Интеллектуальные системы управления, анализ данных

© 2019 г. Ю.А. ДУБНОВ (yury.dubnov@phystech.edu)
(Институт системного анализа Федерального исследовательского центра "Информатика и управление" РАН, Москва;
Национальный исследовательский университет "Высшая школа экономики", Москва;
Московский физико-технический институт (Государственный университет))

ЭНТРОПИЙНОЕ ОЦЕНИВАНИЕ В ЗАДАЧАХ КЛАССИФИКАЦИИ[1]

Рассматривается задача бинарной классификации, предлагается алгоритм ее решения, базирующийся на методе энтропийного оценивания параметров решающего правила. Приведено подробное описание энтропийного метода оценивания и алгоритма классификации, описаны преимущества и недостатки такого подхода, приведены результаты численных экспериментов и сравнения с классическим методом опорных векторов по точности классификации и степени зависимости от объема обучающей выборки.

Ключевые слова: машинное обучение, классификация, метод максимума энтропии.

DOI: 10.1134/S0005231019030097

1. Введение

Задача бинарной классификации является одной из основополагающих в теории машинного обучения и одной из самых распространенных задач прикладного анализа данных [1]. Существует множество алгоритмов решения задачи классификации, различающихся способами работы с обучающей выборкой, моделями решающего правила и методами решения задачи оптимизации. В качестве наиболее распространенных и универсальных алгоритмов можно выделить решающие деревья (Decision tree) [2], наивный байесовский классификатор (NB) [3] и логистическую регрессию (Logistic regression) [4]. Среди метрических классификаторов наиболее распространенными являются метод ближайших соседей (kNN) [5] и машина опорных векторов (SVM), в том числе ядровые модификации (Kernel SVM) для линейно неразделимых классов [6]. Еще две отдельные группы образуют методы, основанные на построении нейронных сетей и различные композиционные алгоритмы обучения, такие как стекинг и бустинг [7]. В данной статье будет представлен альтернативный метод классификации, основанный на рандомизации модели решающего правила и принципе максимума энтропии [8].

[1] Работа выполнена при финансовой поддержке Российского фонда фундаментальных исследований (проект № 17-29-02115).

Принцип максимума энтропии уже применяется в различных методах машинного обучения, например для выбора разделяющего правила в решающих деревьях или при отборе информативных признаков [9]. Кроме того, существует группа методов, основанных на оценке вероятностей принадлежности объектов к разным классам. К ним относятся широко распространенный метод под названием maxEnt [10] и обобщение логистической регрессии на случай нескольких классов (multinomial logistic regression) [11]. Подчеркнем, что в данных методах принцип максимума энтропии применяется в отношении вероятностей принадлежности классам, т.е. ответам классификатора, а не к параметрам математических моделей, описывающих решающее правило.

Под энтропийным оцениванием параметров решающего правила понимают применение классического метода максимума энтропии (General Maximum Entropy) [12], который был разработан преимущественно для задачи регрессии в случаях, когда классический подход, основанный на методе наименьших квадратов, оказывается неэффективным. В [13], например, продемонстрирована эффективность энтропийного оценивания для линейной регрессии при наличии несимметричных и негауссовских шумов, а также в условиях малой обучающей выборки.

Основная идея метода энтропийного оценивания заключается в представлении параметров модели случайными величинами и подборе для них таких функций распределения вероятности, которые обладали бы наибольшей энтропией, но и согласовывались бы с данными обучающей выборки. Такой подход позволяет получить наиболее робастное решение в условиях наибольшей неопределенности [14].

2. Классический линейный классификатор

Линейным классификатором называют алгоритм классификации, основанный на построении линейной разделяющей поверхности в пространстве признаков. Для бинарной классификации двумерных объектов разделяющей поверхностью будет прямая, для трехмерных — плоскость, а в общем случае — гиперплоскость [1].

Пусть объекты из множества X описываются числовыми признаками $f_j : X \to R$, $j = 1, \ldots, d$, в пространстве $\mathbb{R}^d$. Пусть Y — множество меток классов, для случая бинарной классификации $Y = \{-1, 1\}$. Тогда линейный классификатор определяется так:

$$a(x, w) = \operatorname{sign}\left(\sum_{j=1}^{d} w_j f_j(x) - w_0\right) = \operatorname{sign}\langle x, w\rangle\,, \tag{2.1}$$

где $\langle x, w\rangle$ – скалярное произведение признакового описания объекта на вектор параметров $w = (w_0, w_1, \ldots, w_d)$. Далее в качестве признакового описания объектов будут использоваться компоненты матрицы $X = \{x_{ij},\ i = \overline{1,n},\ j = \overline{0,d}\}$, дополненной единичным столбцом.

