Оптимизация, системный анализ и исследование операций

© 2020 г. В.В. ЗЕНКОВ, канд. техн. наук (zenkov-v@yandex.ru)
(Институт проблем управления им. В.А. Трапезникова РАН, Москва)

ПРИМЕНЕНИЕ АППРОКСИМАЦИИ ДИСКРИМИНАНТНОЙ ФУНКЦИИ АНДЕРСОНА И МЕТОДА ОПОРНЫХ ВЕКТОРОВ ДЛЯ РЕШЕНИЯ НЕКОТОРЫХ ЗАДАЧ КЛАССИФИКАЦИИ

Дискриминантная функция Андерсона имеет ряд свойств, полезных для решения задач классификации и для оценки апостериорных вероятностей классов. В качестве математического аппарата используется один и тот же взвешенный метод наименьших квадратов для аппроксимации дискриминантной функции Андерсона в области нулевых значений как при решении задачи классификации, так и при оценке апостериорных вероятностей классов в заданной точке пространства признаков. В методе опорных векторов задача классификации решается методом квадратичного программирования с количеством ограничений, равным количеству строк обучающей выборки, а для оценки апостериорных вероятностей классов используется дополнительная надстройка – калибратор Платта, преобразующий величину отступа точки от границы в апостериорную вероятность класса, с определением параметров калибратора методом максимального правдоподобия. На нескольких примерах решения задач классификации проведено сравнение эффективности методов по критерию эмпирического риска. Результаты оказались в пользу метода аппроксимации дискриминантной функции Андерсона в области нулевых значений.

Ключевые слова: машинное обучение, классификация, дискриминантная функция Андерсона, метод опорных векторов, SVM, аппроксимация дискриминантной функции Андерсона в области нулевых значений.

DOI: 10.31857/S0005231020010109

1. Введение

Дискриминантную функцию Андерсона (ДФА) здесь получаем из представленного Теодором Андерсоном метода решения байесовой задачи классификации в случае нескольких классов [1] в виде разности средних потерь от отнесения точки в пространстве признаков классов в один из двух конкурирующих классов. ДФА есть функция регрессии. Обучающая выборка с учителем просто преобразуется в выборку регрессионного анализа заменой в выборке номеров (меток) классов на соответствующие разности заданных стоимостей ошибок классификации. В случае двух классов в точках на их границе в пространстве признаков апостериорная вероятность (АпоВ) первого

(для определенности) класса зависит только от определяющих эту ДФА стоимостей двух ошибок классификации. Задача со многими классами сводится к совокупности задач с двумя классами по принципу один против остальных.

Для аппроксимации ДФА по обучающей выборке с учителем предложен эвристический алгоритм [2, 3]. Аппроксимация выполняется в области нулевых значений ДФА, поскольку для классификации важен лишь знак ДФА, а не ее изящные изгибы. Тождественная связь ДФА, АпоВ первого класса и стоимостей ошибок, для которых построена ДФА, лежит в основе методов оценивания АпоВ в заданной точке [4, 5].

Популярная, мощная, широко используемая в машинном обучении разновидность метода опорных векторов (SVM – support vector machine) для случая разделения гиперплоскостью двух перекрывающихся между собой множеств точек также является эвристикой [6]. Это обстоятельство побудило выполнить сравнение на нескольких обучающих выборках двух эвристических методов решения задачи классификации: метода аппроксимации ДФА и метода SVM. Критерием качества классификации является доля неправильно классифицированных точек обучающей выборки – эмпирический риск.

2. Дискриминантная функция Андерсона, ее свойства и связь с апостериорными вероятностями классов

Используя [1], запишем определение ДФА $f_{rs}(x)$, разделяющей классы r и s в d-мерном евклидовом пространстве признаков $x \in R^d$, использующее АпоВ классов и заданные стоимости ошибок классификации, в виде

$$f_{rs}\left(x, C\right) \equiv G_r\left(x\right) - G_s\left(x\right) \equiv \mathrm{M}_{k|x}\left(C_{rk} - C_{sk}\right), \tag{1}$$

где $G_r\left(x\right) = \sum_k C_{rk} p\left(k|x\right)$, $G_s\left(x\right) = \sum_k C_{sk} p\left(k|x\right)$ — средние потери по k в точке x, если точку отнести к классу r или, соответственно, к классу s; C — матрица стоимостей ошибок, C_{ij} — стоимость ошибки, когда точка из класса j ошибочно относится в класс i, $0 \le C_{ij} < \infty$, $C_{ii} = 0$; $p(k|x)$ — АпоВ класса k в точке x, $p\left(k|x\right) = P_k p\left(x|k\right) / p\left(x\right)$, $p\left(x\right) = \sum_k P_k p\left(x|k\right)$, P_k — априорные вероятности классов, $p(x|k)$ — условные распределения признаков классов; k — номера классов от 1 до K, K — количество классов; $M_{k|x}\left(\cdot\right)$ — математическое ожидание по k в точке x. ДФА по определению есть функция регрессии от x. В точке x случайная по k дискретная величина $C_{rk|x} - C_{sk|x} = f_{rs}(x, C) + \varepsilon_{rsk|x}$ имеет распределение АпоВ классов $p(k|x)$, $\sum_k p(k|x) = 1$, среднее $f_{rs}(x, C)$ и случайное отклонение от него $\varepsilon_{rsk|x}$. Дискретная случайная величина $\varepsilon_{rsk|x}$ принимает K значений с вероятностями $p(k|x)$ и имеет нулевое среднее.

