Об энтропийных критериях отбора признаков в задачах анализа данных*

Ю.А. Дубнов

Федеральное государственное учреждение "Федеральный исследовательский центр "Информатика и управление" Российской академии наук", г. Москва, Россия

Аннотация. В работе рассматривается задача понижения размерности пространства признаков для описания объектов в задачах анализа данных на примере бинарной классификации. В статье приводится обзор существующих подходов к решению данной задачи и предлагается несколько модификаций, в которых понижение размерности рассматривается как задача извлечения наиболее релевантной информации из признакового описания объектов и решается в терминах Шеноновской энтропии. Для выявления наиболее значимых признаков используются такие информационные критерии, как условная энтропия (conditional entropy), взаимная информация (mutual information) и расстояние Кульбака-Ляйблера (Kullback-Leibler divergence).

Ключевые слова: понижение размерности, отбор признаков, классификация, энтропия.

DOI 10.14357/20718632180205

Введение

С развитием науки и технологий машинного обучения, а также с ростом объемов исследуемых массивов данных становится все более актуальной задача понижения размерности признакового пространства. Решение задачи понижения размерности зависит от целей исследования, и единственно верного подхода к решению данной задачи, по-видимому, не существует [1-3].

Традиционно при решении прикладных задач машинного обучения, таких как распознавание изображений, прогнозирование, задачи скоринга и пр., объекты из обучающей коллекции характеризуются множеством признаков, числовых или категориальных. Например, в задачах регрессии определение статистической взаимосвязи между признаками (предикторами) и целевой переменной позволяет сделать выводы о значимости того или иного признака для предсказания целевой переменной. С другой стороны, в задачах классификации изображений размерность пространства, описывающего объекты-изображения, может достигать нескольких десятков тысяч. В таком случае построение статистических выводов по данным становится не только трудоемким, но и менее точным в силу возможного наличия шумовых признаков.

Помимо уменьшения вычислительной сложности и повышения точности алгоритмов машинного обучения, актуальность задачи понижения размерности продиктована еще и необходимостью визуализации данных. Так, при решении задачи кластеризации оказывается весьма наглядным отображение объектов на дву- или трехмерных диаграммах, особенно если в выбранных переменных кластеры оказываются визуально различимы.

Перейдем к формальной постановке задачи и рассмотрим следующую формулировку для задачи понижения размерности (Dimentionality Reduction). Пусть коллекция обучающих объектов задана матрицей X размерности $n \times d$, где каждый объект характеризуется несколькими числовыми признака-

* Работа выполнена при поддержке Российского фонда фундаментальных исследований (проект 17-07-00286).

ми. Требуется перейти от изначальной выборки к совокупности векторов меньшей размерности k, сохранив при этом структуру исходных данных.

$$X = \begin{pmatrix} x_{11} & \cdots & x_{1d} \\ \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{nd} \end{pmatrix} \overset{D.R.}{\rightarrow} Y = \begin{pmatrix} y_{11} & \cdots & y_{1k} \\ \vdots & \ddots & \vdots \\ y_{n1} & \cdots & y_{nk} \end{pmatrix} \quad (1)$$

В такой формулировке X и Y представляют собой матрицы объекты-признаки, где каждая строчка характеризует один объект, а столбцы представляют общие признаки, свойственные всем объектам коллекции. Новая матрица Y с меньшей размерностью может как состоять из некоторых столбцов исходной матрицы X, в этом случае говорят об отборе признаков (feature selection), так и содержать новые числовые значения, характеризующие объекты тем или иным образом, в этом случае говорят об извлечении информации из признаков (feature extraction) [4].

В данной статье приводится обзор наиболее распространенных методов понижения размерности, как для извлечения информации (раздел 2), так для отбора признаков на основе информационных критериев (раздел 3). В разделе 4 приводятся результаты компьютерных экспериментов по сравнению эффективности приведенных критериев на примере задачи классификации.

1. Методы извлечения информации

Методы извлечения информации нацелены на то, чтобы сконцентрировать всю имеющуюся информацию об объекте в ограниченном количестве переменных, и применяются, как правило, в приложениях, где объекты характеризуются тысячами признаков, например, при классификации небольших изображений и текстов.

Обзору существующих методов извлечения информации посвящено множество статей и отчетов [5-7]. Некоторые методы являются универсальными и приводятся в качестве базовых алгоритмов для анализа данных (PCA, LDA) [4], некоторые являются более специфическими и применяются для решения определенных прикладных задач. Так, для анализа цифровых сигналов используется метод независимых компонент (ICA) [8], для обработки аудиосигналов и систем компьютерного зрения – метод неотрицательного матричного разложения (NMF) [9], для наглядной визуализации многомерных данных – метод tSNE (t-distributed stochastic neighbor embedding) [10].

К этой категории относятся все методы снижения размерности, приводящие к генерации новых числовых векторов на основе изначального признакового описания объектов. Это могут быть как проекции на некоторое подпространство меньшей размерности, так и различные производные от изначальных признаков (например, логарифмические и степенные функции, линейные комбинации и пр.).

1.2. Метод главных компонент (PCA)

Метод главных компонент (Principal Component Analysis, PCA) является одним из самых распространенным методов анализа многомерных данных. Первые работы по этому методу датируются началом 20-го века (Pearson, 1901 г. [11]), а сам метод стал повсеместно использоваться с развитием технологий анализа данных и машинного обучения [3, 12].

