Инженерный вестник Дона, №4 (2019)
ivdon.ru/ru/magazine/archive/n4y2019/5873

Алгоритм поиска искажений в данных при оценке параметров множественной линейной регрессии

Е.А. Питухин[1], О.А. Зятева[1], П.В. Питухин[2]

[1]Петрозаводский государственный университет
[2]Филиал "Протвино" государственного университета "Дубна"

Аннотация: В статье представлены результаты исследования случаев намеренного искажения «объективных» рейтингов, построенных на основе частных показателей. Целью исследования стало построение алгоритма определения несправедливо расставленных мест в рейтингах вузов в результате ручной корректировки. Основным инструментом является идентификация весовых коэффициентов частных рейтингов при известном виде функциональной зависимости общего рейтинга. Проведен анализ методик построения одного из популярных российских рейтингов, получены математические модели зависимости общего рейтинга вуза от его частных рейтингов. Было выявлено наличие субъективизма в построении рейтинга, который проявляется в виде несправедливой бальной оценке. Стандартные методы анализа не позволяют выявить такие неслучайные «выбросы» и получить объективную оценку. Предложенный алгоритм позволяет находить такие «выбросы», исключать их из выборки и определять справедливые значения. Предлагаемый алгоритм может быть полезен руководителям вузов для проверки корректности занимаемого места в рейтинге их организации и понимания своего реального положения среди других вузов.
Ключевые слова: статистический анализ, множественная линейная регрессия, метод наименьших квадратов, аппроксимация, рейтинги, показатели, вузы.

В последнее время рейтингованию подвергаются практически все сферы деятельности, в частности, высшего образования [1–4]. Инициаторами комплексных оценок являются федеральные органы власти, информационные агентства и известные организации. Большое внимание со стороны научного сообщества уделяется показателям, по которым происходит оценка деятельности вузов и построение их рейтинга, а также поиску путей их увеличения [5–7]. Общий рейтинг вуза в большинстве случаев строится на основе линейной комбинации частных показателей, взятых с определенными априори весовыми коэффициентами. Таким образом, зная результаты частных рейтингов, можно рассчитать общий, если весовые коэффициенты известны. Но часто «правила игры» неизвестны для непосвященных, а методика расчета общего рейтинга обычно остаётся

скрытой и нигде не публикуется. Попытки построить линейную функцию аппроксимации официального рейтинга от частных традиционными методами (например, методом наименьших квадратов или его модификацией с итеративным пересчётом весов) приводит к некорректным результатам, поскольку места в официальном рейтинге уже подверглись субъективному вмешательству «экспертным путем» [8]. В таком случае, при идентификации параметров модели рейтинговой функции, значения весовых коэффициентов получаются смещенными, поскольку значения модельного рейтинга находятся между исходными и скорректированными значениями официального рейтинга. В связи с этим, возникает задача разработки алгоритма, который бы мог находить скорректированные места рейтинга, исключать их из выборки и определять, в итоге, близкие к априорным значения параметров модели рейтинговой функции.

Алгоритм поиска неслучайных «выбросов» в рейтингах

При моделировании общего рейтинга на основе частных, возникла проблема определения его корректности, для чего стало необходимо проанализировать этот рейтинг и выявить, подвергался ли он «экспертным» корректировкам, или нет. Задачу облегчало то, что имелась достоверная «инсайдерская» информация о значениях весовых коэффициентов частных рейтингов. Поэтому не составило труда построить истинную модель рейтинга и обнаружить, что существуют точки значительного расхождения значений официального и модельного рейтингов, которые не объяснить случайностью.

Вначале для нахождения «неслучайных» выбросов была предпринята попытка использовать существующие пакеты интеллектуального анализа данных (Data Mining). В частности, для поиска выбросов был использован пакет Analysis Services, встроенный в СУБД MS SQL Server 2012.

