Классификация динамических объектов в задаче статистического оценивания*

А.Л. Чернявский, А.А. Дорофеюк, И.В. Покровская

Аннотация. Рассматривается задача сглаживания траекторий, описывающих изменение некоторого показателя во времени. Необходимость сглаживания связана с тем, что значения показателя имеют большой случайный разброс. Предлагается метод сглаживания, основанный на объединении «близких» траекторий. В качестве меры близости используется коэффициент корреляции.

Ключевые слова: *динамические объекты, автоматическая классификация, сглаживание траекторий.*

Введение

В прикладных задачах статистического анализа данных часто встречается ситуация, когда каждый объект (например, предприятие, территориальное образование) характеризуется множеством показателей (параметров), причем их значения меняются во времени. Таким образом, каждый объект описывается некоторой многомерной траекторией. В этом случае исходный материал о функционировании исследуемой системы представляет собой куб данных «объект-показатель-время», т.е. трехмерную матрицу:

$$\left\| y_{t,\ i}^{(j)}, i = 1,...,N; j = 1,...,K; t = 1,...T \right\|,$$

где N– число объектов; K – число показателей; T – число моментов времени.

На практике чаще всего возникает задача построения классификации таких многомерных объектов. Ее обычно сводят к классификации траекторий. Пусть, например, мы имеем данные об объектах за ряд лет. Поскольку стандартные алгоритмы классификации работают с двухмерными матрицами «объект-показатель», такое сведение можно осуществить двумя способами: «умножением количества объектов», когда каждый объект формально рассматривается как совокупность T независимых объектов, или «умножением количества показателей», когда каждый объект описывается KT параметрами [1, 2].

Существуют, однако, задачи, в которых построение классификации не является самоцелью, а используется для решения других задач. Рассмотрим важный для практики случай, когда объектами являются территории, а их показатели определяются с помощью выборочных обследований проживающего на этих территориях населения. Такого рода обследования требуют больших затрат, поэтому выборки, как правило, оказываются недостаточно представительными. В результате значения показателей имеют большой случайный разброс и не позволяют непосредственно судить о реальной динамике оцениваемого показателя. Поэтому полученные путем выборочного обследования траектории объектов нуждаются в коррекции («сглаживании»). В настоящей работе предлагается метод сглаживания, основанный на построении классификации специального типа.

Идея заключается в том, что для повышения точности прогнозирования исследуемый объект включается в группу объектов, близких по динамике исследуемого показателя, после чего выборки, по которым определяется этот показатель для каждого объекта, объединяются, и прогнозирование ведется по всей группе. За счет увеличения размера выборки траектории получаются более «гладкими». Однако для построения такой группы используется не разбиение множества всех объектов на классы и включение интересующего нас объекта в один из этих классов (как в обычной классификации), а формирование группы с использованием исследуемого объекта в качестве эталона.

1. Постановка задачи

Далее для простоты изложения предлагаемый подход иллюстрируется на конкретном примере, в котором динамическими объектами являются регионы Российской Федерации, а в качестве измеряемого показателя выбран уровень безработицы в регионе, определяемый путем ежемесячных выборочных обследований.

Пример такого динамического объекта приведен на рис. 1.

* Работа выполнена при частичной поддержке РФФИ, гранты 17-07-00857-а, 15-07-06713-а, 16-07-00896-а, 16-07-00895-а, 16-29-12880-офи, 16-29-12895-офи, 16-29-12943-офи.

Рис. 1. Динамика оценок измеряемого показателя и скользящего среднего

Пусть $x^j = (x_1^j,..., x_k^j)$ - вектор оценок значений измеряемого показателя j-го объекта ($j = 1,...,N$) в интервалы времени $t_1,..., t_k$ (на рис. 1 это месяцы), полученный с помощью выборочных обследований (на рис. 1 эти оценки – точки, соединенные тонкими линиями). Очевидно, что уровень безработицы не может столь резко меняться от месяца к месяцу, т.е. результаты выборочных обследований нуждаются в сглаживании.

Простейшим методом сглаживания является метод скользящего среднего. Он заключается в том, что данные выборочных обследований за три последовательных интервала времени объединяются в одну выборку, по этой укрупненной выборке рассчитывается средний за этот период показатель, и это значение показателя условно относится к среднему интервалу. На рис. 1 оценки, полученные методом скользящего среднего, – это точки, соединенные жирными линиями.

Оценки скользящего среднего существенно ближе к реальным значениям оцениваемого показателя. Однако метод скользящего среднего имеет один существенный недостаток: его нельзя использовать для оценки текущего значения показателя (в интервале времени t_k), поскольку для этого необходимы данные выборочного обследования в интервале t_{k+1}, которых еще нет.