Если $f_{rs}(x, C) \le 0$, то точка x относится в класс r и класс s исключается из дальнейшего процесса сравнения, иначе — в класс s и класс r исключается. Так обеспечивается минимум средней стоимости ошибок классификации — байесов критерий решающего правила [1].

В случае двух классов, $K = 2$, имеем $C_{1k|x} - C_{2k|x} = f_{12}(x, C) + \varepsilon_{12k|x}$. Дискретная случайная величина $\varepsilon_{12k|x}$ в точке x принимает два значения: $-C_{21} - f_{12}(x, C)$ с вероятностью $p(1|x)$ и $C_{12} - f_{12}(x, C)$ с вероятностью $1 - p(1|x)$, с нулевым средним и дисперсией $(C_{12}+C_{21})^2 p(1|x)(1-p(1|x))$.

2.1. Свойства ДФА

Перечислим свойства ДФА.

Утверждение 1. Стоимости ошибок классификации можно выбирать при условии, что их сумма равна единице.

Доказательство следует из (1). На результат попарного сравнения $G_r(x)$ и $G_s(x)$ не влияет умножение их на одно и то же положительное число.

Утверждение 2. ДФА есть ограниченная функция регрессии.

Доказательство следует из (1), так как стоимости ошибок классификации ограничены.

Следствие 1. В случае $K = 2$ имеем $-C_{21} < f_{12}(x) < C_{12}$.

Следствие 2. Чтобы аппроксимировать ДФА как функцию регрессии, следует преобразовать обучающую выборку с учителем задачи классификации в выборку задачи регрессионного анализа, заменив номера классов в выборке следующим образом: первый класс на $-C_{21}$, а второй класс — на C_{12}.

Следствие 3. Факт регрессионной зависимости ДФА от признаков позволяет, в частности, выполнять отбор признаков, используемых для решения задачи аппроксимации, по коэффициентам корреляции признаков со столбцом, в котором номера классов заменены стоимостями ошибок классификации. Учитывать при этом необходимо и коэффициенты корреляции признаков между собой [7].

2.2. Связь ДФА с AnoB классов

Утверждение 3. При $K = 2$ для AnoB первого класса и ДФ Андерсона, полученной для заданных C_{12} и C_{21}, имеет место тождество

$$p(1|x) \equiv (C_{12} - f_{12}(x))/(C_{12} + C_{21}). \tag{2}$$

Доказывается по (1) с использованием равенства $p(1|x) + p(2/x) = 1$.

Следствие 4. Задавая стоимости ошибок при условии равенства единице их суммы, тождество (2) *можно представить в виде*

$$p(1|x) \equiv p^* - f_{12}(x, p^*), \tag{3}$$

где скалярный параметр p^* определяет недиагональные элементы матрицы ошибок классификации:

$$C_{12} = p^*, \quad C_{21} = 1 - p^*. \tag{4}$$

Из (3) следует, что в точках на границе классов, если она существует, $f_{12}(x, p^*) = 0$, AnoB первого класса равна p^*, а в точках, относимых в первый класс, где $f_{12}(x, p^*) \leqslant 0$, из (3) следует, что $p(1|x) \geqslant p^*$. Если границы между классами в пространстве признаков классов нет, то p^* задает лишь недиагональные элементы матрицы стоимостей ошибок для ДФА и может находиться в пределах (0, 1), чтобы сумма стоимостей ошибок равнялась единице.

2.3. Условие неразличимости классов

Условие неразличимости классов в задаче с двумя классами для заданного p^* имеет вид

$$\left(\min_x f_{12}(x, p^*) > 0\right) \vee \left(\max_x f_{12}(x, p^*) < 0\right), \tag{5}$$

при выполнении которого все точки надо относить в один соответствующий класс, а параметр p^* есть величина, лишь определяющая стоимости ошибок классификации (4).

2.4. Способы оценивания АпоВ класса

Из тождества (3) вытекают по крайней мере три способа оценивания АпоВ первого класса в задаче с двумя классами.

По одному способу [4] для серии заданных значений параметра p^* строятся аппроксимации ДФА, по которым путем интер- и экстраполяции находится АпоВ класса в заданной точке по одной–двум соседним с точкой аппроксимациям ДФА, или по всем аппроксимациям с использованием аналога ядерных функций. Условия неразличимости классов (5) влияют на выбор предельных значений параметра p^*. Результат естественно зависит от удачного выбора вида аппроксимирующих ДФА функций.

По второму способу [5] величина параметра p^* подбирается итерационно так, чтобы в заданной точке получить нулевое значение аппроксимации ДФА. При этом АпоВ класса будет равна найденному p^*, как это следует из (3). Для аппроксимации ДФА в точке не обязательно использовать аппроксимирующие зависимости сложнее линейных. Но нужно иметь в виду, что в некоторых точках вследствие (5) решение может не существовать.

По третьему способу для произвольно заданного p^*, величина которого из интервала (0, 1) не влияет на результат оценивания АпоВ класса и не влияет так же различимость или неразличимость классов (5), строится аппроксимация ДФА в заданной точке и затем по ней и по p^* вычисляется по (3) оценка АпоВ класса. Для аппроксимации ДФА в точке используется линейная аппроксимирующая функция.

2.5. Аппроксимация ДФА в области нулевых значений

Область нулевых значений не известна. Для аппроксимации ДФА используется прием последовательных приближений к области нулевых значений, использующий взвешенный метод наименьших квадратов [2, 3].

