

В случае, если значения признаков образов делятся на 2 одинаковые группы и принимают взаимно противоположные граничные возможные значения, расстояние между значениями признаков образов в классе i равно

$$S_i = \frac{n_i * 1}{2} - \frac{n_i * 0}{2} = \frac{n_i}{2}. \quad (10)$$

В случае, если значения признаков образов равномерно «размазываются» в диапазоне возможный значений, значения признаков образов элементов класса являются равномерно распределенной дискретной величиной, имеющие номера от 0 до $n_i - 1$. Допустим, значение признака объекта равняется его номеру. Такое допущение возможно, поскольку условие равномерности распределения значений признаков образов не нарушается. В этом случае могут лишь поменяться номера элементов в классе и увеличиться их значения признаков образов в $n_i - 1$ раз. В результате этого допущения перейдем к дискретному множеству значений признаков образов имеющему диапазон от 0 до $n_i - 1$, что необходимо в дальнейшем.

В таком случае сумма расстояний между 0-й точкой и остальными $n_i - 1$ точками класса равна $\frac{n_i * (n_i - 1)}{2}$ или количеству комбинаций из n_i элементов по 2 — $C_{n_i}^2$. В общем случае расстояние между i-й точкой и остальными $n_i - 1$ точками можно представить как площадь двух треугольников, представленных на рис.2.

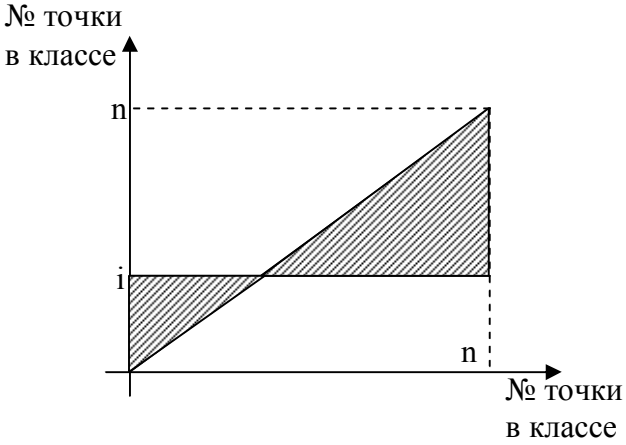


Рисунок 2 — Суммарное расстояние между i-й и остальными точками класса

В аналитическом виде с учетом дискретности пространства площадь треугольников может быть представлена как [7]:

$$C_{n_i - i}^2 = \frac{(n_i - i) * (n_i - i - 1)}{2} \quad (11)$$

– площадь верхнего треугольника, и

$$C_i^2 = \frac{i * (i - 1)}{2} \quad (12)$$

– площадь нижнего треугольника.

Соответственно, суммарное расстояние между i-й точкой и остальными точками класса можно определить как:

$$S_i = C_{n_i - i}^2 + C_i^2, \quad (13)$$

а общее расстояние между всеми точками класса — как

$$S = \sum_{i=0}^{n_i} S_i \,. \tag{14}$$

Общее расстояние между всеми точками класса можно проиллюстрировать рис.3. На рисунке можно видеть две кривыем (выделены жирным). Одна из них — суммарное расстояние между i-й точкой и точками (0..i–1) класса, а вторая — суммарное расстояние между i-й точкой и точками (i+1..n) данного класса. Заштрихованная фигура показывает суммарное общее расстояние между всеми элементами в классе.

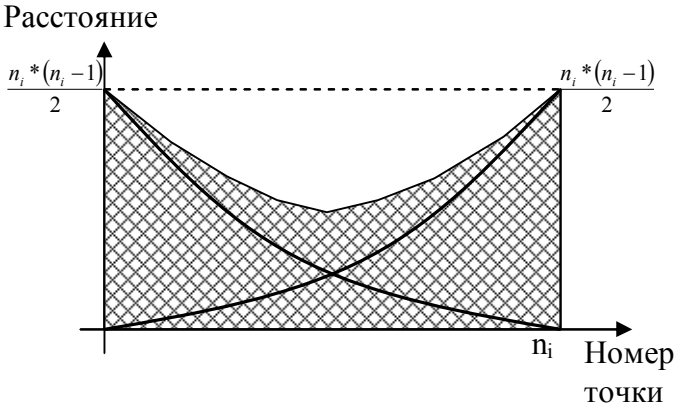


Рисунок 3 — Общее расстояние между всеми точками класса

Для перехода к диапазону [0..1] значений признаков объекта достаточно разделить формулы (10) и (11) на $n_i - 1$. Тогда формулы (13) и (14) принимают вид:

$$S_i' = \frac{C_{n_i-i}^2 + C_i^2}{n_i - 1}, \tag{15}$$

$$\text{и} \quad S' = \sum_{i=0}^{n_i} S_i' \,. \tag{16}$$

S_i' является максимальной мерой расстояния между значениями признаков объектов внутри одного класса. Для перехода от величины S_i' к величине противоречивости необходимо найти, согласно формуле (3), отношение действительного расстояния между значениями признаков объектов к их максимальному значению.

