АНАЛИЗ СУЩЕСТВУЮЩИХ МЕТОДОВ СНИЖЕНИЯ РАЗМЕРНОСТИ ВХОДНЫХ ДАННЫХ

DOI: 10.36724/2072-8735-2022-16-1-30-37

Ерохин Сергей Дмитриевич,
Московский технический университет связи и информатики, Москва, Россия, ***esd@mtuci.ru***

Борисенко Борис Борисович,
Московский технический университет связи и информатики, Москва, Россия,
fepem@yandex.ru

Мартишин Иван Дмитриевич,
Московский технический университет связи и информатики, Москва, Россия,
martishinid@gmail.com

Фадеев Александр Сергеевич,
Московский технический университет связи и информатики, Москва, Россия,
aleksandr-sml@mail.ru

Manuscript received 28 October 2021;
Accepted 07 December 2021

Ключевые слова: *системы обнаружения вторжений (СОВ), набор данных (датасет), отбор признаков, селекция признаков,генерация признаков, методы отбора признаков, анализ соответствий, компьютерные атаки (КА)*

взрывной рост объема массивов данных, как по количеству записей, так и по атрибутам (признакам), вызвал разработку ряда платформ для работы с большими данными (Amazon Web Services, Google, IBM, Infoworks, Oracle и др.), а также параллельных алгоритмов анализа данных (классификации, кластеризации, ассоциативных правил). В свою очередь, это подтолкнуло к использованию методов снижения размерности. Выбор признаков, как стратегия предварительной обработки данных, доказал свою эффективность и действенность в подготовке данных (особенно высокоразмерных данных) для решения различных задач сбора данных и машинного обучения. Снижение размерности не только полезно для ускорения выполнения алгоритмов, но и может помочь в конечной точности классификации/кластеризации. Слишком шумные или даже ошибочные входные данные часто приводят к менее чем желаемой производительности алгоритма. Удаление неинформативных или малоинформативных столбцов данных может действительно помочь алгоритму найти более общие области и правила классификации и в целом достичь лучших показателей. В статье рассмотрены широко используемые методы снижения размерности данных, их классификация. Преобразование данных состоит из двух шагов: генерация признаков и отбор признаков. Различают скалярный отбор признаков и векторный (оберточные методы, методы фильтрации, встроенные методы и гибридные методы). Каждый метод обладает своими достоинствами и недостатками, которые изложены в статье. Описано применение одного из наиболее эффективных методов снижения размерности – метода анализа соответствий, для датасета CSE-CIC-IDS2018. Проверена эффективность данного метода в задачах снижения размерности указанного датасета при обнаружении компьютерных атак.

Информация об авторах:
Ерохин Сергей Дмитриевич, *к.т.н, доцент, ректор, Московский технический университет связи и информатики, Москва, Россия*
Борисенко Борис Борисович, *к.т.н., доцент, ведущий научный сотрудник, Московский технический университет связи и информатики, Москва, Россия*
Мартишин Иван Дмитриевич, *научный сотрудник, Московский технический университет связи и информатики, Москва, Россия*
Фадеев Александр Сергеевич, *научный сотрудник, Московский технический университет связи и информатики, Москва, Россия*

Для цитирования:
Ерохин С.Д., Борисенко Б.Б., Мартишин И.Д., Фадеев А.С. Анализ существующих методов снижения размерности входных данных // T-Comm: Телекоммуникации и транспорт. 2022. Том 16. №1. С. 30-37.

For citation:
Erokhin S.D., Borisenko B.B., Martishin I.D., Fadeev A.S. (2022) Analysis of existing methods to reduce the dimensionality of input data. *T-Comm*, vol. 16, no.1, pp. 30-37. (*in Russian*)

По мере повышения уровня цифровизации общества в геометрической прогрессии растут объемы передаваемых/получаемых данных. Вместе с тем растёт и число угроз и атак (вторжений) на информационные системы. Для их защиты государственные и коммерческие структуры уделяют повышенное внимание вопросам развития применяемых методов и систем обнаружения вторжений (СОВ).

СОВ – это программное обеспечение и/или аппаратные средства, которые автоматически сканируют события, происходящие в сети, в поисках вторжения. К собираемым первичным данным СОВ предъявляются требования [1]:

– полнота, как обеспечение сбора всех значений требуемых данных, например, всех системных вызовов без пропусков, возникающих из-за несовершенства применяемого метода и особенностей операционной системы;

– достоверность, как обеспечение неискаженности данных, например, из-за отказов оборудования и действий злоумышленников;

– своевременность, как возможность получения доступа к данным в реальном времени с целью выработки адекватной ответной реакции.

Любое действие, направленное на нарушение работы информационной системы или на то, чтобы сделать ресурс недоступным или получить несанкционированный доступ, является вторжением [2]. Основными методами обнаружения вторжений являются сигнатурный и эвристический [3].

Независимо от используемой модели в основе применяемых подходов в задачах классификации сетевого трафика основной проблемой является снижение количества атрибутов, функций или входных переменных набора данных – размерности данных. Целью является извлечение подмножества данных из массивного набора данных с сохранением свойств и характеристик.

Снижение размерности – это преобразование данных из многомерной выборки к вектору меньшей размерности путем удаления неинформативных признаков, с максимальным сохранением структуры данных и информации в них содержащихся [4].

