Вычислительные технологии, 2020, том 25, № 3, с. 119–129. © ИВТ СО РАН, 2020 ISSN 1560-7534
Computational Technologies, 2020, vol. 25, no. 3, pp. 119–129. © ICT SB RAS, 2020 eISSN 2313-691X

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

DOI:10.25743/ICT.2020.25.3.013

Разработка модификации метода главных проекций Торгерсона с применением анализа кумулятивных кривых в задаче выявления выбросов в данных больших размерностей

Н. С. Олейник†, В. Ю. Щеколдин
Новосибирский государственный технический университет, Новосибирск, Россия
†Контактный автор: Олейник Никита С., e-mail: olejnik.2015@stud.nstu.ru

Поступила 10 марта 2020 г., доработана 23 марта 2020 г., принята в печать 16 апреля 2020 г.

Рассмотрена задача выявления аномальных наблюдений в данных больших размерностей на основе метода многомерного шкалирования с учетом возможности построения качественной визуализации данных. Предложен алгоритм модифицированного метода главных проекций Торгерсона, основанный на построении подпространства проектирования исходных данных путем изменения способа факторизации матрицы скалярных произведений при помощи метода анализа кумулятивных кривых. Построено и проанализировано эмпирическое распределение F_1-меры для разных вариантов проектирования исходных данных.

Ключевые слова: выбросы, многомерные данные, метод главных проекций Торгерсона, кумулятивные кривые, CC-ABOD, мера качества классификации.

Цитирование: Олейник Н.С., Щеколдин В.Ю. Разработка модификации метода главных проекций Торгерсона с применением анализа кумулятивных кривых в задаче выявления выбросов в данных больших размерностей. Вычислительные технологии. 2020; 25(3):119–129.

Введение

В разнообразных приложениях технического, экономического, социального характера приходится иметь дело с данными больших размерностей, поскольку изучаемые объекты и процессы оказываются, как правило, весьма сложной природы, а их функционирование и эволюция зависят от большого числа различных факторов. При выборе тех или иных методов исследования не последнюю роль играет возможность получения корректных и хорошо объясняемых результатов, в частности, допускающих построение различных графических интерпретаций, удобных для восприятия как профильными специалистами, так и людьми, не имеющими глубоких знаний в области прикладной статистики, но использующими ее подходы в своей деятельности.

За последние пару десятилетий в связи с широким и глубоким развитием вычислительной техники все большее внимание исследователей привлекают методы, позволяющие строить удобные графические визуализации, в том числе и в ситуациях, когда анализируемые данные отличаются большой размерностью. Можно отметить такие подходы, как широко известный метод главных компонент и многочисленные его модификации [1], методы многомерного шкалирования [2], особенно часто применяемые

для решения задач экономики, социологии и психологии, метод кривых Эндрюса [3], метод радиальной визуализации (RadViz) [4], его развитие — "свободная" визуализация (FreeViz) [5] и т. д. Возрастающее число публикаций на эту тему, особенно в зарубежных научных изданиях, свидетельствует об актуальности и востребованности подобных методов.

В настоящей работе в свете разработки методов визуализации многомерных данных предлагается рассмотреть решение одной из популярных задач прикладной статистики — выявление аномальных наблюдений (выбросов), которая часто возникает при решении как теоретических, так и практических задач в самых различных областях. В одно- и двумерном случаях проблем с визуализацией, как правило, не появляется, поскольку исследователь может наглядно изобразить анализируемые данные на прямой или плоскости и, проводя визуальный, а затем и статистический анализ, выявить те или иные резко различающиеся наблюдения, которым можно согласно определенным критериям придать статус выбросов. Однако ситуация серьезно усложняется при анализе данных, размерность которых превышает два, поскольку даже трехмерные изображения на плоскости могут вызывать неверное восприятие и приводить к ошибочным выводам.

Среди перечисленных выше подходов особый класс занимают методы многомерного шкалирования, поскольку начиная с середины прошлого века их аппарат непрерывно развивается и совершенствуется, а результаты, получаемые на их основе, регулярно публикуются и обсуждаются. Привлекательность этого типа методов состоит в том, что они, как правило, достаточно просты в реализации и удобны в восприятии, а также универсальны с точки зрения возможности применения в различных приложениях и адаптации к тем или иным требованиям предметной области.

В статье предлагается модификация одного из классических методов многомерного шкалирования — метода главных проекций Торгерсона [6] — с привлечением аппарата анализа кумулятивных кривых, а также геометрического подхода при выявлении аномальных наблюдений.