Идея метода главных компонент состоит в представлении облака многомерных точек одним многомерным эллипсоидом, полуоси которого ортогональны и совпадают с собственными векторами ковариационной матрицы для данного многомерного вектора. И поскольку направление максимальной дисперсии для любого случайного вектора совпадает с направлением собственного вектора, соответствующего максимальному собственному значению, то проекции точек на направления, задаваемые собственными векторами, обладают максимально возможной дисперсией. Другими словами, проецирование на собственные вектора обеспечивает наибольший разброс точек, что оказывается эффективным при разделении точек по группам.

Представим ковариационную матрицу изначальной таблицы данных в виде

$$\Sigma = \{c_{ij}, i = \overline{1,d}, j = \overline{1,d}\}, \quad (2)$$

где

$$c_{ij} = cov(X_i, X_j) = \mathbb{E}[X_i X_j] - \mathbb{E}[X_i]\mathbb{E}[X_j]. \quad (3)$$

Без ограничения общности можно полагать, что данные центрированы, т.е. $\mathbb{E}[X_i] = 0$. В противном случае предварительно проводится процедура центрирования данных, никак не влияющая на разброс точек по какому-либо направлению. Таким образом, диагональные элементы ковариационной матрицы – это дисперсии признаков, а остальные элементы – ковариация между соответствующими парами признаков.

$$\Sigma = XX^T \quad (4)$$

В результате преобразования по методу главных компонент ковариационные составляющие обнуляются и матрица принимает диагональный вид. В матричном представлении преобразование метода главных компонент записывается следующим образом:

$$\Sigma_{PCA} = YY^T, \qquad Y = AX, \quad (5)$$

где матрица преобразования A состоит из собственных векторов ковариационной матрицы Σ из выражения 2.

Понижение размерности признакового пространства достигается путем отбрасывания наименее значимых собственных компонент и соответствующих им проекций. Например, если всего рассматривается d признаков, и в результате преобразования было получено d ортогональных проекций, то каждая из них описывает некоторую часть общего разброса в облаке точек. Тогда доля объясненной дисперсии при выборе первых k точек составит:

$$\delta = \frac{\lambda_1+\cdots+\lambda_k}{\lambda_1+\lambda_2+\cdots+\lambda_d}, \quad (6)$$

где $\lambda_i, i = 1, \dots, d$ – собственные числа ковариационной матрицы Σ, они же диагональные элементы матрицы Σ_{PCA}. И поскольку собственные числа λ_i убывают от большего к меньшему, то, как правило, достаточно выбрать первые k проекций с суммарной объясняющей дисперсией не менее $90-95\%$.

Метод главных компонент является крайне универсальным и эффективным для большинства задач анализа данных небольшой размерности. В приложениях, где речь идет о тысячах признаков, таких как распознавание изображений, вычисление собственных чисел становится крайне трудоемкой задачей. Для многомерных данных применяют модификацию метода главных компонент под названием LSA (Latent Semantic Analysis), основанную на сингулярном разложении матрицы признаков [13].

Основным недостатком метода главных компонент является его линейность, что приводит к смешиванию групп (классов) точек в задачах со сложной топологией. В таких случаях используют ядровые модификации метода главных компонент (Kernel PCA), основанные на использовании различных нелинейных и степенных функций вместо линейного преобразования в классическом PCA [14].

Еще одной ключевой особенностью и вместе с тем недостатком метода главных компонент является то, что и преобразование, и отбор значимых проекций производится без учета распределения самих точек по классам. Поэтому возможна ситуация, когда преобразование PCA лишь ухудшает разделение по классам, что особенно актуально в случае нескольких близких классов (Рис. 1 [1]). В таком случае, при использовании метода PCA будет выбрана проекция наибольшего разброса, и классификация точек окажется невозможной.

1.2. Линейный дискриминантный анализ

В отличие от метода главных компонент, при понижении размерности методом линейного дискриминантного анализа (Linear Discriminant Analysis, LDA) используется не только матрица

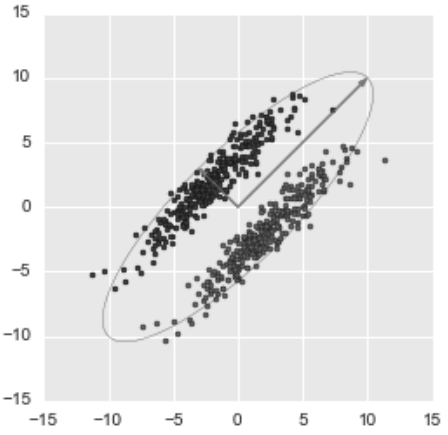

Рис. 1. Пример данных, образующих два близко расположенных класса

данных, но и метки классов для этих объектов. Метод был изобретен Фишером в 1936 году и иногда называется линейным дискриминантом Фишера [15]. Идея метода заключается в нахождении таких направлений, в проекциях на которые максимизируется метрика межклассового расстояния.

В общем случае каждому объекту $\bar{x}^i = \{x_{i1}, \dots, x_{id}\}, i = 1, \dots, N$ из обучающей коллекции соответствует метка класса $y_i \in \mathcal{C}$ из множества всех возможных классов, например, в случае бинарной классификации $\mathcal{C} = \{-1, +1\}$. Тогда по обучающей коллекции можно рассчитать вектора средних значений признаков отдельно для первого класса $\bar{\mu}_{y=-1}$ и для второго $\bar{\mu}_{y=+1}$, а также их ковариационные матрицы $\Sigma_{y=-1}$ и $\Sigma_{y=1}$ соответственно.

