УДК 004.93

А.В. Бармак, Ю.В. Крак, Э.А. Манзюк, В.С. Касьянюк

ИНФОРМАЦИОННАЯ ТЕХНОЛОГИЯ СИНТЕЗА РАЗДЕЛЯЮЩИХ ГИПЕРПЛОСКОСТЕЙ ДЛЯ ЛИНЕЙНЫХ КЛАССИФИКАТОРОВ

Ключевые слова: многомерное шкалирование, многомерная классификация, визуализация данных.

Введение и постановка задачи

Рассмотрим достаточно широкий класс моделей, используемых для проектирования систем классификации информации (временные сигналы, изображения и т.д.). Несмотря на то, что результатов современных исследований для указанных систем достаточно много [1–3], остается актуальной проблема повышения качества классификации в задачах распознавания образов, анализе текстовых контентов и других. Исследования, приведенные в данной работе, коррелируются с исследованиями, в которых используются методы группового учета аргументов МГУА [1], метод опорных векторов (SVM) [2], оптимального синтеза линейных и нелинейных преобразований [3], регрессионного анализа и тому подобное. В отличие от этих важных и хорошо известных методов, в данной работе предлагается подход, реализованный в виде некоторой технологии как совокупности вычислительных и информационных процедур, который позволяет: 1) визуализировать рассматриваемые объекты в двумерном пространстве для оценки их распределения; 2) в полученном двумерном пространстве эмпирически задавать линии, разделяющие объекты на нужные классы; 3) по заданным линиям строить гиперплоскости-аналоги во входном n-мерном пространстве, которые в дальнейшем используются для классификации.

Пусть $x = (x_1, x_2, ..., x_n)^{\mathrm{T}}$, где T — знак транспонирования, вектор признаков, характеризующий объект, который нужно классифицировать. Примем гипотезу об идентичности или похожести объектов по расстоянию между ними в пространстве характеристических признаков. Для учебной последовательности $\Omega_x = \{x : x(1), ..., x(m)\}$, где $x(j) = (x_{j1}, \ldots, x_{jn})^{\mathrm{T}}$, $j = \overline{1, m}$, в пространстве признаков, элементы которой соответствуют различным состояниям объектов в предметной области, которая исследуется, сформулируем задачу классификации следующим образом: необходимо найти такую гиперплоскость $w^{\mathrm{T}} x + b = 0$, $x \in R^n$, чтобы при $x(j) \in \Omega_x(1)$ имело место неравенство $w^{\mathrm{T}} x + b > 0$, а при $x(j) \in \Omega_x(2)$ — соответственно $w^{\mathrm{T}} x + b < 0$. Здесь $\Omega_x(1) \subset \Omega_x$ — подмножество векторов признаков, которые соответствуют объектам первого класса, $\Omega_x(2) \subset \Omega_x$ — подмноже-

© А.В. БАРМАК, Ю.В. КРАК, Э.А. МАНЗЮК, В.С. КАСЬЯНЮК, 2019

ство признаков для второго класса, $\Omega_x = \Omega_x(1) \cup \Omega_x(2)$, $w = (w_1, ..., w_n)^{\mathrm{T}}$ — вектор коэффициентов, b — некоторое число. Исходя из дихотомии, без потери общности будем рассматривать только два класса.

Информационная технология синтеза разделяющей гиперплоскости

При практической реализации проблем, которые формализуются приведенным выше образом, очень важен процесс валидации, т.е. определение соответствия предложенной модели реальному физическому объекту в пределах предметной области. Поэтому на первом этапе информационной технологии предлагается визуализировать учебную последовательность Ω_x, чтобы убедиться в ее способности представлять данные в виде двух раздельных классов.

Под визуализацией необходимо понимать такой способ представления векторов обучающей последовательности n-мерного пространства признаков в двумерном пространстве, при котором качественно отображались бы основные закономерности, присущие начальному распределению — его кластерная структура, топологические особенности, информация о расположении данных в исходном пространстве и т.д. В качестве критерия сходства/различия примем расстояние между векторами в пространстве: будем считать, что два вектора идентичны, если расстояние между ними равно нулю, и векторы подобны, если расстояние между ними меньше некоторого наперед заданного значения $\varepsilon \geq 0$, и отличны, если расстояние больше ε. Предложенный способ отображения n-мерного пространства признаков в пространство меньшей размерности коррелирует с методом многомерного шкалирования (Multidimensional scaling — MDS) [4, 5].