Для аппроксимации ДФА возьмем линейную комбинацию заданных функций от признаков $\lambda'\varphi(x)$, первая компонента – единица, с вектором коэффициентов λ, который находится по обучающей выборке с учителем $\{x_n, k_n\}$, $n = 1 \div N$, k_n — номер класса в строке n, $k_n = \{1, 2\}$, $x_n \subset R^d$ — вектор действительных значений признаков размерности d в строке n. Решается последовательность задач взвешенным методом наименьших квадратов, в которой

на каждом шаге минимизируется по λ_i критерий

$$Q(\lambda_i) = \min_{\lambda_i} \sum_{n=1}^{n=N} \left\{ \left[C_{1k_n} - C_{2k_n} - \lambda_i' \varphi(x_n) \right]^2 \exp\left(-W_i |\lambda_{i-1}' \varphi(x_n)|^k \right) \right\}, \tag{6}$$

где i — номер итерации, $i = 1 \div I$. На первом шаге задача решается без весовой функции, на последующих шагах весовая функция придает больший вес точкам, более близким к нулевой области предыдущей аппроксимации ДФА. I — заданное количество итераций, $I < \infty$. W_i — заданный весовой коэффициент на шаге i, $W_i > 0$, k — заданный показатель степени. Размерность обращаемой матрицы, равная размерности искомого вектора параметров, не зависит от количества строк N в обучающей выборке. Вид весовой функции — не обязательно экспонента. Лучшим значением λ является тот вектор, которому соответствуют меньшие средние по выборке потери (эмпирический риск)

$$R = N^{-1} \left(C_{12} N_2 + C_{21} N_1 \right), \tag{7}$$

где N_1 — количество точек первого класса, ошибочно отнесенных во второй класс, N_2 — количество точек второго класса, ошибочно отнесенных в первый класс.

3. Постановка задачи

Цель работы – сравнение по величине эмпирического риска результатов решения нескольких задач классификации двумя методами: вышеописанного, использующего аппроксимацию ДФА в области нулевых ее значений, и широко известного, мощного и популярного в машинном обучении метода опорных векторов (SVM) в варианте с линейным ядром классификатора.

3.1. Процедуры метода SVM

Для сравнения с методом аппроксимации ДФА использованы три имеющиеся процедуры библиотеки scikit-learn инструментального средства Питон (Python) [8, 9]. Одна из процедур (LinearSVC) исключительно ориентирована на функцию ядра линейного типа, другие (SVC, NuSVC) позволяют задавать вид функции ядра, являясь в этом смысле более универсальными. ДФА используется в линейном относительно искомых коэффициентов виде, поэтому и процедуры SVM использованы в линейном варианте. Процедура, реализующая метод аппроксимации ДФА в области нулевых значений, также написана на Питоне.

Проблема регуляризации не затрагивалась, сравнивалось качество разделения двух наборов точек гиперплоскостью в пространстве признаков, а также в пространстве с координатами – произведениями и степенями признаков не выше второго порядка.

3.2. Исходные данные

Исходные данные – обучающие выборки с учителем – взяты в основном из примеров, ранее использованных в работах автора [2–5]. Чтобы иметь возможность проверить некоторые из приведенных результатов, в качестве одного из примеров использованы данные из репозитория UCI [10], описания которых приведены в [11]. Доступны в интернете и конкурсные данные ТМШ [12]. Данные из репозитория и конкурсные данные являются реальными. Остальные были сгенерированы в качестве модельных примеров с заданными нормальными законами условных распределений двух признаков и с заданными априорными вероятностями двух классов.

Данные из репозитория были проверены. Из них были удалены строки с отсутствующими величинами некоторых из 9 признаков.

Данные из репозитория и из конкурсной задачи представлялись в двух вариантах: с полным набором признаков и с подмножеством признаков, отобранных по коэффициентам корреляции признаков с искомой величиной — оценками ДФА, полученными путем замены номеров классов в выборке стоимостями ошибок классификации, в данном случае значениями 0,5 и −0,5. При отборе учитывались и корреляции признаков между собой. Если уменьшение количества признаков до 5 из 9 в задаче из репозитория (пример № 8, таблица) практического смысла не имело (цель — получить еще один вариант данных для сравнения методов классификации), то в конкурсной задаче выбор 3 из 216 исходных признаков, пример № 2, имело практический смысл — удобство использования полученной аппроксимации ДФА и снижение переобучения. В конкурсной задаче со всеми 216 признаками, пример № 6, не имело смысла решать полиномиальный вариант из-за гигантского количества членов. Метод аппроксимации ДФА попросту не смог бы решить такую задачу в лоб, в то время как SVM задачи, в которых количество признаков или функций от признаков больше количества строк обучающей выборки, решать может.

В задаче из репозитория отбор 5 функций от признаков по корреляции выполнялся из 54 членов полинома второго порядка, полученного по 9 исходным признакам (пример № 8). Из наиболее коррелированных с номером класса членов полинома были отброшены те, которые имели корреляцию между собой более 0,8. Осталось 5 из 54 членов. Корреляция 0,8 задавалась исключительно из соображения получить не слишком много членов.

