Оценка информативности признаков в задачах распознавания символов с помощью разреживающих эластичных сетей*

И. А. Дубнов, А. Б. Мерков

Аннотация. В статье представлен метод максимизации обоснованности для автоматического подбора параметров регуляризации эластичных сетей [1], т. е. обобщенных линейных вероятностных моделей, в которых априорное распределение является промежуточным между лапласовым и гауссовым. Этот метод использован при оценке полезности разных наборов признаков для задачи распознавания рукописных символов.
Ключевые слова: *распознавание символов; признаковая модель; эластичная сеть.*

Введение

В задачах распознавания символов одним из важнейших аспектов является подбор признаков для обучения. От выбора признаков зависят как временные характеристики работы распознавателя, так и его точность, поэтому оптимальным набором является наименьший набор признаков, дающий, по возможности, наибольшую точность. Однако поиск такого набора далеко не тривиален и зависит не только от конкретной решаемой задачи, но и от самих данных, предоставленных для обучения.

На сегодняшний день существует отдельное направление теории статистического обучения, посвященное проблеме отбора признаков (Feature selection, [2]). Главной предпосылкой использования методов отбора признаков является наличие у обучающих данных избыточных признаков, которые могут быть удалены. Процесс отбора признаков может являться как отдельной частью предварительной обработки данных (энтропийный и корреляционный анализ), так и частью непосредственно процесса обучения (например, в случаях обучения с l_1-регуляризацией, таких как sparse regression, LASSO, и l_1-SVM) [3].

Обучение с l_1-регуляризацией замечательно тем, что, как правило, приводит к разреженной модели, не зависящей от многих признаков. Обучение с наиболее распространенной l_2-регуляризацией разреженность модели не обеспечивает, но обеспечивает более точное предсказание. В работе [1] был предложен промежуточный способ обучения распознавателей $F(x,w)$, параметризованных вещественным вектором w, минимизацией регуляризованной суммы потерь $J(F(x,w)),y)$ на обучающем наборе $\{(x_i,y_i),i=1,...,N\}$

$$\sum_{i=1}^{N} J(F(x_i,w),y_i)+\lambda\,|\,w\,|+\frac{\mu}{2}\|\,w\,\|^2\to\min_w, \qquad (1)$$

где $\|\cdot\|$ и $|\cdot|$ — l_2 и l_1-нормы соответственно, а λ и μ — неотрицательные параметры регуляризации. Экспериментально показано, что, варьируя параметры λ и μ, можно балансировать между разреженностью и качеством предсказания обученной модели. Это двухпараметрическое семейство методов обучения (1) получило название «*эластичная сеть*».

В настоящей статье эластичные сети используются для оценки полезности различных дополнительных наборов признаков в задаче распознавания символов. Обучая эластичную сеть на разных наборах признаков, являющихся пиксельными значениями изображения и некоторыми функциями от них, мы увидим, добавление каких признаков улучшает распознавание, а какие признаки избыточны.

В первой главе приведены математическая модель и постановка задачи обучения эластичной сети. Во второй главе описан алгоритм обучения и применение метода максимизации обоснованности для подбора параметров эластичной сети. В третьей главе описаны эксперименты по обучению эластичной сети для распознавания символов на разных наборах признаков. В четвертой главе обсуждаются результаты экспериментов.

* Работа выполнена при частичной поддержке гранта РФФИ № 13–07–12178.

1. Математическая модель

Пусть имеется задача классификации $x \mapsto y$, где $x \in \mathbb{R}^d$ — многомерный вектор признаков, а ответ y — один из q классов, и набор $\mathbf{T} = \{\mathbf{X}, \mathbf{Y}\}$ из N пар (x_i, y_i). Вместо детерминированной классификации используется классическая многомерная логистическая регрессия, которая основана на построении вероятностных моделей вида

$$p(y \mid x, \vec{w}) = \frac{e^{\vec{w}^y \vec{x}}}{\sum_{l=1}^{q} e^{\vec{w}^l \vec{x}}}, \qquad (2)$$