После снижения размерности эффективными методами сокращенные наборы данных часто показывают лучшие результаты классификации (коэффициент точности) [5]. Это связано с уменьшением количества элементов, которые в исходном наборе данных проявляются в виде избыточных параметров и неинформативных значений. При снижении размерности входного вектора достигается ряд преимуществ [6]:

– увеличивается общая производительность алгоритмов машинного обучения за счет уменьшения времени на обучение/классификацию и вычислительных ресурсов;

– исключается проблема переобучения;

– достигается лучшая визуализация данных, отображаемых на 2D- или 3D-графике;

– устраняется мультиколлинеарность. В регрессии мультиколлинеарность возникает, когда независимая переменная сильно коррелирует с одной или несколькими другими независимыми переменными. Снижение размерности объединяет такие сильно коррелированные переменные в набор некоррелированных;

– осуществляется поиск скрытых переменных (факторов), которые не измеряются непосредственно одной переменной, а выводятся из других переменных в наборе данных;

– повышается точность модели за счет удаления шумов в данных;

– появляется возможность преобразовывать нелинейные данные в линейно разделяемую форму.

Поэтому в СОВ применение алгоритмов снижения размерности является важным этапом. Преобразование данных состоит из двух шагов: генерация признаков и отбор признаков.

Генерация признаков – выявление признаков, которые наиболее полно описывают объект [7].

Отбор признаков – выявление признаков, которые имеют наилучшие классификационные свойства для конкретной задачи.

$X \in R^m$ – множество признаков,

$Y \in R^l$ – множество признаков, которые нужно выбрать в процессе отбора, причем $l<m$.

Задача отбора задается следующим образом: $X \to Y$.

Различают скалярный и векторный отбор признаков. При скалярном отборе рассматривается отдельно каждый признак из данного множества, что позволяет выбирать оптимальные комбинации признаков. При векторном отборе одновременно исследуются свойства группы признаков, основываясь на взаимной корреляции между ними. Скалярный отбор имеет преимущество в упрощении вычислений, однако может быть неэффективным для набора данных с взаимно коррелированными признаками [8].

Отбору признаков предшествует предобработка, позволяющая привести их в единый масштаб измерений и провести некоторые дополнительные улучшения.

Основные операции предобработки:

– удаление выбросов;
– нормализация;
– пропуск данных (потери).

Пусть X – множество признаков, X_r – подмножество из r признаков, $C(X_r)$ – мера отделимости классов на множестве признаков X_r.

Тогда задача выглядит следующим образом:

$$C(X_r) \to \max_{X_r \le X}.$$

Стратегия сокращения вектора признаков.

Пусть X – множество признаков.

Шаг алгоритма: для набора признаков X_r, выполняются условия, что $\max_{X_i \le X_r} C(X_r \setminus \{x_i\})$ и $X_{r-1}=X_r \setminus \{x_i\}$.

Условие остановки: $|C_{t+1} - C_t| < \varepsilon$, либо $r=m$.

Перед удалением избыточных признаков их ранжируют по релевантности с помощью методов векторного отбора, которые классифицируют следующим образом [8,9].

Оберточные методы (wrappers)

Основываются на прогностической эффективности заданного алгоритма обучения для оценки качества выбранных признаков. Учитывая конкретный алгоритм обучения, типичный оберточный метод состоит из двух этапов:

– поиск подмножества признаков;
– оценка выбранных признаков.

Оберточные методы повторяют поиск подмножества признаков и оценки выбранных признаков до тех пор, пока не будут удовлетворены некоторые критерии останова. Компонент поиска подмножества признаков сначала генерирует подмножество признаков, а затем алгоритм обучения действует как «черный ящик» для оценки качества этих признаков на основе результатов обучения. То есть, весь процесс работает итеративно до тех пор, пока не будет достигнута наивысшая эффективность обучения или не будет получено желаемое количество отобранных признаков. Затем подмножество признаков, обеспечивающее наивысшую эффективность обучения, возвращается в качестве выбранных признаков. Известной проблемой оберточных методов является то, что пространство поиска для d признаков составляет 2^d, что неприемлемо, когда d очень велико. В задачах выбора оптимального набора признаков используют три типа поиска: экспоненциальный, последовательный и рандомизированный [10].

Примерами оберточных методов являются генетические алгоритмы (Genetic Algorithms), последовательный выбор признаков (Sequential Forward Selection), обратное исключение (Sequential Backward Elimination) и др.

Методы фильтрации (filters)

Данные методы не зависят от каких-либо алгоритмов обучения. Для оценки важности признаков они опираются на характеристики данных. Методы фильтрации обычно более эффективны с вычислительной точки зрения, чем оберточные методы. Однако из-за отсутствия конкретного алгоритма обучения, управляющего фазой выбора признаков, выбранные признаки могут быть неоптимальными для целевых алгоритмов обучения. Обычно метод фильтрации состоит из двух этапов. На первом этапе важность признаков ранжируется в соответствии с некоторыми критериями оценки признаков. Процесс оценки важности признаков может быть как одномерным, так и многомерным. На втором этапе типичного метода фильтрации отфильтровываются признаки с низкими весами.

К группе методов фильтрации относятся [9]:

– методы, основанные на сходстве (критерий Фишера (Fisher Score), критерий коэффициента трассировки (Trace Ratio Criterion), алгоритм ReliefF, критерий Лапласа (Laplacian Score), коэффициент Джини (Gini Index) и др.);

– методы, основанные на статистиках (критерий T-Score, критерий F-Score, критерий Хи-квадрат, отбор признаков на основе меры корреляции (Correlation Based Feature Selection) и др.);

– методы, основанные на разреженном обучении (эффективный и надежный отбор признаков (Efficient and Robust Feature Selection), многокластерный отбор признаков (Multi-Cluster Feature Selection), выбор признаков с помощью неотрицательного спектрального анализа (Feature Selection Using Nonnegative Spectral Analysis) и др.);

– методы, основанные на теории информации (прирост информации (Information Gain), критерий минимальной избыточности/максимальной релевантности (Minimum Redundancy Maximum Relevance), быстрый фильтр на основе корреляции (Fast Correlation Based Filter) и др.).

