УДК 519.68
DOI: 10.15827/0236-235X.131.439-448
Дата подачи статьи: 28.01.20
2020. Т. 33. № 3. С. 439–448

Разработка модификации метода опорных векторов для решения задачи классификации с ограничениями на предметную область

И.С. Михайлов [1], *к.т.н., доцент, fr82@mail.ru*
Зеар Аунг [1], *аспирант, zayaraung53@gmail.com*
Йе Тху Аунг [1], *аспирант, yethuaungg55@gmail.com*

[1] *Национальный исследовательский университет «Московский энергетический институт», кафедра прикладной математики и искусственного интеллекта, г. Москва, 111250, Россия*

Метод опорных векторов (support vector machine, SVM) является одним из методов интеллектуального анализа данных для решения задачи классификации. Его основная идея заключается в переводе исходных векторов в пространство более высокой размерности с применением метода ядра для обеспечения линейной разделимости классов и в поиске разделяющей гиперплоскости с максимальным зазором между гиперплоскостью и опорными векторами в этом пространстве.

Несмотря на высокую точность работы, метод имеет и недостатки – отсутствие общего подхода к автоматическому выбору ядра, а также высокую вычислительную сложность. В связи с этим разработана модификация алгоритма опорных векторов (алгоритм FS-SVM) для решения задачи классификации с ограничениями на предметную область. Сформулировано ограничение функциональной разделимости классов, накладываемое на предметную область, которое позволит применить алгоритм FS-SVM. Проведено теоретическое исследование обоснованности данных предположений.

Введены формальные определения функциональной разделимости на основе непрерывности и монотонности разделяющей функции и производной разделяющей функции. Проведено сравнение понятий функциональной разделимости классов и выпуклости классов.

В статье рассмотрены основные блоки предложенного алгоритма FS-SVM: поиск опорных элементов, определение точек разделяющей гиперповерхности, построение разделяющей гиперповерхности как кусочно-линейной функции в проекции на рассматриваемые оси координат. В дальнейшем указанную функцию алгоритма предлагается аппроксимировать многочленом Чебышева для получения гладкой кривой.

В качестве примера предметной области, в которой может применяться алгоритм FS-SVM, приводится задача классификации режимов течения нефте-водо-газового потока по исходных данным, регистрируемым на устье нефтяной скважины.

Ключевые слова*: машинное обучение, интеллектуальный анализ данных, машина опорных векторов, модификация метода опорных векторов, нефтяные скважины.*

Одним из вариантов практического применения алгоритмов машинного обучения является решение задачи классификации объектов [1]. Каждый объект данных представляется как вектор (точка) в p-мерном пространстве (упорядоченный набор p чисел). Каждая из этих точек принадлежит одному из двух классов. Необходимо определить, возможно ли разделение точек гиперплоскостью размерности $p - 1$, то есть, являются ли эти два класса линейно разделимыми. Искомых гиперплоскостей может быть много, поэтому полагают, что максимизация зазора между классами способствует более точной классификации. Необходимо найти такую гиперплоскость, чтобы расстояние от нее до ближайшей точки было максимальным. Это эквивалентно тому, что сумма расстояний до гиперплоскости от двух ближайших к ней точек, лежащих по разные стороны от нее, максимальна. Если такая гиперплоскость существует, она называется оптимальной разделяющей гиперплоскостью, а соответствующий ей линейный классификатор – оптимально разделяющим классификатором. Поиск данной гиперплоскости осуществляется с помощью алгоритма метода опорных векторов SVM (support vector machine) [2].

Метод SVM – это алгоритм обучения с учителем, используемый для решения задач классификации [3]. Его основная идея – перевод исходных векторов в пространство более высокой размерности (спрямляющее пространство)

для обеспечения линейной разделимости классов и поиск разделяющей гиперплоскости с максимальным зазором в этом пространстве. Алгоритм работает в предположении, что, чем больше разница или расстояние между этими параллельными гиперплоскостями, тем меньше средняя ошибка классификатора.

Постановка задачи классификации

Определим точки обучающей выборки следующим образом: $\{(x_1, c_i), (x_2, c_2), \ldots, (x_n, c_n)\}$, где c_i принимает значение 1 или −1 в зависимости от того, к какому классу принадлежит точка x_i [4]. Каждое x_i – это p-мерный вещественный вектор, обычно нормализованный значениями [0, 1] или [−1, 1]. Если точки не будут нормализованы, то точка с большими отклонениями от средних значений координат точек слишком сильно повлияет на классификатор. В обучающей выборке для каждого элемента задан класс, к которому он принадлежит. Необходимо, чтобы алгоритм метода опорных векторов классифицировал их таким же образом. Для этого строится разделяющая гиперплоскость, имеющая вид $wx - b = 0$.

Вектор w – перпендикуляр к разделяющей гиперплоскости. Параметр $\frac{b}{\|w\|}$ равен по модулю расстоянию от гиперплоскости до начала координат. Если параметр b равен нулю, гиперплоскость проходит через начало координат, что ограничивает решение.

Так как необходимо найти оптимальное разделение, следует определить опорные векторы и гиперплоскости, параллельные оптимальной и ближайшие к опорным векторам двух классов. Можно показать, что эти параллельные гиперплоскости могут быть описаны следующими уравнениями (с точностью до нормировки): $wx - b = 1$, $wx - b = -1$.