Использовался инструмент Highlight Exceptions (выделение исключения), который использует алгоритм кластеризации Microsoft [9]. Модель кластеризации определяет группы строк со сходными характеристиками. Данный инструмент выделяет подсветкой ячейки в исходной таблице данных с подозрительными значениями. В результате работы этого сервиса были найдены исключения, но, к сожалению, они совершенно не совпали с теми, которые были заранее известны. Универсальные методы выявления выбросов в Data Mining не подходят в силу специфики их алгоритма, т.к. не позволяют устанавливать функциональную зависимость между входящими частными рейтингами и выходным итоговым рейтингом. Они используют для выявления выбросов методы, основанные на кластеризации, которые не предусматривают выявление функциональных связей между факторами. Исследуемый же класс задач содержит функциональную зависимость, поэтому корректное решение не может быть найдено с помощью кластеризации. Попытка использовать инструменты аппроксимации нелинейных зависимостей, например, нейронные сети, не принесли ожидаемого результата. Это связано с тем, что исходные данные, которые можно было бы использовать для обучения сети, изначально содержат искажения. Поэтому было решено протестировать другие известные алгоритмы, чтобы выбрать из них пригодный для решения поставленной задачи, либо, за их отсутствием, разработать собственный алгоритм.

Известные авторам стандартные методы либо не подходят для решения поставленной задачи, либо расходятся, либо дают оценки, отличные от истинных. В связи с этим, авторами был предложен свой собственный алгоритм оценки параметров модели множественной линейной регрессии с правилом логического выбора.

Первый шаг алгоритма заключается в том, что методом МНК находятся коэффициенты модели, с которыми частные рейтинги входят в общий

официальный. Далее рассчитывается модельный рейтинг, после чего для каждого места рейтинга вычисляется модуль относительного отклонения модельного рейтинга от официального. Учитывая, что значения рейтинга вузов должны располагаются в порядке убывания, далее рейтинг исследуется с целью поиска точек, подозрительных на «выброс».

Для проверки места рейтинга на наличие выброса используется специальный индикатор, который, в свою очередь, является произведением двух вспомогательных индикаторов. Первый – индикатор направления «выброса» – находит те места в рейтинговом массиве, которые выбиваются из общей убывающей последовательности. Он равен ±1, в зависимости от того, выше или ниже должно быть значение на соответствующем месте. Второй – индикатор силы «выброса». Он равен 1, если абсолютное значение относительного отклонения модельного и общего рейтинга конкретного вуза больше удвоенного аналогичного среднего значения по всем вузам, иначе – 0. Место является подозрительным на «выброс», если оба вспомогательных индикатора дают ненулевое значение.

После этого массив значений рейтинга, вместе с итоговым индикатором сортируется по убыванию модуля относительного отклонения модельного рейтинга от общего. Далее отбрасывается группа вузов с ненулевым индикатором, которая расположена в начале отсортированного списка. Данная процедура повторяется нужное количества раз, пока не будет достигнуто условие останова. Это условие наступит, когда группа с ненулевым значением индикатора в начале списка будет отсутствовать.

При реализации каждого шага данного алгоритма вычисляется ряд показателей, таких, как средняя абсолютная ошибка (*MAE*) и средняя абсолютная процентная ошибка (*MAPE*), нормированный R^2 и оценки качества аппроксимации в виде среднеквадратической ошибки *RMSE_1* *RMSE_2*, учитывающих штраф за исключение значений с выбросами из

общей выборки. Величина штрафа при оценке *RMSE_1* равна числу исключенных на текущем шаге «выбросов» из рейтинга. В случае *RMSE_2* величина штрафа удваивается [10].

Описание алгоритма оценки параметров модели множественной линейной регрессии с правилом логического выбора приведено ниже.