Встроенные методы (embedded).

Такие подходы наделены достоинствами оберточных методов и методов фильтрации:

– включают взаимодействие с алгоритмом обучения;

– намного эффективнее оберточных методов, поскольку не требуется итеративно оценивать наборы признаков.

Наиболее широко используемыми встроенными методами являются модели регуляризации, которые нацелены на подгонку модели обучения путем минимизации ошибок подгонки и приведения коэффициентов признаков к малым значениям (или абсолютному нулю). После этого модель регуляризации и выбранные наборы признаков возвращаются в качестве окончательных результатов.

К данным методам относятся рекурсивное исключение признаков для метода опорных векторов (Recursive Feature Elimination for Support Vector Machine), отбор признаков с помощью персептронов (Feature Selection-Perceptron) и др.

Гибридные методы (hybrid)

Гибридные методы можно рассматривать как комбинацию нескольких алгоритмов отбора признаков. Основная цель – решить проблемы нестабильности и пертурбации ряда существующих алгоритмов отбора признаков.

Например, для небольших высокоразмерных данных небольшое изменение в обучающих данных может привести к совершенно другим результатам отбора признаков. Объединение подмножеств признаков, отобранных разными методами, позволяет повысить надежность результатов и, следовательно, достоверность отобранных признаков.

Обзор существующих исследований показал, что разработано большое количество различных методов, позволяющих ранжировать признаки по релевантности. При этом большой популярностью среди исследователей пользуются следующие методы отбора признаков: критерий хи-квадрат, прирост информации (Information Gain), индекс Джини, алгоритм ReliefF.

А. Хи-квадрат (Chi-Square,χ^2)

Критерий χ^2 – один из распространенных статистических методов, который оценивает независимость двух событий. Выбор признака по критерию хи-квадрат позволяет получить новый набор данных. Для определения того, является ли признак независимым от метки класса, критерий хи-квадрат использует тест независимости, который измеряет степень корреляции между признаком и классом. Учитывая конкретный признак f_i с r различными значениями признака, показатель хи-квадрат этого признака может быть вычислен по формуле:

$$\chi^2(f_i) = \sum_{j=1}^{r}\sum_{s=1}^{c}\frac{(n_{js}-\mu_{js})^2}{\mu_{js}},$$

где n_{js} – количество экземпляров с j-м значением для данного признака f_i, $\mu_{js} = \frac{n_{*s}n_{j*}}{n}$, n_{j*} обозначает число экземпляров данных с j-м значением признака f_i, а n_{*s} обозначает количество экземпляров данных в классе r. Чем больше показатель хи-квадрат, тем важнее признак.

Основным ограничением применения критерия хи-квадрат является предположение независимости признаков. Поскольку взаимодействие релевантных признаков не учитывается, фильтрация по сильно коррелированным признакам может ухудшить производительность классификатора.

B. Алгоритм ранжирования атрибутов на основе информационного усиления так же известный, как прирост информации (InfoGain) измеряет уменьшение энтропии до и после включения признаков. Признак с высоким значением прироста информации предпочтительнее других, то есть происходит ранжирование признаков по их значимости, при этом избыточные признаки не удаляются. Информационное усиление класса X, предоставляемое признаком Y, вычисляется следующим образом:

$$InfoGain(X|Y) = H(X)-H(X|Y),$$

где $H(X) = -\sum_{x_i \in X} P(x_i)\log_2 P(x_i)$ – энтропия рассматриваемого класса X до наблюдения признака Y,

$$H(X|Y) = -\sum_{y_j \in Y} P(y_j)\sum_{x_i \in X} P(x_i|y_j)\log_2(P(x_i|y_j)) -$$

энтропия после наблюдения признака Y.

Прирост информации или информационное усиление между X и Y используется для измерения количества информации, совместно используемой X и Y:

$$InfoGain(X;Y) = H(X) - H(X|Y) =$$

$$= \sum_{x_i \in X}\sum_{y_j \in Y} P(x_i,y_j)\log_2 \frac{P(x_i,y_j)}{P(x_i)P(y_j)},$$

где $P(x_i,y_j)$ – совместная вероятность x_i и y_j . Информационное усиление симметрично, так что $InfoGain(X;Y) = InfoGain(Y;X)$ и равно нулю, если дискретные переменные X и Y независимы.

C. Индекс Джини является широко используемой статистической мерой для количественной оценки того, способен ли признак разделять экземпляры из разных классов. Учитывая признак f_i с r различными значениями, предположим, что W и $\overline{W}$ обозначают множество экземпляров со значением признака меньше или равным j-му значению и большим, чем j-е значение, соответственно. Другими словами, j-е значение признака может разделить набор данных на W и $\overline{W}$, и тогда оценка индекса Джини для признака f_i определяется следующим образом:

$$GiniIndex(f_i) = \min_W \left(p(W)(1-\sum_{s=1}^{c} p(C_s|W)^2) + p(\overline{W})(1-\sum_{s=1}^{c} p(C_s|\overline{W})^2) \right),$$

где $p(.)$ обозначает вероятность. Например, $p(C_s|W)$ – это условная вероятность класса s с учетом W. Для бинарной классификации индекс Джини может принимать максимальное значение 0,5, также данный индекс может использоваться в задачах многоклассовой классификации. Чем ниже значение индекса Джини, тем более значимым является признак.