1. Постановка задачи

С точки зрения решения вопроса о построении качественной визуализации многомерных данных основную задачу можно сформулировать следующим образом. Исходные объекты необходимо разместить в определенном метрическом пространстве, по возможности двумерном, при выполнении условия сохранения их взаимного расположения: похожие элементы должны находиться в этом пространстве близко друг к другу, а различающиеся — далеко. При этом под "схожестью" объектов будем понимать величину, обратную выбранной в метрическом пространстве мере расстояния между ними.

Пусть имеется симметричная матрица $D = \|D_{ij}\|_{nxn}$ различий D_{ij} между объектами, расположенными в n-мерном пространстве. Необходимо построить пространство меньшей размерности $r < n$, в котором матрица различий $d = \|d_{ij}\|_{nxn}$ между анализируемыми объектами была бы в смысле некоторого критерия близка к исходной матрице различий.

В качестве критерия, определяющего свойства визуализации, будем использовать оценку классификатора наблюдений, строящегося на основе ABOD-подхода. Основная идея этого метода, применяемого для выявления выбросов, состоит в том, что при анализе геометрии расположения исходных данных может быть получена более полная

информация, которая описывала бы их внутреннюю структуру. Вычислительная схема ABOD-подхода будет представлена ниже.

Оценка качества классификации выборочных наблюдений, строящейся на основе ABOD-подхода, будет осуществляться при помощи так называемой F_1-меры [7], определяемой как

$$F_1 = \frac{2 \cdot P \cdot R}{P + R}, \tag{1}$$

где P (precision) — точность классификации — доля элементов выборки, которые алгоритм классификации считает выбросами и которые при этом действительно ими являются, а R (recall) — полнота классификации — доля выбросов среди элементов выборки, выявленных алгоритмом. Эти величины вычисляются по следующим формулам:

$$P = \frac{TP}{TP + FP}, \quad R = \frac{TP}{TP + FN}.$$

Здесь TP (true positive) — число элементов выборки, которые алгоритмом были верно отнесены к выбросам; FP (false positive) — число элементов выборки, которые неверно отнесены к выбросам; FN (false negative) — число элементов выборки, которые неверно определены как регулярные наблюдения. При этом если N — объем анализируемой выборки, то верно $TP+TN+FP+FN = N$, где TN (true negative) — число элементов выборки, которые верно распознаны как регулярные наблюдения.

Таким образом, наилучшей классификацией наблюдений выборки будет считаться та, которая доставляет максимальное значение величине (1).

2. Методы решения и схема вычислительного эксперимента

При построении классификации наблюдений, обеспечивающей не только статистическую эффективность метода, но и одновременно возможность удобной и понятной визуализации как самих данных, так и процесса выявления аномальных наблюдений, важно корректно выбрать метод выявления выбросов и способ представления исходных данных. Методы многомерного шкалирования имеют достаточно простой и широко исследованный математический аппарат для получения требуемого условиями задачи представления данных, но при этом обладают рядом существенных недостатков, ограничивающих область их применения.

Одной из существенных проблем методов шкалирования является выбор начала системы координат, на которую проводится проектирование исходных многомерных данных. Уоррен С. Торгерсон в [8] предложил одно из наиболее разумных решений — поместить начало координат в центр тяжести анализируемых данных, что позволит получить единственное решение задачи шкалирования и уменьшить дисперсию каждого отдельного наблюдения за счет усреднения, что, в свою очередь, повысит эффективность получаемых оценок.

Алгоритм классического метода главных проекций Торгерсона выглядит следующим образом.

- По заданной матрице различий вычисляется симметричная матрица скалярных произведений векторов с началом в центре тяжести с элементами

$$b^*_{jk} = d_{jk} - \frac{1}{n}\sum_{i=1}^{n} d_{ik} - \frac{1}{n}\sum_{l=1}^{n} d_{jl} - \frac{1}{n^2}\sum_{i=1}^{n}\sum_{l=1}^{n} d_{il}.$$

- Определяются два наибольших собственных значения λ_1, λ_2 матрицы и соответствующие им собственные векторы $\mathbf{e}_1$, $\mathbf{e}_2$.
- Вычисляется матрица проекций $X = E_2\Lambda_2^{\frac{1}{2}}$, где E_2 — матрица, состоящая из $\mathbf{e}_1$, $\mathbf{e}_2$; Λ_2 — диагональная матрица, на диагонали которой λ_1, λ_2.