Задача состоит в разработке такого метода сглаживания, который был бы свободен от указанного недостатка.

2. Метод группировки объектов с использованием эталона

Далее для краткости объект, для которого ищется оценка измеряемого показателя, будем называть расчетным объектом. Метод включает 3 этапа:

Этап 1. Производится сглаживание данных выборочного обследования с помощью процедуры скользящего среднего.

Этап 2. Формируется группа объектов, близких по динамике измеряемого показателя к расчетному объекту. Выборки, по которым оценивается значение показателя для вошедших в эту группу объектов, объединяются. Полученная группа объектов рассматривается как один виртуальный объект, ассоциируемый с расчетным объектом.

Этап 3. На основе объединенной выборки виртуального объекта с помощью процедуры масштабирования находится искомая оценка текущего значения измеряемого показателя расчетного объекта.

2.1. Формирование виртуального объекта

Алгоритм формирования виртуального объекта представляет собой пошаговую процедуру, на каждом шаге которой в группу, образующую виртуальный объект, вводится объект, наиболее близкий к расчетному. В качестве меры близости в алгоритме используется коэффициент корреляции между соответствующими временными рядами выборочных оценок (траекториями).

Для удобства описания алгоритма объектам присваиваются номера в том порядке, в котором они вводятся в группу: расчетному объекту присваивается номер 1, следующему введенному в группу объекту – номер 2 и т.д. Рассмотрим (**i**+1)-й шаг алгоритма. К началу (i+1)-го шага группа (виртуальный объект) включает i объектов, введенных в нее на предыдущих шагах, и представлена следующей информацией:

а) Временной ряд оценок скользящего среднего измеряемого показателя расчетного объекта $y^1 = (y_1^1,..., y_{k-1}^1)$. Далее этот временной ряд будем называть эталоном.

б) Временной ряд выборочных оценок измеряемого показателя для сформированного к (i+1)-му шагу виртуального объекта (эти оценки получены на базе объединенной выборки всех объектов, включенных в виртуальный объект) $y^i = (y_1^i,..., y_k^i)$.

в) Коэффициент корреляции между временным рядом выборочных значений измеряемого показателя для сформированного к (**i**+1)-му шагу виртуального объекта и эталоном: $r_i = r(y^i, y^1)$ (при подсчете r_i используются первые k-1 точек временного ряда y_i).

На (i+1)-м шаге алгоритма из всех еще не вошедших в группу объектов выбирается такой l-й объект, добавление которого к группе доставляет максимум коэффициенту корреляции r_{i+1}:

$$r_{i+1}(l) = \max_j \left(r_{i+1}(j)\right), \text{ где}$$
$$r_{i+1}(j) = r\left(y^{i+1}(j)\ y^1\right),$$

$y^{i+1}(j)$ – временной ряд оценок измеряемого показателя, сформированного на (i+1)-м шаге виртуального объекта в результате добавления к нему l-го объекта.

Если $r_{i+1} \geq r_i$ (коэффициент корреляции после включения в виртуальный объект выбранного l-го региона не уменьшился), то этот объект вводится в виртуальный объект, ему присваивается номер (i+1) и алгоритм переходит к следующему шагу. Если $r_{i+1} < r_i$ (коэффициент корреляции уменьшился), то работа алгоритма заканчивается.

Проиллюстрируем работу алгоритма на примере, в котором динамическими объектами являются регионы Российской Федерации.

К началу работы алгоритма виртуальный объект состоит из одного расчетного объекта – Вологодской области. Коэффициент корреляции временного ряда оценок измеряемого показателя с эталоном равен 0,280.

На первом шаге алгоритма производится поочередное пробное добавление в виртуальный объект одного объекта из списка всех остальных объектов – это все регионы за исключением Вологодской области. При каждом таком добавлении для образовавшегося виртуального объекта (Вологодская область и добавленный регион) подсчитываются: временной ряд выборочных оценок измеряемого показателя $y^2 = \left(y_1^2, ..., y_k^2\right)$ и коэффициент корреляции *r(i)* между этим временным рядом и эталоном (в рассматриваемом примере эталон – это временной ряд скользящих средних для Вологодской области).

По результатам всех пробных добавлений находится объект с номером l, добавление которого доставляет максимум коэффициенту корреляции *r(j)*, т.е. $r(l) > r(j)$ для всех $j \neq l$. Этот объект и вводится на первом шаге в группу (виртуальный объект). В рассматриваемом примере таким объектом оказалась Липецкая область. Ее добавление в виртуальный объект увеличивает коэффициент корреляции между измеряемым показателем и эталоном до значения 0,642 (рис. 2).