Если обучающая выборка линейно разделима, то можно выбрать гиперплоскости таким образом, чтобы между ними не лежала ни одна точка обучающей выборки, а затем максимизировать расстояние между гиперплоскостями. Ширина полосы между ними равна $\frac{2}{\|w\|}$, таким образом, поставленная задача сводится к минимизации $\|w\|$. Чтобы исключить все точки из полосы, необходимо убедиться, что для всех i

$$\begin{cases} wx_i - b \geq 1, & c_i = 1, \\ wx_i - b \leq 1, & c_i = -1. \end{cases}$$

Указанные соотношения могут быть также записаны в виде

$$c_i(wx - b \geq 1),\ 1 \leq i \leq n. \quad (1)$$

Метод опорных векторов (SVM)

Основная идея метода SVM заключается в классификации объектов с помощью перевода исходных векторов, описывающих объекты, в пространство более высокой размерности и в поиске разделяющей гиперплоскости с максимальным зазором в этом пространстве. Две параллельные гиперплоскости строятся по обеим сторонам гиперплоскости, разделяющей классы. Это будет гиперплоскость, максимизирующая расстояние до двух параллельных гиперплоскостей. Алгоритм работает в предположении, что, чем больше разница или расстояние между этими параллельными гиперплоскостями, тем меньше средняя ошибка классификатора [5].

Случай линейной разделимости. Проблема построения оптимальной разделяющей гиперплоскости сводится к минимизации $\|w\|$ при условии (1). Это задача квадратичной оптимизации, имеющая вид

$$\begin{cases} \|w\|^2 \to \min, \\ c_i\left(wx - b \geq 1\right),\ 1 \leq i \leq n. \end{cases} \quad (2)$$

По теореме Куна–Таккера эта задача эквивалентна двойственной задаче поиска седловой точки функции Лагранжа:

$$\begin{cases} L\left(w, b; \lambda\right) = \frac{1}{2}\|w\|^2 - \sum_{i=1}^{n} \lambda_i \left(c_i\left(w \cdot x_i\right) - b - 1\right) \to \\ \to \min_{w,b} \max_{\lambda}, \\ \lambda_i \geq 0, \quad 1 \leq i \leq n, \end{cases}$$

где $\lambda = (\lambda_1, \ldots, \lambda_n)$ – вектор двойственных переменных.

Задача сводится к эквивалентной задаче квадратичного программирования, содержащей только двойственные переменные:

$$\begin{cases} -L\left(\lambda\right) = \sum_{i=1}^{n} \lambda_i + \sum_{i=1}^{n} \sum_{j=1}^{n} \lambda_i \lambda_j c_i c_j \left(x_i \cdot x_j\right) \to \min_{\lambda}, \\ \lambda_i \geq 0, \quad 1 \leq i \leq n, \\ \sum_{i=1}^{n} \lambda_i c_i = 0. \end{cases} \quad (3)$$

После решения данной задачи w и b можно найти по формулам: $w = \sum_{i=1}^{n} \lambda_i c_i x_i$, $b = wx_i - c_i$, $\lambda_i > 0$.

В итоге алгоритм классификации может быть записан в виде

$$a(x) = \sin\left(\sum_{i=1}^{n} \lambda_i c_i x_i \cdot x - b\right). \quad (4)$$

В данном случае суммирование идет только по опорным векторам, для которых $\lambda_i \neq 0$, а не по всей выборке.

Случай линейной неразделимости. Для работы алгоритма в случае, если классы линейно неразделимы, следует допустить ошибку классификации на обучающей выборке. Вводится набор дополнительных переменных $\xi_i \geq 0$, характеризующих величину ошибки на объектах x_i, $1 \leq i \leq n$. В качестве основы используется постановка задачи (2). В данные неравенства вводятся погрешность классификации, а также штраф за суммарную ошибку в минимизируемый функционал:

$$\begin{cases} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n} \varepsilon_i \to \min_{w,x,\varepsilon}, \\ c_i\left(w \cdot x_i - b\right) \geq 1 - \varepsilon_i,\ 1 \leq i \leq n,\ \varepsilon_i \geq 0,\ 1 \leq i \leq n. \end{cases}$$

Коэффициент C – параметр настройки метода, который позволяет регулировать отношение между максимизацией ширины разделяющей полосы и минимизацией суммарной ошибки.

Аналогично по теореме Куна–Таккера задача сводится к поиску седловой точки функции Лагранжа:

$$\begin{cases} L(w,b,\varepsilon,\lambda,\eta) = \frac{1}{2}\|w\|^2 - \\ -\sum_{i=1}^{n} \lambda_i\left(c_i\left(\left(w \cdot x_i\right) - b\right) - 1\right) - \sum_{i=1}^{n} \varepsilon_1\left(\lambda_i + \eta_i - C\right) \to \\ \to \min_{w,x,\varepsilon} \max_{\lambda,\eta}, \\ \begin{cases} \varepsilon_1 \geq 0,\ \lambda_i \geq 0,\ \eta_i \geq 0, & 1 \leq i \leq n, \\ \begin{cases} \lambda_i \geq 0, \\ c_i\left(w \cdot x_i - b\right) = 1 - \varepsilon_i, \end{cases} & 1 \leq i \leq n, \\ \begin{cases} \eta_i = 0, \\ \varepsilon_1 \geq 0, \end{cases} & 1 \leq i \leq n. \end{cases} \end{cases}$$