где $\vec{x} = (1, x) \in \mathbb{R}^{d+1}$ — расширенный вектор признаков, а матрица параметров модели $\vec{w}$ состоит из q $(d+1)$-мерных векторов-строк $\vec{w}^l = (w_0^l, w_1^l, ..., w_d^l)$. Подматрицу $\vec{w}$ без нулевого столбца $w_0 = (w_0^l, w_0^2, ..., w_0^q)$, т. е. без свободных членов, обозначим w. Нулевой столбец w_0 выделен отдельно, потому что обычно априорное распределение $p_0(\vec{w})$ берут не зависящим от w_0 (т. е. $p_0(w)$), чтобы получить несмещенные оценки вероятностей $p(y \mid x)$. В этом пространстве моделей ищется модель с максимальной апостериорной вероятностью при условии априорного распределения $p_o(\vec{w})$ и обучающего набора $\mathbf{T}$. Поскольку

$$p(\vec{w} \mid \mathbf{T}) = \frac{p_0(\vec{w}) p(\mathbf{T} \mid \vec{w})}{p(\mathbf{T})} = \frac{p_0(\vec{w}) p(\mathbf{Y} \mid \mathbf{X}, \vec{w})}{p(\mathbf{Y} \mid \mathbf{X})}, \qquad (3)$$

и знаменатель от модели $\vec{w}$ не зависит, максимизация апостериорной вероятности модели равносильна максимизации числителя или, что удобнее, его логарифма

$$\ln(p_0(\vec{w}) p(\mathbf{Y} \mid \mathbf{X}, \vec{w})) = \ln(p_0(\vec{w})) + \\ + \sum_{i=1}^{N} \ln p(y_i \mid x_i, \vec{w}) \to \max_{\vec{w}}. \qquad (4)$$

Второе слагаемое в (4) — это логарифм правдоподобия модели $L(\vec{w}; \mathbf{T})$, а первое слагаемое зависит от выбора априорного распределения, в качестве которого в простейших случаях берут сферическое гауссово распределение или лапласово (и получают оптимизационную задачу с l_2- или l_1-регуляризацией соответственно). Априорное распределение вероятностей $p_0(\vec{w})$ для эластичных сетей имеет вид:

$$p_0(\vec{w}) = \frac{1}{Z(\lambda, \mu)} e^{-\lambda|w| - \mu \frac{\|w\|^2}{2}}, \qquad (5)$$

где нормировочный множитель

$$Z(\lambda, \mu) = \int e^{-\lambda|w| - \mu \frac{\|w\|^2}{2}} dw = \left(\int_{-\infty}^{\infty} e^{-\lambda|t| - \mu \frac{t^2}{2}} dt \right)^{qd} = \\ = \left(\frac{2 e^{\frac{\lambda^2}{2\mu}}}{\sqrt{\mu}} \int_{\frac{\lambda}{\sqrt{\mu}}}^{\infty} e^{-\frac{\tau^2}{2}} d\tau \right)^{qd} = \left(\frac{2 e^{\frac{\lambda^2}{2\mu}}}{\sqrt{\mu}} \sqrt{2\pi} \Phi\left(-\frac{\lambda}{\sqrt{\mu}} \right) \right)^{qd}$$

(напоминаем, что w пробегает qd-мерное пространство), а через $\Phi(\cdot)$ обозначена кумулятивная функция стандартного одномерного гауссова распределения

$$\Phi(t) = \int_{-\infty}^{t} \frac{1}{\sqrt{2\pi}} e^{-\frac{\tau^2}{2}} d\tau.$$

Для удобства вычислений в дальнейшем вместо $\Phi(\cdot)$ используется функция

$$\Psi(t) = e^{\frac{t^2}{2}} \int_{-\infty}^{t} e^{-\frac{\tau^2}{2}} d\tau = \sqrt{2\pi} e^{\frac{t^2}{2}} \Phi(t).$$