4. Решение задачи

Коэффициенты весовой функции при обращении к процедуре аппроксимации ДФА (6) изменялись в итерационном процессе по правилу

$$W_i = w \times i, \quad i = 0 \div I, \tag{8}$$

где шаг изменения w подбирался вручную из нескольких значений, чтобы получить поменьше величину эмпирического риска (7) в итерационном процессе, см. рис. 1–5. Автоматический перебор значений шага w для (8) из некоторого диапазона для поиска наилучшего значения не выполнялся, потому что итерационные процессы в несколько десятков шагов занимали мало времени,

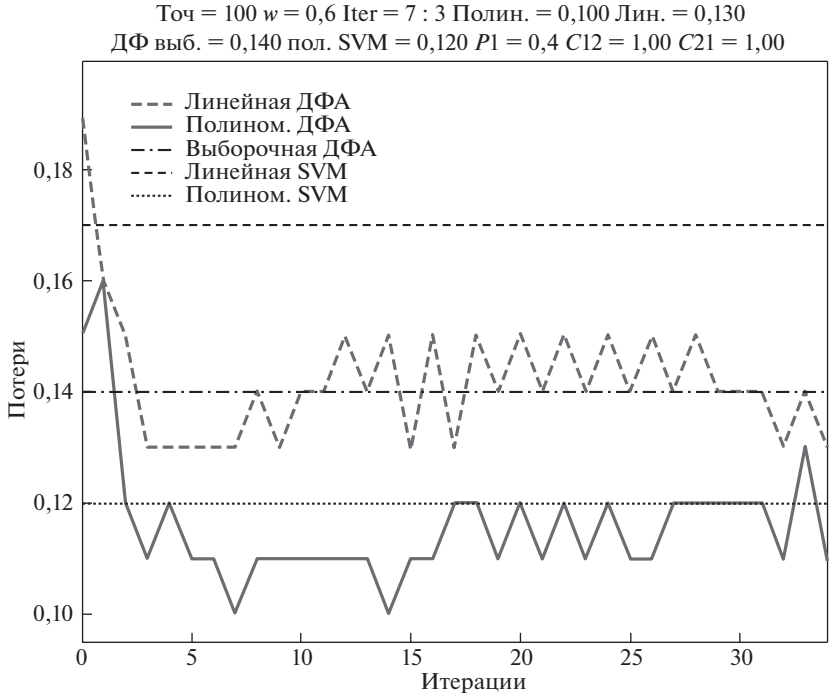

Рис. 1. Пример № 1.

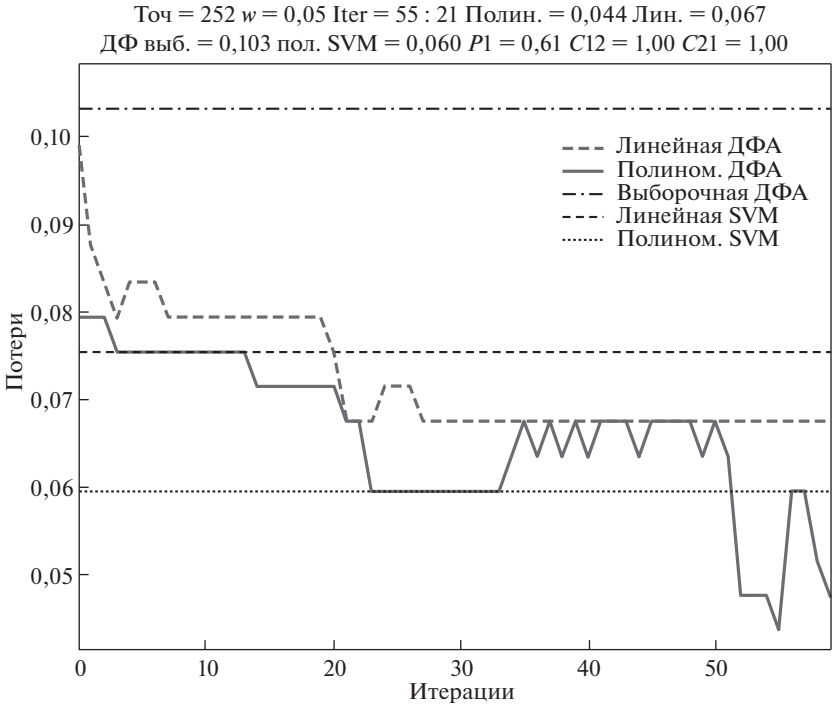

Рис. 2. Пример № 2.

