УДК 519.8 doi:10.15421/321617

А.Д. Фирсов
Университет таможенного дела и финансов, г. Днепропетровск

КЛАСТЕРИЗАЦИЯ УЧАСТКОВ ВРЕМЕННОГО РЯДА ПО ПРИЗНАКАМ – КОЭФФИЦИЕНТАМ ПОЛИНОМОВ

Предложен метод извлечения дополнительной информации о временном ряде с применением кластеризации признаков, полученных методом наименьших квадратов для участков исходного ряда.

Запропоновано метод одержання додаткової інформації про часовий ряд із застосуванням кластеризації ознак, одержаних за допомогою методу найменших квадратів для ділянок початкового ряду.

The method of extracting additional information on the time series using clustering features proposed. Features obtained by least squares method for the source series parts.

Ключевые слова: временной ряд, нелинейная динамика, кластеризация.

Введение. В работе проведены исследования по извлечению дополнительной информации о временном ряде, основанные на поиске схожих участков ряда. Целью исследований являлась проверка возможности применения метода *k*-средних для кластеризации участков временного ряда, описывающего систему, имеющую нелинейную динамику либо квазипериодические процессы.

Проведённые ранее исследования показали хорошие результаты применения кластеризации для прогнозирования в случае рассмотрения участков ряда как векторов признаков [2]. При этом концептуальным недостатком такого подхода является упрощение, связанное с неявным предположением об отсутствии функциональной зависимости между соседними значениями элементов временного ряда. Также в работе основным направлением стало прогнозирование, а не результат кластеризации. Изучение характеристик нелинейного временного ряда при помощи кластеризации его частей требует учёта свойств кластеров, их числа, особенностей организации вычислительного процесса. Улучшение прогноза подразумевает ещё и способы выбора прогнозной последовательности из кластера в рамках постановки конкретных задач прогнозирования.

В работе [1] была описана реализация алгоритма прогнозирования нелинейного временного ряда при помощи муравьиных колоний. В [3] приведены результаты вычислительного эксперимента. В нашем случае интерес представляет не сам прогноз, а способ усреднения значений временного ряда. Была применена процедура деления отрезка [0,1] на k равных частей, где значениям на отрезке соответствовали все возможные значения элементов временного ряда после нормировки. То есть производилась группировка, которую можно назвать кластеризацией, и из такой группы, впоследствии выбирались элементы для построения прогнозируемых данных.

Постановка задачи. Одним из ключевых способов исследования нелинейных временных рядов является вычислительный эксперимент. При этом объёмы вычислений могут быть достаточно большими [1 – 3] , а влияние малых изменений в параметрах вычислительных алгоритмов на результат и его динамику – существенным.

Задачей описанного далее исследования стала постановка вычислительного эксперимента по кластеризации участков нелинейного временного ряда по коэффициентам полинома, полученным для участков различной длины методом наименьших квадратов, изучение свойств этих кластеров, получение дополнительных сведений о самом временном ряде.

Метод решения. Пусть дан временной ряд A(i) i=1,…,N, предположительно сгенерированный нелинейной динамической системой, где N – число элементов ряда. Зафиксируем длину k участка ряда, выбрав ее из расчёта k>2, что позволит построить полином степени k по его значениям и k<<N, а это даст возможность рассмотреть периодически повторяющиеся подобные участки ряда. Пронумеруем все неповторяющиеся участки ряда длины k. Таких участков будет не более L·k, где L – число элементов исходного ряда A.

Далее для каждой последовательности из k элементов при помощи метода наименьших квадратов получим коэффициенты полинома соответствующей степени. При этом область значений аргумента будет одинаковой для всех последовательностей длины k ряда A(i). Для текущего вычислительного эксперимента будем применять полином второй степени, что позволит учесть нелинейность участка кривой, а его выпуклость – однозначно трактовать результаты последующего сравнения. Для однозначности в качестве значений аргумента будем применять последовательность 1,..,k. После вычислений получим набор из номера последовательности и трёх коэффициентов, которые и будут выступать в роли кластеризируемых признаков.

Для кластеризации применим метод *k*-средних и его реализацией в пакете R, а именно функцию – kmeans(d,n). Подготовка данных для кластеризации согласно описанной выше идее реализована при помощи циклически выполняемого кода для исходного временного ряда:

```
x<-c(1,2,3)#c(i,i+1,i+2)
y<-c(data[i,2],data[i+1,2],data[i+2,2])
parabola<-data.frame(x,y)
model<-lm(y~poly(x,2,raw=TRUE),dat=parabola)
summary(model)$coefficients[,1]
coefs<-list(a=summary(model)$coefficients[3,1],b=summary(model)
$coefficients[2,1],c=summary(model)$coefficients[1,1])
ifelse(i==1,d<-coefs,d<-rbind(d,coefs))
```

Приведённый пример кода реализует построение трех рядов признаков для случая применения интерполяции полиномом второй степени, то есть получения трёх коэффициентов-признаков для каждого участка ряда длины три.

