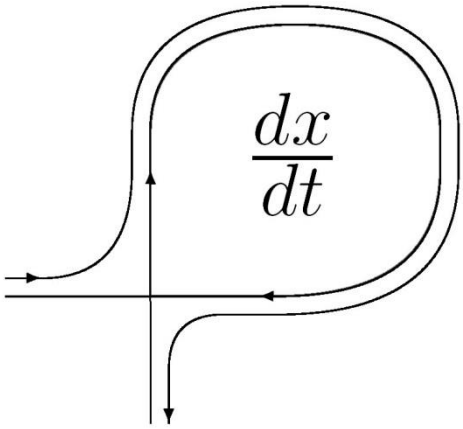

ДИФФЕРЕНЦИАЛЬНЫЕ УРАВНЕНИЯ И ПРОЦЕССЫ УПРАВЛЕНИЯ
N. 4, 2024
Электронный журнал,
рег. Эл № ФС77-39410 от 15.04.2010
ISSN 1817-2172

http://diffjournal.spbu.ru/
e-mail: jodiff@mail.ru

Машинное обучение и искусственный интеллект в процессах управления

Визуализация композиций многомерных объектов с локальными описаниями в метрических алгоритмах машинного обучения

Дюк В.А.

Институт проблем транспорта им. Н.С. Соломенко Российской академии наук

v_duke@mail.ru

Аннотация. В различных областях все чаще используются модели искусственного интеллекта (ИИ) для принятия тех или иных решений, основанные на машинном обучении. В метрических методах машинного обучения объекты рассматриваются как прецеденты и используется только одна операция – определение сходства (различия) этих прецедентов с неизвестным объектом. Основное ограничение эффективности известных метрических методов связано с представлением об общем для всех объектов пространстве признаков и, соответственно, о единой мере для измерения расстояний между объектами. Это ограничение снимается путем конструирования для любого объекта собственного локального пространства признаков и нахождения индивидуальной меры, определяющих иерархию его сходства с другими объектами, релевантную заданному контексту. В статье рассматривается проблема анализа множества объектов с локальными описаниями и пути её решения с использованием $d^{(S)}$-метрик, которые отражают различия рядов удаленностей одних и тех же объектов, но в разных локальных пространствах. Введение $d^{(S)}$-метрик позволяет использовать для дальнейшего визуального анализа композиций объектов с локальными описаниями методы многомерного метрического шкалирования. В статье приведен практический пример такого анализа в задаче распознавания типов транспортных средств по геометрическим признакам их силуэтов.

Ключевые слова: машинное обучение, искусственный интеллект, контекстно-зависимые локальные метрики.

1. Введение

В различных областях все чаще используются модели искусственного интеллекта (ИИ) для принятия тех или иных решений, основанные на машинном обучении (Machine Learning - ML). Разные компании применяют различные методы ML, от простых, таких как наивный байесовский классификатор, до более изощренных, таких как метод опорных векторов (Support Vector Machine - SVM), и более сложных, таких как современные нейросети с миллиардами настраиваемых параметров.

В метрических методах машинного обучения объекты рассматриваются как прецеденты и используется только одна операция – определение сходства (различия) этих прецедентов с неизвестным объектом. Сходство (различие) выражается геометрически через расстояние в пространствах признаков. В зависимости от условий конкретной задачи роль отдельного прецедента может меняться в широких пределах от главной до весьма косвенного участия. Этим, с одной стороны, обусловлено дальнейшее разделение метрических методов на подгруппы. С другой стороны, это разделение связано с различными мерами для определения сходства-различия объектов.

Простейшим является метод сравнения с прототипом. Он применяется тогда, когда классы ω_i объектов $\mathbf{x}$ отображаются в многомерном пространстве признаков компактными геометрическими группировками. В таком случае обычно в качестве точки – прототипа выбирается центр геометрической группировки класса (или ближайший к центру объект). Для классификации неизвестного объекта $\mathbf{x}$ находится ближайший к нему прототип, и объект относится к тому же классу, что и этот прототип.

Другими более продвинутыми являются метод k-ближайших соседей, в котором классифицируемый объект относится к классу большинства ближайших к нему объектов обучающей выборки [1], алгоритмы вычисления оценок, где сходство между классифицируемым и эталонными объектами определяется через так называемую «обобщенную близость», представленную комбинацией близостей, вычисленных на множестве частичных описаний [2], метод, основанный на функции конкурентного сходства (FRiS-функции) [3], ядерные методы, преобразующие данные в пространство большей размерности с помощью функций ядра, что позволяет работать с нелинейными отношениями (самый известный сегодня представитель – метод опорных векторов – SVM [4]).

Наиболее популярным из группы метрических является метод k-ближайших соседей. Первоначально он рассматривался как непараметрический метод оценивания отношения правдоподобия в окрестности $\mathbf{x}$. Для этого метода получены теоретические оценки его эффективности в сравнении оптимальным байесовским классификатором. Так, для случая $k = 1$ в [5] была доказана следующая теорема.

Пусть P_N — вероятность сделать по правилу первого ближайшего соседа (1-БС) в выборке $\mathbf{X}$ объема N. Тогда при распознавании двух классов в предположении, что из $\mathbf{X}$ делаются независимые случайные выборки с возвращением

$$P^* \le P_\infty \le 2P^*(1-P^*), P_\infty = \lim_{N\to\infty} P_N, \tag{1}$$

где P^* — риск ошибочной классификации любого случайным образом выбранного объекта при использовании байесовского метода оптимальной классификации.

В работе [6] приведен аналогичный результат для K классов

$$P^* \le P_\infty \le P^*\left(2 - \frac{K}{K-1}P^*\right). \tag{2}$$

Приведенные выражения показывают, что асимптотические вероятности ошибки для правила 1-БС превышают ошибки правила Байеса не более чем в два раза.