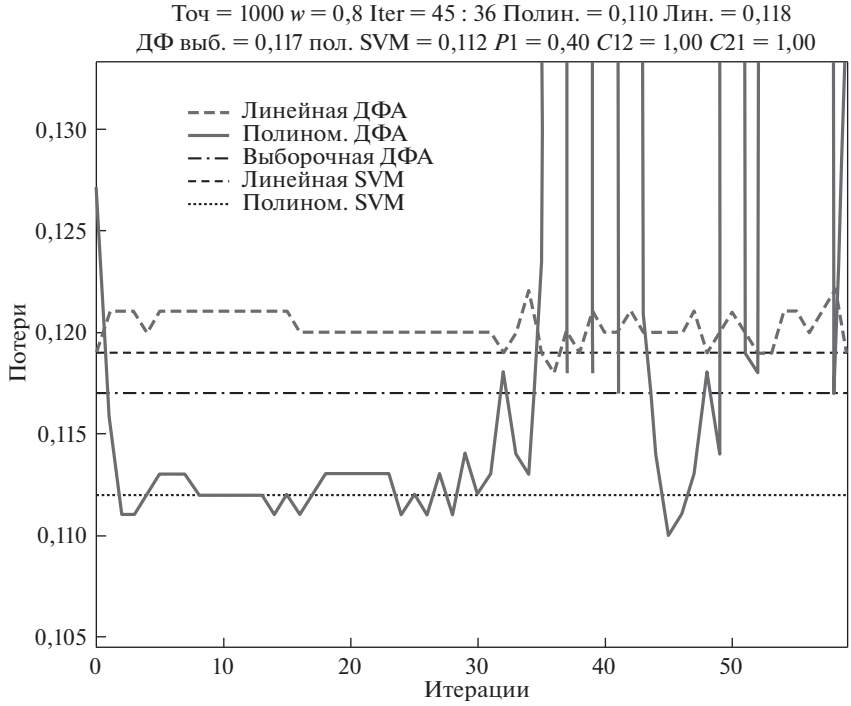

Рис. 3. Пример № 3.

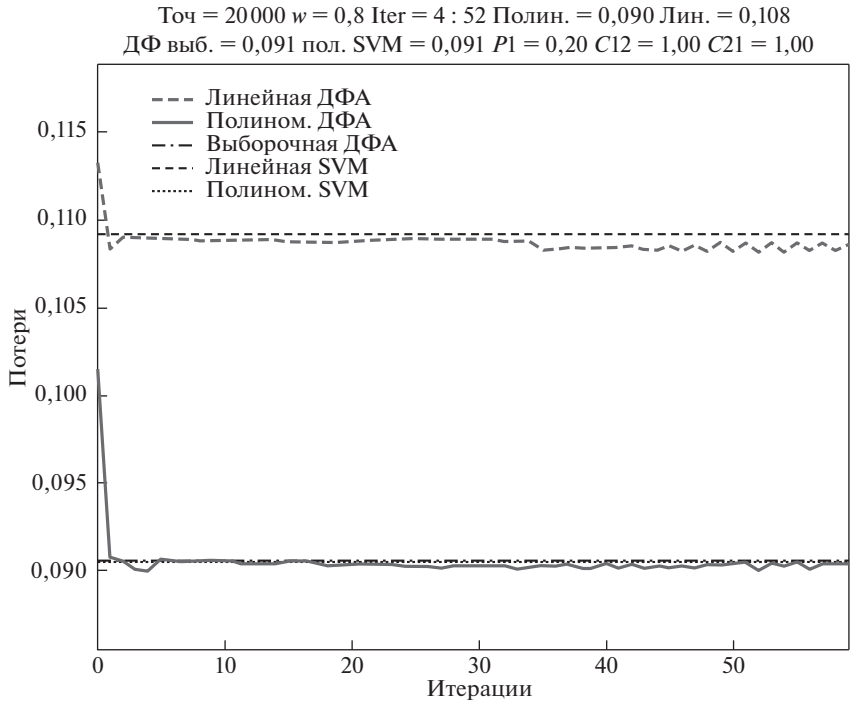

Рис. 4. Пример № 4.

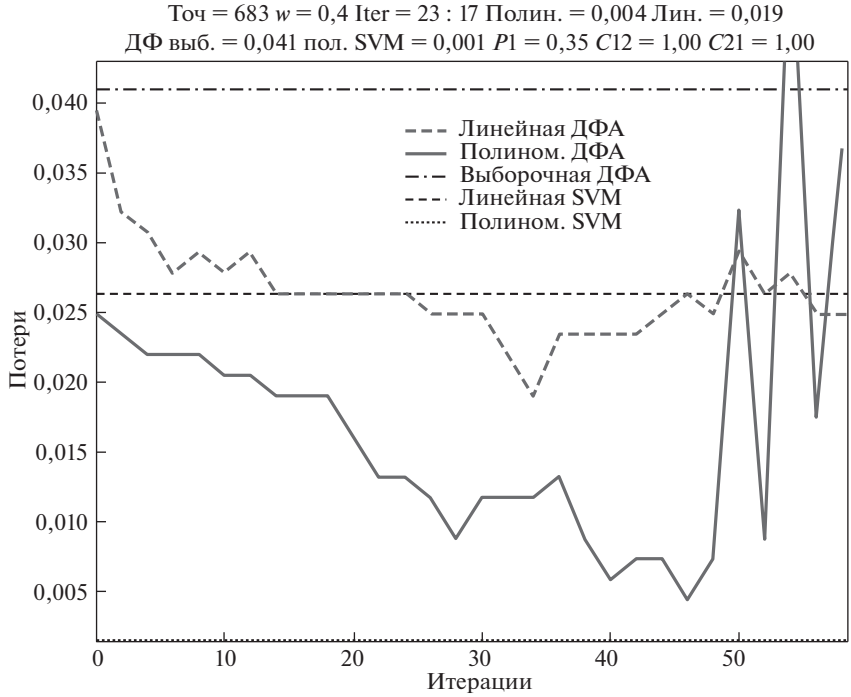

Рис. 5. Пример № 7.