Такая процедура гарантирует, что в случае линейной отделимости множеств регулярных и аномальных наблюдений будет получено единственное решение. Однако очевидно, что такой вариант в реальности может возникать крайне редко, тем более при анализе многомерных данных. Поэтому логично не только определять проекции на координатные оси, соответствующие двум наибольшим собственным значениям матрицы B^*, но и рассматривать другие варианты, которые в случае существенной нелинейности решающей функции классификатора могут оказаться более эффективным выбором для решения задачи выявления аномальных наблюдений.

Для проверки качества результатов визуализации будем пользоваться модифицированной процедурой нахождения выбросов CC-ABOD (Cumulative Curves for Angle Based Outlier Detection) на основе метода анализа кумулятивных кривых, предложенной авторами в [9]. Она основана на оценке углов, под которыми из каждой точки пространства, соответствующей определенному наблюдению, видны остальные наблюдения выборки. Оценка дисперсии всевозможных углов для наблюдения A осуществляется на основе следующего выражения [10]:

$$ABOF(A) = \underset{B,C\in\Omega}{VAR}\left(\frac{(\mathbf{AB},\mathbf{AC})}{||\mathbf{AB}||^2 * ||\mathbf{AC}||^2}\right), \qquad (2)$$

где $ABOF(A)$ — функция, оценивающая степень аномальности наблюдения в точке A; $VAR(.)$ — функция дисперсии; B, C — точки многомерного пространства, выбираемые из базы данных; $(.\,,.)$ — скалярное произведение двух векторов, $\|.\|^2$ — норма (длина) соответствующего вектора в многомерном пространстве.

Наблюдения, величины типа (2) которых минимальны в выборке, потенциально считаются выбросами. Выброс, согласно Д. Хоукинсу [11], — это наблюдение, "которое так сильно отличается от остальных, что может возникнуть предположение, что оно появилось в выборке принципиально другим способом".

Поскольку для каждого анализируемого набора данных нельзя заранее задать количество наблюдений, являющихся выбросами (например, считать, что к ним относятся 10 % наблюдений, наиболее отличающихся от центра тяжести данных), то необходимо определять конкретные величины дисперсий углов (2), начиная с которых соответствующие наблюдения признаются выбросами. В предыдущих работах авторов [9, 12] рассматривался синтез метода ABOD и аппарата анализа кумулятивных кривых, параметры которых оцениваются на основе метода наименьших квадратов [13]. Также было установлено, что такой модифицированный метод, называемый CC-ABOD, позволяет получать статистически более корректные результаты, поскольку он не использует конкретное распределение, а опирается только на внутреннюю структуру исходных данных.

При построении модификации классической схемы метода Торгерсона важно учитывать, что необходимо, с одной стороны, разработать статистически корректную схему определения оптимального числа осей проектирования, а с другой — обеспечить вычислительную эффективность этой схемы, позволяющую получать требуемые результаты за разумное время. При этом, естественно, для обеспечения корректности требуется привлечение методов статистического моделирования. Проблема определения

оптимального числа осей проектирования качественно может быть сопряжена с проблемой определения достаточного числа выделяемых (латентных) факторов в факторном анализе. Ранее авторами в [14] рассматривались подобные постановки. Так, для решения рассматриваемых задач была показана эффективность привлечения аппарата кумулятивных кривых. Рассматривая вариационный ряд собственных значений корреляционной матрицы и строя по ним соответствующую кумулятивную кривую, а затем применяя интегральный метод для ее разбиения на части, можно определить, какое количество главных компонент следует выделять для адекватного описания рассматриваемых объектов. Важно отметить, что если в факторном анализе определение оптимального числа выделяемых факторов является достаточным для решения задачи, то при решении визуализации многомерных данных необходимо установить, какие именно из выделенных осей проектирования следует выбрать для наилучшего решения задачи определения выбросов.

Предлагаемая в работе вычислительная схема может быть представлена в виде следующего алгоритма.

1^o Задать n-мерные данные об изучаемом объекте (процессе) и провести их стандартизацию.

2^o Определить полную систему координат представления данных на основе классического метода проекций Торгерсона.

3^o Построить множество Ψ наиболее информативных осей системы координат на основе кумулятивной кривой, соответствующей собственным значениям матрицы скалярных произведений B^*, мощность множества $|\Psi| = r < n$.

4^o Выбрать из множества осей Ψ очередную пару $(\psi_i, \psi_j) \in \Psi^2,\ i < j,\ i, j \in \{1, ..., r\}$.