Основной задачей машинного обучения в данном случае является "настройка" вектора весов, т.е. по заданной обучающей выборке $X^n =$

$= \{(x_1, y_1), \ldots, (x_n, y_n)\}$ построить алгоритм указанного вида, минимизирующий функционал эмпирического риска, т.е. число неправильно классифицированных объектов:

$$Q(w) = \sum_{i=1}^{n} [a(x_i, w) \neq y_i] \to \min_w . \tag{2.2}$$

Для решения оптимизационной задачи (2.2) вводится понятие отступа (margin)

$$M(x_i) = y_i \langle x_i, w \rangle . \tag{2.3}$$

Отступ представляет собой числовое значение, характеризующее степень погружения объекта в соответствующий класс. Чем меньше значение $M(x_i)$, тем ближе объект x_i находится к разделяющей поверхности между классами и тем выше становится вероятность ошибки классификации. Поскольку величина отступа $M(x_i)$ становится отрицательной лишь для тех примеров, на которых алгоритм ошибается, то функционал эмпирического риска можно переписать в виде

$$Q(w) = \sum_{i=1}^{n} [M(x_i) < 0] . \tag{2.4}$$

Поскольку в таком виде функционал эмпирического риска является недифференцируемой функцией от вектора параметров w, то градиентные методы оптимизации оказываются неприменимыми. Существует несколько подходов для численного решения данной задачи. Так, в методе опорных векторов (SVM) задача минимизации эмпирического риска заменяется задачей построения оптимальной разделяющей гиперплоскости. Еще одним подходом для численного решения задачи оптимизации является замена пороговой функции потерь некоторыми непрерывными аппроксимациями, после чего минимизируется не сам функционал эмпирического риска, а его верхняя оценка:

$$Q(w) \leqslant \tilde{Q}(w) = \sum_{i=1}^{n} L\left(M(x_i)\right) \to \min_w . \tag{2.5}$$

Наиболее распространенными аппроксимациями функции потерь являются: квадратичная (дискриминант Фишера), сигмоидная (однослойный перцептрон), логарифмическая, экспоненциальная и кусочно-линейная с L_2-регуляризацией [15].

3. Линейный классификатор с энтропийным оцениванием параметров

От классических методов машинного обучения энтропийное оценивание отличается другой интерпретацией модели решающего правила. Вектор параметров w, определяющий разделяющую поверхность, полагается многомерной случайной величиной, реализации которой будут описывать не одну разделяющую поверхность, а ансамбль возможных поверхностей. Задачей энтропийного оценивания является вычисление энтропийно-оптимальной функции

плотности распределения вероятности для параметров решающего правила. Такой подход был предложен в [14, 16], однако в настоящей статье рассмотрена его модификация, основанная на использовании дискретных случайных величин вместо непрерывных, что существенно снижает вычислительные затраты и ускоряет работу алгоритма.

Итак, пусть дискретные случайные величины w_j, $j = \overline{0,d}$, определены на множествах значений $W_j = \{w_{j1}, w_{j2}, \ldots, w_{jk}\}$ и имеют дискретные функции распределения $p^{(j)} = \{p_{j1}, p_{j2}, \ldots, p_{jk}\}$. Задачей оценивания является восстановление значений вероятностей для всех возможных исходов случайных величин p_{jl}, $j = \overline{0,d}$, $l = \overline{1,k}$, обладающих наибольшей энтропией:

$$H(p) = -\sum_{j=0}^{d}\sum_{l=1}^{k} p_{jl} \ln p_{jl} \to \max_{p}. \tag{3.1}$$

Нетрудно видеть, что максимальной энтропией обладают равномерно распределенные случайные величины, поэтому дополнительно накладываются ограничения на значения функций распределения

$$\sum_{l=1}^{k} p_{jl} = 1, \quad j = \overline{0,d}, \tag{3.2}$$

и балансовые ограничения на средние значения случайных величин

$$\mathbb{E}^w\left[M(x_i)\right] > 0, \quad i = \overline{1,n}. \tag{3.3}$$

Ограничение (3.3) является аналогом условия неотрицательности отступов (2.3) для рандомизированных параметров w и обеспечивает согласованность с обучающей выборкой. Распишем более подробно выражение (3.3):

$$\begin{aligned}\mathbb{E}^w\left[M(x_i)\right] = \mathbb{E}^w\left[y_i \langle x_i, w\rangle\right] = \mathbb{E}^w\left[\sum_{j=0}^{d} y_i x_{ij} w_j\right] = \\ = \sum_{j=0}^{d} y_i x_{ij} \mathbb{E}^w\left[w_j\right] = \sum_{j=0}^{d}\sum_{l=1}^{k} y_i x_{ij} w_{jl} p_{jl} > 0.\end{aligned} \tag{3.4}$$