Поскольку проекция вектора признаков $\bar{x}' = \bar{v}\bar{x}$ является линейной комбинацией его компонент, то вектор средних значений и ковариационная матрицы для проекции изменятся следующим образом:

$$\bar{\mu}'_{y=c} = \bar{v}\bar{\mu}_{y=c}, \quad \Sigma'_{y=c} = \bar{v}^T\Sigma_{y=c}, \bar{v}, \quad c \in \mathcal{C}. \quad (7)$$

Фишер предложил как метрику качества при поиске направлений проецирования использовать отношение межклассовой дисперсии к внутриклассовой:

$$S(\bar{v}) = \frac{\left(\bar{\mu}'_{y=-1} - \bar{\mu}'_{y=+1}\right)^2}{\Sigma'_{y=-1} + \Sigma'_{y=+1}} = \frac{\left(\bar{v}(\bar{\mu}_{y=-1} - \bar{\mu}_{y=+1})\right)^2}{\bar{v}^T(\Sigma_{y=-1} + \Sigma_{y=+1})\bar{v}}. \quad (8)$$

Величина $S(\bar{v})$ определяется направлением проецирования и характеризует степень разделимости классов в проекциях на данное направление. Оптимальным выбором $\bar{v}$ является вектор, максимизирующий метрику (8):

$$\bar{v}_{LDA} = \operatorname{argmax} S(\bar{v}). \quad (9)$$

Таким образом, в результате преобразований по методу LDA будут получены проекции изначальных

[1] Источник: https://habrahabr.ru/post/304214/

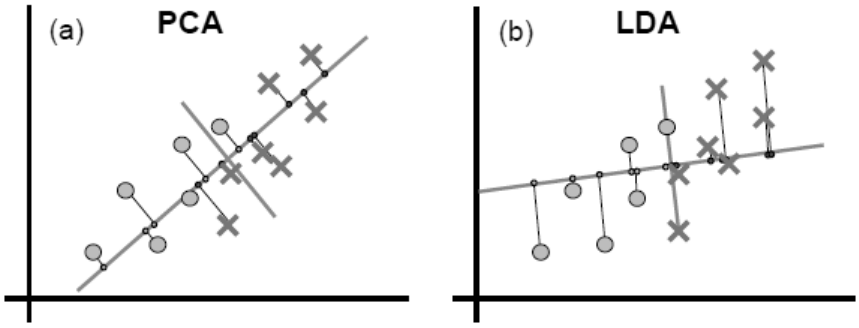

Рис. 2: Метод главных компонент (PCA) и метод линейного дискриминантного анализа (LDA)

точек, наилучшим образом разделяющие точки по классам. Понижение размерности, как и в методе главных компонент, достигается путем отбрасывания наименее значимых проекций. Однако в отличие от PCA, где проекции строились по принципу максимизации дисперсии, в данном случае проекции построены с учетом меток классов, что предотвращает случай возможного отбрасывания значимых проекций.

На Рис. 2 схематично продемонстрирована разница при построении проекций для методов PCA и LDA [2]. Метод главных компонент (PCA) выбирает проекции, максимизирующие разброс точек, что не гарантирует высокой точности классификации. Метод линейного дискриминантного анализа (LDA) выбирает проекции, максимизирующие специальную метрику разделимости классов.

Как и для метода главных компонент, для линейного дискримнантного анализа существуют ядровые расширения, необходимые для построения нелинейных разделяющих поверхностей [16]. На практике использование дискриминантного анализа оказывается особенно эффективным в задачах распознавания лиц и маркетинговых исследованиях при разделении пользователей по группам [17].

1.3. Случайные проекции (RP)

Оба рассмотренных выше метода широко распространены при работе с небольшими матрицами данных, однако в задачах с огромным количеством объектов ($N > 10^6$) и признаков ($d > 10^4$) необходимые для методов PCA и LDA вычисления становятся крайне трудоемкими и требуют множество вычислительных ресурсов. Поэтому в задачах, связанных с обработкой больших объемов данных, таких как распознавание по видеопотокам, их сегментирование, классификация коллекций документов и пр., для понижения размерности используют метод случайных проекций (Random Projections) [18].

Понижение размерности на основе случайных проекций базируется на простой идее, известной как лемма Джонсона-Линденштрасса [19]. Если точки векторного пространства проецируются на некоторое выбранное случайным образом подпространство достаточно большой размерности, то в среднем расстояния между точками сохраняются. В матричном виде это преобразование можно представить в виде

$$Y = XR, \tag{10}$$

где матрица R имеет размерность $[d \times k]$ и определяет выбранное подпространство размерности $k << d$, а X и Y – матрицы исходных и преобразованных данных с размерностями $[N \times d]$ и $[N \times k]$ соответственно.

Таким образом, для вычисления матрицы Y необходимо лишь перемножить матрицы X и R, следовательно, это преобразование имеет сложность $\mathcal{O}(kdN)$, в то время как в методах PCA и LDA получаемая после преобразования матрица имеет размерность $[d \times d]$. То есть преобразование к проекциям имеет сложность $\mathcal{O}(Nd^2)$, что и объясняет вычислительную эффективность метода случайных проекций для понижения размерности многомерных данных при $d > 10^4$.