Данная задача аналогично сводится к эквивалентной:

$$\begin{cases} -L(\lambda) = -\sum_{i=1}^{n} \lambda_i + \sum_{i=1}^{n}\sum_{j=1}^{n} \lambda_i \lambda_j c_i c_j \left(x_i \cdot x_j\right) \to \min_{\lambda}, \\ 0 \leq \lambda_i \leq C, \quad 1 \leq i \leq n, \\ \sum_{i=1}^{n} \lambda_i c_i = 0. \end{cases}$$

На практике при работе SVM решают именно эту задачу, а не (3), так как гарантировать линейную разделимость точек на два класса в общем случае не представляется возможным. Данный алгоритм называется алгоритмом с мягким зазором (soft-margin SVM). В случае линейной разделимости точек алгоритм называется алгоритмом с жестким зазором (hard-margin SVM).

Для алгоритма классификации сохраняется формула (4), однако теперь ненулевыми λ_i обладают не только опорные, но и неверно классифицированные объекты. В определенном смысле это недостаток, поскольку нарушителями часто оказываются шумовые выбросы и построенное на них решающее правило по сути опирается на шум. Константу C обычно выбирают по критерию скользящего контроля. При каждом значении C задачу приходится решать заново, что повышает трудоемкость работы алгоритма. Если есть основания полагать, что выборка почти линейно разделима и лишь объекты-выбросы классифицируются неверно, то можно применить фильтрацию выбросов. В этом случае вначале задача решается при некотором C и из выборки удаляется небольшая доля объектов, имеющих наибольшую величину ошибки . После этого задача решается заново по усеченной выборке. Вероятно, что для решения задачи будет необходимо проделать несколько таких итераций, пока оставшиеся объекты выборки не окажутся линейно разделимыми.

Ядра. Для ускорения работы метода SVM существует способ создания нелинейного классификатора, в основе которого лежит переход от скалярных произведений к произвольным ядрам, так называемый kernel trick, позволяющий строить нелинейные разделители [6]. Отличием данного алгоритма является то, что каждое скалярное произведение в приведенных выше формулах заменяется нелинейной функцией ядра (скалярным произведением в пространстве с большей размерностью). В этом пространстве уже может существовать оптимальная разделяющая гиперплоскость. Так как размерность получаемого пространства может быть больше размерности исходного, то преобразование, сопоставляющее скалярные произведения, будет нелинейным, а значит, функция, соответствующая в исходном пространстве оптимальной разделяющей гиперплоскости, также будет нелинейной. Необходимо отметить: если исходное пространство имеет достаточно высокую размерность, то велика вероятность, что выборка в нем окажется линейно разделимой [7].

Особенности SVM.

Преимущества:

– SVM является наиболее быстрым методом нахождения решающих функций;

– метод сводится к решению задачи квадратичного программирования в выпуклой области, всегда имеющей единственное решение;

– метод находит разделяющую полосу максимальной ширины, что в дальнейшем позволяет осуществлять более уверенную классификацию.

Недостатки:

– метод чувствителен к шумам и стандартизации данных;

– не существует общего подхода к автоматическому выбору ядра (и построению спрямляющего подпространства в целом) в случае линейной неразделимости классов;

– метод имеет высокую вычислительную сложность [8].

Метод FS-SVM – модификация метода SVM

Актуальность. Необходимо отметить, что работа алгоритма опорных векторов в общем случае связана с громоздкими вычислениями, что может привести к неоправданно высоким вычислительным затратам (поиск седловой точки функции Лагранжа, применение метода ядра и т.д.). Также существенным недостатком метода SVM является отсутствие общего подхода к автоматическому выбору ядра и построению спрямляющего подпространства в целом в случае линейной неразделимости классов.

В связи с этим необходима модификация метода опорных векторов, позволяющая обеспечить высокую точность решения задачи классификации (как и стандартный метод SVM в нелинейном случае с использованием функции ядра) и имеющая меньшую вычислительную и алгоритмическую сложность.

Для этого предлагается наложить некоторые ограничения на рассматриваемую задачу, что позволит разработать и применить модифицированный метод опорных векторов.

Функциональная разделимость классов. В качестве ограничения следует принять рассмотрение не общей задачи классификации классов произвольной формы, а задачи классификации при условии функционально разделимых классов.

Определение 1. Функциональная разделимость классов.

Классы являются функционально разделимыми, если существует функциональная зависимость между каждым выходным параметром и набором входных параметров. При этом данная функция является непрерывной и монотонной на множестве значений входных параметров.

