

Экспериментальное исследование методов прогнозирования, базирующихся на алгоритмах сжатия данных *

Б.Я. Рябко, В.А. Монарев

Предлагается и экспериментально исследуется метод построения алгоритмов прогноза на основе методов сжатия данных (или так называемых архиваторов). На примере задачи прогноза курсов валют показано, что получаемые методы обладают относительно высокой точностью.

1 Введение

В теории информации известна глубокая взаимная связь между случайностью, или предсказуемостью, некоторой последовательности символов и возможной степенью ее "сжатия", открытая А.Н. Колмогоровым [1] и развитая его учениками и коллегами [2, 3]. В конце 80-ых годов, после открытия оптимальных универсальных кодов, близкие идеи были использованы при построении оптимальных методов прогноза для источников, порождающих символы из конечных алфавитов [4]. В дальнейшем подход из [4] был использован для построения асимптотически оптимальных методов прогнозирования для довольно многочисленных классов случайных процессов [5, 6, 7, 8]. Однако большая часть этих методов хотя и обладает хорошим асимптотическим поведением, мало применима к решению реальных задач прогнозирования при небольших объемах наблюдений.

В данной заметке делается попытка использовать алгоритмы сжатия данных для построения практически применимых методов прогноза. В качестве модельной выбрана задача прогнозирования курса валют (доллар США/евро и доллар США/рубль), представляющая и некоторый практический интерес, а в качестве методов сжатия данных используются практически применяемые архиваторы. Здесь стоит отметить, что эти компьютерные программы базируются как на многочисленных глубоких результатах и конструкциях теории информации, так и на практическом опыте

*Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (грант 03-01-00495) и INTAS (Grant 00-738).

и изобретательности многочисленных разработчиков архиваторов. Стоит также отметить, что архиваторы оказались довольно эффективным средством при распознавании авторства текстов [9] и находят применение при оценивании близости генетических последовательностей.

Приведенные в работе экспериментальные результаты показывают, что методы прогноза, построенные на основе архиваторов, обладают довольно высокой точностью.

2 Описание построения методов прогноза по алгоритмам сжатия данных

Сначала введём необходимые обозначения. Пусть A – некоторое множество и $\{X_i\}_{i=0}^{\infty}$ – случайный процесс, принимающий значения из A .

Рассматривается следующая задача прогнозирования: по известным значениям X_0, \dots, X_{n-1} предсказать среднее значение процесса в момент времени n , то есть $E(X_n | X_0, \dots, X_{n-1})$, в случае, когда статистические характеристики процесса заранее не известны. Отметим, что эта постановка задачи является одной из наиболее распространенных, см. [6 - 10].

Приведем теперь необходимые сведения о неискажающем кодировании сообщений. Пусть D некоторый конечный алфавит, D^n – множество всех слов длины $n, n \geq 1, D^* = \bigcup_{n=1}^{\infty} D^n$. Отображение $\phi : D^n \rightarrow \{0, 1\}^*$ называется разделимым (блоковым) неискажающим кодом, если для любых слов $x_1, x_2, \dots, x_m \in D^n, m > 1, n \geq 1$ последовательность $\phi(x_1)\phi(x_2)\dots\phi(x_m)$ может быть однозначно декодирована как $x_1x_2\dots x_m$. В теории информации хорошо известно, что для разделимого кода ϕ выполняется неравенство Крафта [11]:

$$\sum_{u \in D^n} 2^{-|\phi(u)|} \leq 1.$$

(Здесь и ниже $|x|$ – длина x , если x слово, и количество элементов в x , если x множество.)