Основными преимуществами метода *k*-ближайших соседей являются:

- Строго доказанная точность, близкая к теоретически достижимому пределу при неограниченном увеличении объема выборки.
- Свобода от априорных предположений о структуре данных.
- Устойчивость к аномальным выбросам.
- Алгоритм хорошо распараллеливается.
- Результат работы алгоритма интерпретируем.
- Естественным образом реализуется технология «обучения с подкреплением».
- Модели, основанные на методе *k*-ближайших соседей слабо реагируют на ошибки разметки данных (ошибки учителя) [7].

Вместе с тем, у метода *k*-ближайших соседей отмечают следующие недостатки:

- Считается, что для работы метода требуется хранить в памяти всей обучающей выборки.
- Практически все исследователи говорят о непростой проблеме выбора метрики для измерения расстояния между объектами.
- Отмечается трудоемкость поиска ближайших соседей при больших объемах и размерностях данных.
- Нет теоретических оснований выбора определенного числа соседей.

2. Контекстно-зависимые локальные метрики

Основным недостатком, который ограничивает эффективность применения известных метрических методов, по нашему мнению, является представление об общем для всех объектов пространстве признаков и, соответственно, единой метрике для измерения расстояний между объектами. В задачах анализа данных, когда мы имеем дело с системами надкибернетического уровня сложности [8, 9], каждый объект следует рассматривать как самостоятельный информационный факт (совокупность событий), имеющий ценные уникальные особенности [10]. Указанные особенности раскрываются путем конструирования для любого объекта собственного локального пространства признаков и нахождения индивидуальной меры, определяющих иерархию его сходства с другими объектами, релевантную заданному контексту (контекстно-зависимые локальные метрики). Без такого раскрытия описания объектов нивелированы, могут содержать много ненужных, шумящих, отвлекающих и даже вредных деталей, и «сферы действия» объектов как представителей своих классов являются суженными.

Индивидуально сконструированные локальные метрики обеспечивают каждому объекту, как представителю своего класса, максимально возможную «сферу действия», которой нельзя достигнуть при построении общего пространства признаков и использовании одинаковой метрики для всех объектов. Описание каждого эмпирического факта оказывается полностью избавленным от неинформативных элементов, что позволяет в дальнейшем иметь дело с чистыми «незашумленными» структурами данных. В этом описании остается только то, что действительно важно для отражения сходства и различия эмпирического факта с другими фактами в контексте решаемой задачи.

Модели машинного обучения в этом случае представляют собой ансамбли объектов с привязанными к ним локальными контекстно-зависимыми метриками, что, вообще говоря, эквивалентно ансамблям линейных решающих правил [11]. Эти модели свободны от ряда вышеупомянутых недостатков метода *k*-ближайших соседей.

В свете представлений о контекстно-зависимых локальных метриках очевидно, что один и тот же объект может поворачиваться разными гранями своего многомерного описания сообразно заданному контексту. К любому объекту, запечатленному в памяти как целостная многомерная структура, "привязан" набор различных локальных метрик, каждая из которых оптимизирует

иерархию его сходства (различия) с другими объектами соответственно целям определенной задачи отражения отношений между объектами реального или идеального мира.

Представление о контекстно-зависимых локальных метриках позволяет объяснить, в частности, некоторые феномены психического отражения у человека, которые выглядят как нарушения метрических отношений между элементами матрицы близостей при применении техники парных сравнений. Например, в [12] описан эксперимент, где респондент, сравнивая "активную деятельную жизнь" ($\mathbf{x_1}$), "жизненную мудрость" ($\mathbf{x_2}$) и "здоровье" ($\mathbf{x_3}$), дал следующие оценки парных различий этих объектов: $d_{12} = 2, d_{13} = 1, d_{23} = 7$ Содержательно это означает, что респондент считает близкими ценности "активная деятельная жизнь" и "жизненная мудрость", а также "активная деятельная жизнь" и "здоровье". Однако считает далекими "здоровье" и "жизненную мудрость". Тем самым, хотя данные оценки (каждая по отдельности) являются интуитивно приемлемыми, их нельзя интерпретировать как геометрические расстояния между ценностями (нарушено неравенство треугольника $d_{23} > d_{12} + d_{13}$) и, соответственно, невозможно изобразить исследуемые объекты в виде точек в некотором статическом субъективном семантическом пространстве ценностных ориентаций.

Отмеченный факт мы объясняем существованием у респондента не одного, а нескольких субъективных подпространств с различными свойствами (локальными метриками). Так как внешние условия эксперимента являются постоянными, то смена локальных метрик может происходить вследствие изменения контекста, инициируемого различными парами сравниваемых объектов. Это влечет за собой разнокачественное восприятие сходства объектов и выражается в нарушении метрической аксиомы неравенства треугольника, которого бы не произошло, если бы субъективное пространство оставалось неизменным в ходе всего эксперимента.

При реализации представлений о контекстно-зависимых метриках на практике необходимо дополнительно решать задачу о выборе алгоритмов определения этих метрик и использовать специальные методы построения композиции прецедентов по матрицам близости с нарушением метрических отношений. Один из возможных вариантов определения контекстно-зависимых метрик описан нами в [13].

В следующем разделе рассмотрен подход к анализу композиций многомерных объектов с собственными локальными описаниями методами исследования многомерных структур, опирающимися на геометрическую метафору экспериментальной информации.

3. $d^{(s)}$-метрики в машинном обучении

В результате построения локальных метрик $d_i(\mathbf{x_i}, \mathbf{x_j}) = d_{ij}^{(L)}$ различных объектов, отношения между этими объектами описываются матрицей удаленностей $\boldsymbol{D}^{(\boldsymbol{L})} = \left(d_{ik}^{(L)}\right), i, k = \overline{1, N}$. Так как локальные метрики у разных объектов могут не совпадать, то для элементов матрицы $\boldsymbol{D}^{(\boldsymbol{L})}$ могут не выполняться требования симметричности и неравенства треугольника. Поэтому данная матрица, хотя и отражает отношения различия между объектами, не может истолковываться как матрица расстояний и в таком виде не пригодна для анализа совокупности таких объектов, с привязанными к каждому из них собственными локальными метриками, методами исследования многомерных структур, опирающимися на геометрическую метафору экспериментальной информации.