Формальное определение:

$$\begin{pmatrix} \forall C_i \in M \ \exists (x_{i1}, ..., x_{ik}) \in X : \\ \exists f(x_{i1}, ..., x_{ik}) = y_j, \ y_j \in C_i \end{pmatrix},$$

&

$$\begin{pmatrix} \exists q \in \{-1, 1\} \ \forall (x_{i1}, ..., x_{ip}^{*}, ..., x_{ik}), (x_{i1}, ..., x_{ip}^{**}, ..., x_{ik}), \\ x_{ip}^{*} < x_{ip}^{**} \\ q \cdot \left(f(x_{i1}, ..., x_{ip}^{**}, ..., x_{ik}) - f(x_{i1}, ..., x_{ip}^{*}, ..., x_{ik}) \right) \geq 0 \end{pmatrix},$$

где M – множество классов; C_i – i-й класс из m классов; X – множество n входных параметров; $(x_{i1}, \ldots, x_{ik})$ – подмножество k входных параметров, $1 \leq k \leq n$; y_i – значение функции; $R = \{\geq, \leq\}$, $1 \leq p \leq k$.

Данное определение проиллюстрировано на рисунке 1.

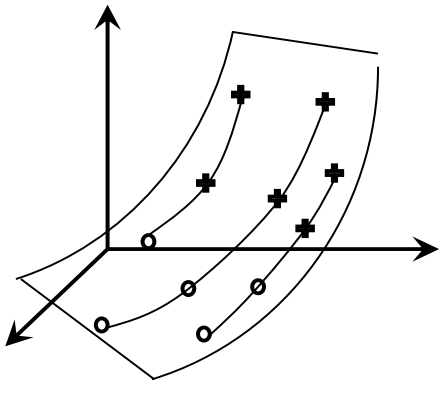

Рис. 1. Пример функциональной монотонной зависимости между параметрами

Fig. 1. Example of a functional monotonic dependence between parameters

В качестве определения функциональной разделимости классов также можно рассматривать следующее.

Определение 2. Функциональная разделимость классов.

Классы являются функционально разделимыми, если существует функциональная зависимость между каждым выходным параметром и набором входных параметров. При этом производная данной функции не меняет знак или не определена в некоторых точках на множестве значений входных параметров.

Формальное определение:

$$\left(\forall C_i \in M \exists \left(x_{i1},...,x_{ik}\right) \in X : \exists f\left(x_{i1},...,x_{ik}\right) = y_j, y_j \in C_i\right),$$

&

$$\exists q \in \{-1,1\} \forall \left(x_{i1},...,x_{ik}\right):$$

$$\begin{cases} qf'\left(x_{i1},...,x_{ik}\right) \geq 0, \\ f'\left(x_{i1},...,x_{ik}\right) \text{ не определена.} \end{cases}$$

Производная функции f'может оказаться неопределенной в случае, если функциональная зависимость f между входными и выходными параметрами будет представлена кусочной функцией. Тогда в точках излома данной функции производная будет не определена. Таких точек будет конечное множество, поскольку обучающая выборка, по которой может быть построена функция, является конечной последовательностью объектов, следовательно, и количество изломов может оказаться конечным.

Функциональная разделимость классов и выпуклые классы. В качестве ограничения, накладываемого на задачу классификации, также может рассматриваться условие выпуклости классов, однако оно является более жестким, чем условие функциональной разделимости (рис. 2).

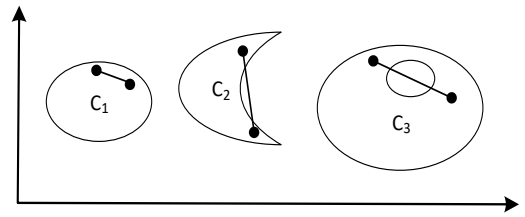

Рис. 2. Выпуклое множество C_1 и невыпуклые множества C_2 и C_3

Fig. 2. Convex set C_1 and non-convex sets C_2 and C_3

Определение.

Множество $M \subset R^d$ выпукло, если для любых точек A, $B \in M$ отрезок AB лежит в M.

На рисунке 3 представлено отличие данных условий.

При условии выполнимости ограничения функциональной разделимости классов для рассматриваемой задачи исключаются случаи, представленные на рисунке 3в, при которых в одной из проекций выходных переменных на множество входных переменных возможна ситуация включения одного класса в другой.

Описание блоков алгоритма FS-SVM. Рассмотрим подробно работу модифицированного

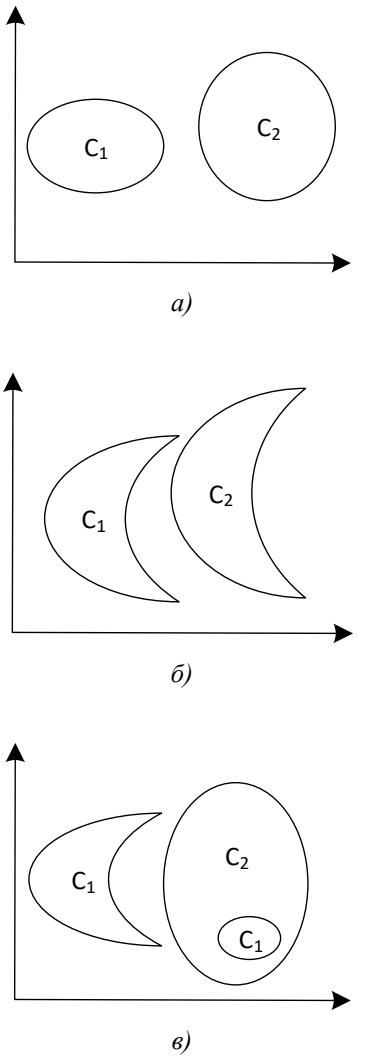

Рис. 3. Функциональная разделимость классов C_1 и C_2: а) функционально разделимые и выпуклые классы, б) функционально разделимые и невыпуклые классы, в) функционально неразделимые и невыпуклые классы

Fig. 3. Functional separability of classes C_1 and C_2: a) functionally separable and convex classes, б) functionally separable and non-convex classes, в) functionally inseparable and non-convex classes

метода опорных векторов – Functionally Separable – Support Vector Machine (FS-SVM).