Визуальное представление структуры многомерного набора данных может быть отображено в одно-, дву- и трехмерных пространствах отображений. Отметим, что наиболее информативным для восприятия человеком геометрических структур для извлечения информации и проведения анализа является двумерное отображение.

Далее визуализация данных используется как способ отображения многомерного распределения данных на двумерной плоскости, при котором качественно отображаются основные закономерности, присущие исходному распределению. При этом необходимо минимизировать потерю информативности и ее проявлений в кластерной структуре, топологических особенностях, зависимостях между признаками расположения данных в исходном пространстве. Визуальное отображение позволяет при малом количестве данных определить наличие информационных связей, которые слабо проявляются при совокупности использования методов. В этом случае информационные связи сложно определимы при подходах, использующих разную природу образования моделей [6, 7].

Исходная информация подается не в виде таблицы типа «объект-признак», а в виде квадратной симметричной матрицы D взаимных расстояний объектов. На пересечении i-й строки j-го столбца в матрице находится значение расстояния от i-го к j-му объекту.

Таким образом, сначала каждому объекту присваиваются координаты в многомерном пространстве. Задача многомерного шкалирования — сконструировать набор данных в обычном трехмерном пространстве или на плоскости таким образом, чтобы расстояния между объектами максимально отвечали изначально заданным в матрице расстояниям [4]. Вводимые координатные оси могут быть интерпретированы как некоторые неявные факторы, значения которых определяют отличия объектов между собой. Если предоставить каждому объекту пару координат, то в результате получим образ визуализации данных.

Шкалирование многомерного пространства

Исходными данными для шкалирования является матрица попарных расстояний между объектами. Расстояние между i- и j-м объектами обозначается $\delta_{ij} = d(X_i, X_j)$. Объекты определяются многомерными точками $X_i = \{x_{i1}, x_{i2},...$ $..., x_{in}\}$, $i = 1...n$. Расстояние вычисляется таким образом:

$$d(X_i, X_j) = \left(\sum_{k=1}^{n} (x_{ik} - x_{jk})^2 \right)^{1/2}. \quad (1)$$

Расстояние между точками в пространстве меньшей размерности (редуцированном пространстве) по аналогичной (1) формуле и будем обозначать $d(Y_i, Y_j)$.

Цель шкалирования — поиск точек пространства $Y_i = \{y_{i1}, y_{i2},..., y_{in}\}$, $i = \overline{1, n}$, таким образом, чтобы расстояние между точками в редуцированном пространстве было наиболее близко к расстоянию в многомерном исходном пространстве. Соответственно определяется мера качества отображения σ-стресс (stress), которая может быть обозначена функцией наименьших квадратов [8, 9]:

$$\sigma = \sum_{i<j} w_{ij} (d(Y_i, Y_j) - \delta_{ij})^2, \quad (2)$$

где w_{ij} — неотрицательные веса.

При нормализации стресс определяется следующим образом:

$$\sigma = \frac{\sum_{i<j} w_{ij} (d(Y_i, Y_j) - \delta_{ij})^2}{\sum_{i<j} w_{ij} \delta_{ij}^2}. \quad (3)$$

Нормализация позволяет улучшить интерпретацию качества визуализации и уменьшает зависимость от количества объектов и их расположения.

Для сформированной матрицы попарных расстояний D в многомерном пространстве осуществим следующие предварительные действия:

1) двойное центрирование матрицы одним из известных методов;

2) определим на базе размерности выхода n собственные векторы e_1, $e_2,..., e_n$ из полученной матрицы;

3) вычислим матрицу $X = E_n \Lambda_n^{0,5}$, E_n — матрица из собственных векторов $e_1, e_2,..., e_n$, Λ_n — диагональная матрица собственных значений.

Тогда координатную матрицу, которая используется в многомерном шкалировании, получим путем разложения по собственным значениям матрицы $B = XX^{\mathrm{T}}$.

Отметим, что функции ошибок при различных вариантах проекции данных довольно обширны и базируются на интерпретациях многомерного шкалирования и алгоритмов оптимизации. При неметрических методах многомерного шкалирования используются не количественные меры сходства объектов, а только их относительный порядок. Минимизация функции стресса σ соответствует нахождению наиболее оптимального согласования матрицы исходных расстояний с матрицей результирующих расстояний.