Для устранения нарушений метрических отношений между элементами матрицы $\boldsymbol{D}^{(\boldsymbol{L})}$ вводится специальный класс $d^{(S)}$-метрик, который был нами впервые предложен в [10,14]. Он определяется следующим образом.

$$d^{(S)}(\mathbf{x_i}, \mathbf{x_j}) = a \cdot S\left[\varphi\left(d_{ik}^{(L)}\right), \varphi\left(d_{jk}^{(L)}\right)\right] + b, k = \overline{1, N}, \tag{3}$$

где $d_{ik}^{(L)}$ и $d_{jk}^{(L)}$ – элементы i-й и j-й строк матрицы $\boldsymbol{D}^{(\boldsymbol{L})}$;

$\varphi\left(d_{ik}^{(L)}\right)$ – монотонное преобразование $d_{ik}^{(L)}$, либо преобразование в классификационный показатель $\varphi\left(d_{ik}^{(L)}\right) = \omega_m(k)$, где $m = rank\left(d_{ik}^{(L)}\right)$ и $\omega(k) = K_k$ – номер класса, к которому принадлежит $\mathbf{x_k}$;

$S[\cdot,\cdot]$ – мера подобия или различия двух последовательностей $\varphi\left(d_{ik}^{(L)}\right)$ и $\varphi\left(d_{jk}^{(L)}\right)$;

a и b – константы, значения которых подбираются с целью масштабирования и выполнения метрической аксиомы неравенства треугольника (так называемая модель с аддитивной константой [15, 16].

Расстояние между объектами $\mathbf{x_i}$ и $\mathbf{x_j}$, измеренное в $d^{(S)}$-метрике имеет следующий смысл. Образно говоря, если окинуть взором множество объектов с точки, занимаемой объектом $\mathbf{x_i}$, в пространстве, специально сконструированном для $\mathbf{x_i}$, то для такого взора объекты выстроятся в специфический ряд по степени удаленности от данной точки. С другой точки $\mathbf{x_j}$ и в другом пространстве ряд удаленностей тех же самых объектов будет иметь свой специфический вид. Мера сходства (различия) этих рядов S, подвергнутая линейному преобразованию с целью выполнения метрической аксиоматики – есть $d^{(S)}$ -расстояние между объектами $\mathbf{x_i}$ и $\mathbf{x_j}$.

Семейство $d^{(S)}$ -метрик отличается большим разнообразием, которое определяется множеством преобразований φ и мер подобия S.

Выбор конкретного преобразования φ зависит от того, на каком аспекте структуры данных исследователь решает сделать акцент. Например, может использоваться преобразование $d_{ik}^{(L)}$ в ранговую величину $\varphi\left(d_{ik}^{(L)}\right) = rank\left(d_{ik}^{(L)}\right)$. Это следует делать тогда, когда интерес представляет порядок удаленностей изучаемых объектов от $\mathbf{x_i}$.

Другой вариант – преобразование $d_{ik}^{(L)}$ в классификационный показатель. В этом случае все объекты, проранжированные по удаленности от $\mathbf{x_i}$, заменяются идентификатором своего класса, образно говоря, «окрашиваются» в цвета своего класса.

Выбор меры S зависит, с одной стороны, от вида преобразования φ и, с другой стороны, от того, какие особенности рядов $\varphi\left(d_{ik}^{(L)}\right)$ и $\varphi\left(d_{jk}^{(L)}\right)$, $(k = \overline{1,N})$ имеется намерение оттенить при определении их сходства (различия).

Прямой способ основан на вычислении расстояния (например, евклидова) между $\varphi\left(d_{ik}^{(L)}\right)$ и $\varphi\left(d_{jk}^{(L)}\right)$. В данном случае не требуется дальнейшего подбора констант a и b для соблюдения метрических требований, так как они выполняются автоматически, и $d^{(S)}$-метрика выглядит следующим образом:

$$d^{(S)}(\mathbf{x_i}, \mathbf{x_j}) : d^{(d)}(\mathbf{x_i}, \mathbf{x_j}) = \sum_{k=1}^{N} [\varphi(d_{ik}) - \varphi(d_{jk})]^2. \qquad (4)$$

Метрика $d^{(d)}(\mathbf{x_i}, \mathbf{x_j})$ может применяться, когда $\varphi(d_{ik})$ представлено количественной или ранговой величиной. Ее привлекательность, как отмечалось выше, обусловлена тем, что при переходе к ней автоматически выполняются требования метрической аксиоматики. Однако $d^{(d)}$-метрика нивелирует важные особенности рядов $\varphi(d_{ik})$ и $\varphi(d_{jk})$, для сравнения которых часто ценность представляет не столько сумма различий, сколько иерархия близостей объектов выборки к $\mathbf{x_i}$ и $\mathbf{x_j}$, измеренных в соответствующих локальных метриках. Поэтому бывает более целесообразно использовать в качестве меры S тот или иной коэффициент связи, например, коэффициент корреляции Пирсона, τ-Кендалла и др. Если преобразование $\varphi\left(d_{ik}^{(L)}\right)$ дает классификационную переменную, то мерой подобия может служить какой-либо коэффициент сопряженности для номинальных переменных.