Алгоритм FS-SVM состоит из следующих блоков.

1. Поиск опорных элементов, то есть элементов классов, которые находятся на границе класса (рис. 4а).

1.1. Элементы упорядочиваются по одной из координат при фиксированных значениях другой координаты.

1.2. Пара элементов, для которых значение класса переходит с C_i на C_j, отмечаются как опорные элементы.

1.3. Процедура повторяется для оставшихся координат элементов.

2. Поиск точек для разделяющей гиперповерхности (рис. 4б).

2.1. Вычисляются векторы, соединяющие опорные элементы каждого класса.

2.2. Для каждого элемента класса C_i определяется элемент класса C_j, для которого расстояние до рассматриваемого элемента C_i будет минимальным.

2.3. Вычисляется середина вектора, соединяющего данные опорные элементы классов C_i и C_j. Полученная точка определяется как точка, через которую должна будет проходить разделяющая гиперповерхность.

2.4. Указанная процедура повторяется для всех опорных элементов классов C_i и C_j.

3. Построение разделяющей гиперповерхности как кусочно-линейной функции в проекции на рассматриваемые оси координат (рис. 4в). В дальнейшем развитии алгоритма указанную функцию можно аппроксимировать многочленом Чебышева для получения гладкой кривой (рис. 4г).

При выполнении алгоритма осуществляются проверка контролируемых параметров и определение стороны относительно разделяющей гиперповерхности, в которой они находятся, для получения результата классификации.

Условие применимости метода FS-SVM. Необходимо отметить, что принципиальным ограничением данного алгоритма является необходимость выполнения условия, что все элементы класса должны находиться с одной из сторон от разделяющей функции (недопустимы случаи, представленные на рисунке 3в).

Докажем следующую теорему.

Теорема (FS).

В случае выполнения условия функциональной разделимости классов возможно применение модифицированного метода опорных векторов FS-SVM.

Доказательство.

Согласно определению функциональной разделимости классов, существует функциональная зависимость f между наборами исходных данных и результирующими данными задачи. Данная зависимость непрерывна и монотонна.

Монотонность функции f означает, что производная данной функции не меняет знак на области определения набора входных параметров.

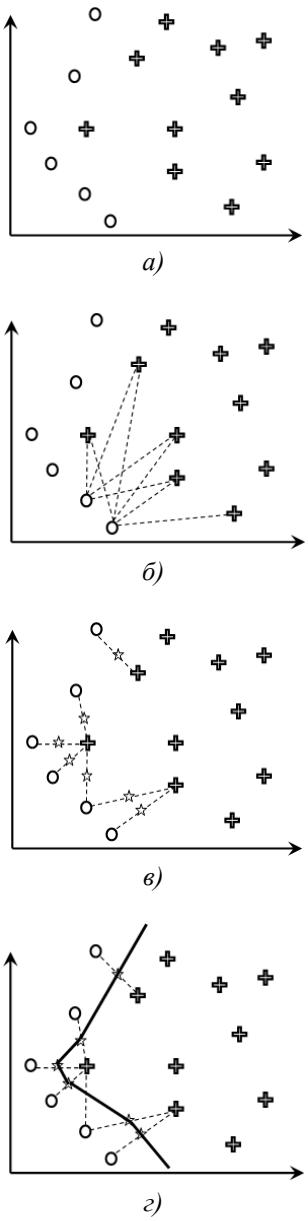

Рис. 4. Алгоритм FS-SVM построения функции, разделяющей классы C_1 и C_2

Fig. 4. FS-SVM algorithm for constructing a function that separates classes C_1 and C_2

Следовательно, пересечение данной гиперповерхности f пространством наборов исходных данных всегда будет представлять собой кривую незамкнутую линию.

Рассмотрим любую проекцию функциональной зависимости f на набор данных x. Тогда из условия монотонности следует, что она пересечет любой построенный перпендикуляр на ось Y не более одного раза: в случае строгой

монотонности в точке или в случае нестрогой монотонности на отрезке.

Рассмотрим эту лемму.

Лемма (FS).

Монотонная функция f пересекает поверхность не более одного раза: в случае строгой монотонности в точке или в случае нестрогой монотонности на отрезке (рис. 5).

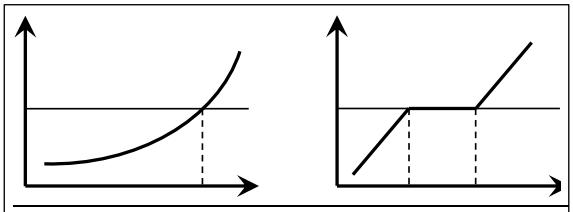

Рис. 5. Примеры пересечения поверхности строго монотонно возрастающей функцией (слева) и нестрого монотонно возрастающей функцией (справа)

Fig. 5. Examples of the intersection of a surface with a strictly monotonically increasing function (on the left) and a non-strict monotonically increasing function (on the right)

Доказательство.