Пусть: N_0 – начальная длина массива

M_0 – длина первой ненулевой итерации

$j = 0$ – количество итераций алгоритма

$n = 6$ – число частных рейтингов

$R(N_0) = \{r_k\}, k \in [1, N_0]$ – массив значений официального рейтинга

$P_i(N_0)$ – массивы значений частных рейтингов для каждого $i \in [1,n]$

Пусть $Condition(j) = true$

Пока истинно $Condition(j)$

- Оцениваются МНК коэффициенты множественной линейной регрессии $k_i, i \in [1,n]$ для $Y = R(N_j)$ и $X_i = P_i(N_j)$
- Вычисляются $\hat{R}(N_j)$: $\hat{r_k} = \sum_{i=1}^{n} k_i P_i(N_j)$ для каждого $k \in [1, N_j]$
- Вычисляются $MAE, MAPE, R^2$
- Вычисляется *RMSE_1*: $RMSE\_1_j = \frac{\sum_{k=1}^{N_j}(r_k - \hat{r_k})^2}{N_j - M_j}$
- Вычисляется *RMSE_2*: $RMSE\_2_j = \frac{\sum_{k=1}^{N_j}(r_k - \hat{r_k})^2}{N_j - 2M_j}$

- Вычисляется относительная ошибка: $rer_k = \frac{r_k - \widetilde{r_k}}{r_k}$ для каждого $k \in [1, N_j]$
- Для каждого $k \in [1, N_j]$ вычисляем:
  - указатель направления указатель направления $Id_k = f_d(R(N_j), rer, k) \in \{-1,0,1\}$
  - амплитудный индикатор $Ia_k = f_a(rer, k) \in \{0,1\}$
  - интегральный индикатор $I_k = Id_k \wedge Ia_k$
- Сортируется $R(N_j), rer, I$ массив по убыванию коэффициента rer:

  $Rs(N_j) = \{rs_k\}, k \in [1, N_0] = sort\_desc\left(rer, R(N_j)\right)$
- Ищется ведущая последовательность с ненулевым интегральным индикатором в верхней части $Rs(N_j)$:
  - Пусть $M_j = 0$ – длина ненулевой ведущей последовательности
  - Пусть $Condition(M_j) = true$
  - Пока истинно $Condition(M_j)$:

    Если $I_{M_j+1} = 0$,

    тогда $Condition(M_j) = false$,

    иначе $M_j = M_j + 1$
- Если $M_j = 0$, тогда $Condition(M_j) = false$, иначе:

– $N_j = N_j - M_j$ – уменьшаем длину рейтинга $R(N_j)$ для M_j элементов

– Для каждого $k \in [1, N_j]$:

$r_k = rs_{k+M_j}$ составляем уменьшенный рейтинг без исключений

– $j = j + 1$

Отличие представленного алгоритма от известных состоит в том, что к алгоритму аппроксимации добавляется правило логического выбора, которое заключается в вычислении и сравнении двух вспомогательных индикаторов. Итоговым критерием исключения точек выброса из исходного множества стало нахождение кортежей ненулевой длины, состоящих из результатов конъюнкций вспомогательных индикаторов, в самом начале отсортированного по убыванию перечня элементов [10].

Предложенный алгоритм был реализован в среде MathCAD и апробирован на серии искусственно смоделированных рейтингов, при построении которых веса частных рейтингов были известны заранее. В рейтинги искусственно были внесены корректировки, которые являются моделью неслучайных «выбросов». В результате работы алгоритма были получены весовые коэффициенты, отличающиеся, в среднем, от истинных не более, чем на 0,73%. Алгоритм продемонстрировал высокую сходимость (до 10 шагов).

Заключение

В результате данного исследования был разработан и реализован алгоритм поиска неслучайных «выбросов» в рейтингах вузов, основанный на пошаговом исключении групп участников рейтинга с максимальными значениями критериев отбора. Это позволяет дать оценку весовым

коэффициентам частных рейтингов, с которыми они входят в общий официальный рейтинг. С практической точки зрения такие знания могут быть полезны руководителям вузов для проверки корректности занимаемого места в рейтинге их организации и понимания своего реального положения среди других участников.