$$\delta_i = \frac{\sum_{j=1}^{n_c} \left| c_i - c_j \right|}{S_i} \,. \tag{17}$$

Формула (4) принимает вид:

$$\delta_l = \frac{1}{n_l} \sum_{i=1}^{n_l} \delta_i \,. \tag{18}$$

Для практического расчета величины противоречивости удобнее рассчитывать величину противоречивости для каждой точки отдельно, пользуясь формулами (15) и (17), т.к. для данной точки легко найти точки из того же класса.

Формулы (17) и (18) выведены для случая равномерного распределения значений признаков объектов в пределах от 0 до 1. Покажем, что данные формулы справедливы и для случая, когда значения признаков образов делятся на 2 одинаковые группы и принимают взаимно противоположные граничные возможные значения (0 и 1)

В случае i=0 S_i' принимает значение

$$\frac{\frac{(n_i-0)*(n_i-0-1)}{2}+\frac{0*(0-1)}{2}}{n_i-1}=\frac{n_i}{2}, \tag{19}$$

а в случае i= n_i S_i' также принимает значение

$$\frac{\frac{(n_i-n_i)*(n_i-n_i-1)}{2}+\frac{n_i*(n_i-1)}{2}}{n_i-1}=\frac{n_i}{2}, \tag{20}$$

что соответствует формуле (9).

Практические результаты

Сравнение ARMA и ARIMA моделей. Описание задачи приведено в предыдущей работе [8], там же есть список входных факторов, используемых при прогнозировании. Среди 20 входных факторов есть и временные лаги входных переменных. При изменении структуры сети введены дополнительные входы сети, на которые поданы ошибки, взятые с выхода НС. Можно сказать, что при прогнозировании использовалась модель ARMA(10,1). Было произведено сравнение результатов прогнозирования на обычных нормализованных к диапазону [0..1] данных с теми же данными, но проинтегрированными. Во втором случае используемая при прогнозировании модель – ARIMA(10,1,1). Результаты приведены в таблице 1. Однозначно результаты, полученные с помощью ARIMA модели точнее (на 47–67%), чем полученные с помощью аналогичной ARMA модели. Это объясняется тем, что в данных присутствует циклическая компонента с периодом 24 часа, а сами временные ряды имеют длину намного меньшую. Хотя существует и сезонная ARIMA-модель, но она учитывает влияние сезонности на рядах с длиной намного больших, чем период циклической компоненты. Т.О., в данном случае дает лучшие результаты именно интегрирование.

Таблица 1 — Результаты исследований

	Ошибка,%	Прот.	Повт.	Ошибка,%	Прот.	Повт.
	ARMA			ARMA, равномерный вид		
ОВ, 1 задача	4.44	0.002	0.051	4.682	0	0
ОВ, 2 задача	7.098	0.004	0.027	6.998	0	0
ТВ, 1 задача	5.02	0.008	0.16	5.936	0	0
ТВ, 2 задача	8.772	0.015	0.097	8.81	0	0
	ARIMA			ARIMA, равномерный вид		
ОВ, 1 задача	2.19	0.027	0.386	1.624	0	0
ОВ, 2 задача	3.692	0.06	0.202	3.132	0	0
ТВ, 1 задача	2.458	0.018	0.201	1.948	0	0
ТВ, 2 задача	4.212	0.036	0.096	3.764	0	0
	Упрощенное представление			Упрощенное предст., равн. вид		
ОВ, 1 задача	1.914	0.103	0.583	1.762	0.873	0.06
ОВ, 2 задача	4.05	0.234	0.271	3.918	0.926	0.052
ТВ, 1 задача	2.056	0.112	0.549	1.86	0.877	0.063
ТВ, 2 задача	4.394	0.247	0.26	4.352	0.896	0.052

Преобразование данных к равномерному виду. Было выполнено сравнение результатов, полученных на обычных и преобразованных к равномерному виду данных. Для ARIMA модели, получено улучшение результатов на 10–25%. Для ARMA модели преобразование дало ухудшение. Анализ стационарности ОВ и ТВ объясняет это. Как уже показано выше, в данных присутствует циклическая компонента. Поэтому данные из ОВ и ТВ являются взаимно нестационарными. Отсюда неправильное преобразование ТВ к равномерному виду, поскольку для преобразования использовалась плотность вероятности ОВ. Как уже было показано, интегрирование убирает влияние циклической компоненты, так что для ARIMA-модели ОВ и ТВ стационарны.