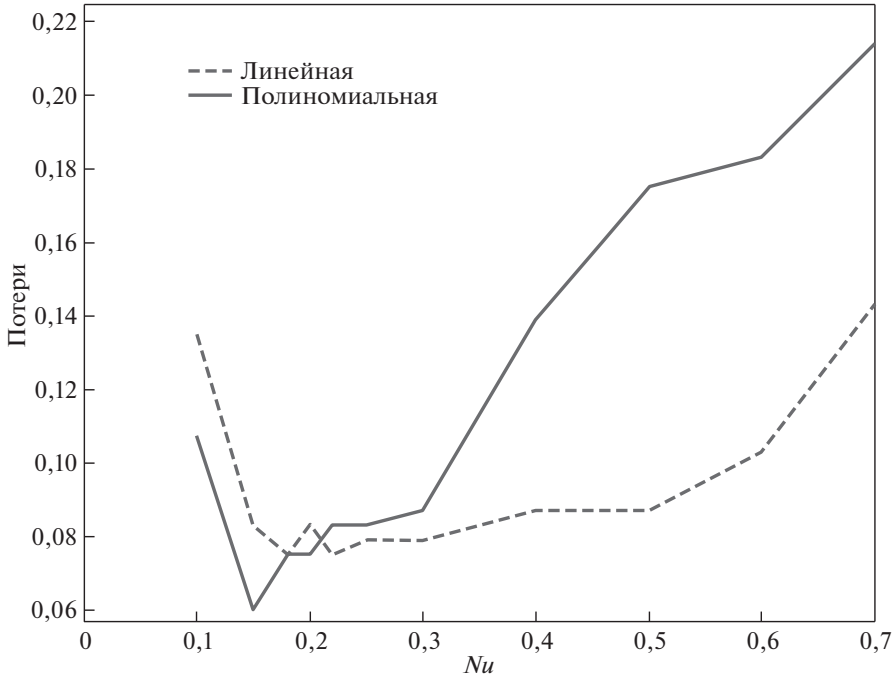

Рис. 6. Зависимость потерь от параметра Nu.

ручной выбор не был обременителен и визуальный контроль итерационного процесса облегчал подбор w и I. Показатель степени k в (6) равен единице. Лишь в одном случае (пример № 7, полином), когда аппроксимация ДФА не была лучше SVM, пытались долго и безуспешно искать и лучшее значение w и k в (6).

Количество шагов итераций I выбирается достаточно большим, но не слишком, потому что график изменения потерь от итераций с некоторого момента становится хаотическим из-за уменьшения части выборки, выделяемой весовой функцией с ростом W_i по правилу (8) и влияющей на результат,

Характер изменения по итерациям весового коэффициента в виде арифметической прогрессии в (8) хорошо зарекомендовал себя при решении автором предыдущих задач.

Процедуры метода SVM также настраивались в каждом примере параметрами: C в SVC и LinearSVC; Nu в NuSVC. И если параметр C (коэффициент штрафной функции) изменял эмпирический риск (7) сравнительно в малых пределах, то параметр Nu изменял (7) в широких, см. рис. 6. Параметр Nu — верхняя граница ошибок обучения. Он должен находиться в интервале (0, 1]. По умолчанию равен 0,5.

Остальные параметры процедур в примерах не изменялись. Для LinearSVC они были установлены отличными от значений по умолчанию на уровнях: dual=False. Параметр dual определяет функцию потерь. Значение ‘hinge‘ – это стандартная потеря SVM (используемая, например, классом SVC для решения задач классификации), в то время как значение ‘squared_hinge‘ – это использование квадрата потерь. Параметр penalty, задающий норму штрафа, установлен в значение ‘l1‘. Норма ‘l2‘ – это стандарт, используемый в SVC. Параметр max_iter = 100000 (максимальное количество итераций, которое можно выполнять). Для SVC и NuSVC установлены: kernel = ’linear’ – определяет тип ядра, который используется в алгоритме.

Проблема регуляризации не затрагивалась, сравнивалось качество разделения двух наборов точек гиперплоскостью в пространстве признаков, а также в пространстве с координатами – произведениями и степенями признаков не выше второго порядка.

Результаты решения задач разными методами представлены в таблице.

Графики итерационных процессов получения аппроксимаций ДФА, на которых горизонтальными линиями изображены решения задач процедурами SVM, представлены на рис. 1–4, соответствующих примерам № 1–4. Рисунок 5 соответствует примеру № 7. Рисунок для примера № 6 не представлен, так как графики вырождаются в горизонтальные линии. Рисунок для примера № 8 не представлен для экономии места.

Рисунок 6 показывает зависимости потерь (7) от параметра Nu, полученных при ручном поиске лучших значений для процедуры NuSVC в случаях линейной и полиномиальной дискриминантных функций для условий примера № 2. Из рисунка видно, что полагаться на установленное по умолчанию значение параметра $Nu = 0{,}5$ не стоит, если стремиться к получению более точного решения задачи классификации методом SVM, реализованным в про-