Минимизация функции стресса на основе алгоритма мажоризации

Для минимизации функции стресса предлагается подход, который заключается в нахождении матрицы близости и итеративного использования алгоритма SMACOF (Scaling by MAjorizing a COmplicated Function — шкалирование мажорированием сложной функции) до заданного значения стресса. Алгоритм SMACOF базируется на стратегии, использование которой обеспечивает хорошую конвергенцию модели, минимизируя влияния данных [10, 11]. Цель, в соответствии с принципом мажоризации, — нахождение более простой и управляемой функции $g(x, y)$, которая мажорирует целевую функцию $f(x)$. При этом для всех x имеем $g(x, y) \geq f(x)$, здесь y — фиксированные значения опорной точки. Опорная точка является точкой касания поверхности $g(y, y) = f(y)$, при этом минимизирующая точка x_* удовлетворяет неравенству $f(x_*) \leq g(x_*, y) \leq g(y, y) = f(y)$, образуя таким образом слоистую структуру.

В общем случае мажорирование представляет собой итеративную процедуру, состоящую из последовательности шагов:

- определение опорной точки $y = y_0$;
- вычисление x_* исходя из условия $g(x_*, y) \leq g(y, y)$;
- переход на предыдущий шаг с установкой $y = x_*$, если не достигнуто условие $f(y) - f(x_*) \leq \varepsilon$.

Этот подход успешно обобщается многомерными пространствами при условии соблюдения неравенства и используется для минимизации целевой функции.

На функцию мажорирования $g(x, y)$ накладывается ряд условий, которые и обуславливают преимущество ее использования. Она должна:

- минимизировать проще, чем $f(x)$;
- быть не меньше в исходном поле, чем изначальная функция $f(x) \leq g(x, y)$;
- быть касательной к $f(x)$ в опорой точке $f(y) = g(y, y)$.

Набор $Y = \{Y_1, Y_2, ..., Y_m\}$ m точек итеративно вычисляется с помощью преобразований Гутмана [11]

$$Y_{k+1} = V^{+} B(Y_k) Y_k, \tag{4}$$

где k — номер итерации; V^{+} — псевдообратная матрица для матрицы весов V с элементами

$$v_{ij} = -w_{ij}, \quad i \neq j; \; v_{ii} = \sum_{i=1, j \neq i}^{m} w_{ij}. \tag{5}$$

Матрица $B(Y_k)$ содержит элементы:

$$b_{ij} = \begin{cases} -\dfrac{w_{ij}\delta_{ij}}{d(Y_i, Y_j)}, & i \neq j \,\&\, d(Y_i, Y_j) \neq 0, \\ 0, & i \neq j \,\&\, d(Y_i, Y_j) = 0; \end{cases} \tag{6}$$

$$b_{ii} = -\sum_{i=1, j \neq i}^{m} b_{ij}. \tag{7}$$

При условии, что в формуле (4) веса $w_{ij} = 1$, получим

$$Y_{k+1} = \frac{1}{m} B(Y_k) Y_k. \tag{8}$$

ISSN 0572-2691

Отсюда для построения процедуры мажорирования необходимо осуществить следующие действия:

- задать начальные значения редуцированного пространства Y_0;
- записать функцию стресса $\sigma = \sum_{i<j} w_{ij}(d(Y_i, Y_j) - \delta_{ij})^2$;
- найти значения $Y_{k+1} = V^{+}B(Y_k)Y_k$;
- вычислить функцию стресса для $\sigma(Y_{k+1})$;
- задать итерационный инкремент $k++$;
- проверить условие конвергенции $\sigma(Y_{k-1}) - \sigma(Y_k) < \varepsilon$, иначе переход на функцию стресса.

Таким образом, на этом этапе процесс информационной технологии будет состоять из таких шагов:

- формирование матрицы попарных расстояний на базе входных данных;
- нахождение квадрата расстояний матрицы расстояний;
- использование двойного центрирования матрицы;
- определение собственных значений и собственных векторов матрицы;
- оптимизация карты алгоритмом SMACOF.

В результате получим набор объектов с парой координат, которые можно отобразить. Объекты являются отображением размеченных данных с многомерного пространства. Поскольку объекты размечены, необходимо обозначить классы на результирующей плоскости для формирования дерева решений на базе линейного классификатора.

