электронный журнал
МОЛОДЕЖНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК
Издатель ФГБОУ ВПО "МГТУ им. Н.Э. Баумана". Эл No. ФС77-51038.

УДК 004.021

Прогнозирование экономических параметров объекта на базе метода группового учета аргументов

Липкин В. Н., *студент*
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана,
кафедра «Программное обеспечение ЭВМ и информационные технологии»

Научный руководитель: Рудаков И. В., к.т.н., зав.каф
Россия, 105005, г. Москва, МГТУ им. Н.Э. Баумана,
кафедра «Программное обеспечение ЭВМ и информационные технологии»
irudakov@bmstu.ru

Введение

В современном мире информационные технологии находят все более широкое применение, в том числе и в область прогнозирования экономических величин: цены на акции, курсы валют, прогнозирование спроса и пр.

Одним из возможных применений информационных технологий в области прогнозирования экономических величин является прогнозирование успеха медиа-объектов, основанное на анализе поведения пользователей интернета.

Перечислим факторы, отражающие «интерес» пользователей к объекту в интернете:

- количество упоминаний объекта в социальных сетях и блогах и их эмоциональная окрашенность;
- количество поисковых запросов, относящихся к объекту;
- количество скачиваний или просмотра демонстрационной версии объекта (к примеру, трейлера фильма);
- посещаемость страницы объекта на различных сайтах с обзорами и, к примеру, сайтах кинотеатров;
- количество показов рекламы объекта и переходов по ней.

Постановка задачи

Каждый из вышеперечисленных факторов можно рассматривать в виде временного ряда величин от начала наблюдений до момента, предшествующего старту продаж. Обозначим множество временных рядов как

$$X = \{x_{ij} |\ i = 1, N, j = 1, M\}, \text{ где}$$

N — количество рассматриваемых факторов;

M — количество элементов временных рядов.

Обозначим множество статических параметров объекта как

$$K = \{k_i | i = 1, L\}, \text{ где}$$

L количество статических параметров объекта.

В случае прогнозирования сборов фильма такими параметрами могут быть:

- количество копий фильма (иными словами, количество кинотеатров, в которых будет прокат фильма);
- бюджет фильма;
- рейтинг актеров и режиссеров.

Объединив множества X и K, получим множество параметров объекта

$$P = X \cup K$$

Пусть y — функция прогнозируемого параметра объекта (количество продаж, сборы в кинотеатре)

$$y = f(P)$$

Таким образом, задача прогнозирования экономического параметра объекта сводится к задаче нахождения модели регрессии. Саму модель и ее параметры найдем с помощью метода группового учета аргументов (далее МГУА).

Отметим особенности задачи, благодаря которым МГУА является хорошим выбором для ее решения:

- сравнительно небольшие размеры выборок, которые можно использовать для нахождения модели — порядка сотен;
- перечисленные параметры и факторы обладают неизвестной взаимной корреляцией, выяснение которой является крайне трудоемкой задачей;
- МГУА позволяет найти модель с ограниченным количеством параметров, что упростит сбор данных для ее использования.

Заметим, что если в качестве исходных параметров модели использовать один-два временных ряда, можно получить модель удовлетворительной точности, что позволяет снизить трудозатраты на сбор обучающих данных.

В качестве альтернативы рассматривалось прогнозирование с использованием нейронных сетей. С помощью нейронных сетей можно добиться усреднения обучающих данных, в то время как нам необходимо выявить параметры, действительно влияющие на успех, учитывая взаимную корреляцию параметров.

Разработанный алгоритм

Входными данными разработанного алгоритма являются:

- Выборка объектов

  $Q = \{q_i | i = 1, S\}$, где S — количество объектов.
- список значений прогнозируемого параметра для выборки Q

  $Y_r = \{y_i | i = 1, S\}$
- список параметров

  $P = \{p_i | i = 1, N_p\}$, где $N_p = N * M + L$ — количество параметров каждого объекта
- максимальное количество параметров N в искомой модели
- список значений n степенной функции $f(p_i) = y^n$

  $V = \overline{2, D} \vee \emptyset$, где D — максимальная требуемая степень.

Разделим выборку Q на три равные части:

- TR — обучающая выборка;
- T — тестовая выборка;
- CTR — контрольная выборка.