Таблица

№	Кол-во точек	Кол-во призн., вид модели		ДФА				LinearSVC		SVC		NuSVC		Выб, ДФ
				R	t	I	w	R	t	R	t	R	t	R
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
1	100	2	Линейн.	**0,130**	+	35	0,6	0,180	0,02	**0,170**	0,000	**0,170**	0,000	
		5	Полином.	**0,100**	0,32	35	0,6	0,140	0,00	**0,120**	0,000	0,150	0,000	0,140
2	252	3	Линейн.	**0,067**	+	60	0,05	**0,075**	0,00	**0,075**	0,016	**0,075**	0,000	
		9	Полином.	**0,044**	1,06	60	0,05	**0,060**	0,05	0,067	0,781	**0,060**	0,078	0,103
3	1000	2	Линейн.	**0,118**	+	60	0,8	**0,119**	0,00	**0,119**	0,047	**0,119**	0,047	
		5	Полином.	**0,110**	3,72	60	0,8	0,114	0,06	**0,112**	0,047	0,116	0,062	0,117
4	20000	2	Линейн.	**0,1082**	+	60	0,8	0,1092	0,11	**0,1091**	22,57	0,1099	22,25	
		5	Полином.	**0,0900**	59,1	60	0,8	0,0907	0,72	**0,0905**	28,93	0,0924	37,28	0,0906
5	50000	2	Линейн.	**0,1405**	+	60	0.8	0,14108	0,25	0,1411	98,00	**0,1409**	152	
		5	Полином.	**0,1205**	142	60	0,8	**0,12120**	1,72	0,1213	89,14	0,1214	217	0,1213
6	252	216	Линейн.	**0**	7,5	60	0,8	**0**	0,91	**0**	0,016	**0**	0,03	
7	683	9	Линейн.	**0,019**	+	30	0.4	0,0278	0,03	**0,0264**	0,109	0,0278	0,016	
		54	Полином.	0,0044	1,91	30	0,4	**0,0015**	27,9	**0,0015**	26,78	0,0029	0,031	0,041
8	683	5	Линейн.	**0,023**	+	30	0,4	0,0322	0,03	**0,0293**	0,258	**0,0293**	0,016	
		20	Полином.	**0,019**	1,31	30	0,4	0,0307	0,23	**0,0264**	433,0	0,0278	0,047	0,042

цедуре NuSVC. Аналогичное замечание справедливо и в отношении реализаций SVM в процедурах LinearSVC и SVC, хотя в них изменения параметра C не столь существенно сказываются на результате.

Пояснения к таблице. Столбцы: 1 – номер примера; 2 – количество точек выборки; 3 – количество признаков и количество использованных членов в полиноме второго порядка; 4 – вариант дискриминантной функции; 5 – эмпирический риск по итерациям в двух видах аппроксимаций ДФА; 6 – суммарное время на выполнение заданных итераций по двум видам аппроксимации ДФА; 7 – количество заданных итераций в каждом виде аппроксимаций ДФА; 8 – коэффициент весовой функции; 9 – эмпирический риск метода SVM LinearSVC; 10 – время решения SVM LinearSVM в секундах; 11–14 – аналогично 9 и 10, но для SVC и NuSVC; 15 – эмпирический риск в предположении, что обучающая выборка подчиняется двум нормальным условным законам распределения признаков и параметры их получены по выборке (для примеров 1, 3–5 предположение справедливо по построению. Дискриминантная функция для нормально распределенных признаков двух классов является полиномом второго порядка).

В столбцах 5, 9, 11 и 13 полужирным шрифтом отмечены лучшие эмпирические риски сравниваемых методов, причем в столбцах 9, 11 и 13 отмечены и лучшие риски среди реализаций метода SVM. В столбце 5 отмечены лучшие результаты для метода аппроксимации ДФА по сравнению с лучшим результатом среди реализаций SVM.

В одном из 15 примеров, в примере № 7, рис. 5, в полиномиальном варианте все реализации метода SVM дали лучший результат, чем метод аппроксимации ДФА. Причем в отличие от других примеров в примере № 7 выполнялся тщательный поиск лучшего решения не только в широком диапазоне коэффициента w в (8), но испытывались и другие весовые функции с разными степенями k в весовой функции.

Работа выполнена на ноутбуке SMARTBOOK 116C Prestigio, CPU x64, 144GHz 144GHz. Оперативная память 2 ГБ. Дисковая память 32 ГБ. Операционная система Windows 10 домашняя 32-разрядная.

4.1. Способ проверить некоторые результаты

Чтобы можно было проверить результаты сравнения методов на общедоступных наборах данных [10, 12], приводим две аппроксимации ДФА. Для сравнения с ними решений, получаемых методом SVM, можно использовать доступные в разных инструментальных средствах реализации метода SVM подобно взятым из Питона [8, 9].

Линейная модель для примера № 7, рак легких, репозиторий [10, 11]:

$$\begin{aligned} f_{12}(x) = {} & -0{,}45946108 + 0{,}02133231x1 + 0{,}01406552x2 + \\ & + 0{,}02060665x3 + 0{,}01399266x4 - 0{,}00209362x5 + 0{,}0264621x6 + \\ & + 0{,}01320144x7 + 0{,}01350233x8 + 0{,}02748957x9, \end{aligned} \tag{9}$$

где переменные $x1$–$x9$ обозначают признаки, перечисленные в списке атрибутов под номерами 2–10 в [11]. Если $f_{12}(x) < 0$, то точку следует отнести в класс с меткой 2 — доброкачественный (benign), иначе — в класс с мет-

кой 4 — злокачественный (malignant). Количество ошибочно классифицированных точек — 13 из 683 точек обучающей выборки. Из 699 строк выборки [10] были предварительно исключены 16 строк с неполными данными.

Полиномиальная модель для примера № 2, диагностика заболевания, конкурсная задача [12]:

$$f_{12}(x) = -0{,}81685233 - 0{,}54083596x1 - 0{,}34128288x2 - 0{,}1861543x3 + \\ + 0{,}01519183x1^{**}2 + 0{,}30754489x1x2 + 0{,}42900778x1x3 + \\ + 0{,}0584966x2^{**}2 - 0{,}06766405x2x3 + 0{,}01869679x3^{**}2, \tag{10}$$

где $x1$, $x2$, $x3$ — соответственно элементы в столбцах 22, 104 и 115 обучающей выборки. В первом столбце 0 — метка здорового пациента, 1 — больного. Три признака отобраны из 216 по коэффициентам корреляции с первым столбцом и коэффициентам корреляции между собой.