Задача поиска условного экстремума (3.1) с ограничениями равенствами (3.2) и ограничениями неравенствами (3.4) удовлетворяет всем условиям теоремы Куна–Таккера и решается методом неопределенных множителей Лагранжа. Функция Лагранжа имеет вид

$$\begin{aligned}L(\bar{\mu}, \bar{\lambda}) = \sum_{j=0}^{d}\sum_{l=1}^{k} p_{jl} \ln p_{jl} + \sum_{j=0}^{d} \mu_j \left(\sum_{l=1}^{k} p_{jl} - 1\right) - \\ - \sum_{i=1}^{n} \lambda_i \left(\sum_{j=0}^{d}\sum_{l=1}^{k} y_i x_{ij} w_{jl} p_{jl}\right),\end{aligned} \tag{3.5}$$

тогда в стационарной точке ее частные производные обращаются в 0:

$$\frac{\partial L(\bar{\mu},\bar{\lambda})}{\partial p_{jl}} = \ln p_{jl} + 1 + \mu_j - \sum_{i=1}^{n} \lambda_i y_i x_{ij} w_{jl} = 0. \tag{3.6}$$

Следовательно, решение для p_{jl} имеет вид:

$$p_{jl} = \exp\left(-1 - \mu_j + \sum_{i=1}^{n} \lambda_i y_i x_{ij} w_{jl}\right). \tag{3.7}$$

Подставляя это выражение в ограничения равенства (3.2), получим:

$$p_{jl} = \frac{\exp\left(\sum\limits_{i=1}^{n} \lambda_i y_i x_{ij} w_{jl}\right)}{\sum\limits_{l=1}^{k} \exp\left(\sum\limits_{i=1}^{n} \lambda_i y_i x_{ij} w_{jl}\right)}, \qquad j = \overline{0,d}, \quad l = \overline{1,k}. \tag{3.8}$$

Выражение (3.8) определяет вид энтропийно-оптимальных функций распределения, параметризованных множителями Лагранжа $\lambda_1, \ldots, \lambda_n$, которые восстанавливаются из условий неотрицательности ($\lambda_i \geqslant 0$) и дополнительной нежесткости:

$$\lambda_i \mathbb{E}^w \left[M(x_i)\right] = \sum_{j=0}^{d} \sum_{l=1}^{k} \lambda_i y_i x_{ij} w_{jl} p_{jl} = 0, \quad i = \overline{1,n}. \tag{3.9}$$

Система уравнений (3.9) решается численными методами, например, посредством минимизации квадратичной функции невязки с использованием градиентных методов оптимизации.

Таким образом, были вычислены значения функций распределения для параметров модели линейного классификатора (2.1). В качестве точечной оценки параметров используется среднее значение по полученным согласно выражению (3.8) распределениям вероятности, т.е.:

$$w_j^* = \mathbb{E}^w \left[w_j\right] = \sum_{l=1}^{k} w_{jl} p_{jl}, \qquad j = \overline{0,d}. \tag{3.10}$$

Параметры (3.10) задают разделяющую поверхность между точками разных классов аналогично с классическим линейным классификатором. С другой стороны, полученные распределения вероятностей открывают более широкие возможности для дальнейшей настройки и тестирования классификатора.

4. Алгоритм рандомизированной классификации

Рандомизированная классификация базируется на идее семплирования полученных в результате энтропийного оценивания распределений случайных

величин. В таком случае каждый семпл параметров будет определять некоторую отдельную разделяющую поверхность, а весь набор случайных величин — ансамбль линейных классификаторов.

В отличие от использования точечной оценки семплирование случайных величин позволяет получить некоторую выборку значений и применить аппарат математической статистики; например, в задаче регрессии это позволит рассчитать дисперсию для оценки параметров модели и построить доверительные интервалы. Преимуществом такого подхода в задаче классификации является возможность определения вероятностей принадлежности объектов к различным классам, даже для методов, предполагающих только бинарную, но не вероятностную классификацию. Например, описанный алгоритм линейной классификации аналогично методу опорных векторов позволяет определить лишь класс объекта, но не вероятность принадлежности к данному классу, в отличие, например, от метода логистической регрессии для задачи классификации.