Более того, в работе [20] было показано, что для генерации матрицы случайных чисел R может использоваться стандартное нормальное распределение, а при некоторых ограничениях даже равномерное.

2. Информационные критерии отбора

Стоит отметить, что процедура предварительного снижения размерности не всегда является обязательной при решении задачи анализа данных. Некоторые алгоритмы машинного обучения способны

[2] Источник: P. Cunningham. Dimension Reduction. Technical Report UCD-CSI-2007-7, University College Dublin, 2007.

выбирать наиболее значимые признаки в процессе обучения, к таким алгоритмам относятся решающие деревья (Decision Trees), модели с L_1 регуляризацией и набирающие особую популярность в последнее время нейронные сети с автоэнкодерами для глубокого обучения. Эта категория методов снижения размерности объединяет так называемые встроенные методы (embedded methods) и в данной работе не рассматривается, более подробную информацию о них можно найти, например, в работах [21, 22].

Методы отбора признаков используются для выбора наиболее значимых признаков и, вместе с тем, обнаружения признаков, не влияющих на целевую переменную, так называемых избыточных данных (от англ. redundant). Согласно широко распространенной терминологии, наряду со встроенными методами для отбора признаков различают методы фильтрации (filter/selection methods, [23]) и оберточные методы для генерации оптимального набора признаков на основе целевой функции потерь (wrapper methods, [24]).

Под генерацией оптимального набора признаков понимается выбор определенного набора исходных признаков, обеспечивающего наилучший из возможных результатов в терминах выбранного функционала качества. Причем, поскольку в общем случае задача выбора оптимального подпространства признаков имеет порядок сложности 2^d, то для ее решения, как правило, применяют различные эвристические, жадные и генетические алгоритмы, нацеленные на уменьшение перебора.

Методы фильтрации базируются на вычислении различных мер взаимосвязи между признаками и ответом. Например, в классической статистике значимые переменные определяются на основе коэффициентов корреляции и факторного анализа. С точки зрения теории информации мерой близости случайных векторов могут выступать такие понятия, как условная энтропия, взаимная информация и расстояние Кульбака-Ляйблера.

2.1. Энтропия и условная энтропия (CE)

Идея использования энтропии и связанных с ней понятий из теории информации является естественной, если формулировать задачу понижения размерности в терминах выбора наиболее значимых, то есть информативных признаков. Традиционно понятие энтропии используется для определения количества информации, заложенной в цифровом сигнале [25], поэтому, вычисляя энтропию отдельно для каждого признака, можно определить, какие признаки являются наиболее информативными в терминах энтропии.

Согласно обозначениям в выражении (1), разным признакам соответствуют столбцы матрицы данных. Пусть каждый признак $x_j, j = 1, \dots, d$ имеет ограниченный набор возможных значений $\mathcal{X}_j = \{x_1, \dots, x_{m_j}\}$, тогда энтропия для каждого из них будет рассчитываться следующим образом:

$$H(x_j) = -\sum_{x \in \mathcal{X}_j} p(x) \ln p(x), \qquad (11)$$

где $p(x)$ - эмпирическая функция распределения вероятности для данного признака, которая может быть рассчитана, например, гистограмным методом.

Энтропийный критерий показывает, сколько информации содержится в каждом признаке данного набора данных, но с другой стороны, никак не учитывает значения меток классов для обучающей выборки. Для определения влияния того или иного признака на отношение объекта к определенному классу может быть использовано значение условной энтропии (Conditional entropy, CE):

$$H(c|x_j) = \sum_{x \in \mathcal{X}_j} p(x) H(c|x) = \\ = -\sum_{x \in \mathcal{X}_j} p(x) \sum_{c \in \mathcal{C}} p(c|x) \ln p(c|x). \qquad (12)$$

Условная энтропия показывает вклад выбранного признака в определение класса c из множества доступных меток $\mathcal{C}$ и принимает значение 0, когда значения признака x_j полностью определяет класс объекта. С другой стороны, максимальное значение условной энтропии совпадает с энтропией переменной класса в случае независимых случайных величин, то есть если класс объекта никак не зависит от выбранного признака. Таким образом:

$$0 \le H(c|x_j) \le H(c), \qquad j = 1, \dots, d, \qquad (13)$$

или

$$0 \le \frac{H(c|x_j)}{H(c)} \le 1, \qquad j = 1, \dots, d. \qquad (14)$$

Отбор признаков в этой схеме производится по уменьшению значимости признака для определения класса, то есть в первую очередь выбирается признак с наименьшим значением, затем следующее минимальное значение и т.д.

Преимуществом данного метода является простота вычислений и скорость выполнения, что достигается за счет вычисления эмпирических функций плотности для каждого признака по отдельности. Следовательно, задача анализа многомерных данных сводится к нескольким задачам анализа одномерных случайных векторов.