Для того, чтобы привести употребленный коэффициент связи S к мере различия $d^{(S)}(\mathbf{x_i}, \mathbf{x_j})$, интерпретирующейся как расстояние между $\mathbf{x_i}$ и $\mathbf{x_j}$, достаточно применить дополнительное

преобразование вида $f(S) = b - S$. Функция $f(S)$ должна быть неотрицательной, и для нее должно выполняться неравенство треугольника $f(S_{ij}) \le f(S_{ik}) + f(S_{kj})$. Первое условие неотрицательности $f(S)$ легко выполнимо. Так как большинство коэффициентов связи изменяется в пределах от –1 до +1, можно применить, например, следующее преобразование

$$d^{(S)}(\mathbf{x_i}, \mathbf{x_j}) = \frac{1 - S_{ij}}{2}. \qquad (5)$$

В этом преобразовании осуществляется также масштабирование для того, чтобы $d^{(S)}(\mathbf{x_i}, \mathbf{x_j})$ принимало значение от 0 до 1.

В таком виде выражение (5) удовлетворяет первым двум аксиомам метрики – минимального различия объекта с самим собой и симметричности. Однако при использовании в качестве меры подобия S, например, коэффициента корреляции Пирсона в отдельных случаях возможны нарушения неравенства треугольника. В то же время, как показывает опыт, эти нарушения обычно незначительны. Поэтому подстановка коэффициента корреляции Пирсона в формулу (5) дает на практике достаточно хорошие результаты.

Если при использовании какого-либо коэффициента связи в матрице $d^{(S)}$-расстояний замечаются существенные нарушения неравенства треугольника, то они устраняются путем перехода к модели с аддитивной константой. Это выглядит как добавление к выражению (5) некоторой постоянной величины, представляющей собой минимально возможное постоянное слагаемое, при котором неравенство треугольника выполняется для всех троек объектов из анализируемой совокупности. Особенности решения задачи определения аддитивной константы, как уже указывалось, рассматриваются в [16].

Особое место среди различных $d^{(S)}$-метрик занимает метрика $d^{(\tau)}$, в которой $\varphi\left(d_{ik}^{(L)}\right) = rank\left(d_{ik}^{(L)}\right)$, а мерой подобия служит τ - Кендалла

$$d^{(\tau)}(\mathbf{x_i}, \mathbf{x_j}) = \frac{1 - \tau_{ij}}{2}, \qquad (6)$$

где $\tau_{ij} = \tau\left[\varphi\left(d_{ik}^{(L)}\right), \varphi\left(d_{jk}^{(L)}\right)\right], k = \overline{1, N}$ – коэффициент ранговой корреляции между переменными $\varphi\left(d_{ik}^{(L)}\right)$ и $\varphi\left(d_{jk}^{(L)}\right)$.

Значения $d^{(\tau)}$ изменяются в пределах от 0 до 1 и для данной меры различия объектов $\mathbf{x_i}$ и $\mathbf{x_j}$ в собственных локальных пространствах всегда выполняются метрические требования симметричности и неравенства треугольника.

Для доказательства данного утверждения напомним, что τ-Кендалла вычисляется по формуле

$$\tau_{ij} = \frac{4P_{ij}}{N(N-1)} - 1, \qquad (7)$$

где P_{ij} — количество пар, для которых совпадает порядок на ранговой переменной x_i с порядком на другой ранговой переменной x_j.

Примем во внимание, что величина $p_{ij} = 2P_{ij}/ N(N–1)$ трактуется как вероятность того, что два случайно выбранных объекта из N объектов имеют одинаковый порядок на i-й и j-й переменных. С учетом (7) и введенного обозначения p_{ij} выражение (6) примет вид

$$d^{(\tau)}(\mathbf{x_i}, \mathbf{x_j}) = 1 - p_{ij}. \qquad (8)$$

Таким образом неравенство треугольника запишется следующим образом

$$(1 - p_{ij}) \le (1 - p_{il}) + (1 - p_{jl}) \quad (9)$$

Или

$$p_{ij} \ge p_{il} + p_{jl} - 1. \quad (10)$$

Так как количество пар объектов, у которых совпадает порядок на x_i и x_j равно количеству пар объектов, у которых этот порядок одновременно совпадает на двух парах переменных (x_i и x_l) и (x_j и x_l) плюс количество пар объектов этот порядок одновременно не совпадает на этих же двух парах переменных, то справедливо следующее соотношение

$$p_{ij} = p_{il}p_{jl} + (1 - p_{il})(1 - p_{jl}). \quad (11)$$

Подставив (11) в (10), нетрудно убедиться, что неравенство треугольника выполняется для всех значений $p_{ij}, p_{il}, p_{jl} \le 1$. Следовательно, принимая во внимание справедливость для $d^{(\tau)}(\mathbf{x_i}, \mathbf{x_j})$ метрических требований рефлексивности и симметричности, выражение (6) удовлетворяет всем требованиям, предъявляемым к метрикам.

4. Практический пример

Теоретический и практический интерес представляет задача распознавания транспортных средств по сильно загрубленным их изображениям – силуэтам. В нашем примере используется экспериментальный материал из репозитория данных UCI (UCI Machine Learning Repository) [17].

Задача исследования

Требуется построить алгоритм автоматического распознавания типов транспортных средств, используя набор геометрических признаков, характеризующих их силуэты.

Экспериментальные данные

Для выделения признаков, описывающих форму силуэтов, использовалась специальная система HIPS (Hierarchical Image Processing System). С помощью этой системы силуэт описывался набором параметров, основанных на измерениях моментов, и на других измерениях типа отношения максимального радиуса к минимальному, статистических характеристик этих измерений (дисперсия, асимметрия, эксцесс). Кроме того, форма силуэтов описывалась рядом эвристических характеристик, отражающих компактность изображения, округлость, прямоугольность, впадины и др.

В эксперименте участвовали 4 транспортных средства: двухэтажный автобус (bus), микроавтобус Cheverolet (van), легковые автомобили Saab 9000 и Opel Manta 400.

Изображения объектов были получены камерой, расположенной на возвышении (34,2; 37,5 и 30,8 градусов). В данном эксперименте для упрощения этапа предобработки изображений объектов размещались на поверхности, подсвеченной рассеянным светом. Изображения имели слабое разрешение 128×128 пикселей, и были получены в серой шкале с 64 уровнями градаций серого.