Приведем один из возможных алгоритмов использования ансамбля случайных величин для решения задачи бинарной классификации.

- Генерируется выборка значений $w^{(s)} = \{w_0^{(s)}, \ldots, w_d^{(s)}\}$, $s = \overline{1,m}$, объемом m. Тогда каждый набор параметров $w^{(s)}$ определяет линейный классификатор вида

$$a(x, w^{(s)}) = \operatorname{sign}\left\langle x, w^{(s)}\right\rangle = \operatorname{sign}\left(\sum_{j=0}^{d} x_j w_j^{(s)}\right). \tag{4.1}$$

- Для каждого объекта из обучающей коллекции вычисляется ансамбль ответов:

$$a(x, w^{(s)}) : x_i \to y_i^{(s)}, \qquad i = \overline{1,n}, \quad s = \overline{1,m}. \tag{4.2}$$

- Вычисляются эмпирические вероятности принадлежности объектов к разным классам, равные доле ответов соответствующего класса среди всего ансамбля классификаторов:

$$\begin{aligned} p_1^{(i)} &= \tfrac{1}{m}\sum_{s=1}^{m}\left[y_i^{(s)} = -1\right], \\ p_2^{(i)} &= \tfrac{1}{m}\sum_{s=1}^{m}\left[y_i^{(s)} = 1\right], \end{aligned} \qquad i = \overline{1,n}. \tag{4.3}$$

Эмпирические вероятности (4.3) могут использоваться в качестве ответов в задачах, в которых требуется определить лишь вероятность принадлежности к классам, а для задач с требованием четкого определения класса для каждого нового объекта эмпирические вероятности должны быть преобразованы в метки классов, например посредством выбора некоторого порогового

значения $t \in (0,1)$:

$$y_i = \begin{cases} -1, & p_1^{(i)} \geqslant t, \\ 1, & p_1^{(i)} < t. \end{cases} \tag{4.4}$$

Пороговое значение t определяет уровень достоверной классификации и является гиперпараметром алгоритма. Как и другие гиперпараметры алгоритмов классификации, оптимальное пороговое значение может быть определено в результате серий экспериментов над обучающей выборкой. В данной статье использовался метод перекрестной проверки по пяти блокам (5-fold). Другим известным подходом к подбору гиперпараметров является ROC-анализ, основанный на построении ROC-кривых (ROC-curve) [17].

5. Случай линейной неразделимости классов

Несмотря на широкую распространенность линейных классификаторов, большинство практических примеров обучающих коллекций являются линейно неразделимыми. В таких случаях применяют ядровые модификации, основанные на использовании нелинейных функций похожести точек обучающей выборки [1, 6]. Приведем аналог ядровой модификации для линейного классификатора с энтропийным оцениванием параметров.

Предположим, что исходное признаковое пространство проецируется в пространство большей размерности с помощью отображения $\phi : \mathbb{R}^d \to \mathbb{R}^n$, в котором также определено скалярное произведение между образами точек $\langle \phi(x_i), \phi(x_j) \rangle$. Запишем решающее правило линейного классификатора в новом пространстве аналогично (2.1):

$$\begin{gathered} a(x_i, w) = \operatorname{sgn}\left(s(x_i, w)\right) = \operatorname{sgn}\left(\langle \phi(x_i), \phi(w) \rangle\right), \\ \text{где} \quad \phi(w) = \sum_{j=1}^{n} w_j y_j \phi(x_j). \end{gathered} \tag{5.1}$$

Тогда

$$s(x_i, w) = \sum_{j=1}^{n} w_j y_j \left\langle \phi(x_i), \phi(x_j) \right\rangle, \quad i = \overline{1,n}. \tag{5.2}$$

Прием, получивший в публикациях название “Kernel trick”, позволяет в качестве скалярного произведения между образами точек использовать функцию ядра, обладающую свойствами симметричности и неотрицательной определенности. Например, гауссовское ядро выглядит так:

$$K(x_i, x_j) = \exp\left(-||x_i - x_j||^2\right), \qquad i, j = \overline{1,n}. \tag{5.3}$$

Таким образом, решающее правило линейного классификатора принимает вид

$$s(x_i, w) = \sum_{j=1}^{n} w_j y_j K(x_i, x_j), \quad i = \overline{1,n}. \tag{5.4}$$