D. Алгоритм ReliefF выбирает объекты для разделения экземпляров из разных классов. Предположим, что l объектов данных выбираются случайным образом среди всех n объектов, тогда вес признака f_i в ReliefF определяется следующим образом:

$$ReliefF(f_i) = \frac{1}{c}\sum_{j=1}^{l}(-\frac{1}{m_j}\sum_{x_r \in NH(j)} d(X(j,i)-X(r,i)) +$$

$$+\sum_{y \neq y_j}\frac{1}{h_{jy}}\frac{p(y)}{1-p(y)}\sum_{x_r \in NM(j,y)} d(X(j,i)-X(r,i))),$$

где $NH(j)$ и $NM(j,y)$ – ближайшие объекты x_j в том же классе и в классе y, соответственно. Их размеры равны m_j и h_{jy}, соответственно. $p(y)$ – вероятность объектов в классе y. $X \in \mathbb{R}^{nxd}$ – матрица данных с n объектами и d признаками.

E. Отбор признаков на основе корреляции (Correlation-based Feature Selection, CFS) позволяет найти признаки, которые сильно коррелируют с целевой переменной, но имеют низкую интеркорреляцию между признаками, с помощью коэффициента корреляции. Для отбора признаков на основе корреляции вычисляется корреляция каждой пары признаков. Коэффициенты корреляции располагаются по убыванию. Оценка отбора признаков на основе корреляции имеет вид:

$$M_s = \frac{k\overline{r}_{cf}}{\sqrt{k+k(k-1)\overline{r}_{ff}}},$$

где M_s – эвристическая оценка подмножества S, содержащего k признаков, $\overline{r}_{cf}$ – среднее значение корреляции между признаками и классом, $\overline{r}_{ff}$ – средняя корреляция между признаками.

Числитель указывает на предсказательную силу набора признаков, а знаменатель показывает, насколько избыточным является набор признаков. Основная идея заключается в том, что лучшее для классификации подмножество признаков должно иметь сильную корреляцию с метками класса и слабую взаимосвязь. Для получения корреляции признаков с классом и корреляции признака с признаком, метод отбора признаков на основе корреляции использует симметричную неопределенность. Поскольку поиск глобально оптимального подмножества требует больших вычислительных затрат, используется стратегия поиска наилучшего варианта для поиска локально оптимального подмножества признаков. Сначала вычисляется значимость каждого признака с учетом корреляции между признаком и классом и признака с признаком.

В начале имеется пустой набор, который пополняется за счет признаков с наибольшей значимостью до тех пор, пока не будет удовлетворен некоторый критерий останова.

F. Фильтр, основанный на быстрой корреляции (Fast Correlation-Based Filter, FCBF), использует одновременно корреляцию между классами признаков и корреляцию рассматриваемого признака с признаками.

Алгоритм работает следующим образом:

1. Учитывая заданный порог δ, выбирается подмножество признаков S, которые сильно коррелируют с метками класса с $SU \geq \delta$, где SU – симметричная неопределенность. Симметричная неопределенность между набором признаков X_S и меткой класса Y задается следующим образом:

$$SU(X_S,Y) = 2\frac{I(X_S;Y)}{H(X_S)+H(Y)}.$$

Признак X_k называется преобладающим, если $SU(X_k,Y) \geq \delta$ и не существует признака $X_j \in S$ $(j \neq k)$ такого, что $SU(X_j,X_k) \geq SU(X_k,Y)$.

Признак X_j считается избыточным для признака X_k, если $SU(X_j,X_k) \geq SU(X_k,Y)$.

2. Набор избыточных признаков обозначается как S_{Pi}, который далее разбивается на $S^+_{P_i}$ и $S^-_{P_i}$, которые, в свою очередь, содержат избыточные признаки для признака X_k с $SU(X_j,Y) > SU(X_k,Y)$ и $SU(X_j,Y) < SU(X_k,Y)$, соответственно.

3. Различные эвристические методы применяются к S_P, $S^+_{P_i}$ и $S^-_{P_i}$ для удаления избыточных признаков и сохранения признаков, которые являются наиболее важными для определения класса.

Изложенные методы позволяют присваивать признакам веса и ранжировать их по релевантности.

К другим часто применяемым методам относятся: анализ главных компонентов (PCA) [11,12,29], факторный анализ (FA) [13], линейный дискриминантный анализ (LDA) [13,29], сингулярное разложение (SVD) [14,15], анализ основных компонентов ядра (Kernel PCA) [17], стохастическое вложение соседей с t-распределением (t-SNE) [17,18], многомерное масштабирование (MDS) [16,27], изометрическое отображение (Isomap) [17,27], анализ независимых компонент (ICA) [19,20], общий дискриминантный анализ (GDA) [13], канонический корреляционный анализ (CCA) [9,21], матричная факторизация [22,23], анализ соответствий [24,25], дискриминантный анализ Фишера с использованием ядра (KFDA) [16], локально линейное встраивание (LLE) [17], Laplacian Eigenmaps [17,26], отображение(проекция) Саммона [27], автоэнкодеры [28], метод опорных векторов с рекурсивным отбором признаков и перекрестной проверкой (SVM+RFECV) [30].

В качестве эталонных данных для оценки СОВ используются различные наборы данных (датасеты). CSE-CIC-IDS2018 [31] является одним из наиболее полных и применимых на практике датасетом. Достоинства данного набора [32]:

– данные размечены по различным классам атак;
– большой размер набора записей (440 ГБ);
– большой спектр атак;
– корректная архитектура сети;
– наличие возможности для выделения признаков из pcap-файлов.

Датасет CSE-CIC-IDS2018 включает в себя несколько различных сценариев атак: брутфорс, ботнет, DoS, DDoS, веб-атака, проникновение в сеть изнутри. Некоторые из них делятся на подгруппы, всего 15 классов, включая чистый трафик. В датасете учитываются 78 признаков, полученных CICFlowMeter-V3.