5^o Спроектировать исходные данные на систему координат (ψ_i, ψ_j).

6^o Выявить выбросы на основе алгоритма CC-ABOD.

7^o Вычислить значения TP, TN, FP, FN и меры F_1.

8^o Повторять шаги 4–7 до перебора всех пар элементов множества Ψ.

9^o Определить оптимальную систему осей $(\psi^*_{(1)}, \psi^*_{(2)})$, соответствующую максимальному значению меры F_1.

10^o Построить визуализацию исходных данных согласно выбранной оптимальной системе координат $(\psi^*_{(1)}, \psi^*_{(2)})$.

Отметим, что под стандартизацией данных понимается их нормирование на значение среднеквадратического отклонения с целью исключения влияния так называемого эффекта масштаба. Вследствие чувствительности метода кумулятивных кривых к параметру сдвига [15] нормирование данных на среднее значение не производится. Для обеспечения статистической корректности разработанного алгоритма, в частности, для определения того, какие варианты построения новой системы координат наиболее предпочтительны с точки зрения задачи выявления выбросов, можно рекомендовать выполнение шагов $1^o - 10^o$ для большого набора различных выборок, моделируемых согласно условиям проводимого эксперимента. Эмпирическое распределение значений меры F_1, а также соответствующие выбираемым осям графические представления исходных данных позволят сделать заключение о том, какие варианты проекций выгоднее всего использовать на практике.

В рамках проводимой работы использовалась классическая схема статистического моделирования многомерных данных, которые представляли собой совокупность двух типов наблюдений: наблюдения, равномерно распределенные в пятимерном эллипсоиде с заданными осями, и выбросы, в среднем удаленные в 1.5–2 раза от регулярных

вдоль выбранных осей. В этом случае для моделирования данных удобно использовать гиперсферические координаты (в пятимерном пространстве), задаваемые следующим образом:

$$x_1 = \rho \cdot \cos\alpha_4 \cdot \cos\alpha_3 \cdot \cos\alpha_2 \cdot \cos\alpha_1, \quad x_2 = \rho \cdot \cos\alpha_4 \cdot \cos\alpha_3 \cdot \cos\alpha_2 \cdot \sin\alpha_1,$$

$$x_3 = \rho \cdot \cos\alpha_4 \cdot \cos\alpha_3 \cdot \sin\alpha_2, \quad x_4 = \rho \cdot \cos\alpha_4 \cdot \sin\alpha_3, \quad x_5 = \rho \cdot \sin\alpha_4,$$

где ρ — полярный радиус, а $\alpha_1 \in [0; 2\pi), \alpha_2, \alpha_3, \alpha_4 \in [0; \pi]$ — углы поворота полярного радиуса. При этом объем моделируемой выборки был выбран $N = 100$, регулярные наблюдения составляли 90 % объема, а выбросы — 10 %. Для обеспечения статистической корректности получаемых результатов проводилась серия экспериментов с числом репликаций $R = 1000$. Полученные результаты усреднялись и служили основой для построения и анализа эмпирических распределений показателей качества классификатора наблюдений.

3. Обсуждение результатов

В результате реализации разработанного алгоритма были определены наилучшие варианты осей проектирования в модифицированном методе Торгерсона при решении задачи обнаружения аномальных наблюдений при помощи подхода CC-ABOD. На рис. 1 представлена графическая интерпретация результатов проектирования на наиболее ин-