Кусочно-линейные области ограничения

Если дискриминантная функция линейна, то классификатор $d(\bar{x})$ определяется соотношением

$$d(\bar{x}) = \overline{W}^{\mathrm{T}}\bar{x} + w_n, \tag{9}$$

где $\bar{x} = (x_0, x_1, ..., x_{n-1})^{\mathrm{T}}$ — вектор признаков, определяющий образ классифицируемого объекта; $W = (w_0, w_1, ..., w_{n-1})^{\mathrm{T}}$ — вектор весовых коэффициентов классификатора; w_n — пороговое значение.

Принадлежность к одному из двух классов — $\Omega(1), \Omega(2)$ — определяется правилом

$$d(\bar{x}) = \sum_{i=0}^{n-1} w_i x_i + w_N \begin{matrix} < \\ \\ > \end{matrix} 0 \rightarrow \begin{cases} \Omega(1), \\ \Omega(2). \end{cases} \tag{10}$$

Отсюда для формирования линейного классификатора необходимо найти вектор коэффициентов $\overline{W}$ и пороговое значение w_n.

Отметим, что для практических применений получить разделение на два класса, используя только линейный классификатор (10), достаточно сложно, при этом не представляется возможным разделить данные кривой или ломаной линией. Один из способов решения данной проблемы — построение классификатора с помощью комбинации линейных, образуя, таким образом, кусочно-линейный набор с необходимой степенью дискретизации. Этот подход обладает преимуществом визуализации в редуцированном пространстве и позволяет демонстрировать управляемость классификации данных. При использовании линейного классификатора в многомерном пространстве ищется гиперплоскость, которая и будет разделяющим критерием соответствия классу. Далее ищется вектор y_i, для нового элемента, представленного точкой x_i, и для некоторого граничного значения b из условия

$$y_i = \begin{cases} +1, & wx_i > b; \\ 0, & wx_i = b; \\ -1, & wx_i < b. \end{cases} \tag{11}$$

Уравнение (11) при равенстве нулю описывает гиперплоскость. При этом известно, что вектор w, перпендикулярный искомой разделительной линии с соответствующими свойствами: лучшая разделительная линия, максимально далекая от ближайших к нему точек классов разделения [12, 13]. Отметим, что расстояние между этими точкам задает полосу разделения, которая соответствует условию $-1 < wx_i - b < 1$, и в границах полосы точки элементов отсутствуют, при этом ширина разделительной полосы равна $2/|w|$. Отметим, что при разделении классов с помощью разделяющей полосы важное значение имеют только граничные точки, поскольку полоса состоит из параллельных линий, проходящих по границам классов. Эти линии уже не представляют собой разделение классов (эту функцию берет на себя полоса разделения), а обозначают границы классов, таким образом, эти линии ограничивают их. Отсюда задача преобразуется не в нахождение линий разграничения, а в нахождение границ классов, которые представляют собой линии ограничения. Итак, гиперпространство разделяется на некоторые ограниченные гиперобъемы, внутри которых представлены классы.

Отметим, что при существовании нескольких классов и при построении линейного классификатора происходит пересечение линий и построение отрезков, образуя кусочно-линейную структуру, которая в общем случае нелинейная. Используя линии ограничения классов, получим некоторую геометрическую структуру, ограничивающую класс. Элемент, который попал в это ограничение, принадлежит данному классу.

Для улучшения задачи линейной классификации используется увеличение (расширение) размерности пространства. Пространство расширяется с помощью функции отображения $\phi(x)$ на новое пространство. Для расширения двумерного пространства в трехмерное функция отображения представляется так:

$$\phi(x) = \phi(x_1, x_2) = (x_1^2, x_2^2, \sqrt{2}x_1x_2). \tag{12}$$

Повышение размерности пространства позволяет, благодаря изгибанию пространства, найти гиперплоскость линейной классификации. Таким образом, гиперплоскость позволяет линейно выделить разделяющиеся классы, что возможно при условии выпуклости целевой функции. Далее, при обратном понижении пространства линию разделения классов можно приемлемым образом описать кусочно-линейным способом линиями ограничения. Это позволяет использовать многократное повышение мерности пространства при обратной проекции для определения гиперплоскостей. Как следствие этого, используя подходы редуцирования пространства методами шкалирования, можно определить границы классов методами визуализации с последующей их проекцией в многомерное пространство. В этом случае функция отображения в n-мерное пространство будет иметь вид