2.2. Взаимная информация (MI)

Еще одно понятие из теории информации, учитывающее не только разброс значений одной переменной и относительное влияние одной переменной на другую, но и совместное распределение случай-

ных величин - взаимная информация (Mutual Information, MI). Взаимная информация характеризует количество информации, содержащееся в одной случайной величине относительно другой и определяется следующим образом:

$$MI(c, x_j) = \sum_{c\in\mathcal{C}} \sum_{x\in X_j} p(c,x)\ln\left(\frac{p(c,x)}{p(c)p(x)}\right), \quad (15)$$

где $p(c,x)$ - функция совместной плотности распределения вероятности для переменной класса и выбранного признака. Взаимная информация - неотрицательная величина, принимающая значение 0 для независимых случайных величин. Следовательно, критерием отбора признаков является максимизация взаимной информации значений признаков и значений зависимой переменной - метки класса:

$$MI(c, x_j) \to \max_{j=1,\dots,d}. \quad (16)$$

В некоторых работах этот критерий называют приростом информации (Information Gain, IG), поскольку, как и условная энтропия, он характеризует информационный вклад признаков в определение класса объекта [26].

Кроме того, полезным оказывается не только вычисление взаимной информации между переменной класса и признаками, но и между парами признаков. Низкие, близкие к 0, значения будут показывать независимость признаков друг от друга, а высокие – потенциальное дублирование информации, так называемые избыточные (redundant) признаки. Именно на этой идее основан метод mRMR (Max-Relevance and Min-Redundancy), изложенный в работе [27].

Иногда применяют нормализованный вариант описанного выше критерия – NMI (Normalized Mutual Information):

$$NMI(c, x_j) = \frac{MI(c,x_j)}{H(c)}, \quad 0 \le NMI(c, x_j) \le 1. \quad (17)$$

Равенство 0 означает независимость случайных величин, в то время как равенство 1 свидетельствует о том, что выбранный признак полностью определяет значения класса для объектов обучающей коллекции.

2.3. Расстояние Кульбака-Ляйблера (KL)

Расстояние Кульбака-Ляйблера (Kullback-Leibler divergence, KL) используется в теории информации и математической статистике для определения похожести функций распределения различных случайных величин. Метрика KL также используется для отбора признаков в задачах классификации [3, 28] и для преобразования по методу t-SNE с целью визуализации многомерных данных [10]. Для понижения размерности с помощью расстояния Кульбака-Ляйблера критерием отбора признаков будет уже не информационный вклад признаков, а различие условных эмпирических функций плотности распределения вероятности.

Пусть, как и ранее, некоторый признак x_j определен на множестве значений $\mathcal{X}_j = \{x_1, \dots, x_{m_j}\}$. Наличие обучающей выборки с метками классов для всех обучающих объектов позволяет вычислить условные распределения значений признака для объектов первого класса $p_1(x_j|c=-1)$ и второго класса $p_2(x_j|c=+1)$. Тогда критерий отбора признаков KL определяется следующим образом:

$$KL(p_1||p_2) = \sum_{x\in X_j} p_1(x)\ln\left(\frac{p_1(x)}{p_2(x)}\right). \quad (18)$$

Значение KL характеризует близость условных распределений значений признака для разных классов. Соответственно, чем эта величина выше, тем меньше похожи эти распределения, и наоборот, чем значение KL ниже, тем больше сходство между распределениями, что свидетельствует о неинформативности данного признака для разделения по классам.

Отметим, что метрика KL в выражении (18) – несимметричная, т.е.:

$$KL(p_1||p_2) \ne KL(p_2||p_1). \quad (19)$$

Поэтому на практике более универсальной является симметричная модификация для расстояния Кульбака-Ляйблера (Symmetrical KL, SKL):

$$SKL(p_1||p_2) = KL(p_1||p_2) + KL(p_2||p_1). \quad (20)$$

Данная метрика для отбора признаков оказывается особенно эффективной для задач классификации, в которых объекты разных классов могут иметь одинаковые значения признаков, но их распределение по объектам разных классов существенно отличается. К таким задачам относятся маркетинговые исследования, социологические и психологические опросы, а также спам-фильтры.

3. Результаты экспериментов

Приведенные выше методы понижения размерности были протестированы на примере задачи бинарной классификации. В качестве базового классификатора используется классический метод SVM (Support Vector Machine) с линейной разделяющей поверхностью [29].

3.1. Схема экспериментов

Для экспериментов использовались наборы данных из открытого репозитория данных для задач машинного обучения лаборатории KEEL (Knowledge Extraction based on Evolutionary Learning) [30].

В Табл. 1 приведены сведения об использующихся наборах данных: объём коллекции (N), число признаков (d) и средняя точность классификации без понижения размерности (acc). Точность классификации вычисляется традиционным методом кроссвалидации по 10 блокам (10-fold) и усредняется по результату 500 испытаний Монте-Карло.

В первую очередь стоит отметить, что получение наилучшей точности классификации не является целью настоящих экспериментов. Очевидно, что для каждого из приведенных наборов данных можно получить точность классификации выше указанной в Табл. 1 посредством оптимизации классификатора, например, используя предварительную стандартизацию данных и/или подбор оптимальной ядерной функции (Kernel function) для разделяющей поверхности, или используя классификаторы другого типа, например, логистическую регрессию или решающие деревья. С другой стороны, используя абсолютно одинаковые базовые классификаторы, можно сравнить эффективность разных методов понижения размерности.

Эксперименты проведены на персональном компьютере с 4-х ядерным процессором Intel(R) Core(TM) i7 CPU 920 @ 2.67 GHz. и 12 Gb оперативной памяти в програмной среде MATLAB R2017b с использованием инструментов статистического обучения (Statistics and Machine Learning Toolbox). Благодаря усреднению значений, все полученные результаты являются воспроизводимыми в пределах указанного среднеквадратического отклонения.