В ходе исследования применительно к датасету CSE-CIC-IDS2018 для снижения размерности был применен метод анализа соответствий [24].

Было получено разложение исходной матрицы $A(78\times n)$ в виде $= U\Sigma V$, где Σ – сингулярная матрица, то есть диагональная матрица (частный случай), на главной диагонали которой расположены корни из собственных значений матрицы A^T в порядке убывания. Матрицы U и V являются ортогональными. В матрице Σ выделяются первые r строк и столбцов, а оставшиеся исключаются. Первые r самых значимых сингулярных чисел называются главными компонентами. Реконструируем исходную матрицу с использованием меньшего объема входной информации:

$A(78 \times n) = U(78 \times r)\Sigma(r \times r)V\ (r \times n)$

Критерием качества восстановления матрицы А служит близость к единице коэффициента детерминации, который вычисляется по формуле: $Q(r) = \frac{\sum_{k=1}^{r}\lambda_k}{\sum_{k=1}^{n}\lambda_k}$, где λ_k – собственные значения A^T. Зависимость коэффициента детерминации от числа главных компонент позволяет оценить эффективность алгоритма [33].

В ходе исследования было выявлено, что из 78 оставшихся признаков целесообразно оставить 36 без существенных потерь точности.

Так для двух из 78 признаков коэффициент детерминации составляет 0,42, для 36 из 78 признаков коэффициент детерминации составляет 0,99.

Таблица 1

Коэффициенты детерминации признаков CSE-CIC-IDS2018

Число признаков	Коэффициент детерминации	Число признаков	Коэффициент детерминации
1	0,28	26	0,96
2	0,41	27	0,96
3	0,52	28	0,96
4	0,6	29	0,97
5	0,67	30	0,97
6	0,7	31	0,97
7	0,73	32	0,98
8	0,76	33	0,98
9	0,78	34	0,98
10	0,8	35	0,98
11	0,82	36	0,99
12	0,84	37	0,99
13	0,85	38	0,99
14	0,87	39	0,99
15	0,88	40	0,99
16	0,89	41	0,99
17	0,9	42	0,99
18	0,91	43	0,99
19	0,91	44	0,99
20	0,92	45	0,99
21	0,93	46	0,99
22	0,94	47	0,99
23	0,94	48	0,99
24	0,95	49	0,99
25	0,95	50	0,99

В таблице 1 приведены коэффициенты детерминации для различного числа признаков. Исходя из роста коэффициента, предлагается взять 36 признаков, так как дальнейший прирост незначительный.

Из недостатков данного подхода необходимо выделить следующие:

– данные необходимо нормализовать;

– полученные признаки не несут смысловую нагрузку, то есть невозможно качественно определить, какие признаки можно было бы удалять на момент первичной обработки.

Достоинство заключается в положительном применении указанного датасета в качестве обучающего для дальнейшего использования в рамках комплекса СОВ.

Выводы

Отбор признаков эффективен при предварительной обработке данных и снижении их размерности. Целью отбора признаков является построение более простых и полных моделей, повышение эффективности обработки данных. За последние несколько лет было разработано существенное количество методов отбора признаков. В данной статье представлены основные методы отбора признаков и показана важность применения отбора признаков для решения практических задач. В частности, классифицированы традиционные методы отбора признаков как методы, основанные на сходстве, информационно-теоретические методы, методы, основанные на разреженном обучении, статистические методы и другие в зависимости от используемой технологии. В заключение проверена эффективность метода анализа соответствий для датасета CSE-CIC-IDS2018. Данный метод позволил сократить входной вектор более, чем в 2 раза.