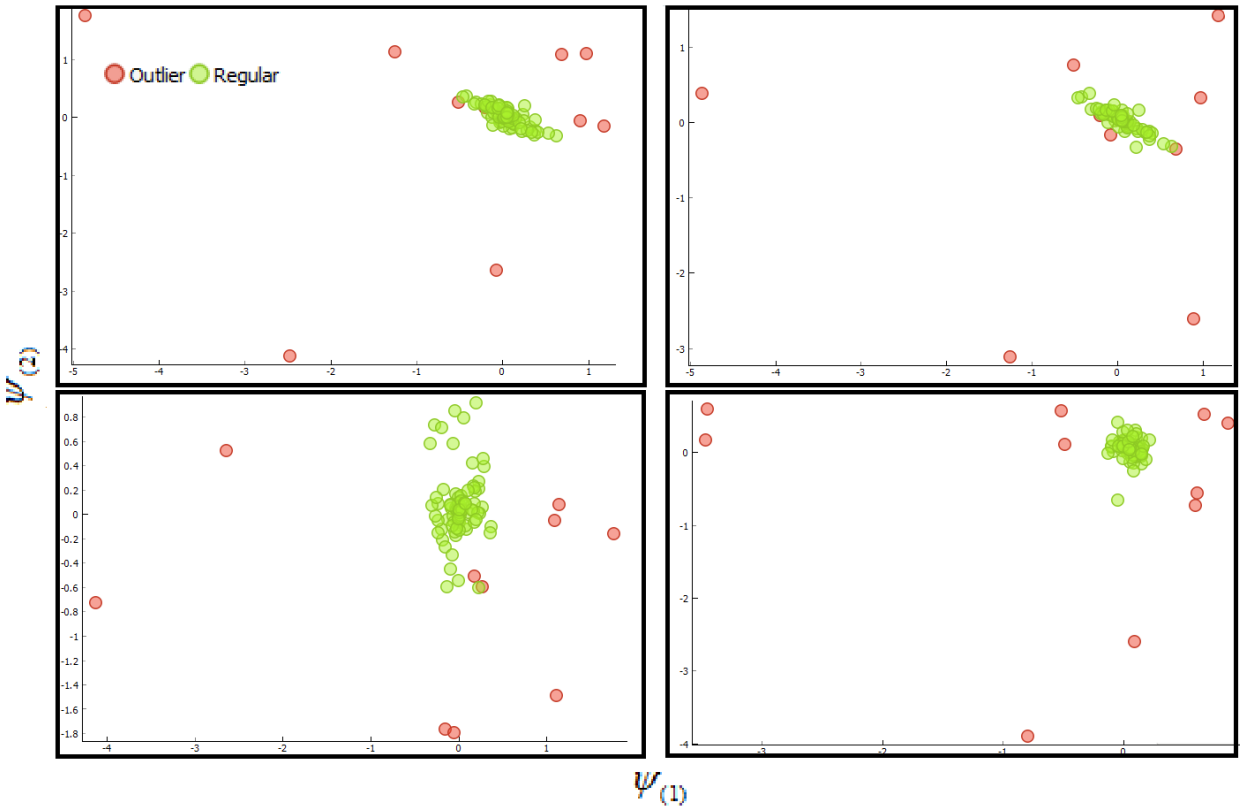

Рис. 1. Визуализация результатов проектирования на различные оси
Fig. 1. Results of design visualization for different axes

формативные с точки зрения используемого классификатора пары осей $(\psi_{(1)}, \psi_{(2)})$. В рассматриваемом примере это оси, соответствующие собственным значениям матрицы скалярных произведений B^* с номерами $(1, 2)$ — левый верхний график, $(2, 8)$ — левый нижний, $(1, 6)$ — правый верхний и $(3, 4)$ — правый нижний.

Отметим, что для одной и той же выборки ее проекции на разные пары осей приводят к существенно различным изображениям даже с учетом симметрии исходных данных. Заметно, что аномальные наблюдения не всегда располагаются на периферии множества спроецированных данных в двумерном пространстве, что, безусловно, сказывается на качестве работы классификатора. Можно также заметить, что визуально легко определить, какой вариант проектирования оказывается наиболее эффективным для решения задачи выявления выбросов, в данном конкретном случае это, очевидно, случай $(3, 4)$, где существует вполне четкое различие между регулярными и аномальными наблюдениями. С другой стороны, поскольку в методе CC-ABOD могут появляться так называемые промежуточные элементы, для которых решение вопроса их классификации затруднено, в каждом случае проектирования такие наблюдения могут возникать естественным образом, находясь не так близко к общей массе регулярных наблюдений, но не уходя далеко на границы изображения.

Выбор наилучшего варианта проектирования, как было отмечено ранее, осуществлялся на основе анализа значений F_1-меры. На рис. 2 представлено эмпирическое распределение этих значений для всевозможных способов проектирования в виде мозаичного графика. По его осям располагаются варианты выбранных наилучшими с точки