Применение равномерности и противоречивости как индикаторов качества выборки. Применение интегрирования дало, как и следовало ожидать, увеличение и повторяемости, и противоречивости. В [4] описано, что оптимальные параметры для прогнозирования: противоречивость=[0;0.2], повторяемость=[0.4;0.7]. Исходя из этого, ARIMA модель использовала близкие к оптимальным параметры обучающей выборки. Однако приведение выборки к равномерному виду дало неожиданные результаты: при увеличении точности прогнозирования повторяемость и противоречивость стали очень близки к 0. Это можно объяснить тем, что после приведения выборки к равномерному виду, поскольку приведение каждого влияющего фактора выполнялось независимо друг от друга, ситуации, ранее принадлежащие к одному классу, стали принадлежать к различным. Таким образом, благодаря большому количеству влияющих факторов количество данных в ОВ и ТВ является недостаточным для нормального разделения данных на классы. Пока распределение данных было близким к нормальному закону, некоторые классы обладали достаточным количеством данных для расчета повторяемости и противоречивости. Приведение данных к равномерному виду убрало места «сгущения» данных, и количество данных во всех классах стало малым. Чтобы проверить это, было выполнено прогнозирование на упрощенной выборке с 1 влияющим фактором. Результатом стало резкое повышение противоречивости при уменьшении повторяемости. Точность прогнозирования при этом повысилась. Можно сделать вывод о необходимости дальнейшего исследования этих параметров как индикаторов качества обучающей выборки.

Выводы

В результате проведенных исследований можно сделать следующие выводы:

- Доказана необходимость интегрирования данных в случае наличия циклической компоненты при длине ряда меньшей, чем период этой компоненты
- Проверено влияние равномерности плотности распределения ОВ и ТВ на точность прогнозирования. Точность увеличивается при стационарности ТВ относительно ОВ.
- Модификация параметра противоречивости позволяет более гибко учитывать неоднородности данных внутри одного класса. Кроме разделения данных на 2 группы учитывается также равномерное «размазывание» данных внутри одного класса.
- Не доказан факт того, что повторяемость и противоречивость являются индикаторами оценки качества выборки. Необходимы дальнейшие исследования.

Литература

1. Alexessander da Silva Couto Alves. Internet traffic Engineering. An Artifical Intelligence Approach. // http://paginas.fe.up.pt/~alx/downloads/alves04internet.pdf
2. Nancy K. Groschwitz, George C. Polyzos. A Time Series Model of Long-Term NSFNET Backbone Traffic. // http://www.caida.org/outreach/papers/1994/tsm
3. Dudul S.V., Ghatol A.A. Application of multilayer perceptron neural network in distant predictions of a typical nonstationary time series //http://journal-ci.csse.monash.edu.au/edit/uploads/vasant01/internet_ci.doc
4. Тарасенко Р.А., Крисилов В.А. Предварительная оценка качества обучающей выборки для нейронных сетей в задачах прогнозирования временных рядов. — //Тр. Одесского политехн. ун-та. — Одесса, 2001. — Вып. 1.
5. Крисилов В.А., Кондратюк А.В. Преобразование входных данных нейросети с целью улучшения их различимости. // http://neuroschool.narod.ru
6. Родионов П.Е. Краткосрочное прогнозирование котировок ОГВВЗ с использованием аппарата нейронных сетей. Сборник «Интеллектуальные технологии и системы» под ред. Ю.Н.Филипповича; Изд-во МГТУ им.Баумана, Москва, 1998г.
7. Чернова Н.И. Теория вероятностей. — // http://text.marsu.ru/books_edu/11/lec.html
8. Хмелевой С.В., Скобцов Ю.А., Панченко З.В. Некоторые аспекты предварительной обработки данных в задачах нейросетевого прогнозирования и классификации.