$$\phi(x) = \phi(x_1, x_2, ..., x_n). \tag{13}$$

Приведем способ построения кусочно-линейного классификатора с помощью алгоритма дерева решений на основе системы визуализации данных. Дерево решений — это способ отображения правил в иерархической, последовательной структуре, где каждому объекту соответствует единственный узел, дающий решение [14]. Под правилом понимается логическая конструкция, представленная в виде «если ... то ...». Правила задаются кривыми, которые разделяют группы объектов на плоскости системы визуализации. Кривые задаются в виде кусочно-линейной структуры с разной необходимой степенью дискретизации, образуя таким образом кривые в первом приближении. В результате при добавлении нового объекта можно четко указать, к какому классу он принадлежит.

Определим несколько ситуаций из множества возможных T при построении дерева решений.

1. Множество T содержит элементы, которые относятся к одному классу. В этом случае дерево решений определяет класс. Если множество T не содержит элементов, дерево решений определяет ветку и класс, ассоциированный с этой веткой, извлекается из другого множества, отличного от T, например узла-предка.

2. Если множество T содержит элементы, которые относятся к различным классам, то множество разделяется на подмножества. Для этого определяется признак, который содержит больше двух отличающихся значений $O_1, O_2, ..., O_n$. Множество T разделяется на подмножества, при этом каждое подмножество T_i содержит элементы, которые имеют значение O_i для выбранного признака. Процесс рекурсивный, конечным условием которого является формирование подмножеств, которые состоят из элементов одного класса.

При построении дерева на каждом внутреннем узле необходимо найти условие разделения множества на этом узле на подмножества. В качестве условия принимается один из атрибутов. Общее правило состоит в следующем: атрибут разделяет множество таким образом, что результирующие подмножества состоят из объектов, которые принадлежат одному классу или максимизированы по этому признаку. Для нахождения атрибутов используется алгоритм C4.5 [15], где атрибут $Gain(\Theta)$ множества Θ выбирается по следующему критерию:

$$Gain(\Theta) = Info(T) - Info_x(T)\,Gain(\Theta) = nfo(T) - Info_x(T), \tag{14}$$

где $Info(T)$ — энтропия множества T;

$$Info_x(T) = \sum_{i=1}^{n} \frac{|T_i|}{|T|} Info(T_i). \tag{15}$$

Подмножества $T_1, T_2, ..., T_n$ получены из исходного множества T при проверке множества Θ. Выбирается атрибут, который дает максимальное значение по критерию (15). При этом для уменьшения количества подмножеств необходимо минимизировать количество узлов и веток.

Синтез разделяющих гиперплоскостей

Нелинейный (кусочно-линейный) классификатор изначально работает в многомерном пространстве. Для формирования разделяющих ограничений в этом пространстве необходимо преобразовать ограничения (линии) кусочно-линейного классификатора редуцированного пространства в ограничения гиперплоскостями многомерного пространства. Для этого расширяется размерность редуцированного пространства до изначального.

После расширения пространства и формирования гиперплоскостей определяются их уравнения. Для построения гиперплоскостей в n-мерном пространстве необходимо соответственно n точек, которые были получены добавлением дополнительных $n-2$ точки на линейных отрезках. Таким образом, получаем систему линейных уравнений:

$$\begin{cases} wX_1 + b_1 = 0, \\ \ldots \\ wX_n + b_n = 0, \end{cases} \tag{16}$$

которая в общем случае решается методом Гаусса. Здесь w — неизвестные коэффициенты гиперплоскостей.

Класс в многомерном пространстве определяется ограничительными гиперплоскостями. Для классификации новых данных определяется их местоположение в многомерном пространстве путем определения их положения относительно гиперплоскостей. Подставляя координаты данных в уравнение гиперплоскости, определяем их относительное местоположение из множества $\{-1, 0, 1\}$. Если результат меньше нуля, элемент находится условно «справа» относительно плоскости, если результат больше нуля, элемент находится «слева» плоскости и если равняется нулю, элемент находится на разделяющей плоскости соответственно.