Функция $f(x)$ непрерывна, то есть

$\forall y\ \exists(x_{i1}, \ldots, x_{ik})$: $y = f(x_{i1}, \ldots, x_{ik})$.

Функция $f(x)$ монотонна:

$$\exists q \in \{-1,1\}\ \forall \left(x_{i1},...,x_{ip}^{*},...,x_{ik}\right), \left(x_{i1},...,x_{ip}^{**},...,x_{ik}\right),$$

$$x_{ip}^{*} < x_{ip}^{**},$$

$$q\left(f\left(x_{i1},...,x_{ip}^{**},...,x_{ik}\right) - f\left(x_{i1},...,x_{ip}^{*},...,x_{ik}\right)\right) \geq 0.$$

Зафиксируем $q = 1$, то есть допустим, что функция f не убывает. Для $q = -1$ доказательство проводится аналогично:

$$\forall \left(x_{i1},...,x_{ip}^{*},...,x_{ik}\right), \left(x_{i1},...,x_{ip}^{**},...,x_{ik}\right),$$

$$x_{ip}^{*} < x_{ip}^{**}:$$

$$-f\left(x_{i1},...,x_{ip}^{*},...,x_{ik}\right) \leq f\left(x_{i1},...,x_{ip}^{**},...,x_{ik}\right).$$

Проведем доказательство от противного. Допустим, что

$$\exists x_{ip}^{***} < x_{ip}^{*}: f\left(x_{i1},...,x_{ip}^{***},...,x_{ik}\right) \geq$$

$$\geq f\left(x_{i1},...,x_{ip}^{*},...,x_{ik}\right).$$

Однако, если $x_{ip}^{***} < x_{ip}^{*}$, то $x_{ip}^{***} < x_{ip}^{**}$, тогда по определению

$$f\left(x_{i1},...,x_{ip}^{***},...,x_{ik}\right) \leq f\left(x_{i1},...,x_{ip}^{**},...,x_{ik}\right).$$

Поскольку $x_{ip}^{***} < x_{ip}^{*} < x_{ip}^{**}$, в предыдущей формуле заменим x_{ip}^{**} на x_{ip}^{*}. Получим:

$$f\left(x_{i1},...,x_{ip}^{***},...,x_{ik}\right) \leq f\left(x_{i1},...,x_{ip}^{*},...,x_{ik}\right).$$

Получено противоречие предположению доказательства. Следовательно, верно обратное утверждение. Аналогично проводится доказательство для $x_{ip}^{*} < x_{ip}^{***}$.

Теорема Больцано–Коши о промежуточном значении. Пусть функция f непрерывна на отрезке $[a, b]$ и C – произвольное число, находящееся между значениями функции на концах отрезка: $A = f(a)$ и $B = f(b)$. Тогда существует точка $z \in [a, b]$, для которой $f(z) = C$.

Пусть для $x_{ip}^{*} < x_{ip}^{**}$, $f\left(x_{i1}, ..., x_{ip}^{*}, ..., x_{ik}\right) = A$ и $f\left(x_{i1}, ..., x_{ip}^{**}, ..., x_{ik}\right) = B$.

Тогда по теореме Больцано–Коши о промежуточном значении получим, что

$$\exists z_{ip}, x_{ip}^{*} < z_{ip} < x_{ip}^{**}: f\left(x_{i1},...,z_{ip},...,x_{ik}\right) = C.$$

В данном случае C – неопределенность значения класса.

Лемма доказана.

Таким образом, можно расширить вывод леммы FS на каждый параметр x_{ik}, тогда

$$\begin{cases} \forall f\left(x_{i1},...,x_{ik}\right) \exists!\left(z_{i1},...,z_{ik}\right) \\ \quad \text{в случае строгой монотонности } f, \\ \forall f\left(x_{i1},...,x_{ik}\right) \exists\left(z_{i1},...,z_{ik}\right): \\ z_{ip} \in \left[a_{ip}, b_{ip}\right],\ 1 \leq p \leq k, \\ \quad \text{в случае нестрогой монотонности } f. \end{cases}$$

Поскольку функция f является разделительной между классами и незамкнутой, можно определить, с какой стороны от функции лежит конкретный пример из обучающей выборки, а значит, применить алгоритм FS-SVM.

Теорема доказана.

Практическое применение метода FS-SVM. Задача классификации режимов течения нефте-водо-газового потока.

Рассмотренное ограничение функциональной разделимости классов не является критическим для большого количества задач. В частности, это условие будет верно для задач, в которых исходные данные получены в результате наблюдения физических явлений с установленной монотонностью их происхождения.

Одной из таких задач является определение режимов течения нефте-водо-газового потока. Рассмотрим ее более подробно.

Нефтяное месторождение состоит из набора кустов скважин, каждый из которых содержит порядка десяти нефтедобывающих скважин, имеющих свои физические характеристики

(плотность нефти, средние значения дебита продукции, количество примесей и другие). На нефтяных скважинах установлены насосы различных типов и производительности.