В частности, нормировочный множитель $Z(\lambda, \mu)$ выражается через $\Psi(t)$:

$$Z(\lambda, \mu) = \left(\frac{2}{\sqrt{\mu}} \Psi\left(-\frac{\lambda}{\sqrt{\mu}} \right) \right)^{qd}. \qquad (6)$$

Таким образом, подставляя (5) в (4), получаем оптимизационную задачу для эластичной сети:

$$-\sum_{i=1}^{N} \ln p(y_i \mid x_i . \vec{w}) + \lambda \mid w \mid + \frac{\mu}{2} \| w \|^2 \to \min_{\vec{w}}. \qquad (7)$$

Выписанное выше априорное распределение (5) и соответствующие ему регуляризационные слагаемые при обучении (7) симметричны по всем признакам. Однако сами признаки могут быть неравноправны по природе своего происхождения. Для устранения этого неравноправия разобьем их на K групп признаков одинаковой природы: например, все пиксели изображения имеют одинаковую природу и должны быть в одной группе признаков, а некоторые вычисляемые по ним признаки — в других группах. Т. е. зафиксируем разбиение множества индексов

$$\{1, ..., d\} = \sqcup_{k=1}^{K} D_k \qquad (8)$$

на подмножества D_k мощностей $d_k = \#D_k$, и для каждой группы индексов поддерживаются свои параметры регуляризации. Тогда вместо эластичной сети (7) получается задача

$$-\sum_{i=1}^{N} \ln p(y_i \mid x_i . \vec{w}) + \\ + \sum_{k=1}^{K} \left(\lambda_k \sum_{j \in D_k} \mid w_j \mid + \frac{\mu_k}{2} \sum_{j \in D_k} \| w_j \|^2 \right) \to \min_{\vec{w}},$$

которая для линейной логистической регрессии (2) имеет вид

$$-\sum_{i=1}^{N}\left(w^{y_i}\vec{x}_i-\ln\sum_{l=1}^{q}e^{w^l\vec{x}_i}\right)+$$
$$+\sum_{k=1}^{K}\left(\lambda_k\sum_{j\in D_k}|w_j|+\frac{\mu_k}{2}\sum_{j\in D_k}\|w_j\|^2\right)\to\min_{\vec{w}}. \quad (9)$$

Легко проверить, что минимизационная задача (9) выпукла по $\vec{w}$.

2. Алгоритм обучения и максимизация обоснованности

2.1. Негладкая выпуклая оптимизация

Так как в состав минимизируемых функций (7) и (9) входят недифференцируемые слагаемые $|w|$ или $|w_j|$, стандартные градиентные методы неприменимы. Поэтому для обучения эластичной сети используется ускоренный алгоритм оптимизации негладких выпуклых функций, предложенный Ю. Е. Нестеровым в [4] для минимизации суммы гладкой функции и достаточно простой функции, каковыми, в частности, являются эластичные сети (7) и (9). Алгоритм Нестерова замечателен тем, что дает наибыстрейшую сходимость к решению среди всех методов негладкой оптимизации, известных на сегодняшний день [5].

Полное описание алгоритма и оценки скорости сходимости можно найти в работе [4] (Accelerated method, p. 15). В ней также показано, что если минимизируемая функция μ-выпукла (сильно выпукла с параметром $\mu>0$), то алгоритм сходится быстрее. Нетрудно заметить, что минимизируемая функция в задаче (9) была бы сильно выпуклой, если бы регуляризация по L_2-норме применялась и к нулевому столбцу параметров модели w_0. Рассмотрим следующую модификацию задачи (9):

1. Нулевой столбец $\hat{w}_0$ оценивается по формуле

$$\hat{w}_0^l=\ln\frac{n_l}{N} \quad \text{для } l=1,\ldots,q, \quad (10)$$

где n_l — количество символов класса l в обучающей выборке. Оценка $\hat{w}_0$ является решением минимизационной задачи

$$-\sum_{i=1}^{N}\ln p(y_i\,|\,w_0)=-\sum_{i=1}^{N}\ln\frac{e^{w_0^{y_i}}}{\sum_{l=1}^{q}e^{w_0^l}}\to\min_{w_0},$$

т. е. максимизации правдоподобия логистической регрессии, не использующей признаков.