Формирование правил нелинейной классификации производится такой последовательностью действий:

1) формирование кусочно-линейных визуальных ограничений класса в редуцированном пространстве;

2) расчет опорных точек-правил для класса;

3) трансформация точек-правил в многомерное пространство;

4) построение гиперплоскостей в многомерном пространстве на базе трансформированных точек;

5) формирование правил для класса в многомерном пространстве на базе ограничительных гиперплоскостей.

Кусочно-линейные ограничительные правила определяют области классов и позволяют визуально определить необходимость увеличения или ограничения площади класса, что важно в пограничных данных. Это обеспечивает хорошую интерпретируемость результатов классификации, управляемость ограничительной областью класса и, фактически, интерактивность системы классификации. Обеспечение наличия визуальной составляющей при классификации важно в сравнении с другими подходами, особенно в условиях пограничных сложных условий классификации. При этом обеспечивается наличие дополнительной информационной составляющей с помощью визуальных интерактивных средств определения ограничений класса. Это обеспечивает инструментарий получения системой дополнительной информации и контролируемость процесса классификации. Результаты работы системы хорошо понимаемы и управляемы вследствие визуального представления и интерактивности ограничительных правил. Область ограничения обеспечивается минимально необходимыми визуальными границами, которые, при необходимости, могут переопределяться. При этом линии ограничения трансформируются в многомерное пространство и представляются в нем гиперплоскостями, образуя ограничительные области. Классификация новых данных происходит в многомерном пространстве на базе расчетных данных и их соответствующего положения относительно ограничивающих гиперплоскостей. Определяется пространственное положение нового элемента относительно всех гиперплоскостей, определяя его нахождение в категориях классов ограниченных объемов. Этот процесс контролируемый, поскольку результаты шкалируются на редуцированное пространство и имеют визуальное представление нового классифицированного элемента данных. Как следствие, результат классификации в многомерном пространстве представляется и оценивается визуально, предоставляя возможность интерпретации и анализа соответствия новых элементов данных относительно категорий классов. В условиях мультиклассовости, если элемент имеет необходимое информационное содержание, согласно которому он принадлежит нескольким классам, проблема интерпретации и качественного соответствия определяется с помощью визуального аспекта.

Практическая реализация информационной технологии

Для построения разделительных линий необходимо понижение мерности пространства к двумерному для отображения на плоскости. Совокупности данных, которые визуально можно отнести к определенному классу, ограничиваются линейными отрезками. Для тестирования предложенной технологии подготовлена выборка из текстовых данных по определенным тематикам (использовано 12 тематик). Из них формируется набор разделительных областей для демонстрации возможностей предложенной информационной технологии. Для построения правил разделения областей визуально определяются границы группировки данных. Выбором любой точки (путем указания мышкой) в этой области формируются правила соответствия нового элемента (рис. 1).

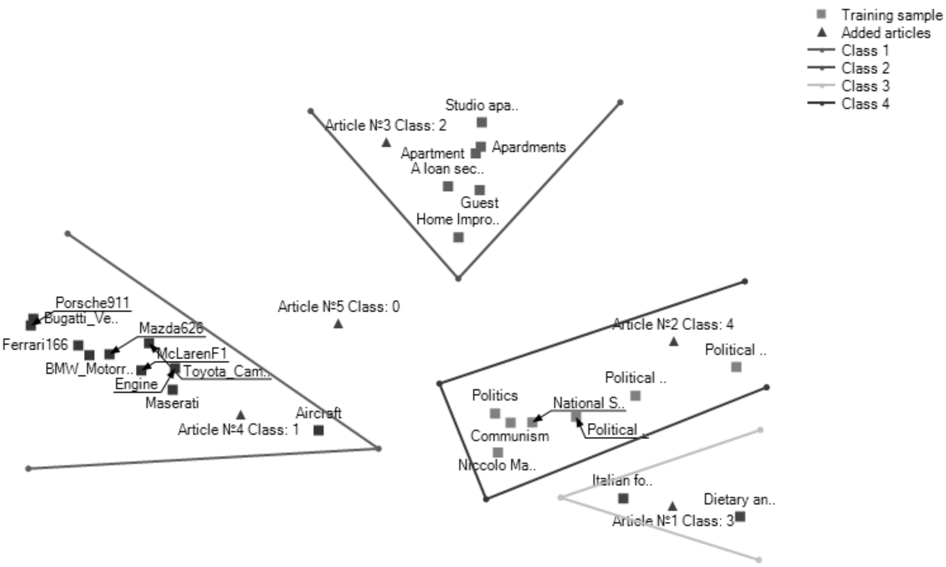

Рис. 1

Рассчитывается положение этой точки относительно отрезков, которые ограничивают область согласно условиям (11) (рис. 2). Таким образом, формируется правило отношения любой точки (элемента данных), которая попала в эту область ограничения. Следуя этим правилам, формируются области классов и правила отношения к ним данных.