Для получения силуэтов автомобилей вводился эвристически подобранный порог, и на выходе экспериментаторы имели бинарное изображение (только оттенки черного и белого цвета). Кроме того, применялась процедура для удаления помех в виде ошибочно белых элементов изображения

и черных точек. Все автомобили во время съемки вращались в координатной сетке от 0 до 360 градусов.

Всего было получено 210 изображений легковых автомобиля Opel, 216 изображений Saab, 218 изображений двухэтажного автобуса и 202 изображений микроавтобуса. Геометрические признаки силуэтов приведены в табл. 1.

Табл. 1. Названия геометрических признаков силуэтов транспортных средств и их обозначения

Оригинальное наименование признака	Тип признака	Обозначение
Compactness	Количественный	x_1
Circularity	Количественный	x_2
Distance Circularity	Количественный	x_3
Radius ratio	Количественный	x_4
Pr.axis aspect ratio	Количественный	x_5
Max.length aspect ratio	Количественный	x_6
Scatter ratio	Количественный	x_7
Elongatedness	Количественный	x_8
Pr.axis rectangularity	Количественный	x_9
Max.length rectangularity	Количественный	x_10
Scaled variance along major axis	Количественный	x_11
Scaled variance along minor axis	Количественный	x_12
Scaled radius of gyration	Количественный	x_13
Skewness about major axis	Количественный	x_14
Skewness about minor axis	Количественный	x_15
Kurtosis about minor axis	Количественный	x_16
Kurtosis about major axis	Количественный	x_17
Hollows ratio	Количественный	x_18

Полное описание и ссылки на таблицу экспериментальных данных приведены на страницах репозитория Statlog [18].

Применение напрямую для классификации транспортных средств метрического метода k-БС приводит к удовлетворительным, но не самым лучшим по сравнению, в частности, с методом «случайный лес» (random forest) результатам [19, 20]. При этом увеличение числа ближайших соседей k только ухудшает точность классификации транспортных средств. Этот феномен хорошо объясняется, если рассмотреть диаграмму рассеяния расстояний от какого-либо объекта до всех остальных объектов выборки. Приведем такую диаграмму, например, для объекта № 152, относящегося к классу легковых автомобилей (рис. 1).

Как следует из рис. 1, в исходном 18-мерном пространстве признаков к выбранному объекту 152 лишь первый ближайший сосед попадает в область точной классификации класса «легковые автомобили». Далее по мере увеличения номера ближайшего соседа разные классы фактически «перемешаны», встречаются со сопоставимыми значениями вероятностей. Аналогичная картина наблюдается и для других объектов выборки.

Для конструирования локальной метрики могут применяться различные критерии и алгоритмы, рассмотрение которых выходит за рамки настоящей статьи. Отметим здесь, что в данном случае в качестве критерия оптимальности локальной метрики в алгоритме поиска

взвешенной метрики Хэмминга использовалась площадь под концентрированной кривой ошибок (Concentrated Receiver Operating Characteristic – CROC) [20].

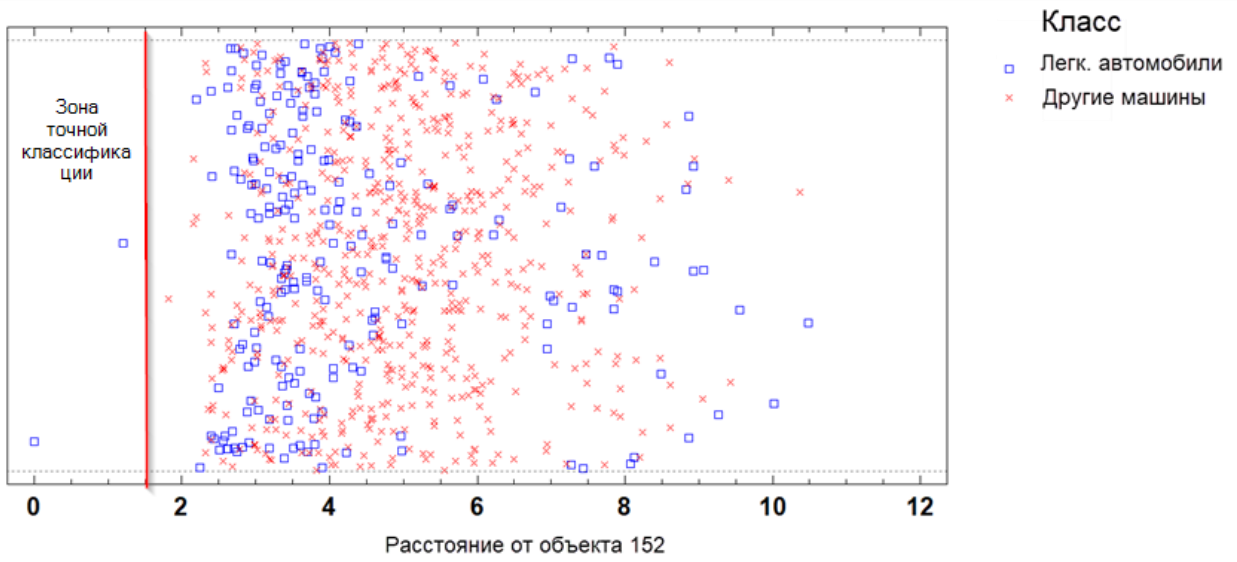

Рис. 1. Диаграмма рассеивания расстояний объектов выборки до объекта 152

В результате построения локальной взвешенной метрики для объекта № 152 из 18-ти исходных признаков эффективными оказались только 2 признака – x_3 с весом 2,1 и x_5, взятый с весом 4,8. Диаграмма рассеивания расстояния объектов выборки до объекта 152 в его собственном локальном пространстве приведена на рис. 2.