Литература

1. Быкова О.Н. О рейтингах в системе высшего образования // Ценности и смыслы. 2014. №2(30). С. 83–86.
2. Петросянц Д., Светцова А. Оптимальный выбор критериев для рейтингов университетов // Проблемы теории и практики управления. 2015. №12. С. 97–107.
3. Кондрашова Н.В., Кондратова А.С. Обзор применяемых рейтингов в оценке образовательной деятельности // Апрельские научные чтения имени профессора Л.Т. Гиляровской. Воронеж, 2015. С. 320–323.
4. Меликян А.В. Показатели мониторинга системы высшего образования в России и за рубежом // Университетское управление: практика и анализ. 2014. № 3(91). С. 58–66.
5. Зятева О.А., Питухин Е.А., Пешкова И.В., Шабалина И.М. Интеллектуальный анализ данных при категоризации преподавателей вуза на основе наукометрических показателей // Инженерный вестник Дона, 2017, №4. URL: ivdon.ru/magazine/archive/n4y2017/4580.
6. Котенко Ю.С., Названова И.А., Подопригора М.Г. Проблемы современного вуза и маркетинговые методы их выявления и оценки // Инженерный вестник Дона, 2013, №2. URL: ivdon.ru/ru/magazine/archive/n2y2013/1631.
7. Рождественская Е.А. Рейтинговая система оценивания деятельности преподавателей вуза // NovaInfo. 2019. Т. 1. № 96. С. 186–191.

8. Zyateva O., Pitukhin E., Peshkova I. Impact of university performance indicators on their position in rankings // EDULEARN16 Proceedings: 8th International Conference on Education and New Learning Technologies. Barcelona. 2016. – pp. 8751–8759.
9. Выделение исключений URL: docs.microsoft.com/ru-ru/sql/analysis-services/highlight-exceptions-table-analysis-tools-for-excel?view=sql-server-2014 (дата обращения 21.12.2018).
10. Zyateva O.A., Pitukhin E.A., Peshkova I.V. Upwards excursion algorithm providing the weight rankings coefficients of universities // Proceedings of the First International Workshop on Stochastic Modeling and Applied Research of Technology. Petrozavodsk. 2018. – pp. 62–70.

References

1. Bykova O.N. Cennosti i smysly. 2014. №2 (30). pp. 83–86.
2. Petrosyants D., Svetsova A. Problemy teorii i praktiki upravleniya. 2015. №12. pp. 97–107.
3. Kondrashova N.V., Kondratova A.S. Aprel'skie nauchnye chteniya imeni professora L.T. Gilyarovskoj. Voronezh. 2015. pp. 320–323.
4. Melikian A. V. Universitetskoe upravlenie: praktika i analiz. 2014. №3 (91). pp. 58–66.
5. Zyateva O.A., Pitukhin E.A., Peshkova I.V., Shabalina I.M. Inženernyj vestnik Dona (Rus), 2017, №4. URL: ivdon.ru/magazine/archive/n4y2017/4580.
6. Kotenko Yu.S., Nazvanova I.A., Podoprigora M.G. Inženernyj vestnik Dona (Rus), 2013, №2. URL: ivdon.ru/ru/magazine/archive/n2y2013/1631.
7. Rozhdestvenskaya E.A. NovaInfo. 2019. V. 1. № 96. pp. 186–191.

8. Zyateva O., Pitukhin E., Peshkova I. EDULEARN16 Proceedings: 8th International Conference on Education and New Learning Technologies. Barcelona, 2016. pp. 8751–8759.
9. Vydelenie isklyuchenij URL: docs.microsoft.com/ru-ru/sql/analysis-services/highlight-exceptions-table-analysis-tools-for-excel?view=sql-server-2014 (data obrashcheniya 21.12.2018).
10. Zyateva O.A., Pitukhin E.A., Peshkova I.V. Proceedings of the First International Workshop on Stochastic Modeling and Applied Research of Technology. Petrozavodsk, 2018. pp. 62–70.