Теперь сформулируем задачу энтропийного оценивания для вычисления параметров w_j, $j = \overline{1,n}$. В данном случае оценивается вектор параметров значительно большей размерности ($n > d$), однако общий подход, изложенный в разделе 3, остается неизменным:

$$H(p) = -\sum_{j=1}^{n}\sum_{l=1}^{k} p_{jl}\ln p_{jl} \to \max_{p} \tag{5.5}$$

при ограничениях

$$\sum_{l=1}^{k} p_{jl} = 1, \quad j = \overline{1,n}, \tag{5.6}$$

$$\mathbb{E}^w\left[M(x_i)\right] > 0, \quad i = \overline{1,n}. \tag{5.7}$$

С учетом определения отступа $M(x_i)$ согласно (2.3) и решающего правила (5.4) ограничение (5.7) примет вид

$$\mathbb{E}^w\left[M(x_i)\right] = \mathbb{E}^w\left[y_i s(x_i, w)\right] =$$

$$= \sum_{j=1}^{n}\sum_{l=1}^{k} y_i y_j K(x_i, x_j) w_{jl} p_{jl} > 0, \quad i = \overline{1,n}. \tag{5.8}$$

Дальнейшая процедура восстановления энтропийно-оптимального распределения вероятностей аналогична разделу 3 с тем отличием, что в данном случае потребуется дополнительно $O(n^2)$ памяти для хранения значений функции ядра, следовательно, увеличиваются необходимые вычислительные ресурсы и время работы алгоритма.

6. Результаты экспериментов

Приведем результаты тестирования алгоритма классификации с использованием энтропийного оценивания (GME) по сравнению с методом опорных векторов как для линейной разделяющей поверхности, так и для гауссовской функции ядра.

6.1. Данные для тестирования

Для экспериментов использовались наборы из открытого репозитория данных для задач машинного обучения лаборатории KEEL (Knowledge Extraction based on Evolutionary Learning) [18]. Все датасеты включают примеры двух классов в различных пропорциях p_1 и p_2, а все признаки являются числовыми. Общее число признаков d и объем коллекции n приведены в табл. 1.

Очевидно, что все приведенные наборы данных имеют свои особенности и свою уникальную структуру, поэтому построение наилучшего классификатора для каждого из них является отдельной задачей. Целью же данных

Таблица 1. Параметры датасетов для тестирования

Название	N	d	p_1, %	p_2, %
hepatitis	80	19	0,162	0,838
appendicitis	106	7	0,802	0,198
sonar	208	60	0,466	0,534
spectfheart	267	44	0,206	0,794
heart	270	13	0,444	0,556
haberman	306	3	0,735	0,265
ionosphere	351	33	0,359	0,641
wdbc	569	30	0,372	0,628
wisconsin	683	9	0,650	0,350
mammographic	830	5	0,514	0,486

экспериментов является не получение наибольшей точности классификации для рассмотренных датасетов, а начальная апробация предложенных алгоритмов энтропийной и рандомизированной классификации и демонстрация их работоспособности.

6.2. Сравнение точности классификаторов

Для первого эксперимента были выбраны датасеты, показывающие высокую точность при классификации линейным алгоритмом. Сравниваются методы оценивания SVM и GME, а также основанный на GME алгоритм рандомизированной классификации (Randomized GME), изложенный в разделе 4. В качестве решения для алгоритма SVM используется базовая реализация из пакета MATLAB R2017a (*fitcsvm*) с значениями гиперпараметров по умолчанию, в то время как алгоритмы GME и RGME реализованы при использовании инструментов оптимизации (Optimization Toolbox).

При реализации энтропийного оценивания использовалось линейное разбиение области допустимых значений параметров решающего правила для выбора допустимых дискретных значений, т.е. весь отрезок от -1 до 1 делится на $k = 5$ равных интервалов, границы которых образуют множество возможных значений дискретной случайной величины. Отметим, что энтропийное оценивание для непрерывных случайных величин является более точным, но и более вычислительно затратным по сравнению с дискретным приближением. Причем количество интервалов k и схема дискретизации также являются гиперпараметрами алгоритма, а их дополнительная настройка с целью получения наилучшей точности классификации в данных экспериментах не проводилась.

Точность классификации вычисляется методом кроссвалидации по пяти блокам (5-fold) и усредняется по результату 100 различных разбиений, что обеспечивает стабильность и воспроизводимость результатов.