2. Вместо решения задачи (9) решается задача

$$-\sum_{i=1}^{N}\left(\vec{w}^{y_i}\vec{x}_i-\ln\sum_{l=1}^{q}e^{\vec{w}^l\vec{x}_i}\right)+\frac{\mu_0}{2}\|w_0-\hat{w}_0\|^2+$$
$$+\sum_{k=1}^{K}\left(\lambda_k\sum_{j\in D_k}|w_j|+\frac{\mu_k}{2}\sum_{j\in D_k}\|w_j\|^2\right)\to\min_{\vec{w}} \quad (11)$$

с параметром $\mu_0>0$.

В постановке (11) задача минимизации для эластичной сети гарантирует сильную выпуклость минимизируемой функции с параметром $\mu=\min\limits_{k=0,1,..,K}\mu_k$.

2.2. Максимизация обоснованности

Для обучения эластичной сети (7) необходимы разумные значения параметров регуляризации λ и μ. В простейшем случае одного или двух гиперпараметров реализуются сеточные алгоритмы поиска (grid search, например для SVM, [2]). Однако, как описано ранее, для эластичной сети в общем случае необходимо $2K+1$ параметров регуляризации. В этом случае разумным способом подбора параметров является метод максимизация обоснованности, применение которого широко известно для оценки параметра регуляризации гребневой регрессии (см. [7]).

Пусть априорное распределение параметров модели зависит от параметров регуляризации λ и μ. Тогда формула апостериорной вероятности (3) с явным указанием этих параметров примет вид

$$p(\vec{w}\,|\,\mathbf{T},\lambda,\mu)=\frac{p(\mathbf{T},\vec{w})p_0(w\,|\,\lambda,\mu)}{p(\mathbf{T}\,|\,\lambda,\mu)}=$$
$$=\frac{p(\mathbf{Y}\,|\,\mathbf{X},\vec{w})p_0(w\,|\,\lambda,\mu)}{p(\mathbf{Y}\,|\,\mathbf{X},\lambda,\mu)}=$$
$$=\frac{L(\vec{w};\mathbf{T})p_0(w\,|\,\lambda,\mu)}{\int L(\vec{w};\mathbf{T})p_0(w\,|\,\lambda,\mu)d\vec{w}}=$$
$$=\frac{L(\vec{w};\mathbf{T})p_0(w\,|\,\lambda,\mu)}{E(\lambda,\mu;\mathbf{T})}.$$

При максимизации апостериорной вероятности (3) по $\vec{w}$ знаменатель в этой формуле игнорировался, как не зависящий от $\vec{w}$. Однако от λ и μ он зависит. Этот знаменатель $E(\lambda,\mu;\mathbf{T})$ называется обоснованностью (evidence) параметров λ и μ относительно обучающего набора $\mathbf{T}$. Несмотря на специальный термин, он является обычным правдоподобием, аналогичным $L(\vec{w};\mathbf{T})$ в формуле (4), только правдоподобием не отдельной модели, а целого вероятностного пространства моделей.

При априорном распределении (5) обоснованность

$$E(\lambda,\mu;\mathbf{T}) = \int L(\vec{w};\mathbf{T}) p_0(w\,|\,\lambda,\mu) d\vec{w} = \\ = \frac{1}{Z(\lambda,\mu)} \int e^{\ln L(\vec{w};\mathbf{T}) - \lambda|w| - \mu\frac{\|w\|^2}{2}} d\vec{w},$$

и ее максимизация эквивалентна минимизации

$$-\ln E(\lambda,\mu;\mathbf{T}) = qd \ln\left(\frac{2}{\sqrt{\mu}}\Psi\left(-\frac{\lambda}{\sqrt{\mu}}\right)\right) - \\ -\ln\int e^{\ln L(\vec{w};\mathbf{T}) - \lambda|w| - \mu\frac{\|w\|^2}{2}} d\vec{w} \to \min_{\lambda,\mu}. \qquad (12)$$