Табл. 1. Данные для тестирования

Назв.	N	d	acc, %
hepatitis	80	19	82.56 ± 2.22
appendicitis	106	7	87.67 ± 0.93
sonar	208	60	74.53 ± 1.52
spectfheart	267	44	79.23 ± 1.39
heart	270	13	83.69 ± 0.75
haberman	306	3	72.53 ± 0.49
bupa	345	6	68.85 ± 0.92
ionosphere	351	33	88.12 ± 0.80
bands	365	19	68.64 ± 0.88
wdbc	569	30	97.13 ± 0.43
wisconsin	683	9	96.72 ± 0.19
pima	768	8	77.06 ± 0.34
mammographic	830	5	82.27 ± 0.58

3.2. Обсуждение результатов

Результаты сравнения представлены в Табл. 2, где для каждого набора данных приведена средняя точность классификации при предварительном понижении размерности. Понижение размерности по методу главных компонент (PCA) производится по уровню объясненной дисперсии $\delta = 0.95$ (п.п. 1.1). В остальных столбцах понижение размерности производится в два этапа: первый этап – стандартное преобразование к проекциям на главные компоненты, аналогично методу PCA, однако последующий отбор признаков производится по критериям нормированной взаимной информации (NMI) и симметрического расстояния Кульбака-Ляйблера (SKL).

Как видно, использование информационных критериев отбора признаков вместо критерия объясненной дисперсии в классическом варианте PCA позволило повысить точность классификации практически во всех примерах. Такой результат объясняется, в первую очередь, тем, что информационные критерии учитывают влияние каждого признака на разделимость по классам, в то время как объясненная дисперсия гарантирует лишь наибольший разброс в терминах дисперсии.

Табл. 2. Точность классификации для разных критериев отбора признаков.

Назв.	PCA	NMI	SKL
hepatitis	83.18 ± 1.09	87.71 ± 1.11	90.04 ± 1.52
appendicitis	87.41 ± 0.64	88.24 ± 0.94	88.52 ± 0.98
sonar	77.31 ± 1.29	79.15 ± 0.50	76.99 ± 1.29
spectfheart	77.97 ± 1.05	81.02 ± 1.03	80.17 ± 1.22
heart	69.80 ± 0.77	85.06 ± 0.60	85.41 ± 0.51
haberman	72.53 ± 0.51	72.55 ± 0.50	72.55 ± 0.50
bupa	56.68 ± 0.81	69.48 ± 0.72	68.95 ± 0.88
ionosphere	86.30 ± 0.71	87.68 ± 0.81	87.38 ± 0.85
bands	63.01 ± 0.36	68.82 ± 0.95	69.26 ± 0.84
wdbc	90.78 ± 0.17	97.03 ± 0.30	96.86 ± 0.26
wisconsin	96.69 ± 0.19	97.11 ± 0.19	97.04 ± 0.20
pima	74.03 ± 0.18	77.01 ± 0.38	77.04 ± 0.35
mammographic	69.28 ± 0.26	82.87 ± 0.14	82.87 ± 0.14

Кроме того, в некоторых случаях разброс по одному из признаков оказывается настолько большим, что покрывает более 95% дисперсии, в этом случае по методу PCA будет отобрана всего одна главная компонента, что снижает точность классификации. Такой эффект наблюдается для датасетов wdbc, и mammographic. Соответственно, в этих примерах переход к энтропийным критериям существенно повышает точность.

Сравнивая критерии NMI и SKL, видно, что они показывают очень близкие значения, зачастую даже превосходящие точность классификации изначального набора данных без какого-либо понижения размерности. Это происходит в силу присутствия в наборах данных неинформативных признаков, значения которых практически не влияют на определение класса объектов. Именно свойство гарантированного избавления от таких признаков является одним из главных преимуществ энтропийных критериев отбора для понижения размерности.

Заключение

Все рассмотренные выше энтропийные критерии отбора признаков основаны на анализе информационного вклада отдельных признаков в определение класса объектов и, следовательно, не требуют существенных вычислительных ресурсов даже для задач больших и сверхбольших размерностей. Однако, в отличии от методов извлечения информации, таких как метод главных компонент и линейный дискриминантный анализ, энтропийные критерии не предполагают решение задач оптимизации, что, вообще говоря, не гарантирует повышения качества классификации при понижении размерности.

Таким образом, наиболее эффективным представляется комбинация традиционных методов понижения размерности и рассмотренных энтропийных критериев. Например, метод главных компонент позволяет преобразовать матрицу данных, избавившись от корреляции между признаками, а последующий отбор с помощью энтропийных критериев позволит выявить наиболее значимые компоненты, основываясь не на разбросе точек по плоскости, а на информационном вкладе каждого признака.