Литература

1. *Васютин С.В., Корнеев В.В., Райх В.В., Синица И.Н.* Принятие обобщенных решений в системах обнаружения вторжений, использующих несколько методов анализа данных мониторинга // Информационное противодействие угрозам терроризма, 2005, №4. С. 54-65.
2. *Borisenko B.B., Erokhin S.D., Fadeev A.S., Martishin I.D.* Intrusion detection using multilayer perceptron and neural networks with long short-term memory // Systems of Signal Synchronization, Generating and Processing in Telecommunications (SYNCHROINFO), 2021, pp. 1-6. DOI: 10.1109/SYNCHROINFO51390.2021.9488416.
3. *Erokhina O.V., Borisenko B.B., Martishin I.D., Fadeev A.S.* Analysis of the multilayer perceptron parameters impact on the quality of network attacks identification // Systems of Signal Synchronization, Generating and Processing in Telecommunications (SYNCHROINFO), 2021, pp. 1-6. DOI: 10.1109/SYNCHROINFO51390.2021.9488344.
4. *Burges C.J.C.* Dimension reduction: A guided tour // Foundations and Trends in Machine Learning, 2010, vol. 2, no. 4, pp. 275-365. DOI: 10.1561/2200000002.
5. *Li X.-B., Varghese S.J.* Adaptive data reduction for large-scale transaction data // European Journal of Operational Research, 2008, vol. 188, pp. 910-924. DOI:10.1016/j.ejor.2007.08.008.
6. Towards Data Science. 11 dimensionality reduction techniques you should know in 2021 URL:https://towardsdatascience.com/11-dimensionality-reduction-techniques-you-should-know-in-2021-dcb9500d388b (дата обращения: 06.08.2021).
7. *Местецкий Л.М.* Математические методы распознавания образов, курс лекций // М.– МГУ, 2004, 85 с.
8. *Ерохин С.Д., Ванюшина А.В.* Выбор атрибутов для классификации IP-трафика методами машинного обучения // T-Comm: Телекоммуникации и транспорт, 2018, Том 12, №9, с. 25-29. URL: https://cyberleninka.ru/article/n/vybor-atributov-dlya-klassifikatsii-ip-trafika-metodami-mashinnogo-obucheniya (дата обращения: 13.08.2021).
9. *Li J., Cheng K., Wang S., Morstatter F., Trevino R.P., Tang J., Liu H.* Feature selection: a data perspective // ACM Computing Surveys, 2017, vol. 50, no. 6, article 94, 45 p. DOI: 10.1145/3136625.
10. *Molina L. C., Belanche L., Nebot A.* Feature selection algorithms: a survey and experimental evaluation // IEEE International Conference on Data Mining, Proceedings, 2002, pp. 306-313. DOI: 10.1109/ICDM.2002.1183917.
11. *Jolliffe I.T.* Principal component analysis // Second Edition, Springer, 2007, 487 p.
12. *Шелухин О.И., Барков В.В., Полковников М.В.* Сравнительный анализ алгоритмов оценки количества и структуры атрибутов в задачах классификации мобильных приложений // Наукоемкие технологии в космических исследованиях Земли, 2019, т. 11, № 2, с. 90–100. DOI: 10.24411/2409-5419-2018-10263.
13. *Mardia K. V, Kent J.T., Bibby J.M.* Multivariate analysis. Probability and mathematical statistics // Academic Press Limited, 1995, 521 p.
14. *Stewart G.W.* On the early history of the singular value decomposition // SIAM Review, 1993, vol. 35, no. 4, pp. 551-566. DOI:10.1137/1035134.
15. *Lambers J.* The SVD algorithm // Lect. 6 Notes, vol. CME335, no. Spring Quarter 2010–11, pp. 1–2.
16. *Cho H.-W.* Nonlinear feature extraction and classification of multivariate data in kernel feature space // Expert Syst. Appl., 2007, vol. 32, no. 2, pp. 534–542. DOI: 10.1016/j.eswa.2005.12.007.
17. *Van der Maaten L., Postma E., Van den Herik J..* Dimensionality reduction: A comparative review // Tilburg University Centre for Creative Computing, Technical Report TiCC-TR 2009–005, 36 p.
18. *Van der Maaten L., Hinton G.* Visualizing data using t-SNE // Journal of Machine Learning Research, 2008, no.9, pp. 2579–2605.
19. *Hyvarinen A., Karhunen J., Oja E.* Independent component analysis // Book, John Wiley & Sons, 2001, 504 p.
20. *Tharwat A.* Independent component analysis: an introduction. // Applied Computing and Informatics, 2021, vol.17, no. 2, pp. 222-249. DOI:10.1016/j.aci.2018.08.006.
21. *Avron H., Boutsidis C., Toledo S., Zouzias A.* Efficient dimensionality reduction for canonical correlation analysis // Proceedings of the 30th International Conference on Machine Learning, in PMLR, 2013, no. 28(1), pp. 347-355.
22. *Snasel V., Horak Z., Kocibova J., Abraham A.* Reducing social network dimensions using matrix factorization methods // Proceedings of the 2009 International Conference on Advances in Social Network Analysis and Mining, pp. 348-351. DOI:10.1109/ASONAM.2009.48.
23. *Koren Y., Bell R., Volinsky C.* Matrix factorization techniques for recommender systems // IEEE Computer, 2009, no. 8, pp. 30–37. DOI: 10.1109/MC.2009.263.
24. *Erokhin S., Borisenko B., Fadeev A.* Reducing the dimension of input data for ids by using match analysis // Conference of Open Innovations Association, FRUCT, 2021, № 28, pp. 96-102. DOI: 10.23919/FRUCT50888.2021.9347629.
25. *Бурлаков М.Е.* Применение метода анализа соответствий для оптимизации комбинаций атрибутов у наборов данных // Вестник ПНИПУ, Электротехника, информационные технологии, системы управления, 2018, №26. UR https://cyberleninka.ru/article/n/primenenie-metoda-analiza-sootvetstviy-dlya-optimizatsii-kombinatsiy-atributov-u-naborov-dannyh (дата обращения: 16.08.2021).
26. *Belkin M., Niyogi P.* Laplacian eigenmaps for dimensionality reduction and data representation // Neural Computation, 2003, vol. 15, no. 6, pp. 1373-1396. DOI: 10.1162/089976603321780317.
27. *Ghojogh B., Ghodsi A., Karray F., Crowley M.* Multidimensional scaling, Sammon mapping, and Isomap: Tutorial and Survey. // arXiv:2009.08136, 2020, pp. 1-15.
28. *Kiarashinejad Y., Abdollahramezani S., Adibi A.* Deep learning approach based on dimensionality reduction for designing electromagnetic nanostructures // npj Computational Materials, 2020, vol. 6, no. 12, pp. 1-12. DOI: 10.1038/s41524-020-0276-y.

29. *Taghanaki S.A., Dehkordi B.Z., Hatam A., Bahraminejad B.* Synthetic feature transformation with RBF neural network to improve the intrusion detection system accuracy and decrease computational costs // International Journal of Information and Network Security (IJINS), 2012, vol. 1, no. 1, pp. 28-36. DOI:10.11591/IJINS.V1I1.339.

30. *Guyon I., Weston J., Barnhill S., Vapnik V.* Gene selection for cancer classification using support vector machines // Machine Learning, 2002, vol. 46, pp. 389–422. DOI:10.1023/A:1012487302797.