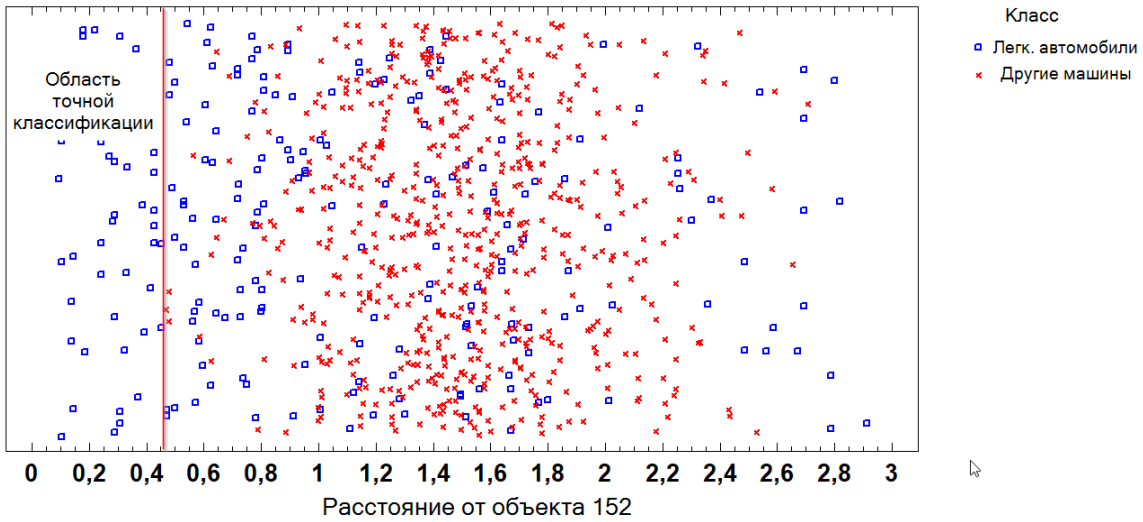

Рис. 2. Диаграмма рассеивания расстояний объектов выборки до объекта 152 в его собственном локальном пространстве.

Из рис. 2 видно, что область точной классификации класса «легковые автомобили» существенно расширилась. Если раньше, в исходном пространстве признаков в эту область попадал всего один объект, то теперь в локальном оптимизированном пространстве сюда вошло 43 объекта. Аналогичным образом была расширена «сфера действия» других объектов выборки, построив для них собственные локальные метрики.

Для достижения точности классификации автомобилей в 95 % предварительно было выделено 52 объекта с оптимизированными локальными метриками (11 объектов из класса «двухэтажный автобус», 24 объекта из класса «легковые автомобили Saab 9000 и Opel Manta 400» и 17 объектов из класса «микроавтобус Cheverolet»). Примеры метрик приведены в табл. 2. Отбор этих объекта

осуществлялся с использованием случайного поиска, и критерием для добавления объекта в композицию служила величина прироста значения правильной классификации для каждого класса в отдельности.

Табл. 2. Примеры формул для вычисления расстояний в оптимизированных локальных пространствах

Номер объекта	Формула для вычисления расстояния в локальном пространстве
1	1,25*Δ_6 + 1,22*Δ_8 + 1,61*Δ_14
5	0,91*Δ_6 + 0,43*Δ_8 + 1,52*Δ_10 + 3,67*Δ_14 + 0,03*Δ_15
10	1,43*Δ_6 + 1,19*Δ_8 + 1,06*Δ_14
20	8,10*Δ_3 + 3,65*Δ_10 + 2,04*Δ_13
23	1,29*Δ_6 + 0,37*Δ_8 + 1,39*Δ_10
30	1,03*Δ_6 + 8,56*Δ_10 + 2,12*Δ_14
50	6,49*Δ_14 + 14,33*Δ_17
51	1,69*Δ_6 + 1,43*Δ_8 + 0,44*Δ_10
60	0,24*Δ_6 + 6,08*Δ_10 + 3,44*Δ_14 + 0,18*Δ_15
63	0,19*Δ_4 + 1,72*Δ_6 + 1,53*Δ_8
100	1,85*Δ_3 + 4,13*Δ_8 + 7,42*Δ_10 + 1,94*Δ_14
103	2,85*Δ_1 + 1,86*Δ_3 + 2,80*Δ_8 + 6,57*Δ_10 + 0,98*Δ_14
113	0,53*Δ_3 + 4,16*Δ_5 + 1,66*Δ_6
118	3,47*Δ_6 + 1,33*Δ_8 + 0,22*Δ_10 + 0,10*Δ_15
200	0,29*Δ_3 + 4,10*Δ_5 + 1,34*Δ_6

Примечание: Здесь **Δ**_J означает расстояние Хэмминга по одному признаку x_J от объекта с соответствующим номером.

Вследствие того, что отбор объектов с локальными метриками (прецедентов) осуществлялся с использованием случайного поиска, результат такого отбора нуждается в дальнейшем исследовании и оптимизации. Эта оптимизация может быть произведена на основе визуального анализа геометрической структуры множества прецедентов с использованием $d^{(\tau)}$-метрик.

Фрагменты матриц коэффициентов ранговых корреляций τ-Кендалла и соответствующих $d^{(\tau)}$-расстояний приведены в табл. 3 и табл. 4.