Результаты первой серии экспериментов представлены в табл. 2. Как видно, энтропийное оценивание параметров модели линейного классификатора показывает конкурентоспособный результат по сравнению с методом SVM, а для некоторых датасетов, таких как “hepatitis”, “sonar”, “heart” и “haberman”, превосходит SVM по точности классификации. Для остальных примеров либо разница в точности невелика, либо энтропийное оценивание существенно

Таблица 2. Сравнение методов оценивания для линейного классификатора

Название	SVM	GME	RGME	RGME vs GME
hepatitis	83,28 ± 2,73	84,08 ± 2,44	85,42 ± 2,69	•
appendicitis	87,43 ± 1,33	85,89 ± 1,72	86,68 ± 1,55	•
sonar	74,87 ± 1,98	78,92 ± 1,91	79,23 ± 1,28	○
spectfheart	78,76 ± 1,59	68,63 ± 2,82	75,61 ± 3,76	•
heart	83,81 ± 0,95	84,13 ± 1,08	84,56 ± 0,87	•
haberman	72,82 ± 0,57	74,97 ± 2,45	75,44 ± 1,41	•
ionosphere	87,84 ± 1,10	87,35 ± 1,01	87,29 ± 0,74	○
wdbc	97,04 ± 0,51	96,87 ± 0,72	95,94 ± 0,69	○
wisconsin	96,72 ± 0,22	96,69 ± 0,15	96,91 ± 0,15	•
mammographic	82,46 ± 0,69	81,24 ± 0,55	81,61 ± 0,77	•

Таблица 3. Сравнение методов оценивания при использовании функции ядра

Название	SVM	GME	RGME	RGME vs GME
hepatitis	83,75 ± 1,07	86,37 ± 1,55	86,51 ± 1,41	○
appendicitis	87,14 ± 0,91	86,12 ± 1,83	86,40 ± 1,17	•
spectfheart	79,40 ± 0,67	71,35 ± 1,24	71,32 ± 1,42	○
haberman	73,26 ± 0,92	73,65 ± 1,17	73,71 ± 1,24	○
wisconsin	95,34 ± 0,26	94,43 ± 0,59	94,56 ± 0,62	○

уступает, например для данных “appendicitis” и “spectfheart”, что может быть связано с особенной структурой признаков или несимметричностью распределения по классам.

В последнем столбце табл. 2 приведены результаты проверки гипотезы о статистической значимости улучшения точности классификации при использовании семплирования (RGME vs GME), где согласно 95-процентному уровню значимости по критерию Стьюдента незакрашенный индикатор (○) означает выполнение нулевой гипотезы ($H_0 : \mu_{RGME} = \mu_{GME}$), а закрашенный (•) соответствует выполнению альтернативной гипотезы ($H_1 : \mu_{RGME} > \mu_{GME}$). В результате экспериментов для большинства рассмотренных датасетов переход к рандомизированной классификации, основанной на семплировании энтропийно-оптимальных распределений вероятностей, позволяет повысить точность линейного энтропийного классификатора, что особенно заметно при малом объеме обучающей выборки до $250 - 300$ объектов. Однако для выборок более 300 объектов энтропийное оценивание становится не столь эффективным и уступает по точности методу SVM. Кроме того, независимо от объема выборки энтропийное оценивание значительно уступает методу SVM по времени исполнения.

Следующая серия экспериментов посвящена сравнению методов оценивания при использовании в классификаторе функции гауссовского ядра (5.3). Для сравнения выбраны наборы данных из табл. 1, показывающие высокую точность при классификации с использованием функции ядра. Результаты сравнения представлены в табл. 3, где все обозначения аналогичны описанным выше.

Как видно из данных табл. 3, эффективность энтропийного оценивания также распространяется и на ядровые модификации линейного классифика-

тора. Однако в отличие от линейной разделяющей поверхности в этом случае для большинства примеров использование ансамбля классификаторов не приводит к статистически значимому повышению точности, что может объясняться высокой размерностью задачи оптимизации (5.5) по сравнению с задачей (3.1).

Кроме того, в данных сериях экспериментов не проводилась настройка гиперпараметров алгоритмов GME и RGME, таких как количество значений для дискретных случайных величин (k в выражениях (3.1) и далее) и объем ансамбля при семплировании (m в выражении (4.2)). Поэтому подбор оптимальных значений гиперпараметров для каждого набора данных позволит дополнительно повысить точность классификации.

6.3. Влияние объема обучающей выборки

Приведем результаты исследования зависимости точности классификаторов SVM, GME и RGME от объема данных в обучающей коллекции. Для этого эксперимента используется метод кросс-валидации с заданным процентом объектов для тестирования (holdout). Например, при значении $h = 0{,}2$ для обучения будет использоваться 80 % выборки, а остальные 20 % — для тестирования, что соответствует одной из пяти итераций по методу 5-fold. Таким образом, при увеличении значения h от 0 до 1 для обучения модели будет использоваться все меньше объектов.