31. Datasets nadian Institute ybersecurity URL:https://www.unb.ca/cic/datasets/index.html (дата обращения: 19.08.2021).

32. *Ерохин С.Д., Журавлев А.П.* Сравнительный анализ открытых наборов данных для использования технологий искусственного интеллекта при решении задач информационной безопасности // Системы синхронизации, формирования и обработки сигналов, 2020. Т.3. №3. С. 12-19.

33. *Афанасьева А.А.* Вычисление сингулярного разложения матриц // Сборник статей. Всероссийская молодежная научная конференция «Все грани математики и механики», Томск, 2017. С. 162-167.

ANALYSIS OF EXISTING METHODS TO REDUCE THE DIMENSIONALITY OF INPUT DATA

Sergey D. Erokhin, *Moscow Technical University of Communications and Informatics, Moscow, Russia,* ***esd@mtuci.ru***
Boris B. Borisenko, *Moscow Technical University of Communications and Informatics, Moscow, Russia,* ***fepem@yandex.ru***
Ivan D. Martishin, *Moscow Technical University of Communications and Informatics, Moscow, Russia,* ***martishinid@gmail.com***
Alexander S. Fadeev, *Moscow Technical University of Communications and Informatics, Moscow, Russia,* ***aleksandr-sml@mail.ru***

Abstract
The explosive growth of data arrays, both in the number of records and in attributes, has triggered the development of a number of platforms for handling big data (Amazon Web Services, Google, IBM, Infoworks, Oracle, etc.), as well as parallel algorithms for data analysis (classification, clustering, associative rules). This, in turn, has prompted the use of dimensionality reduction techniques. Feature selection, as a data preprocessing strategy, has proven to be effective and efficient in preparing data (especially high-dimensional data) for various data collection and machine learning tasks. Dimensionality reduction is not only useful for speeding up algorithm execution, but can also help in the final classification/clustering accuracy. Too noisy or even erroneous input data often results in less than desirable algorithm performance. Removing uninformative or low-informative columns of data can actually help the algorithm find more general areas and classification rules and generally achieve better performance. This article discusses commonly used data dimensionality reduction methods and their classification. Data transformation consists of two steps: feature generation and feature selection. A distinction is made between scalar feature selection and vector methods (wrapper methods, filtering methods, embedded methods and hybrid methods). Each method has its own advantages and disadvantages, which are outlined in the article. It describes the application of one of the most effective methods of dimensionality reduction - the method of correspondence analysis for CSE-CIC-IDS2018 dataset. The effectiveness of this method in the tasks of dimensionality reduction of the specified dataset in the detection of computer attacks is checked.

Keywords: *intrusion detection systems (IDS); dataset; feature generation; feature selection methods, correspondence analysis, computer attacks (CA).*

References

1. S.V. Vasyutin, V.V. Korneev, V.V. Raikh, I.N. Sinitsa (2005). Making generalized decisions in intrusion detection systems using several methods of monitoring data analysis. *Information counteraction to terrorist threats*, 2005, no. 4, pp. 54-65.

2. B.B. Borisenko, S.D. Erokhin, A.S. Fadeev, I.D. Martishin (2021). Intrusion detection using multilayer perceptron and neural networks with Long Short-Term Memory. *2021 Systems of Signal Synchronization, Generating and Processing in Telecommunications (SYNCHROINFO).* IEEE. DOI: 10.1109/synchroinfo51390.2021.9488416

3. O.V. Erokhina, B.B. Borisenko, I.D. Martishin, A.S. Fadeev (2021). Analysis of the multilayer perceptron parameters impact on the quality of network attacks identification. *2021 Systems of Signal Synchronization, Generating and Processing in Telecommunications (SYNCHROINFO)*. IEEE. DOI: 10.1109/synchroinfo51390.2021.9488344

4. C.J.C. Burges (2010). Dimension reduction: A guided tour. *Foundations and Trends in Machine Learning*, vol. 2, no. 4, pp. 275-365. DOI: 10.1561/2200000002