На втором шаге в виртуальный объект вводится Калининградская область, коэффициент корреляции увеличивается до 0,786 и т.д.

На 13-м шаге максимальный коэффициент корреляции, равный 0,971, получается при добавлении в виртуальный объект Чукотского автономного округа. Однако на предыдущем шаге коэффициент корреляции был больше (0,972), поэтому

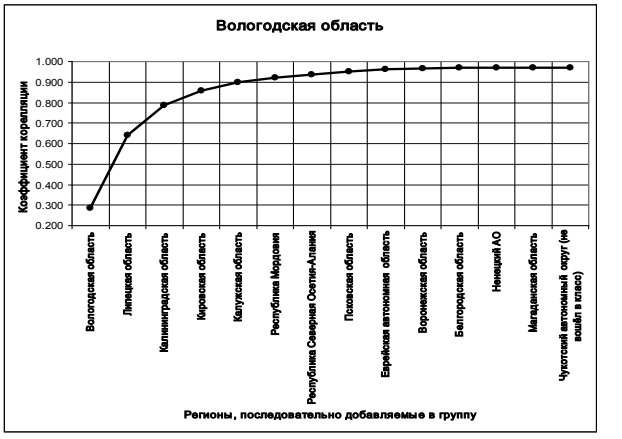

Рис. 2. Иллюстрация работы алгоритма формирования виртуального объекта на примере Вологодской области

алгоритм заканчивает работу, сформировав виртуальный объект как группу из 13 регионов (на рис. 2 они расположены по оси абсцисс в порядке их включения в виртуальный объект).

На рис. 3 показан временной ряд измеряемого показателя виртуального объекта, сформированного для оценки текущего значения измеряемого показателя для Вологодской области, и для сравнения – временной ряд скользящего среднего этого показателя для Вологодской области. На этом примере (который является типичным) видно, что траектория измеряемого показателя виртуального объекта достаточно гладкая. Как правило, она оказывается даже более гладкой, чем скользящее среднее для исходного временного ряда расчетного объекта, потому что объем выборки по виртуальному объекту обычно больше, чем объем выборки по расчетному объекту за три месяца.

2.2. Процедура масштабирования

Хотя траектория измеряемого показателя виртуального объекта по форме почти не отличается от

Рис. 3. Траектории измеряемого показателя виртуального объекта и скользящего среднего на примере Вологодской области

траектории скользящего среднего для расчетного объекта (коэффициент корреляции между соответствующими временными рядами в подавляющем числе случаев больше 0,9), она обычно смещена относительно траектории скользящего среднего и может отличаться от этой последней также и масштабом. Такое смещение и изменение масштаба объясняется тем, что в качестве меры близости временных рядов при формировании виртуального объекта используется значение коэффициента корреляции. А смещение на константу и изменение масштаба не влияет на коэффициент корреляции. Поэтому в группу могут включаться объекты, близкие к расчетному по форме траектории измеряемого показателя, но заметно отличающиеся по абсолютной величине этого показателя и с большей или меньшей амплитудой колебаний этой величины. Чтобы устранить полученное в результате этого смещение и изменение масштаба, производится «масштабирование» полученного в итоге временного ряда.

Цель процедуры масштабирования – с помощью линейного преобразования измеряемого показателя виртуального объекта (т.е. смещением на константу и изменением масштаба) минимизировать его отличие от временного ряда значений скользящего среднего для расчетного объекта. В качестве меры такого отличия (отклонения) в алгоритме используется среднеквадратичная разность соответствующих значений этих временных рядов.

Задача формулируется следующим образом. Обозначим через $y^{во} = \left(y_1^{во},...,y_k^{во}\right)$ вектор значений измеряемого показателя виртуального объекта. Как и прежде, $y^1 = \left(y_1^1,...,y_{k-1}^1\right)$ – вектор оценок скользящего среднего измеряемого показателя расчетного объекта.

Требуется найти такие константы b_0 и b_1 линейной регрессии y^1 на $y^{во}$, чтобы среднеквадратичное отклонение:

$$\Delta y = \sum_{i=1}^{k-1} \left[y_i^1 - \left(b_1 y_i^{во} + b_0 \right) \right]^2$$

было минимальным. Эта задача решается с помощью стандартной процедуры метода наименьших квадратов [3].

Результат решения этой задачи для рассматриваемого примера Вологодской области представлен на рис. 4.