При добыче нефти происходит одновременная добыча нефти, воды, содержащейся в пласте, и попутно нефтяного газа. При этом необходимо постоянно контролировать расход каждой фракции для динамического изменения режима работы насоса скважины.

Например, если своевременно не обнаружить, что большую долю объема добываемой продукции составляет газ, то может произойти «завоздушивание» насоса, что приведет к его холостой работе и остановке скважины.

При существенном увеличении доли воды в добываемой смеси необходимо снизить частоту оборотов насоса или выключить его на некоторое время. В этом случае следует дать нефтяному пласту время на выделение нефти из слоя почвы и продолжить добычу спустя некоторое время.

Данные параметры непрерывно контролируются на устье скважины с помощью многофазного расходомера, состоящего из набора первичных преобразователей и вычислительного блока.

К первичным данным относятся следующие параметры: доплеровская скорость потока (вектор значений), Гц; газосодержание потока (вектор значений), %; скорость звука в смеси (вектор значений), м/с; давление смеси, Мпа; температура, °С.

Расчетные параметры следующие: расход жидкости, $м^3$/сутки; расход газа, $м^3$/сутки; обводненность продукции, %.

Представленный набор данных характеризует режим течения потока, что является важнейшим параметром при работе нефтяной скважины.

Существуют четыре основных класса режимов течения двухфазного потока в вертикальных каналах [9]: пузырьковый, снарядный, эмульсионный и дисперсно-кольцевой [10] (см. http://www.swsys.ru/uploaded/image/2020-3/2020-3-dop/15.jpg).

В данном случае необходимо решить задачу классификации контролируемых параметров нефтяной скважины по режимам течения нефте-водо-газового потока.

Теоретические расчеты, а также результаты многолетних испытаний нефте-водо-газового расходомера на сертифицированных эталонных стендах (в том числе и зарубежных) показали, что рассматриваемые контролируемые параметры монотонно характеризуют режим течения потока (см. http://www.swsys.ru/uploaded/image/2020-3/2020-3-dop/16.jpg, http://www.swsys.ru/uploaded/image/2020-3/2020-3-dop/17.jpg).

Также установлена монотонная зависимость между первичными и расчетными данными, которая легла в основу построения вычислительной модели работы многофазного расходомера (см. http://www.swsys.ru/uploaded/image/2020-3/2020-3-dop/18.jpg, http://www.swsys.ru/uploaded/image/2020-3/2020-3-dop/19.jpg).

Таким образом, показано, что рассматриваемая задача удовлетворяет ограничению функциональной разделимости для задачи классификации и, соответственно, к ней может применяться модифицированный метод опорных векторов.

Заключение

В статье рассмотрен метод SVM для решения задачи классификации, выделены его достоинства и недостатки. Разработан алгоритм FS-SVM (модификация алгоритма опорных векторов) для решения задачи классификации с ограничениями на предметную область.

Кроме того, сформулировано определение функциональной разделимости классов. Предложено использовать функциональную разделимость классов как ограничение, накладываемое на предметную область, которое позволит применить алгоритм FS-SVM.

Авторы вводят теорему (FS) для обоснования применимости алгоритма FS-SVM. Рассмотрено доказательство данной теоремы.

В качестве примера предметной области, в которой может применяться алгоритм FS-SVM, приводится задача классификации режимов течения нефте-водо-газового потока по исходных данным, регистрируемым многофазным расходомером на устье нефтяной скважины.

Работа выполнена при финансовой поддержке РФФИ, проект № 18-01-00459.

Литература

1. Башмаков А.И., Башмаков И.А. Интеллектуальные информационные технологии. М.: Изд-во МГТУ им. Н.Э. Баумана, 2005. 304 с.
2. Вьюгин В.В. Математические основы теории машинного обучения и прогнозирования. М., 2013. 390 с.
3. Statnikov A., Aliferis C.F., Hardin D.P., Guyon I. A Gentle Introduction to Support Vector Machines in Biomedicine: Vol. 1: Theory and Methods. 2011, 200 p. DOI: 10.1142/7922.
4. Rojo-Álvarez J.L., Martinez-Ramon M., Muñoz-Marí J., Camps-Valls G. Support vector machine and kernel classification algorithms. In: Digital Signal Processing with Kernel Methods. 2018, pp. 433–502. DOI: 10.1002/9781118705810.ch10.
5. Pisner D.A., Schnyer D.M. Machine Learning. Methods and Applications to Brain Disorders. 2020, pp. 101–121. DOI: 10.1016/B978-0-12-815739-8.00006-7.
6. Nefedov A. Support Vector Machines: A Simple Tutorial. 2016, 35 p. URL: https://svmtutorial.online (дата обращения: 20.01.2020).
7. Alam S., Kang M., Pyun J.Y., Kwon G.R. Performance of classification based on PCA, linear SVM, and Multi-kernel SVM. Proc. 8th ICUFN, IEEE 2016, pp. 987–989. DOI: 10.1109/ICUFN.2016.7536945.
8. Xiangdong H., Shaoqing W. Prediction of bottom-hole flow pressure in coalbed gas wells based on GA optimization SVM. Proc. 3rd IAEAC, IEEE, 2018, pp. 138–141. DOI: 10.1109/IAEAC.2018.8577488.
9. Баттерворс Д., Хьюитт Г. Теплопередача в двухфазном потоке; [пер. с англ.]. М.: Энергия, 1980. 328 с.
10. Киселев П.Т. Гидравлика: основы механики жидкости. М.: Энергия, 1980. 360 с.