5. X.-B. Li and V.S. Jacob (2008). Adaptive data reduction for large-scale transaction data. *European Journal of Operational Research*. Elsevier BV, 188(3), pp. 910-924. DOI: 10.1016/j.ejor.2007.08.008
6. Towards Data Science. 11 dimensionality reduction techniques you should know in 2021 URL:https://towardsdatascience.com/11-dimensionality-reduction-techniques-you-should-know-in-2021-dcb9500d388b (access date: 06.08.2021).
7. L.M. Mestetsky (2004). Mathematical methods of pattern recognition, a course of lectures. Moscow State University, 85 p.
8. S.D. Erokhin, A.V. Vanyushina (2018). Selecting attributes to classify IP traffic by machine learning methods. *T-Comm*, vol. 12, no. 9, pp. 25-29. URL: https://cyberleninka.ru/article/n/vybor-atributov-dlya-klassifikatsii-ip-trafika-metodami-mashinnogo-obucheniya (access date: 13.08.2021).
9. J. Li, K. Cheng, S. Wang, F. Morstatter, R. . Trevino, J. Tang, H. Liu (2018). Feature selection. ACM Computing Surveys. *Association for Computing Machinery (ACM)*, 50(6), pp. 1-45. DOI: 10.1145/3136625
10. L.C. Molina, L. Belanche, A. Nebot (2002). Feature selection algorithms: a survey and experimental evaluation. *2002 IEEE International Conference on Data Mining, 2002. Proceedings. IEEE Comput. Soc.* DOI: 10.1109/icdm.2002.1183917
11. I.T. Jolliffe (2007). Principal component analysis. Second Edition, Springer, 487 p.
12. O.I. Shelukhin, V.V. Barkov, M.V. Polkovnikov (2019). Comparative analysis of algorithms to estimate the number and structure of attributes in the classification tasks of mobile applications. *Science-intensive Technologies in Space Exploration*, vol. 11, no. 2, pp. 90-100. DOI: 10.24411/2409-5419-2018-10263
13. K.V. Mardia, J.T. Kent, J.M. Bibby (1995). Multivariate analysis. Probability and mathematical statistics. Academic Press Limited, 521 p.
14. G.W. Stewart (1993). On the early history of the Singular Value Decomposition. SIAM Review. *Society for Industrial & Applied Mathematics (SIAM)*, 35(4), pp. 551-566. DOI: 10.1137/1035134
15. J. Lambers (2010). The SVD algorithm. *Lect. 6 Notes*, vol. CME335, no. Spring Quarter 2010-11, pp. 1-2.
16. H.-W. Cho (2007). Nonlinear feature extraction and classification of multivariate data in kernel feature space. *Expert Systems with Applications. Elsevier BV*, 32(2), pp. 534-542. DOI: 10.1016/j.eswa.2005.12.007
17. L. Van der Maaten, E. Postma, J. Van den Herik (2009). Dimensionality reduction: A comparative review. Tilburg University Centre for Creative Computing, Technical Report TiCC-TR 2009-005, 36 p.
18. L. Van der Maaten, G. Hinton (2008). Visualizing data using t-SNE. *Journal of Machine Learning Research,* no.9, pp. 2579-2605.
19. A. Hyvarinen, J. Karhunen, E. Oja (2001). Independent component analysis. Book, John Wiley & Sons, 504 p.
20. A. Tharwat (2020). Independent component analysis: An introduction. *Applied Computing and Informatics*. Emerald, 17(2), pp. 222-249. DOI: 10.1016/j.aci.2018.08.006
21. H. Avron, C. Boutsidis, S. Toledo, A. Zouzias (2013). Efficient dimensionality reduction for canonical correlation analysis. *Proceedings of the 30th International Conference on Machine Learning, in PMLR*, no. 28(1), pp. 347-355.
22. V. Snasel, Z. Horak, J. Kocibova, A. Abraham (2009). Reducing social network dimensions using matrix factorization methods. *2009 International Conference on Advances in Social Network Analysis and Mining.* IEEE. DOI: 10.1109/asonam.2009.48.
23. Y. Koren, R. Bell, C. Volinsky (2009). Matrix factorization techniques for recommender systems. *Computer. Institute of Electrical and Electronics Engineers (IEEE),* 42(8), pp. 30-37. DOI: 10.1109/mc.2009.263.
24. S. Erokhin, B. Borisenko, A. Fadeev (2021). Reducing the dimension of input data for IDS by using match analysis. *2021 28th Conference of Open Innovations Association (FRUCT).* IEEE. DOI: 10.23919/fruct50888.2021.9347629.
25. M.E. Burlakov (2018). Application of correspondence analysis method for optimization of attribute combinations in datasets. *PNRPU Bulletin, Electrical Engineering, Information Technology, Control Systems*, 2018, no. 26. URL: https://cyberleninka.ru/article/n/primenenie-metoda-analiza-sootvetstviy-dlya-optimizatsii-kombinatsiy-atributov-u-naborov-dannyh (access data: 16.08.2021).
26. M. Belkin, P. Niyogi (2003). Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation. MIT Press - Journals,* 15(6), pp. 1373-1396. DOI: 10.1162/089976603321780317
27. B. Ghojogh, A. Ghodsi, F. Karray, M. Crowley (2020). Multidimensional scaling, Sammon mapping, and Isomap: Tutorial and Survey. arXiv:2009.08136, pp. 1-15.
28. Y. Kiarashinejad, S. Abdollahramezani, A. Adibi (2020). Deep learning approach based on dimensionality reduction for designing electromagnetic nanostructures. npj Computational Materials. *Springer Science and Business Media LLC,* 6(1). DOI: 10.1038/s41524-020-0276-y
29. S. Asgari Taghanaki, B., Zamani Dehkordi, A. Hatam, B. Bahraminejad (2012). Synthetic feature transformation with RBF neural network to improve the Intrusion Detection System Accuracy and Decrease Computational Costs. *International Journal of Information and Network Security (IJINS).* Institute of Advanced Engineering and Science, 1(1). DOI: 10.11591/ijins.v1i1.339.
30. I. Guyon, J. Weston, S. Barnhill, V. Vapnik (2002). Machine learning. *Springer Science and Business Media LLC*, 46(1/3), pp. 389-422. DOI: 10.1023/a:1012487302797
31. Datasets Canadian Institute for Cybersecurity URL:https://www.unb.ca/cic/datasets/index.html (access date: 19.08.2021).
32. S.D. Erokhin, A.P. Zhuravlev (2020). Comparative analysis of open data sets for artificial intelligence technologies in solving information security problems. Signal Timing, *Formation and Processing Systems*, 2020, vol. 3, no. 3, pp. 12-19.
33. A.A. Afanasyeva (2017). Computation of singular matrix decomposition. *Collection of articles. All-Russian youth scientific conference "All facets of mathematics and mechanics",* Tomsk, pp. 162-167.

Information about author:
Sergey D. Erokhin, *PhD (technical sciences), associate professor, rector, Moscow Technical University of Communications and Informatics, Moscow, Russia*
Boris B. Borisenko, *PhD (technical sciences), associate professor, lead researcher, Moscow Technical University of Communications and Informatics, Moscow, Russia*
Ivan D. Martishin, *researcher, Moscow Technical University of Communications and Informatics, Moscow, Russia*
Alexander S. Fadeev, *researcher, Moscow Technical University of Communications and Informatics, Moscow, Russia*