point_0	point_2	point_2	point_3
-1	1		
-1	-1		
-1	1		
-1	-1	1	

Рис. 2

Поскольку областей может быть довольно много, новый элемент, попадая в какое-то место пространства, проверяется на соответствие правилам. Если набор правил подходит, указывается, что этот элемент подходит определенной ограниченной области, а значит, относится к классу, который определяет эта область. Как пример, элемент Article №3 Class 2 был определен таким образом, поскольку попал в зону ограниченной области Class 2. Элемент Article №5 Class 0 находится в зоне между областями классов, соответственно был отнесен к Class 0, т.е. не относится ни к одному из четырех наборов правил соответствия обозначенному классу.

Заключение

Предлагается информационная технология, позволяющая реализовать задачи классификации, кластеризации, исследования топологии информационной составляющей данных путем редуцирования многомерного пространства в пространство визуального представления для определения информационного содержания данных с последующим отображением принятых правил, исходя из представленной информации, в многомерное пространство. Это позволяет создать механизм управления данными в многомерном пространстве, передачу правил управления и контроля в это многомерное пространство, а также визуализации работы на базе проекций для повышения интерпретируемости данных в рамках решаемых задач и производимых над ними процессов в целях контроля и повышения информативности.

О.В. Бармак, Ю.В. Крак, Е.А. Манзюк, В.С. Касьянюк

ІНФОРМАЦІЙНА ТЕХНОЛОГІЯ СИНТЕЗУ РОЗДІЛЬНИХ ГІПЕРПЛОЩИН ДЛЯ ЛІНІЙНИХ КЛАСИФІКАТОРІВ

Запропоновано інформаційну технологію, що дозволяє реалізувати завдання класифікації, кластеризації, дослідження топології інформаційної складової даних. Багатовимірний простір ознак редукується у простір візуального представлення для визначення інформаційного змісту даних. Використовується оптимізоване зменшення розмірності простору до двовимірного за допомогою методів багатовимірного шкалування. Візуальне визначення групування даних дозволяє сформувати роздільні області. Наступним етапом є візуальне обмеження категорій класів, використовуючи графічні обмежувачі. Для забезпечення можливості гнучкості обмеження нелінійних областей використовується комбінація лінійних, утворюючи, таким чином, кусочно-лінійний набір з необхідним ступенем дискретизації. Використання кусочно-лінійних обмежувачів дозволяє здійснити проектування в початковий багатовимірний простір ознак. Візуальна побудова обмежувачів дає змогу враховувати поля допусків зміни параметрів ознак, міру поділу класів, нелінійність групування даних. Далі йде зворотне розширення простору з проекцією обмежувачів в n-мірний простір із синтезом розмежувальних гіперплоскостей. Таким чином, формуються обмежувальні області гіперпростору для необхідних категорій класів. При цьому забезпечується візуалізація процесів класифікації в гіперпросторі. Основою інформаційної технології є проектування багатовимірного простору в візуальне (двовимірне), побудова кусочно-лінійних обмежувачів досліджуваних областей, подальше проектування обмежувачів у багатовимірний простір. Таким чином, інформаційна технологія дозволяє синтезувати роздільні гіперплощини, які обмежують категорії класів, у багатовимірному просторі. Описано послідовні етапи застосування технології.

Ключові слова: багатомірне шкалювання, багатомірна класифікація, візуалізація даних.