Соответственно для нескольких рядов различной природы далее выполнена кластеризация признаков, а для каждого из наборов коэффициентов – расчёт первого показателя Ляпунова.

Анализ результатов экспериментов. Для исследования применимости предлагаемого метода в первую очередь был взят хорошо известный нелинейный временной ряд Лоренца, построенный в пакете R, временной ряд биологической природы – кардиограмма человека, полученная из международной базы медицинских данных (www.physionet.org/physiobank/database/mghdb/patient-guide.shtml#mgh001), и финансовый временной ряд, который можно отнести к одному из проявлений функционирования квазипериодической искусственной системы (https://fred.stlouisfed.org/series/DEXUSEU) [4;5;7]. Выбор рядов обусловлен сложностью прогнозирования и квазипериодической структурой, наличие которой и позволяет применять кластеризацию из соображения подобия в данных.

Эксперименты с рядом Лоренца показали заметные отличия в результатах кластеризации для каждого из признаков. Поэтому для каждого ряда признаков был вычислен старший показатель Ляпунова методом Канца при помощи функции R:

```
lyap_k(c.ts, m=3, d=2, s=1000, t=40, ref=9800, k=2, eps=4),
```

где *c.ts* – временной ряд; *m* – размерность вложения; *d* – временная задержка; *s* – число итераций метода; *t* – окно сдвига; *ref* – число рассматри-

ваемых значений; *k* – число соседей; *eps* – радиус поиска ближайших соседей.

Оказалось, что показатель Ляпунова является положительным только для ряда признака коэффициентов полинома степени ноль. Что в свою очередь косвенно подтверждает адекватность подхода, применённого в [1;3], где разбиение на группы соответствовало делению отрезка оси ординат для каждого значения исходного ряда.

Выполненная кластеризация позволяет выбрать интервалы разбиения пропорционально диаметрам кластеров и задать их количество до применения алгоритма прогнозирования.

Вычисление показателя Ляпунова для кардиограммы показало, как и было представлено в [6], отсутствие хаоса в данных по каждому признаку. Тем не менее, сама процедура кластеризации по принадлежности признаков позволила выделить кластеры, а также выявила, что сами кластеры размещаются в трёхмерном пространстве. Формальное определение размерности вложения не выполнялось, так как первый показатель Ляпунова отрицателен. Отношение сторон параллелепипеда, в котором разместились кластеры, – 25:6:1.

В случае финансового ряда показатель Ляпунова оказался отрицательным. Кластеры размещены в плоском слое равномерной глубины, что позволяет учитывать только два признака.

Вариант плоской проекции результатов трёхмерной визуализации результатов кластеризации приведён ниже (рис. 1–3).