В [4] предложено использовать неравенство Крафта для задания распределения вероятностей на множестве кодируемых сообщений

$$P_{\phi}(u) = 2^{-|\phi(u)|} / \sum_{v \in D^n} 2^{-|\phi(v)|}$$

и для предсказания, или оценивания, вероятности появления $d \in D$ в момент $(n + 1)$ по формуле

$$P_{\phi}(X_{n+1} = d | X_1 = x_1, X_2 = x_2, \dots, X_n = x_n) = P_{\phi}(x_1x_2\dots x_nd) / \left(\sum_{e \in D} P_{\phi}(x_1x_2\dots x_ne) \right) \quad (1)$$

и показано, что прогноз, основанный на этом равенстве, асимптотически оптимален, если ϕ универсальный код. Однако, подчеркнем, этот подход применим только в случае, когда случайный процесс принимает значения из конечного множества D . Мы же рассматриваем более общую задачу – построение прогноза в случае, когда множество значений A , вообще говоря, бесконечно и снабжено метрикой, причем, в отличие от [5, 7, 8], исследуются не асимптотические свойства, а возможность практического применения. Оставшаяся часть этого раздела посвящена описанию общей схемы метода прогноза для случая, когда множество значений процесса A – отрезок прямой, однако предлагаемый алгоритм легко обобщается и на случай, когда A часть многомерного пространства. При описании мы рассматриваем различные возможные варианты и указываем на параметры, которые можно менять с целью нахождения метода, дающего наилучшую точность прогноза.

Пожалуй наиболее очевидный способ сведения случая бесконечного "алфавита" A к конечному – разбиение интервал A на k непересекающихся подинтервалов $\{d_1, \dots, d_k\}$ некоторым способом (на способах разбиения мы остановимся ниже). Теперь вместо значений X_0, \dots, X_{n-1} будем рассматривать номера подинтервалов, куда попадают эти значения и оценивать вероятности попадания X_n в эти интервалы по (1). В качестве оценки величины $E(X_n | X_0, \dots, X_{n-1})$ можно взять среднее значение "ступенчатой" функции, порожденной разбиением $\{d_1, \dots, d_k\}$.

Повышения точности прогноза можно достигнуть варьируя количество подинтервалов k , метод разбиения интервала A на k частей и алгоритм сжатия, который будет использован для прогноза. В работе исследовалось два способа разбиения интервала A на k частей. Первый состоит в разбиении A на части равной длины. При втором способе A разбивается на k подинтервалов таким образом, чтобы в каждый из них попадало примерно равное количество выборочных значений и затем это разбиение используется для прогнозирования следующего значения. В качестве алгоритмов сжатия использовались известные архиваторы Rar, arj, pkzip, ha. Наконец, отметим, что предлагаемый метод можно использовать в сочетании с другими подходами и приемами, используемыми при прогнозировании. К таким приемам, рассматриваемым ниже при прогнозировании курсов валют, относится "удаление трендов" и использование для прогноза очередного значения не всего имеющегося временного ряда, а только его последней части, скажем, последних 1000 или 50 значений, часто называемой "окном" или "скользящим окном". (При использовании такой схемы предполагается, что статистические характеристики процесса могут меняться со временем и "старые" данные не несут информации о "новых" статистических характеристиках.)

3 Результаты экспериментов

Мы рассматривали задачи прогнозирования курса доллара США относительно рубля и евро, используя ежедневные данные о стоимости одного доллара в евро и рублях, приведенные на сайтах <http://www.x-rates.com> и <http://www.akm.ru>, соответственно. Как отмечалось, проводились многочисленные эксперименты с различными значениями параметров, методами снятия "трендов" и используемыми архиваторами.

Все эксперименты проводились в два этапа, которые мы условно назовем оцениванием параметров и проверкой. Данные, используемые для оценивания параметров, являлись рядом подряд идущих значений стоимости доллара США, которые мы обозначим через $x_1x_2\dots x_n$. В ходе экспериментов значение x_{n-99} прогнозировалось по данным $x_1x_2\dots x_{n-100}$, значение x_{n-98} - по данным $x_1x_2\dots x_{n-99}$ и т.д., так, что x_n прогнозировалось по данным $x_1x_2\dots x_{n-1}$. Затем вычислялась величина