Здесь для нормировочного множителя $Z(\lambda,\mu)$ использована формула (6). Градиент (12) имеет вид

$$\nabla_\lambda(-\ln E(\lambda,\mu;\mathbf{T})) = -\frac{qd}{\lambda}\left(\frac{\frac{\lambda}{\sqrt{\mu}}}{\Psi\left(-\frac{\lambda}{\sqrt{\mu}}\right)} - \frac{\lambda^2}{\mu}\right) + \\ +\mathbf{E}_{\lambda,\mu}[|\,w\,|], \qquad (13)$$

$$\nabla_\mu(-\ln E(\lambda,\mu;\mathbf{T})) = -\frac{qd}{2\mu}\left(1 - \frac{\frac{\lambda}{\sqrt{\mu}}}{\Psi\left(-\frac{\lambda}{\sqrt{\mu}}\right)} + \frac{\lambda^2}{\mu}\right) + \\ +\frac{1}{2}\mathbf{E}_{\lambda,\mu}\left[\|\,w\,\|^2\right], \qquad (14)$$

где для любой функции $f(w)$ через $\mathbf{E}_{\lambda,\mu}[f]$ обозначено ожидание $f(w)$ по апостериорному распределению, пропорциональному $L(\vec{w};\mathbf{T})p_0(w\,|\,\lambda,\mu)$:

$$\mathbf{E}_{\lambda,\mu}[f] = \frac{\int f(w) e^{\ln L(\vec{w};\mathbf{T}) - \lambda|w| - \mu\frac{\|w\|^2}{2}} d\vec{w}}{\int e^{\ln L(\vec{w};\mathbf{T}) - \lambda|w| - \mu\frac{\|w\|^2}{2}} d\vec{w}}. \qquad (15)$$

Вместо честного градиентного спуска по λ и μ для максимизации обоснованности используется итеративное преобразование в пространстве пар (λ,μ)

$$\lambda \leftarrow \frac{qd\left(\frac{\frac{\lambda}{\sqrt{\mu}}}{\Psi\left(-\frac{\lambda}{\sqrt{\mu}}\right)} - \frac{\lambda^2}{\mu}\right)}{\mathbf{E}_{\lambda,\mu}[|\,w\,|]}, \\ \mu \leftarrow \frac{qd\left(1 - \frac{\frac{\lambda}{\sqrt{\mu}}}{\Psi\left(-\frac{\lambda}{\sqrt{\mu}}\right)} + \frac{\lambda^2}{\mu}\right)}{\mathbf{E}_{\lambda,\mu}[\|w\|^2]}. \qquad (16)$$

Из (13) и (14) следует, что точки максимума обоснованности являются неподвижными точками преобразования (16). Сходимость преобразования (16), увы, не гарантирована, но в экспериментах оно почти всегда позволяет улучшить распознавание.

Для задачи (11) аналог преобразования (16) имеет вид

$$\lambda_k \leftarrow \frac{qd_k\left(\frac{\frac{\lambda_k}{\sqrt{\mu_k}}}{\Psi\left(-\frac{\lambda_k}{\sqrt{\mu_k}}\right)} - \frac{\lambda_k^2}{\mu_k}\right)}{\sum_{j\in D_k}\sum_{l=1}^{q}\mathbf{E}_{\lambda,\mu}[|\,w_j^l\,|]}, \\ \mu_k \leftarrow \frac{qd_k\left(1 - \frac{\frac{\lambda_k}{\sqrt{\mu_k}}}{\Psi\left(-\frac{\lambda_k}{\sqrt{\mu_k}}\right)} + \frac{\lambda_k^2}{\mu_k}\right)}{\sum_{j\in D_k}\sum_{l=1}^{q}\mathbf{E}_{\lambda,\mu}[\|w_j^l\|^2]}, \qquad (17)$$

при $k = 1,\ldots,K$ и

$$\mu_0 \leftarrow \frac{q}{\sum_{l=1}^{q}\mathbf{E}_{\lambda,\mu}[\|\,w_0^l - \hat{w}_0^l\,\|^2]}. \qquad (18)$$