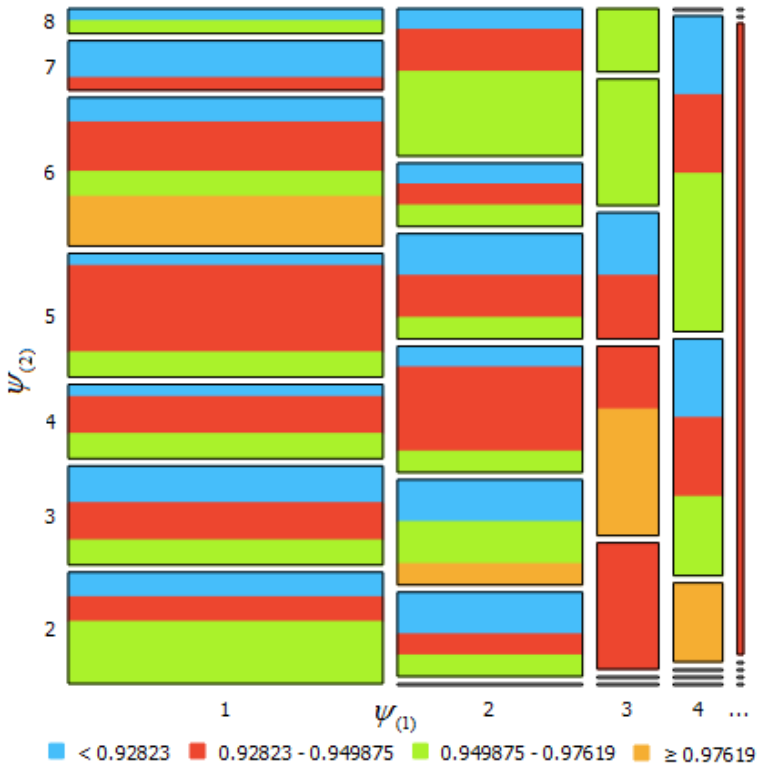

Рис. 2. Мозаичное представление эмпирического распределения значений F_1-меры по наилучшим вариантам осей проектирования

Fig. 2. Mosaic representation for the empirical distribution of F_1-measure according to the best options for the design axes

зрения алгоритма CC-ABOD осей проектирования, размеры прямоугольников соответствуют частоте появления той или иной комбинации осей в качестве решения задачи, а цветовая шкала показывает среднее значение F_1-меры каждого варианта.

Из анализа рис. 2 следует, что, несмотря на то что ось, соответствующая первому собственному значению матрицы B^*, оказываются самой часто выбираемой в качестве "оптимальной" оси проектирования (о чем свидетельствует наибольшая ширина соответствующего прямоугольника), ее выбор не всегда приводит к наилучшим значениям качества классификатора. В целом ряде случаев такой осью оказывается ψ_2, ψ_3 и даже ψ_4. Это означает, во-первых, что не существует оптимального и универсального способа определения наилучшего варианта проектирования, а во-вторых, что на результаты работы алгоритма существенное влияние оказывают нелинейный характер метода классификации (ABOD) и специфика метода анализа кумулятивных кривых, применяемого как в модификации самого ABOD, так и в предложенном в работе алгоритме для определения числа осей проектирования. Кроме того, возможны изменения полученного эмпирического распределения при переходе к другим вариантам моделирования выборочных данных.

Заключение

Рассмотрено применение методов многомерного шкалирования для построения удобных визуализаций многомерных данных и решения на их основе задачи выявления аномальных наблюдений. В качестве базового выбран метод главных проекций Торгерсона, а классификация элементов выборки осуществлялась при помощи метода CC-ABOD, учитывающего геометрию взаимного расположения точек данных в выборочном пространстве и использующего для классификации метод кумулятивных кривых.

Обнаружено, что определение наилучших вариантов проектирования многомерных данных не является однозначной задачей, однако представление результатов проектирования в графическом виде способно упростить исследователю задачу поиска аномальных наблюдений. Построено и проанализировано эмпирическое распределение характеристики классификатора F_1 в зависимости от выбора собственных значений при факторизации матрицы различий. Представляет интерес исследование возможностей предложенного метода, причем не только для более широкого класса модельных выборок, но и для решения практических задач выявления выбросов и визуализации процесса их обнаружения, в том числе и для динамических процессов.

Список литературы

[1] **Тимофеев В.С., Фаддеенков А.В., Щеколдин В.Ю.** Эконометрика: Учебник для академического бакалавриата. Изд. 2-е, перераб. и доп. М.: ЮРАЙТ; 2017: 328.

[2] **Терехина А.Ю.** Методы многомерного шкалирования и визуализации данных (обзор). Автоматика и телемеханика. 1973; (7):80–94.

[3] **Грошев С.В., Пивоварова Н.В.** Использование кривых Эндрюса для визуализации многомерных данных в задачах многокритериальной оптимизации. Наука и образование. 2015; (12):197–214.