Зависимость обычного линейного классификатора протестирована на примере данных “appendicitis”, а классификатора с использованием функции ядра — на примере данных “hepatitis”. Результаты сравнения, усредненные по 1000 различных разбиений, представлены в виде графиков на рис. 1 и 2 соответственно.

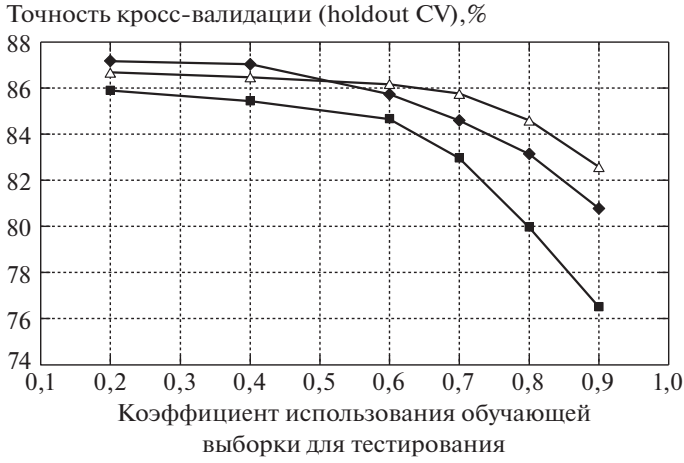

Рис. 1. Зависимость точности линейного классификатора от объема обучающей выборки на примере датасета “appendicitis” (ромбами отмечены точки графика для метода SVM, квадратами и треугольниками — для GME и RGME соответственно).

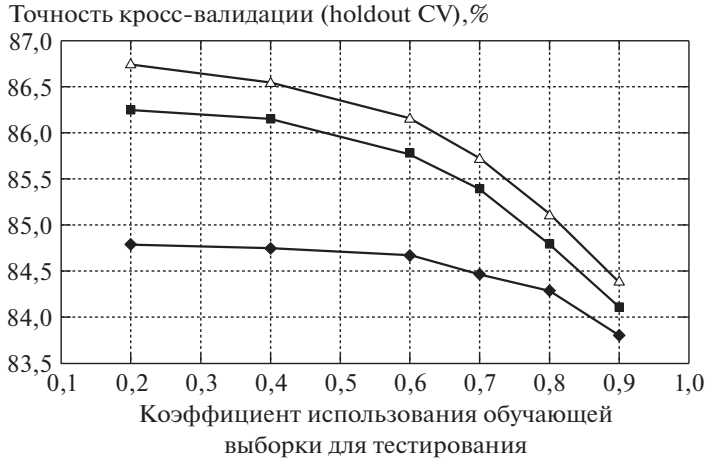

Рис. 2. Зависимость точности классификатора с гауссовским ядром от объема обучающей выборки на примере датасета "hepatitis" (ромбами отмечены точки графика для метода SVM, квадратами и треугольниками — для GME и RGME соответственно).

Результаты исследования показывают, что точность линейного классификатора (рис. 1), основанного на энтропийном оценивании параметров, подвержена влиянию объема обучающей выборки и подобно методу SVM падает при сокращении данных для обучения. С другой стороны, использование семплирования совместно с энтропийным оцениванием в алгоритме RGME позволяет уменьшить падение точности более чем в два раза для рассмотренного датасета. Так, согласно графику на рис. 1 при использовании 40 % и менее данных для обучения рандомизированная классификация позволяет получить точность, превосходящую точность алгоритмов SVM и GME, несмотря на то, что изначально наилучшую точность показывал алгоритм SVM.

С другой стороны, для ядровых модификаций (рис. 1) падение точности классификации при уменьшении объема обучающей выборки оказывается менее существенным, что связано, в первую очередь, с размерностью задачи оптимизации. Если в обычном линейном классификаторе необходимо для d-мерного признакового пространства восстановить d параметров независимо от объема обучающей выборки, то в ядровой модификации происходит переход в n-мерное пространство, где n соответствует объему выборки для обучения. Падение же точности для всех рассматриваемых алгоритмов на рис. 2 меньше, чем на рис. 1, однако именно алгоритм рандомизированной классификации оказывается, хотя и незначительно, но точнее остальных при малом объеме данных для обучения.