Для приближенного вычисления ожиданий $\mathbf{E}_{\lambda,\mu}[|\,w_j^l\,|]$, $\mathbf{E}_{\lambda,\mu}[\|\,w_j^l\,\|^2]$ и $\mathbf{E}_{\lambda,\mu}[\|\,w_0^l - \hat{w}_0^l\,\|^2]$ по апостериорному распределению применяется диагональная аппроксимация Лапласа [8] в точке $w^* = w^*(\lambda,\mu)$, являющейся результатом обучения эластичной сети (11).

2.3. Критерий остановки

Для контроля за процессом обучения и отслеживания эффективности итеративного преобразования (λ,μ) набор $\mathbf{T}$ разбивается на N_{train}-элементный обучающий набор $\mathbf{T}_{\text{train}}$, использующийся непосредственно при решении минимизационной задачи (11), и N_{val}-элементный остановочный набор $\mathbf{T}_{\text{val}}$, который применяется для контроля за переобучением.

Разбиение данных на обучающий и остановочный наборы используется в двух местах: для остановки обучения эластичных сетей алгоритмом Нестерова и для остановки параметров (λ,μ) эластичных сетей.

Целью обучения эластичной сети является увеличение правдоподобия модели на обучающем наборе. Однако в процессе достижения минимума в задаче (11) может произойти переобучение, т. е.

после некоторого момента может начать ухудшаться распознавание на данных, отличных от обучающего набора.

В качестве критерия остановки при обучении каждой эластичной сети взято появление признаков переобучения, а именно уменьшение правдоподобия модели на остановочном наборе $L(\vec{w}, \mathbf{T}_{\text{val}})$ в течение некоторого фиксированного количества итераций H_1 (порядка 30).

Аналогично, итеративная максимизация обоснованности параметров (λ, μ) эластичных сетей прекращается при уменьшении правдоподобия обученной модели на остановочном наборе в течении H_2 (порядка 5) итераций.

3. Поставленные эксперименты

Целью экспериментов являются изучение разреживающих свойств эластичной сети с автоматической настройкой параметров регуляризации с помощью максимизации обоснованности и изучение полезности (информативности) различных наборов признаков для распознавания символов.

Эксперименты проводятся на наборе рукописных цифр MNIST ([9]), представленных в виде монохромных растровых изображений размера $28 \times 28 = 784$ пикселей, принадлежащих одному из $q = 10$ классов. Данные MNIST содержат $N = 60000$ пар (изображение, ответ) для обучения и $M = 10000$ пар для тестирования. Из обучающей выборки выделено 15% для остановочном набора. Таким образом, $N_{\text{train}} = 51000$, $N_{\text{val}} = 9000$.

Значения интенсивностей пикселей на изображении в дальнейшем называются базовыми признаками, а вычисленные по ним функции, например квадраты градиентов, — дополнительными признаками. После вычисления дополнительных признаков все признаки, в том числе и базовые, нормализуются таким образом, чтобы среднее значение каждого признака по всем примерам из обучающей выборки равнялось 0 , а дисперсия — 1 .

Тестирование каждого набора признаков представляет собой серию из 20 экспериментов двухуровневого обучения: обучение эластичной сети (11) алгоритмом Нестерова при фиксированных параметрах регуляризации (λ, μ) и изменении параметров (λ, μ) преобразованием (17)–(18). Базовые признаки и каждая из групп дополнительных признаков являются отдельными группами признаков в разбиении (8), например, в эксперименте с базовыми и одной группой дополнительных признаков имеем $K = 2$. Каждый эксперимент начинается с обучения при значениях $\lambda_k^0 = 1$ и $\mu_k^0 = 1$ для всех k, т. е. признаки после нормализации считаются равноправными.

В экспериментах используются следующие группы дополнительных признаков (в скобках указано количество дополнительных признаков в группе):

1. Дублированные базовые признаки, возмущенные аддитивным гауссовым шумом с ожиданием 0 и дисперсией $0{,}04$ (784).