Табл. 3. Фрагмент матрицы коэффициентов ранговых корреляций τ-Кендалла

	bus_1	bus_2	bus_3	…	car_1	car_2	car_3	…	van_1	van_2	van_3
bus_1	1,00	0,83	0,60	…	-0,25	-0,22	-0,12	…	-0,01	-0,12	-0,14
bus_2	0,83	1,00	0,64	…	-0,45	-0,40	-0,29	…	-0,09	-0,24	-0,07
bus_3	0,60	0,64	1,00	…	-0,25	-0,21	-0,16	…	0,08	-0,03	0,02
…	…	…	…	…	…	…	…	…	…	…	…
car_1	-0,25	-0,45	-0,25	…	1,00	0,82	0,50	…	0,24	0,44	-0,10
car_2	-0,22	-0,40	-0,21	…	0,82	1,00	0,58	…	0,18	0,35	-0,13
car_3	-0,12	-0,29	-0,16	…	0,50	0,58	1,00	…	0,22	0,34	-0,05
…	…	…	…	…	…	…	…	…	…	…	…
van_1	-0,01	-0,09	0,08	…	0,24	0,18	0,22	…	1,00	0,70	0,43
van_2	-0,12	-0,24	-0,03	…	0,44	0,35	0,34	…	0,70	1,00	0,38
van_3	-0,14	-0,07	0,02	…	-0,10	-0,13	-0,05	…	0,43	0,38	1,00
…	…	…	…	…	…	…	…	…	…	…	…

Табл. 4. Фрагмент матрицы $d^{(\tau)}$-расстояний

	bus_1	bus_2	bus_3	…	car_1	car_2	car_3	…	van_1	van_2	van_3
bus_1	0	0,08	0,2	…	0,63	0,61	0,56	…	0,5	0,56	0,57
bus_2	0,08	0	0,18	…	0,73	0,7	0,65	…	0,54	0,62	0,54
bus_3	0,2	0,18	0	…	0,62	0,61	0,58	…	0,46	0,51	0,49
…	…	…	…	…	…	…	…	…	…	…	…
car_1	0,63	0,73	0,62	…	0	0,09	0,25	…	0,38	0,28	0,55
car_2	0,61	0,7	0,61	…	0,09	0	0,21	…	0,41	0,32	0,57
car_3	0,56	0,65	0,58	…	0,25	0,21	0	…	0,39	0,33	0,52
…	…	…	…	…	…	…	…	…	…	…	…
van_1	0,5	0,54	0,46	…	0,38	0,41	0,39	…	0	0,15	0,29
van_2	0,56	0,62	0,51	…	0,28	0,32	0,33	…	0,15	0	0,31
van_3	0,57	0,54	0,49	…	0,55	0,57	0,52	…	0,29	0,31	0
…	…	…	…	…	…	…	…	…	…	…	…

Дальнейшая обработка матрицы $d^{(\tau)}$-расстояний производилась методом метрического многомерного шкалирования [21, 22]. Полученные первые 3 главные компоненты (ГК) объясняют 87,2 % дисперсии значений матрицы $d^{(\tau)}$-расстояний, что считается достаточно информативным для отображения данных в трехмерный объем без существенного искажения их геометрической структуры. Проекция 57 объектов с оптимизированными локальными описаниями в пространство первых трех главных компонент приведена на рис. 3.

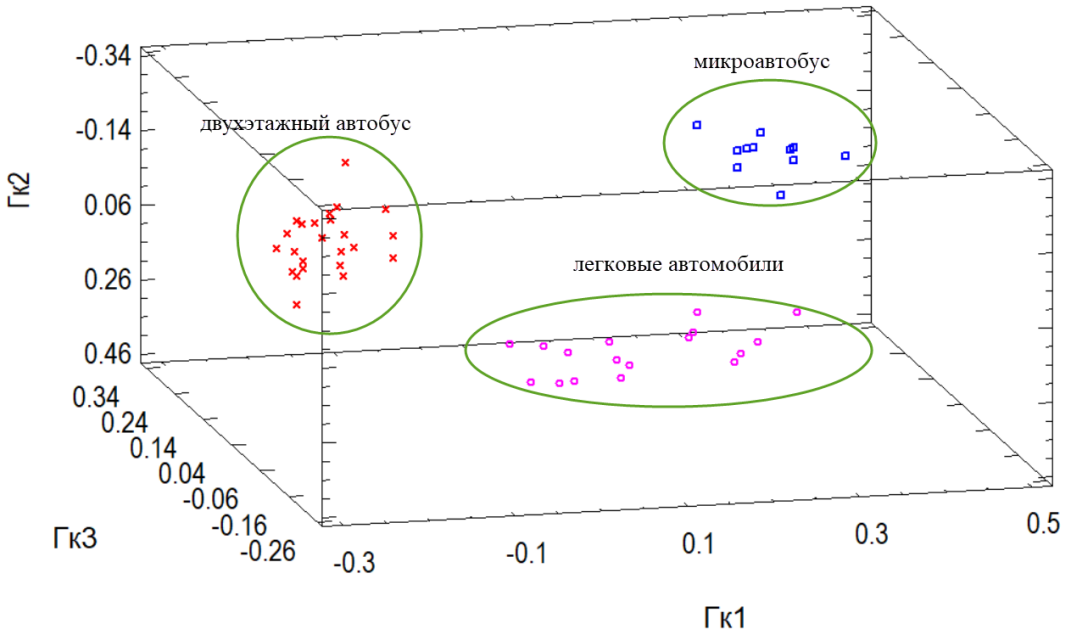

Рис. 3. Проекция объектов с оптимизированными локальными описаниями в пространство первых трех главных компонент

Как следует из рис. 3, объекты разных классов образуют относительно компактные собственные группировки. В классе «легковые автомобили» группировка несколько «размыта», так как в неё входят машины разных марок Saab 9000 и Opel Manta 400. Вместе с тем, как и в этой группировке, так и во всех остальных наблюдаются случаи геометрически очень близких объектов, что дает основание для вывода о дублировании информации и для соответствующего уменьшения количества объектов-прецедентов в том или ином классе автомобилей.

5. Заключение

Метрические методы занимают видное место в области машинного обучения. Они имеют ряд достоинств, в частности, свободу от априорных предположений о структуре данных интерпретируемость модели, устойчивость к ошибкам разметки данных, параллельность операции вычисления расстояний между объектами и др. Вместе с тем, метрические методы имеют недостатки, которые ограничивают их применение: необходимость хранения в памяти всей обучающей выборки, сложность и неопределенность при выборе метрики для измерения расстояния между объектами, а также количества учитываемых в модели ближайших объектов. Использование локальных описаний объектов (контекстно-зависимых локальных метрик) позволяет обойти указанные ограничения, существенно снизить объем запоминаемых объектов и повысить точность моделей классификации данных и прогнозирования.