Software & Systems
DOI: 10.15827/0236-235X.131.439-448

Received 28.01.20
2020, vol. 33, no. 3, pp. 439–448

Support vector method modification development for solving the classification problem with domain restrictions

I.S. Mikhaylov [1], *Ph.D. (Engineering), Associate Professor, fr82@mail.ru*
Zayar Aung [1], *Postgraduate Student, zayaraung53@gmail.com*
Ye Thu Aung [1], *Postgraduate Student, yethuaungg55@gmail.com*

[1] *National Research University "Moscow Power Engineering Institute", Applied Mathematics and Artificial Intelligence Department, Moscow, 111250, Russian Federation*

Abstract. One of the data mining methods for solving the classification problem is the support vector machine (SVM). The method's main idea is to translate the source vectors into a higher-dimensional space using the kernel method to ensure the linear separability of classes and to find a separating hyperplane with the maximum margin between the hyperplane and the reference vectors in this space.

Despite the high accuracy of the method, it also has disadvantages. These include the lack of a general approach to automatic kernel selection, as well as the high computational complexity of the method. In this regard, the authors developed a modification of the support vector algorithm (FS-SVM algorithm) to solve the classification problem with restrictions on the problem domain. The authors formulated the classes "functional separability" restriction, imposed on the problem domain. It will allow applying the FS-SVM algorithm. There is a theoretical study of these assumptions validity in the paper.

The paper introduces formal definitions of "functional separability" based on the separating function continuity and monotony and on the basis of the derivative of discriminant function. The authors show a "functional separability" concepts and classes convexity comparison.

The proposed FS-SVM algorithm main blocks are considered in the paper: the search for support elements, the separating hypersurface points determination, the separating hypersurface construction as a piecewise-linear function in projection onto the coordinate axes under consideration. In further algorithm development, this function is proposed to be approximated by the Chebyshev polynomial to obtain a smooth curve.

The paper presents as an example of a problem domain, in which the FS-SVM algorithm can be applied, the oil-water-gas flow regimes classification problem, based on initial data obtained at the oil well mouth.

Keywords: machine learning, data mining, support vector machine, modification of the support vector method, oil wells.

Acknowledgements. *This work was financially supported by RFBR, project no. 18-01-00459.*

References

1. Bashmakov A.I., Bashmakov I.A. *Intellectual Information Technologies*. Moscow, 2005, 304 p. (in Russ.).
2. Vyugin V.V. *Mathematical Foundations of the Theory of Machine Learning and Forecasting*. Moscow, 2013, 390 p. (in Russ.).
3. Statnikov A., Aliferis C.F., Hardin D.P., Guyon I. *A Gentle Introduction to Support Vector Machines in Biomedicine. Vol. 1: Theory and Methods*. 2011, 200 p. DOI: 10.1142/7922.
4. Rojo-Álvarez J.L., Martinez-Ramon M., Muñoz-Marí J., Camps-Valls G. Support vector machine and kernel classification algorithms. *In: Digital Signal Processing with Kernel Methods*. 2018, pp. 433–502. DOI: 10.1002/9781118705810.ch10.
5. Pisner D.A., Schnyer D.M. *Machine Learning. Methods and Applications to Brain Disorders*. 2020, pp. 101–121. DOI: 10.1016/B978-0-12-815739-8.00006-7.
6. Nefedov A. *Support Vector Machines: A Simple Tutorial*. 2016, 35 p. Available at: https://svmtutorial.online (accessed January 20, 2020).
7. Alam S., Kang M., Pyun J.Y., Kwon G.R. Performance of classification based on PCA, linear SVM, and Multi-kernel SVM. *Proc. 8th ICUFN, IEEE*, 2016, pp. 987–989. DOI: 10.1109/ICUFN.2016.7536945.
8. Xiangdong H., Shaoqing W. Prediction of bottom-hole flow pressure in coalbed gas wells based on GA optimization SVM. *Proc. 3rd IAEAC, IEEE*, 2018, pp. 138–141. DOI: 10.1109/IAEAC.2018.8577488.
9. Butterworth D. and Hewitt G.F. (Eds.). *Two-Phase Flow and Transfer*. Oxford Univ. Press, 1977 (Russ. ed.: Moscow, 1980, 328 p.).
10. Kiselev P.T. *Hydraulics: Fundamentals of Fluid Mechanics*. Moscow, 1980, 360 p.

Для цитирования

Михайлов И.С., Зеар Аунг, Йе Тху Аунг. Разработка модификации метода опорных векторов для решения задачи классификации с ограничениями на предметную область // Программные продукты и системы. 2020. Т. 33. № 3. С. 439–448. DOI: 10.15827/0236-235X.131.439-448.

For citation

Mikhaylov I.S., Zayar Aung, Ye Thu Aung. Support vector method modification development for solving the classification problem with domain restrictions. *Software & Systems*, 2020, vol. 33, no. 3, pp. 439–448 (in Russ.). DOI: 10.15827/0236-235X.131.439-448.