[4] **Dai F., Zhu Y., Maitra R.** Three-dimensional radial visualization of High-dimensional Continuous or Discrete Datasets. ArXiv e-prints; 2019: 20.

[5] **Demsar J., Legan G., Zupan B.** FreeViz — an intelligent multivariate visualization approach to explorative analysis of biomedical data. J. of Biomedical Informatics. 2007; (40):661–671.

[6] **Торгерсон У.С.** Многомерное шкалирование. Теория и метод. В кн.: Статистическое измерение качественных характеристик. М.: Статистика; 1972: 95.

[7] **Powers D.** Evaluation: from precision, recall and F-measure to ROC-informedness markedness and correlation. J. of Machine Learning Technologies. 2011; 2(1):37–63.

[8] **Torgerson W.S.** Theory and methods of scaling. N.Y.: Wiley; 1958: 245.

[9] **Олейник Н.С., Щеколдин В.Ю.** Выявление аномальных наблюдений в данных больших размерностей на основе геометрического ABOD-подхода. Наука. Технологии. Инновации: сб. науч. тр. в 9 ч. Ч. 2. Новосибирск: Изд-во НГТУ; 2018: 253–257.

[10] **Kriegel H., Schubert M., Zimek A.** Angle-based outlier detection in high-dimensional data. Proc. of the 14th ACM SIGKDD Intern. Conf. on Knowledge Discovery & Data Mining. Las Vegas; 2008: 444–452.

[11] **Hawkins D.** Identification of outliers. Chapman and Hall; 1980: 127.

[12] **Oleinik N.S., Shchekoldin V.Yu.** Study of the properties of geometric ABOD-approach modifications for outlier detection by statistical simulation. Applied methods of statistical analysis. Statistical computation and simulation, AMSA'2019: Proc. of the Intern. Conf. Novosibirsk: NGTU; 2019:389–395.

[13] **Shchekoldin V.** Developing the risk classification based on ABC-analysis of possible damage and its probability. Intern. Forum: Proc. of 11th Intern. Forum on Strategic Technology (IFOST-2016). Novosibirsk; 2016:317–319.

[14] **Щеколдин В.Ю.** Выявление потребителей услуг интернет-магазинов на основе ABC-модификации факторного анализа. Ч. 2. Красноярск: Изд-во КГАУ; 2011:186–192.

[15] **Barry C.A., Sarabia J.M.** Majorization and the Lorenz order with application in applied mathematics and economics. 2nd edition. Switzerland: Springer; 2018: 251.

Вычислительные технологии, 2020, том 25, № 3, с. 119–129. © ИВТ СО РАН, 2020 ISSN 1560-7534
Computational Technologies, 2020, vol. 25, no. 3, pp. 119–129. © ICT SB RAS, 2020 eISSN 2313-691X

INFORMATION TECHNOLOGIES

DOI:10.25743/ICT.2020.25.3.013

Development for modification of Torgerson projection method using cumulative curve analysis in outlier detection problem for high-dimensional data

Oleinik Nikita S.*, Shchekoldin Vladislav Yu.
Novosibirsk State Technical University, 630087, Novosibirsk, Russia
*Corresponding author: Oleinik Nikita S., e-mail: olejnik.2015@stud.nstu.ru
Received March 10, 2020, revised March 23, 2020, accepted April 16, 2020

Abstract

Purpose. Purpose of the article. The paper aims at the development of methods for multidimensional data presentation for solving classification problems based on the cumulative curves analysis. The paper considers the outlier detection problem for high-dimensional data based on the multidimensional scaling, in order to construct high-quality data visualization. An abnormal

observation (or outlier), according to D. Hawkins, is an observation that is so different from others that it may be assumed as appeared in the sample in a fundamentally different way.

Methods. One of the conceptual approaches that allow providing the classification of sample observations is multidimensional scaling, representing by the classical Orlochi method, the Torgerson main projections and others. The Torgerson method assumes that when converting data to construct the most convenient classification, the origin must be placed at the gravity center of the analyzed data, after which the matrix of scalar products of vectors with the origin at the gravity center is calculated, the two largest eigenvalues and corresponding eigenvectors are chosen and projection matrix is evaluated. Moreover, the method assumes the linear partitioning of regular and anomalous observations, which arises rarely. Therefore, it is logical to choose among the possible axes for designing those that allow obtaining more effective results for solving the problem of detecting outlier observations. A procedure of modified CC-ABOD (Cumulative Curves for Angle Based Outlier Detection) to estimate the visualization quality has been applied. It is based on the estimation of the variances of angles assumed by particular observation and remaining observations in multidimensional space. Further the cumulative curves analysis is implemented, which allows partitioning out groups of closely localized observations (in accordance with the chosen metric) and form classes of regular, intermediate, and anomalous observations.