Выше нами рассмотрен подход к визуальному анализу композиций многомерных объектов с собственными локальными описаниями. Этот подход основан на применении специальных $d^{(s)}$-метрик, которые отражают различия рядов удаленностей одних и тех же объектов, но в разных локальных пространствах. Введение $d^{(S)}$-метрик позволяет использовать для дальнейшего визуального анализа композиций объектов с локальными описаниями методы многомерного метрического шкалирования. В статье приведен практический пример такого анализа в задаче распознавания типов транспортных средств по геометрическим признакам их силуэтов.

6. Литература

[1] Fix E., Hodges J.L. Discriminatory analysis: nonparametric discrimination: consistency properties. – Rep. N 4. – USAF school of Aviation Medicine. – Texas. – February 1951. – Project 21-49-004. - Contract AF-41-(128)-31.

[2] Журавлев Ю.И., Никифоров В.В. Алгоритмы распознавания, основанные на вычислении оценок // Кибернетика. — 1971. — 1-11 с.

[3] Zagoruiko N.G. , Borisova I.A. , Dyubanov V.V. , Kutnenko O.A. Methods of Recognition Based on the Function of Rival Similarity//Pattern Recognition and Image Analysis, 2008, Vol. 18. No.1. pp.1-6.

[4] Vapnik, Vladimir N. The nature of statistical learning theory.: Springer-Verlag New York, Inc., 1995.

[5] Cover T., Hart P. Nearest neighbour pattern classification//IEEE Trans. Inform. Theory, v. IT 13, 1967. P. 21-27.

[6] Дуда Р., Харт П. Распознавание образов и анализ сцен. – М.: Мир, 1976. 509 с.

[7] Дюк В.А. Экспериментальное исследование реакции алгоритмов машинного обучения на ошибки разметки данных // Дифференциальные уравнения и процессы управления. Электронный журнал. - http://diffjournal.spbu.ru/, 2022. №3. С. 59-72.

[8] Boulding K. E. General Systems Theory – The Skeleton of Science // Management Science, 2, 1956.

[9] Гик Дж., ван. Прикладная общая теория систем. – М.: Мир, 1981.

[10] Дюк В.А. Компьютерная психодиагностика. – СПб: «Братство», 1994. – 364 с.

[11] Таунсенд К., Фохт Д. Проектирование и программная реализация экспертных систем на персональных ЭВМ. – М.: Финансы и статистика, 1990.

[12] Крылов В.Ю. Метод многомерной геометризации психологических данных. Системный подход в математической психологии // Принцип системности в психологических исследованиях. – М.: Наука, 1990. – С. 33-48.

[13] Дюк В.А., Малыгин И.Г., Прицкер В.И. Распознавание транспортных средств по силуэтам -трехкаскадный метод машинного обучения в системах технического зрения // Морские интеллектуальные технологии. – 2022. – № 2-1(56). – С. 162-167.

[14] Dyuk, V. A. Context-dependent local metrics and a geometrical approach to the problem of knowledge formation / V. A. Dyuk // Izvestiya Rossiiskoi Akademii Nauk. Teoriya i Sistemy Upravleniya. – 1996. – No. 5. – P. 36-44.
[15] Справочник по прикладной статистике. В 2-х т. Т. 2 // Под ред. Э. Ллойда, У. Ледермана, С.А. Айвазяна, Ю.Н. Тюрина. – М.: Финансы и статистика. – 1990.
[16] Saito T. The problem of the additive Constante and eigenvalues in metric multidimsional scaling //Psychometrika, v. 43, N 2, 1978.
[17] http://www.ics.uci.edu/~mlearn/MLRepository.html (дата обращения 10.06.2024)
[18] https://archive.ics.uci.edu/ml/datasets/Statlog+(Vehicle+Silhouettes). (дата обращения 10.06.2024)
[19] Дюк В.А., Малыгин И.Г. Сравнение алгоритмов распознавания типов транспортных средств по параметрам их силуэтов // Морские интеллектуальные технологии. – 2018. – № 4-4(42). – С. 197-201.
[20] Swamidass S. J., Azencott C., Daily K., Baldi P. A CROC stronger than ROC: measuring, visualizing and optimizing early retrieval // Bioinformatics (2010) 26 (10): 1348-1356.
[21] Torgerson W.S. Multidimensional Scaling. Theory and Method // Psychometrika, v. 17, № 4, 1952.
[22] Дэйвисон М. Многомерное шкалирование: Методы наглядного представления данных. – М.: Финансы и статистика, 1988.

Visualization of Compositions of Multidimensional Objects with Local Descriptions in Metric Machine Learning Algorithms

Diuk V.A.

Solomenko Institute of Transport Problems of the Russian academy of sciences

v_duke@mail.ru

Abstract. In various fields, artificial intelligence (AI) models are increasingly used for decision-making based on machine learning. In metric machine learning methods, objects are treated as precedents, and only one operation is used: determining the similarity (difference) between these precedents and an unknown object. The main limitation of existing metric methods is related to representing a common feature space for all objects and, consequently, a single measure for measuring distances between objects. This limitation is overcome by constructing a unique local feature space for each object and finding individual measures that determine the hierarchy of its similarity to other objects, relevant to the given context. The article discusses the problem of analyzing sets of objects with local descriptions and proposes a solution using $d^{(S)}$-metrics, which reflect differences in distance series between the same objects but in different local spaces. Introducing $d^{(S)}$-metrics allows for further visual analysis of compositions of objects with local descriptions using multidimensional metric scaling. The article provides a practical example of such analysis in the task of recognizing vehicle types based on geometric features of their silhouettes.

Keywords: machine learning, artificial intelligence, context-dependent local metrics.