В качестве искомой оценки текущего значения измеряемого показателя (в интервале времени t_k) принимается величина $y^{во} = b_1 y_k^{во} + b_0$. В нашем примере оценкой текущего значения измеряемого показателя в Вологодской области будет величина $y^{во} = 1.623 y_k^{во} - 9.268 = 7.410$ (на рис. 4 – последняя справа точка кривой «Уровень безработицы в виртуальном регионе после масштабирования»).

Рис.4. Результат масштабирования

Заключение

Объединение близких по динамике траекторий объектов в одну группу позволяет увеличить размер выборки и за счет этого существенно уменьшить случайный разброс значений измеряемого показателя. При этом, в отличие от метода скользящего среднего, для оценки показателя в момент *t* не требуется знать его значение в момент *t+1*.

Литература

1. *Бауман Е.В., Дорофеюк А.А.* Классификационный анализ данных // Труды Международной конференции по проблемам управления. Том 1. – М.: СИНТЕГ, 1999. - С. 62-67.
2. *Бауман Е.В., Дорофеюк А.А., Дорофеюк Ю.А.* Методы динамического структурного анализа многомерных объектов / Четвертая международная конференция по проблемам управления (МКПУ-IV): Сборник трудов. − М.: ИПУ РАН, 2009. – С. 338-343.
3. *Айвазян С. А., Енюков И. С., Мешалкин Л. Д.* Прикладная статистика: исследование зависимостей. М.: Финансы и статистика, 1985.

Чернявский Александр Леонидович. Старший научный сотрудник ИПУ РАН. Окончил в 1964 г. МФТИ. Кандидат технических наук. Количество печатных работ: 43. Область научных интересов: интеллектуальные методы анализа данных, методы экспертизы и анализа экспертных оценок, методы поддержки принятия решений. E-mail: achern@ipu.ru

Дорофеюк Александр Александрович. Главный научный сотрудник ИСА ФИЦ ИУ РАН и ИПУ РАН. Профессор. Окончил в 1965 г. МФТИ. Доктор технических наук. Количество печатных работ: 239 (в т.ч. 15 монографий). Область научных интересов: математическая статистика, функциональный анализ, интеллектуальные методы анализа данных, методы экспертизы и анализа экспертных оценок, методы поддержки принятия решений, системный анализ. E-mail: daa2@mail.ru

Покровская Ирина Вячеславовна. Старший научный сотрудник ИПУ РАН. Окончила в 1976 г. МГУ им. М.В. Ломоносова. Кандидат технических наук. Количество печатных работ: 64. Область научных интересов: интеллектуальные методы анализа данных, методы экспертизы и анализа экспертных оценок, методы поддержки принятия решений. E-mail: ivp750@mail.ru.

Clustering of dynamic objects in the statistic estimation problem

A.L. Chernyavsky, A.A. Dorofeyuk, I.V. Pokrovskaya

Abstract. The smoothing of time series describing the dynamics of some indicators is considered. The need in smoothing is caused by a considerable random dispersion of indicators magnitude. The smoothing method based on grouping of similar time series samples is proposed. For measure of similarity the correlation factor is used.
Keywords: *dynamic objects, statistic estimation, indicators*.

References

1. *Bauman E.V., Dorofeyuk, A.A.* Cluster-analysis. / Proceedings of the International conference on control science, Vol. 1. -M.: SINTEG, 1999, pp. 62-67. (in Russian).
2. *Bauman E.V., Dorofeyuk, A.A., Dorofeyuk J.A.* Dynamic structural analysis of multidimensional objects. / Proceedings of the IV International conference on control science (MKPU-IV). -M.: ICS RAS, 2009, pp. 338-343. (in Russian).
3. *Aivazyan S.A., Enyukov I.S., Meshalkin L.D.* Applied statistics: dependence analysis. -M.: Finances and statistics, 1985.

Chernyavsky Alexander L. V.A.Trapeznikov's Institute of Control Sciences of RAS, Moscow, Senior Researcher. PhD (Computer Sciences), Associate Professor.

Dorofeyuk Alexander A. Institute for System Analysis of the Federal Research Center «Information and Control», RAS, Moscow, Chief Researcher; V.A.Trapeznikov's Institute of Control Sciences of RAS, Moscow, Chief Researcher. Doctor of Sciences (Computer Sciences), Professor.

Pokrovskaya Irina V. V.A.Trapeznikov's Institute of Control Sciences of RAS, Moscow, Senior Researcher. PhD (Computer Sciences), Associate Professor.