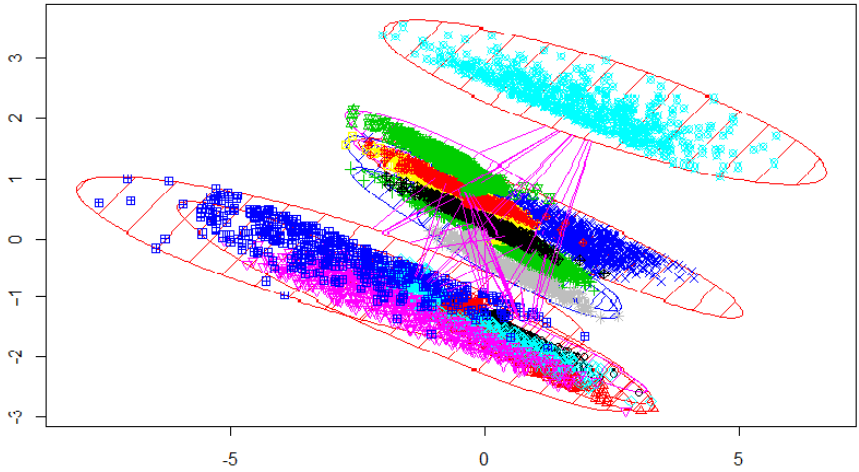

Рис. 1. Результат кластеризации ряда Лоренца

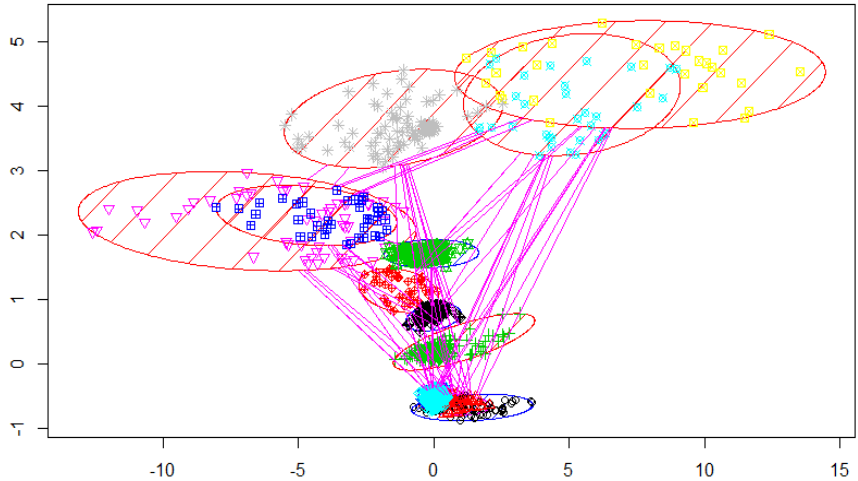

Рис. 2. Результат кластеризации кардиограммы

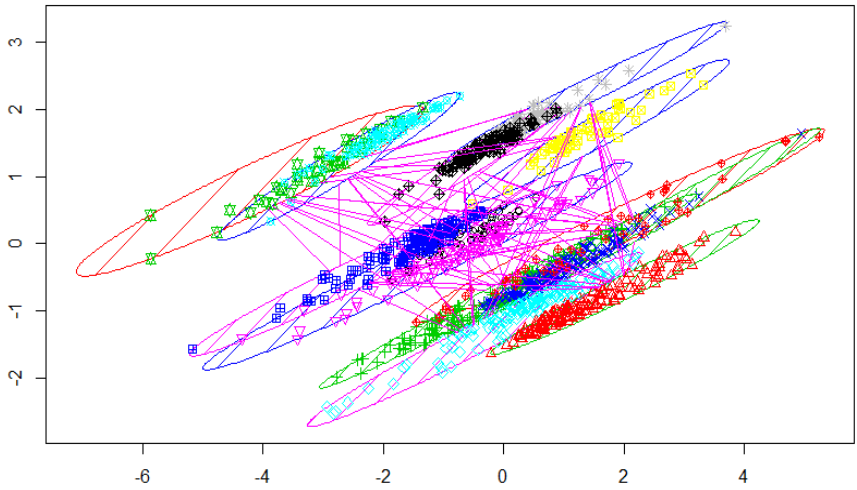

Рис. 3. Результат кластеризации финансового ряда

Выводы. Проведённые вычислительные эксперименты по кластеризации признаков временных рядов продемонстрировали применимость предлагаемого метода, подтвердили идею метода построения мультиграфа по временному ряду и показали возможность учёта размерности пространства признаков.

В тоже время остались нерассмотренными вопросы по применению полиномов более высоких степеней для интерполяции и последующей кластеризации, а также касательно исследования качества прогноза в зависимости от степени полинома.

Дополнительного анализа требует вопрос и о подборе числа кластеров для конкретных рядов в контексте задачи прогнозирования.

Библиографические ссылки

1. **Громов, В.А.** О некоторых алгоритмах построения мультиграфа по временному ряду [Текст] /В.А. Громов, А.Д. Фирсов // Питання прикладної математики і математичного моделювання. – Д., 2014. – С.79–85.

2. **Громов, В.А.** Прогнозирование хаотических временных рядов как задача кластеризации [Текст] /В.А. Громов, А.И. Миняйло // Там же. – 2012. – С.65–73.

3. **Громов, В.А.** Применение метода муравьиных колоний для прогнозирования временных рядов [Текст] /В.А. Громов, А.Н. Шульга // Там же. – С.58–69.

4. **Дроздов, Д.В.** Автоматический анализ ЭКГ: проблемы и перспективы [Текст] / Д.В. Дроздов, В.М. Леванов // Здравоохранение и мед. техника. – 2004. – №1.

5. American National Standard for Testing and Reporting Performance Results of Cardiac Rhythm and ST Segment Measurement Algorithms [Text]. AAMI/ANSI Standard EC57:1998, 1998.

6. **Kaplan, D.T.** Is fibrillation chaos? [Text]/ D.T. Kaplan, R.J. Cohen // Circ Res. 1990. – Vol. 67. – P. 886–892.

7. Latife Ghalayini. Modeling and Forecasting the US Dollar/Euro Exchange Rate [Text] / International J. of Economics and Finance. – 2014. – Vol. 6, №1. – P.194 – 207.

Надійшла до редколегії 29.08.2016