Литература

1. D.L. Donoho. High-dimensional data analysis: The curses and blessings of dimensionality. Lecture delivered at the "Mathematical Challenges of the 21st Century" conference of The American Math. Society, Los Angeles, August 6-11, 2000.
2. J. Friedman, T. Hastie, and R. Tibshirani. Elements of Statistical Learning: Prediction, Inference and Data Mining. Springer, 2001.
3. C. Bishop. Pattern Recognition and Machine Learning (Information Science and Statistics), Springer, 758 p., 2006.
4. E. Alpaydin. Introduction to Machine Learning. MIT Press, 3rd ed., 640 p., 2014
5. M.A. Carreira-Perpinan. A review of dimension reduction techniques. Technical report CS-96-09, Department of Computer Science, University of Sheffield, 1997.
6. Imola K. Fodor. A survey of dimension reduction techniques, Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, 2002.
7. P. Cunningham. Dimension Reduction. Technical Report UCD-CSI-2007-7, University College Dublin, 2007.
8. P. Comon, C. Jutten. Handbook of Blind Source Separation, Independent Component Analysis and Applications. Academic Press, Oxford UK., 2010.
9. Michael W. Berry; et al. Algorithms and Applications for Approximate Nonnegative Matrix Factorization // Computational Statistics & Data Analysis, vol.52, p.155-173, 2007.
10. L. van der Maaten, G. Hinton. Visualizing High-Dimensional Data Using t-SNE // Journal of Machine Learning Research, vol.9, p.2579-2605, 2008.
11. K. Pearson. On lines and planes of closest fit to systems of points in space // Philosophical Magazine, vol.2, p.559-572, 1901.
12. I.T. Jolliffe. Principal Component Analysis, Series: Springer Series in Statistics, 2nd ed., Springer, NY, XXIX, 487p., 2002.
13. S.C. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, and R.A. Harshman. Indexing by latent semantic analysis // Journal of the American Society of Information Science, vol.41(6), p.391-407, 1990.
14. B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear Component Analysis as a Kernel Eigenvalue Problem // Neural Computation, vol.10, no.5, p.1299-1319, 1998.
15. R.A. Fisher. The Use of Multiple Measurements in Taxonomic Problems // Annals of Eugenics, vol.7, p.179-188, 1936.
16. G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach // Neural Computation, vol.12(10), p.2385-2404, 2000.
17. G.J. McLachlan. Discriminant Analysis and Statistical Pattern Recognition. Wiley Interscience, 2004.
18. E. Bingham and H. Mannila. Random projection in dimensionality reduction:Applications to image and text data // Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York: Association for Computing Machinery, p.245-250. 2001.
19. W.B. Johnson and J. Lindenstrauss. Extensions of Lipshitz mapping into Hilbert space // In Conference in modern analysis and probability, vol.26 of Contemporary Mathematics, p.189-206, Amer. Math. Soc., 1984.
20. D. Achlioptas. Database-friendly random projections // Proceeding PODS'01 Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, p.274-281, 2001.
21. Guyon, A. Elisseeff. An Introduction to Variable and Feature Selection // Journal of Machine Learning Research, vol.3, p.1157-1182, 2003.

22. F.R. Bach. Bolasso: model consistent Lasso estimation through the bootstrap // Proceedings of the 25-th international conference on Machine learning, ICML'08, p.33-40, 2008.
23. Blum, P. Langley. Selection of relevant features and examples in machine learning // Artificial Intelligence, vol.97(1-2), p.245-271, 1997.
24. R. Kohavi, G. John. Wrappers for feature subset selection // Artificial Intelligence, vol.97, p.273-324, 1997.
25. T.M. Cover, J.A. Thomas. Elements of information theory. John Wiley and Sons Ltd., New-York, 561 p., 1991.
26. J. Abellán, J.G. Castellano. Improving the Naive Bayes Classifier via a Quick Variable Selection Method Using Maximum of Entropy // Entropy, vol.19, no.6, 247, 2017.
27. H.C. Peng, F. Long, C. Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy // IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.27(8), p.1226-1238, 2005.
28. Y. Zhang, S. Li, T. Wang, Z. Zhang. Divergence-based feature selection for separate classes // Neurocomputing, vol.101, p.32-42, 2013.
29. N. Christianini, J. Shawe-Taylor. An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge, UK: Cambridge University Press, 2000.
30. J. Alcalá-Fdez, A. Fernandez, J. Luengo, J. Derrac, S. Garcia, L. Sánchez, F. Herrera. KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework // Journal of Multiple-Valued Logic and Soft Computing, vol.17:2-3, p.255-287, 2011.

Дубнов Юрий Андреевич. Институт системного анализа Федерального государственного учреждения "Федеральный исследовательский центр "Информатика и управление" Российской академии наук" г. Москва, Россия. Научный сотрудник. Старший преподаватель НИУ ВШЭ г. Москва, Россия. Количество печатных работ: 7. Область научных интересов: динамические системы, машинное обучение, принцип максимума энтропии. E-mail: yury.dubnov@phystech.edu

On Entropic Criteria For Feature Selection In Data Analysis Problems

Yu. A. Dubnov

Federal Research Center “Computer Science and Control” of Russian Academy of Sciences, Moscow, Russia

The paper considers the problem of reducing the dimension of the feature space for describing objects in data analysis problems using the example of binary classification. The article provides a detailed overview of existing approaches to solving this problem and proposes several modifications. In which the dimensionality reduction is considered as the problem of extracting the most relevant information from the characteristic description of objects and is solved in terms of the Shanon's entropy. To identify the most significant features information criteria such as crossentropy, mutual information and Kullback-Leibler divergence are used.

Keywords: dimentionality reduction, feature selection, classification, entropy.