7. Заключение

В статье продемонстрирован новый подход к построению линейного классификатора, базирующийся на энтропийном оценивании его параметров. Под энтропийным оцениванием понимается восстановление распределений веро-

ятности для параметров модели, обладающих наибольшей информационной энтропией и согласующихся с обучающей выборкой в терминах балансовых ограничений. В свою очередь, восстановленные таким образом вероятности могут использоваться как для получения точечной оценки параметров модели, так и для получения ансамбля классификаторов посредством семплирования.

Представленный алгоритм классификации был апробирован на примере нескольких датасетов из открытых источников. Результаты экспериментов показывают высокую, сравнимую с методом опорных векторов, точность классификации. Дополнительно были исследованы статистическая значимость повышения точности при переходе к рандомизированной классификации и влияние объема обучающей выборки данных. В результате исследований была продемонстрирована эффективность энтропийного оценивания и рандомизированной классификации, особенно при малых объемах обучающих выборок.

СПИСОК ЛИТЕРАТУРЫ

1. *Bishop C.* Pattern Recognition and Machine Learning (Information Science and Statistics). Springer, 2006.
2. *Breiman Leo, Friedman J.H., Olshen R.A., Stone C.J.* Classification and Regression Trees. Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software, 1984.
3. *Domingos P, Pazzani M.* On the Optimality of the Simple Bayesian Classifier under Zero-One Loss // Mach. Learn. 1997. No. 29. P. 103–130.
4. *Hosmer D.W., Lemeshow S.* Applied Logistic Regression, 2nd ed. N.Y.: Chichester, Wiley, 2002.
5. *Belur V. Dasarathy* (ed.), Nearest Neighbor (NN) Norms: NN Pattern Classification Techniques. 1991.
6. *Cristianini N., Shawe-Taylor J.* An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods. Cambridge University Press, 2000.
7. *Bühlmann P., Hothorn T.* Boosting Algorithms: Regularization, Prediction and Model Fiting // Stat. Sci. 2007. P. 477–505.
8. *Cover T.M., Thomas J.A.* Elements of Information Theory. N.Y.: John Wiley and Sons Ltd, 1991.
9. *Abellán J., Castellano J.G.* Improving the Naive Bayes Classifier via a Quick Variable Selection Method Using Maximum of Entropy // Entropy. 2017. V. 19. Iss. 6. No. 247.
10. *Phillips Steven J.* A Brief Tutorial on Maxent. Network of Conservation Educators and Practitioners, Center for Biodiversity and Conservation, American Museum of Natural History // Lessons in Conservation. 2009. V. 3. P. 108–135.
11. *Yu Hsiang-Fu, Huang Fang-Lan, Lin Chih-Jen.* Dual Coordinate Descent Methods for Logistic Regression and Maximum Entropy Models // Mach. Lear. 2011. V. 85. P. 41–75.
12. *Golan Amos, Judge George G., Miller Douglas.* Maximum Entropy Econometrics: Robust Estimation with Limited Data. Chichester, U.K.: John Wiley and Sons Ltd., 1996.
13. *Eruygur H. Ozan.* Generalized Maximum Entropy (GME) Estimator: Formulation and a Monte Carlo Study // VII National Sympos. on Econometrics and Statistics, Istanbul, Turkey. 2005. May 26–27.

14. *Popkov Yu.S., Dubnov Yu.A., Popkov A.Yu.* Randomized Machine Learning: Statement, Solution, Applications // Proc. 2016 IEEE 8-th Int. Conf. on Intelligent Systems (IS16). September 4–6, 2016. Hotel Hemus, Sofia, Bulgaria. P. 27–39.
15. *Langford J.* Tutorial on Practical Prediction Theory for Classification // J. Mach. Learn. Research. 2005. V. 6. P. 273–306.
16. *Popkov Yuri S., Volkovich Zeev, Dubnov Yuri A., Avros Renata, Ravve Elena.* Entropy '2'-Soft Classification of Objects // Entropy. 2017. V. 19. Iss. 4. No. 178.
17. *Fawcett T.* An Introduction to ROC Analysis // Pattern Recogn. Lett. 2006. No. 27. P. 861–874.
18. *Alcalá-Fdez J., Fernandez A., Luengo J., Derrac J., Garcia S., Sánchez L., Herrera F.* KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework // J. Multiple-Valued Logic and Soft Computing. 2011. V. 17. No. 2–3. P. 255–287.

Статья представлена к публикации членом редколлегии Б.М. Миллером.

Поступила в редакцию 19.07.2018

После доработки 06.11.2018

Принята к публикации 08.11.2018