$$\delta = \left(\sum_{i=1}^{100} |x_i - x_i^*| \right) / 100, \quad (2)$$

где x_i^* - прогнозируемое значение величины x_i . На основе проведенных расчетов выбирался вариант с тем набором параметров, для которого величина δ была минимальна. На этом этап оценивания заканчивался и мы переходили к этапу проверки, в ходе которого отобранный вариант использовался для прогнозирования новых, хронологически последних, 100 значений, уже известных, но, подчеркнем, не использовавшихся при предшествующих вычислениях. Полученные значения точности прогноза, по-прежнему оцениваемые величиной δ , приведены в таблице для стоимости доллара США в евро и рублях. Там же указаны и значения параметров, подобранных на этапе оценивания. При этом проводилось разбиение на интервалы равной длины и предварительное преобразование данных, направленное на "снятие тренда". При этом исходный ряд x_1, x_2, \dots, x_t преобразовался в последовательность отношений $(x_2/x_1), (x_3/x_2), \dots, (x_t/x_{t-1})$, которая и использовалась при прогнозировании. (Но, естественно, величина δ в (2) вычислялась по абсолютным, а не относительным, значениям.)

В работе использовались данные о стоимости доллара в рублях за период с 03.01.2001 по 07.08.2002 для нахождения параметров, дающих наименьшую погрешность, и с 07.08.2002 по 26.02.2003 для проверки точности прогноза. Аналогично использовались данные по курсу доллар США/ евро с 16.03.2001 по 9.07.2002 и с 9.07.2002 по 2.12.2002, соответственно. Полученные результаты приведены в таблице, а два рисунка с результатами прогноза за 10 дней дают общее представление о точности. В частности, видно, что наибольшие ошибки прогноза происходят при резких измене-

иях курса.

Таблица 1

курс валют	Средняя точность	архиватор	количество интервалов	размер "скользящего окна"
доллар США/евро	0,00479 (евро)	Rar	15	50
доллар США/рубль	1,306 (копейки)	Rar	10	70

Из таблицы мы видим, что средняя ошибка за сто дней для курса доллар США/евро равна 0.00479 евро, для курса доллар США/рубль - 1,306 копейки, что близко к колебаниям курса в течении суток.

Таким образом, по нашему мнению, приведенные данные показывают, что методы сжатия данных (или универсального кодирования) могут служить основой для построения методов прогноза, представляющих практический интерес.

Список литературы

- [1] Колмогоров А.Н. Три подхода к определению понятия "количество информации". // Пробл. передачи информ., т.1, с.3-11.
- [2] Martin- Löff P. *The definition of random sequences*, Information and Control, v.9, 1966, pp.602-619.
- [3] Zvonkin A.K., Levin L.A. *The complexity of finite objects and concepts of information and randomness through the algorithm theory*. Uspehi Math. Nauk, v. 25, n.6. 1970.
- [4] Рябко Б.Я. Прогноз случайных последовательностей и универсальное кодирование.// Пробл. передачи информ. 1988, т.24н.2, СЮЗ-14.
- [5] Morvai G., Yakowitz S. J., Algoet P.H. *Weakly convergent nonparametric forecasting of stationary time series*. IEEE Trans. Inform. Theory, v.43, 1997, pp. 483 - 498
- [6] Kieffer J. *Prediction and Information Theory*, Preprint, 1998. (available at <ftp://oz.ee.umn.edu/users/kieffer/papers/prediction.pdf/>)
- [7] Algoet P., *Universal Schemes for Learning the Best Nonlinear Predictor Given the Infinite Past and Side Information*, IEEE Trans. Inform. Theory, v. 45, pp. 1165-1185, 1999.
- [8] Nobel A.B. On optimal sequential prediction. IEEE Trans. Inform. Theory, v.49, 2003, n.1. pp.83-98.
- [9] Кукушкина О.В., Поликарпов А.А., Хмелев Д.В. Определение авторства текста с использованием буквенной и грамматической информации.// Пробл. передачи информ., 2001, т. 37, н. 2, с. 96 – 109.
- [10] B. Ya. Ryabko, F. Topsøe. On Asymptotically Optimal Methods of Prediction and Adaptive Coding for Markov Source. Journal of Complexity, Vol. 18, No. 1, Mar 2002, pp. 224-241
- [11] Gallager R.G. *Information Theory and Reliable Communication*. Wiley, New York, 1968.
- [12] Ryabko B. Ya. *The complexity and effectiveness of prediction algorithms*. J. of Complexity, v. 10, pp.281-295, 1994.