DOI 10.14357/20718632180205

References

1. D.L. Donoho. High-dimensional data analysis: The curses and blessings of dimensionality. Lecture delivered at the "Mathematical Challenges of the 21st Century" conference of The American Math. Society, Los Angeles, August 6-11, 2000.
2. J. Friedman, T. Hastie, and R. Tibshirani. Elements of Statistical Learning: Prediction, Inference and Data Mining. Springer, 2001.
3. C. Bishop. Pattern Recognition and Machine Learning (Information Science and Statistics), Springer, 758 p., 2006.
4. E. Alpaydin. Introduction to Machine Learning. MIT Press, 3rd ed., 640 p., 2014
5. M.A. Carreira-Perpinan. A review of dimension reduction techniques. Technical report CS-96-09, Department of Computer Science, University of Sheffield, 1997.
6. Imola K. Fodor. A survey of dimension reduction techniques, Center for Applied Scientific Computing, Lawrence Livermore National Laboratory, 2002.
7. P. Cunningham. Dimension Reduction. Technical Report UCD-CSI-2007-7, University College Dublin, 2007.
8. P. Comon, C. Jutten. Handbook of Blind Source Separation, Independent Component Analysis and Applications. Academic Press, Oxford UK., 2010.
9. Michael W. Berry; et al. Algorithms and Applications for Approximate Nonnegative Matrix Factorization // Computational Statistics & Data Analysis, vol.52, p.155-173, 2007.
10. L. van der Maaten, G. Hinton. Visualizing High-Dimensional Data Using t-SNE // Journal of Machine Learning Research, vol.9, p.2579-2605, 2008.

11. K. Pearson. On lines and planes of closest fit to systems of points in space // Philosophical Magazine, vol.2, p.559-572, 1901.
12. I.T. Jolliffe. Principal Component Analysis, Series: Springer Series in Statistics, 2nd ed., Springer, NY, XXIX, 487p., 2002.
13. S.C. Deerwester, S.T. Dumais, T.K. Landauer, G.W. Furnas, and R.A. Harshman. Indexing by latent semantic analysis // Journal of the American Society of Information Science, vol.41(6), p.391-407, 1990.
14. B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear Component Analysis as a Kernel Eigenvalue Problem // Neural Computation, vol.10, no.5, p.1299-1319, 1998.
15. R.A. Fisher. The Use of Multiple Measurements in Taxonomic Problems // Annals of Eugenics, vol.7, p.179-188, 1936.
16. G. Baudat and F. Anouar. Generalized discriminant analysis using a kernel approach // Neural Computation, vol.12(10), p.2385-2404, 2000.
17. G.J. McLachlan. Discriminant Analysis and Statistical Pattern Recognition. Wiley Interscience, 2004.
18. E. Bingham and H. Mannila. Random projection in dimensionality reduction:Applications to image and text data // Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York: Association for Computing Machinery, p.245-250. 2001.
19. W.B. Johnson and J. Lindenstrauss. Extensions of Lipshitz mapping into Hilbert space // In Conference in modern analysis and probability, vol.26 of Contemporary Mathematics, p.189-206, Amer. Math. Soc., 1984.
20. D. Achlioptas. Database-friendly random projections // Proceeding PODS'01 Proceedings of the twentieth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems, p.274-281, 2001.
21. I. Guyon, A. Elisseeff. An Introduction to Variable and Feature Selection // Journal of Machine Learning Research, vol.3, p.1157-1182, 2003.
22. F.R. Bach. Bolasso: model consistent Lasso estimation through the bootstrap // Proceedings of the 25-th international conference on Machine learning, ICML'08, p.33-40, 2008.
23. A. Blum, P. Langley. Selection of relevant features and examples in machine learning // Artificial Intelligence, vol.97(1-2), p.245-271, 1997.
24. R. Kohavi, G. John. Wrappers for feature subset selection // Artificial Intelligence, vol.97, p.273-324, 1997.
25. T.M. Cover, J.A. Thomas. Elements of information theory. John Wiley and Sons Ltd., New-York, 561 p., 1991.
26. J. Abellán, J.G. Castellano. Improving the Naive Bayes Classifier via a Quick Variable Selection Method Using Maximum of Entropy // Entropy, vol.19, no.6, 247, 2017.
27. H.C. Peng, F. Long, C. Ding. Feature selection based on mutual information: criteria of max-dependency, max-relevance, and min-redundancy // IEEE Transactions on Pattern Analysis and Machine Intelligence, vol.27(8), p.1226-1238, 2005.
28. Y. Zhang, S. Li, T. Wang, Z. Zhang. Divergence-based feature selection for separate classes // Neurocomputing, vol.101, p.32-42, 2013.
29. N. Christianini, J. Shawe-Taylor. An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge, UK: Cambridge University Press, 2000.
30. J. Alcalá-Fdez, A. Fernandez, J. Luengo, J. Derrac, S. Garcia, L. Sánchez, F. Herrera. KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework // Journal of Multiple-Valued Logic and Soft Computing, vol.17:2-3, p.255-287, 2011.

Dubnov Yury Andreevich. Institute for Systems Analysis Federal Research Center “Computer Science and Control” of Russian Academy of Sciences, 119333, 44/2 Vavilova str., Moscow, Russia, researcher. Senoir lecturer at Higher School of Economics (HSE), Moscow, Russia. Author of 7 scientific publications. Research interests: dynamic systems, machine learning, principle of maximum entropy. E-mail: yury.dubnov@phystech.edu