Results. A proposed modification of the Torgerson method is developed. The F1-measure distribution is constructed and analyzed for different design options in the source data. An analysis of the empirical distribution showed that in a number of cases the best axes are corresponding to the second, third, or even fourth largest eigenvalues.

Findings. The multidimensional scaling methods for constructing visualizations of multi-dimensional data and solving problems of outlier detection have been considered. It was found out that the determination of design is an ambiguous problem.

Keywords: outliers, multidimensional data, Torgerson's main projection method, cumulative curves, CC-ABOD, classification quality measure.

Citation: Oleinik N.S., Shchekoldin V.Yu. Development for modification of Torgerson projection method using cumulative curve analysis in outlier detection problem for high-dimensional data. Computational Technologies. 2020; 25(3):119–129. (In Russ.)

References

1. Timofeev V.S., Faddeenkov A.V., Shchekoldin V.Yu. Ekonometrika: uchebnik dlya akademicheskogo bakalavriata [Econometrics: A textbook for academic baccalaureate]. Izdanie 2-e, pererabotannoe i dopolnennoe. Moscow: YuRAYT; 2017: 328. (In Russ.)
2. Teryokhina A.Yu. Methods of multidimensional data scaling and visualization (Survey). Automation and Remote Control. 1973; 34(7):1109–1121.
3. Groshev S.V., Pivovarova N.V. Using the Andrews plotss to visualize multidimensional data in multicriteria optimization. Nauka i obrazovanie. 2015; (12):197–214. (In Russ.)
4. Dai F., Zhu Y., Maitra R. Three-dimensional radial visualization of High-dimensional Continuous or Discrete Datasets. ArXiv e-prints; 2019: 20.
5. Demsar J., Legan G., Zupan B. FreeViz — an intelligent multivariate visualization approach to explorative analysis of biomedical data. Journal of Biomedical Informatics. 2007; (40):661–671.
6. Torgerson W.S. Mnogomernoe shkalirovanie. Teoriya i metod. V kn.: Statisticheskoe izmerenie kachestvennykh kharakteristik [Multidimensional scaling. Theory and Method. In book.: Statistical measurement of performance]. Moscow: Statistika; 1972: 95. (In Russ.)
7. Powers D. Evaluation: from precision, recall and F-measure to ROC-informedness markedness and correlation. J. of Machine Learning Technologies. 2011; 2(1):37–63.
8. Torgerson W.S. Theory and methods of scaling. N.Y.: Wiley; 1958: 245.
9. Oleinik N.S., Schekoldin V.Yu. Identification of anomalous observations in large-dimensional data based on the geometric ABOD approach. Science. Technologies. Innovation: collection of scientific papers: in 9 parts. Novosibirsk: NGTU; 2018: 253–257. (In Russ.)

10. Kriegel H., Schubert M., Zimek A. Angle-based outlier detection in high-dimensional data. Proc. of the 14th ACM SIGKDD Intern. Conf. on Knowledge Discovery & Data Mining. Las Vegas; 2008: 444–452.
11. Hawkins D. Identification of outliers. Chapman and Hall; 1980: 127.
12. Oleinik N.S., Shchekoldin V.Yu. Study of the properties of geometric ABOD-approach modifications for outlier detection by statistical simulation. Applied methods of statistical analysis. Statistical computation and simulation, AMSA'2019: Proc. of the Intern. Conf. Novosibirsk: NGTU; 2019: 389–395.
13. Shchekoldin V. Developing the risk classification based on ABC-analysis of possible damage and its probability. Intern. Forum: Proc. of 11th Intern. Forum on Strategic Technology (IFOST-2016). Novosibirsk; 2016:317–319.
14. Shchekoldin V.Yu. Vyyavlenie potrebiteley uslug internet-magazinov na osnove AVS-modifikatsii faktornogo analiza [Identification of consumers of online store services based on ABC-modification of factor analysis]. Pt 2. Krasnoyarsk: KGAU; 2011: 186–192. (In Russ.)
15. Barry C.A., Sarabia J.M. Majorization and the Lorenz order with application in applied mathematics and economics. 2nd edition. Switzerland: Springer; 2018: 251.