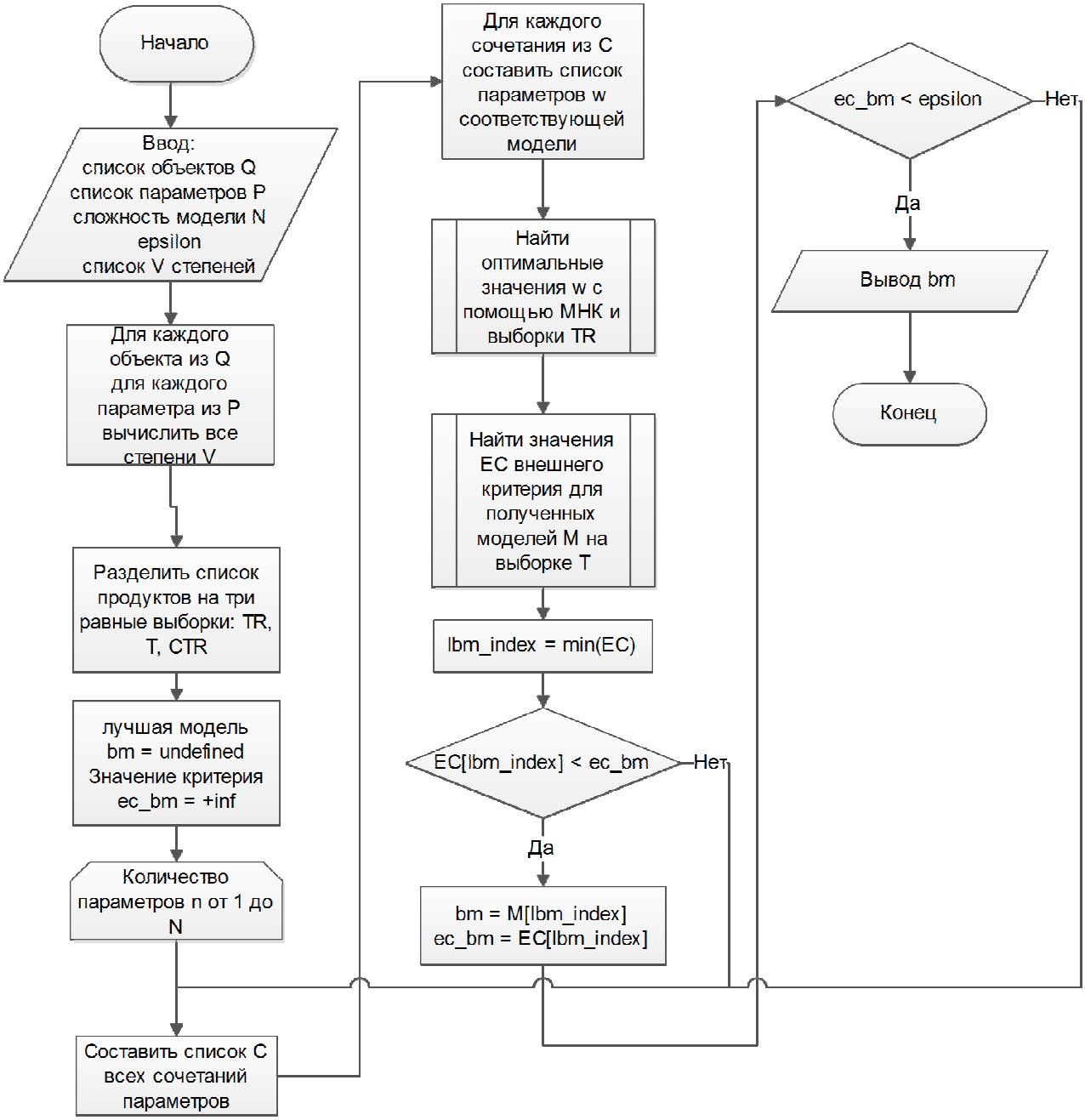

Рис. 1. Общий вид алгоритма

Первые две выборки будут использоваться при нахождении модели, последняя — для оценки адекватности найденной модели.

С помощью списка V расширим списки параметров P степенными функциями от каждого из параметров:

$$P_{ext} = P \cup \{p_i^n | i = 1, N_p; \forall n \in V\}$$

В качестве базовой модели можно выбрать любой полином. Часто в качестве базовой модели выбирают полином Колмогорова-Габора[2], что и было сделано при разработке данного алгоритма:

$$y = w_0 + \sum_{i=1}^{m} w_i x_i + \sum_{i=1}^{m} \sum_{j=1}^{m} w_{ij} x_i x_j + \cdots \quad (1)$$

Перейдем непосредственно к алгоритму, общий вид которого представлен на рисунке 1.