A.V. Barmak, Yu.V. Krak, E.A. Manziuk, V.S. Kasianiuk

INFORMATION TECHNOLOGY OF DIVIDER HYPERPLANE SYNTHESIS FOR LINEAR CLASSIFIERS

Information technology allowing to implement tasks of classification, clustering, researching of topology of the information component of the data is proposed. The multidimensional feature space is reduced to the visual presentation space to determine the information content of the data. Optimized reduction of the space dimension to two-dimensional one applying multidimensional scaling methods is used. Visual definition of grouping data allows separating areas to form. The next stage is visual limitation of categories of classes using graphic separators. To enable flexibility of nonlinear areas limitation a combination of linear ones is used, forming, thus, a piecewise linear set with necessary degree of sampling. Using piecewise linear constraints allows to implement projecting into original multidimensional feature space. Visual construction of restrictive separators makes possible to consider tolerance fields of changing of features parameters, measure separation of classes, nonlinearity of data grouping. This is followed by the reverse expansion of space with the projection of the separators into *n*-dimensional space with the synthesis of separating hyperplanes. Thus, restrictive areas of hyperspace for the necessary categories of classes are formed. At the same time, visualization of classification processes in hyperspace is provided. Information technology base is multidimensional space projecting into visual (two-dimensional) space, constructing piecewise linear constraints of studied areas, subsequent constraints projecting into multidimensional space. Thus, information technology enables to synthesize separating hyperplanes limiting categories of classes in multidimensional space. The technology application successive stages are described.

Keywords: multidimensional scaling, multidimensional classification, data visualization.

1. Ивахненко А.Г. Самоорганизующиеся системы распознавания и автоматического управления. К. : Техніка. 1969. 395 с.
2. Vapnik V.N. Statistical learning theory. New York : Wiley, 1998. 736 p.
3. Kirichenko M.F., Krak Yu.V., Polishchuk A.A. Pseudo inverse and projective matrices in problems of synthesis of functional transformers. *Kibernetika i Sistemnyj Analiz.* 2004. **40**, N 3. P. 116–129.
4. Cox T.F., Cox M.A.A. : Multidimensional scaling, 2nd ed. Chapman and Hall. CRC. 2001. 328 p.
5. Krak I.V., Kudin G.I., Kulias A.I. Multidimensional. Scaling by means of pseudoinverse operations. *Cybernetics and Systems Analysis.* 2019. **55**, N 1. P. 22–29.
6. Barmak O., Krak Y., Manziuk E. Characteristics for choice of models in the ensembles. *Proceedings of the 11th International Conference of Programming UkrPROG* 2018 Kyiv, Ukraine, May 22-24, 2018. **2139**. P 171–179.
7. Manziuk E.A., Barmak O.V., Krak Iu.V., Kasianiuk V.S. Definition of information core for documents classification. *Journal of Automation and Information Sciences.* 2018. **50**, N 4. P. 25–34.
8. Mair P., Borg I., Rusch T. Goodness-of-fit assessment in multidimensional scaling and unfolding. *Multivariate behavioral research.* 2016. **51**, N 6. P. 772–789.
9. Stojkoska B.R. A taxonomy of localization techniques based on multidimensional scaling. *39th International Convention on Information and Communication Technology*, Electronics and Microelectronics (MIPRO). Opatija, 2016. P. 649–654.
10. de Leeuw J., Mair P. Multidimensional scaling using majorization: SMACOF in R. *Journal of Statistical Software.* 2009. **31**, N 3. P. 1–30.
11. Guttman L. A general nonmetric technique for finding the smallest coordinate space for a configuration of points. *Psychometrics*. 1968. **33**, N 4. P. 469–506.
12. Krak I.V., Kryvonos I.G., Barmak O.V., Ternov A.S. An approach to the determination of efficient features and synthesis of an optimal band-separating classifier of dactyl elements of sign language. *Cybernetics and Systems Analysis.* 2016. **52**, N 2. P. 173–180.
13. Kryvonos I.G., Krak I.V., Barmak O.V., Shkilniuk D.V. Construction and identification of elements of sign communication. *Cybernetics and Systems Analysis*. 2013. **49**, N 2. P. 163–172.
14. Kruskal J.B., Wish M. Multidimensional scaling, sage university paper series on quantitative application in the social sciences, 07-011. Beverly Hills and London: Sage Publications. Multidimensional Scaling. 1978. 93 p.
15. Quinlan J.R. C4.5: Programs for machine learning. San Mateo : Morgan Kaufmann Publishers Inc., 1993. 302 p.

Получено 15.03.2019

Статья представлена к публикации членом редколлегии докт. техн. наук Гаращенко Ф.Г.