Если $f_{12}(x) < 0$, то пациент здоров, иначе – болен некоторой болезнью.

Из 252 строк обучающей выборки ошибочно классифицированы 11 строк.

5. Заключение

1. Метод аппроксимации дискриминантной функции Андерсона в области нулевых значений (ДФА) по обучающей выборке с учителем, используемый для решения задач классификации с целью минимизации эмпирического риска, является эвристическим методом. Метод опорных векторов (SVM) также является эвристическим методом. Эвристика этих методов побуждает выполнять их сравнение между собой на модельных примерах и на реальных обучающих выборках. Метод опорных векторов был представлен тремя реализациями, имеющимися в инструментальном средстве Питон. Метод ДФА был написан автором на том же языке. Имеется и вариант метода, написанный автором на МАТЛАБе.

2. Из 15 примеров лишь в одном эмпирические риски, полученные всеми тремя реализациями SVM, оказались меньше, чем эмпирический риск, полученный методом аппроксимации ДФА. В одном примере все реализации SVM и ДФА дали одинаковый, нулевой, результат. В остальных примерах метод аппроксимации ДФА был лучше всех трех реализаций SVM.

3. При сравнении выполнялась ручная настройка параметров методов: w в ДФА, C в SVC и LinearSVC, Nu в NuSVC. Для тех обучающих выборок, которые можно скопировать из интернета, приведены дискриминантные функции, полученные методом аппроксимации ДФА. Для них можно попытаться найти лучшие параметры настроек реализаций методов SVM, чтобы сравнить с результатами работы.

4. Примеры решений задач эвристическими методами не могут дать исчерпывающего ответа на вопрос о том, какой из методов лучше. Так, по показателю эмпирического риска в некоторых случаях один метод оказывается лучше другого. По используемому математическому аппарату (взвешенному методу наименьших квадратов) ДФА проще SVM, использущего квадратичное программирование с количеством ограничений, равным количеству строк в обучающей выборке. Но SVM может решать задачу, когда количество признаков больше количества строк в выборке. В ДФА в таких случаях отбирает-

ся меньшее количество признаков по коэффициентам корреляции. SVM для оценок апостериорных вероятностей классов в точках пространства признаков должен использовать специальные надстройки типа калибратора Платта с оценкой параметров методом максимального правдоподобия, а в ДФА для оценки апостериорных вероятностей классов используется тот же взвешенный метод наименьших квадратов, который используется для аппроксимации ДФА в окрестности нулевых значений.

СПИСОК ЛИТЕРАТУРЫ

1. *Anderson T.W.* An Introduction to Multivariate Statistical Analysis. Third edition. John Wiley & Sons, 2003. 721 p.
2. *Зенков В.В.* Аппроксимация дискриминантных функций в окрестности нулевых значений // Изв. АН СССР. Техн. кибернетика. 1973. № 2. С. 152–156.
3. *Зенков В.В.* Использование взвешенного метода наименьших квадратов при аппроксимации дискриминантной функции цилиндрической поверхностью в задачах классификации // АиТ. 2017. № 9. С. 145–158.

   Zenkov V.V. Using Weighted Least Squares to Approximate the Discriminant Function with a Cylindrical Surface in Classification Problems // Autom. Remote Control. 2017. V. 78. No. 9. P. 1662–1673.
4. *Зенков В.В.* Оценка апостериорной вероятности класса по серии дискриминантных функций Андерсона // АиТ. 2019. № 3. С. 68–71.

   Zenkov V.V. Evaluation of the Posterior Probability of a Class with a Series of Anderson Discriminant Functions // Autom. Remote Control. 2019. V. 80. No. 3. P. 447–458.
5. *Зенков В.В.* Оценка вероятности принадлежности точки классу по аппроксимации одной дискриминантной функции // АиТ. 2018. № 9. С. 46–58.

   Zenkov V.V. Estimating the Probability of a Class at a Point by the Approximation of one Discriminant Function // Autom. Remote Control. 2018. V. 79. No. 9. P. 1580–1590.
6. *Воронцов К.В.* Математические методы обучения по прецедентам (теория обучения машин). http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf
7. *Дрейпер Н., Смит Г.* Прикладной регрессионный анализ, 3-е изд. : Пер. с англ. М.: Изд. дом "Вильямс", 2007. 912 с.
8. Scikit learn. 1.4. Support Vector Machines.

   https://scikit-learn.org/stable/modules/svm
9. Sklearn.svm.LinearSVC.

   https://scikit-learn.org/0.20/modules/generated/sklearn.svm.LinearSVC.html
10. UC Irvine Machine Learning Repository. http://mlr.cs.umass.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.data
11. UC Irvine Machine Learning Repository. http://mlr.cs.umass.edu/ml/machine-learning-databases/breast-cancer-wisconsin/breast-cancer-wisconsin.names
12. Данные для задания на ТМШ 2014.

    http://www.machinelearning.ru /wiki/images/e/e1/School-VI-2014-task-3.rar

Статья представлена к публикации членом редколлегии В.И. Васильевым.

Поступила в редакцию 05.04.2019

После доработки 03.06.2019

Принята к публикации 18.07.2019