Рассмотрим подробнее нахождение значений w: искомые параметры должны давать минимум внутреннему критерию:

$$\| w_j - p_{ext\,i} \| \rightarrow Epsilon^2, \forall p_{ext\,i} \in P_{ext}$$

Отсюда, воспользовавшись методом наименьших квадратов (МНК), найдем значения коэффициентов w [1].

При усложнении модели внутренний критерий не дает минимума для моделей оптимальной сложности, из чего следует невозможность его использования для выбора наилучшей. Поэтому, для выбора наилучшей модели воспользуемся внешним критерием – критерием, вычисленным на данных, не использовавшихся при получении моделей, в нашем случае – на тестовой выборке.

В качестве внешнего критерия используем критерий регулярности, оценивающий устойчивость модели к новым данным:

$$K_{reg} = || w_i - \mathrm{p}_{\mathrm{ext\,i}} ||^2$$

Модель, дающую минимум внешнему критерию назовем наилучшей.

В том случае, если данные сильно зашумлены или среди данных нет необходимых для отыскания модели переменных, минимум внешнего критерия достигнут не будет[3].

Алгоритм заканчивает работу в том случае, если была найдена модель, внешний критерий которой меньше некоторого заранее заданного ε, или же в случае достижения моделей с заданным количеством параметров N.

Рис. 2. Результаты модели на обучающей выборке

Выводы

В процессе экспериментов согласно разработанный алгоритм было разработано программное обеспечение, примененное для прогнозирования сборов кинопремьер. В результате была получена модель, дающая среднюю абсолютную процентную ошибку на тестовой выборке, равную 19%, причем:

- обучающая и контрольная выборки содержали по 24 фильма;
- тестовая выборка содержала 24 фильма;
- в качестве множества переменных модели регрессии использовались бюджет и количество поисковых запросов за каждую из 8 недель перед премьерой фильма;

Результаты аппроксимации данных на обучающей выборке представлены на рис. 2. На рис. 3 показан результат, который дает модель на тестовой выборке, которая никак не участвовала в получении модели.

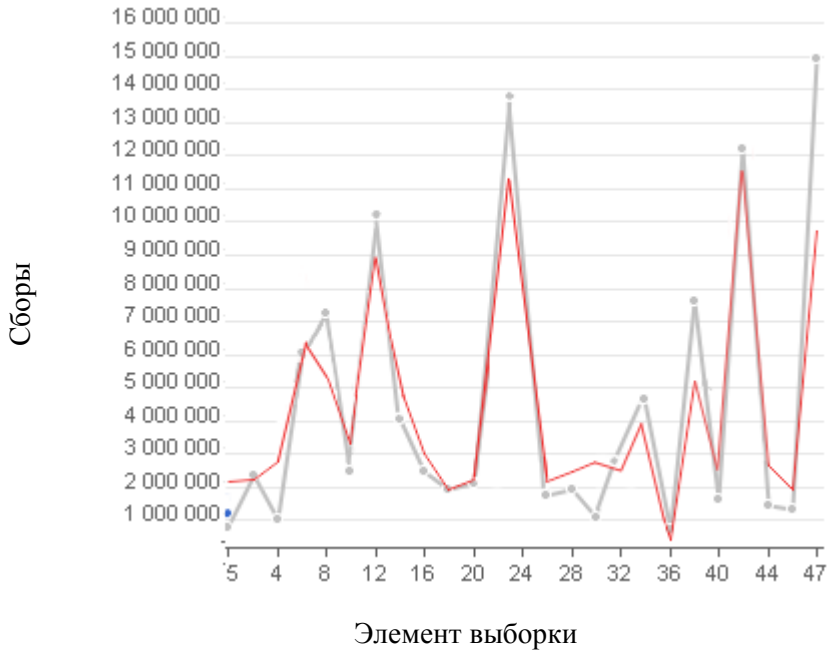

Рис. 3. Результаты модели на тестовой выборке

Конечно, стоит учитывать особенности области применения: в случае с фильмами, может иметь место недостаток рекламы, и интерес к хорошему фильму может проявиться только спустя некоторое время после премьеры. В общем, и особенно в этом случае можно продолжать наблюдение и после премьеры, с каждым новым днем наблюдений повышая точность прогноза общих сборов.

Список литературы

1. Ивахненко А.Г. Индуктивный метод самоорганизации моделей сложных систем. Киев: Наукова думка, 1981. 296 с.
2. Стрижов В.В., Крымова Е.А. Методы выбора регрессионных моделей. М.: ВЦ РАН, 2010. 60 с.
3. Стрижов В.В., Камырова Е.А. Методы выбора регрессивных моделей. М.: ВЦ РАН, 2010. 60 с.