2. Квадраты вертикальных и горизонтальных компонент градиента базовых признаков, вычисленные в каждом пикселе изображения ($784 + 784 = 1568$).

3. Амплитуды и фазы коэффициентов двумерного дискретного преобразования Фурье [10] от исходного изображения ($784 + 784 = 1568$).

4. Проекционные гистограммы [10]: количество ненулевых пикселей в каждой строчке и каждом столбце, а также позиции первого и последнего ненулевых пикселя в каждой строчке и в каждом столбце изображения ($28 + 28 + 28 \cdot 2 + 28 \cdot 2 = 168$).

5. Зонирование (zoning, [10]) — вектор длиной $n \times m$, значения которого соответствуют средним значениям интенсивностей пикселей для каждой из зон сетки $n \times m$, накладываемой на изображение ($n = 7, m = 7$).

6. Матрица угловых точек изображения — матрица, в которой для каждого пикселя исходного изображения ставится в соответствие значение вероятности того, что данный пиксель является угловой точкой на изображении. Матрица угловых точек вычислена функцией MATLAB *cornermetric* [11], которая использует для нахождения угловых точек алгоритм Харриса [12] (784).

7. Бинарная матрица граничных точек изображения — матрица, в которой каждому пикселю исходного изображения ставится в соответствие значение: 1, если данный пиксель является граничной точкой, и 0 иначе. Матрица граничных точек вычислена функцией MATLAB *edge* [11], которая используется для нахождения граничных точек алгоритмом Кэнни [13] (784).

8. Энтропия значений интенсивности, вычисленная для каждого пикселя по соседним 9×9 пикселям функцией MATLAB *entropyfilt* [11] (784).

9. Среднеквадратичное отклонение значений интенсивности, вычисленное для каждого пикселя по соседним 9×9 пикселям функцией MATLAB *stdfilt* [11] (784).

3.1. Максимизация точности

Результаты обучения на объединении базовых и одной группы дополнительных признаков представлены в таблице 1, а результаты обучения на объеди-

Таблица 1

Результат обучения эластичной сети по совокупности базовых и дополнительных признаков

№	Кол-во призн.	Разреж. модели (баз/доп)	Разреж. призн. (баз/доп)	Ост-сь призн.	Кросс-энтропия	Точность (%)
0	784	0,50	0,15	670	0,2694	92,69
1	784 + 784	0,63 / 0,92	0,16 / 0,55	1009	0,2696	92,65
2	784 + 1568	0,31 / 0,26	0,13 / 0,09	2108	0,0927	97,10
3	784 + 1568	0,28 / 0,43	0,13 / 0,00	2249	0,1499	95,75
4	784 + 168	0,73 / 0,32	0,23 / 0,03	769	0,1282	96,11
5	784 + 49	0,61 / 0,98	0,16 / 0,85	665	0,2705	92,67
6	784 + 784	0,29 / 0,25	0,12 / 0,03	1446	0,1646	95,31
7	784 + 784	0,57 / 0,46	0,16 / 0,20	1284	0,2056	94,17
8	784 + 784	0,20 / 0,09	0,10 / 0,00	1488	0,2334	93,61
9	784 + 784	0,27 / 0,14	0,12 / 0,00	1472	0,1572	95,81

Таблица 2

Результат обучения эластичной сети при объединении групп дополнительных признаков 0–9

№	Кол-во призн.	Разреж. модели	Разреж. призн.	Ост-сь призн.	Кросс-энтропия	Точность (%)
0	784	0,50	0,15	670	0,2694	92,69
0, 2	2352	0,28	0,10	2108	0,0927	97,10
0, 2, 4	2520	0,34	0,12	2224	0,0788	97,48
0, 2, 3, 4	4088	0,55	0,11	3644	0,0690	97,68
0, 2, 3, 4, 9	4872	0,56	0,13	4239	0,0624	97,98
0, 2, 3, 4, 6, 9	5656	0,58	0,16	5089	0,0570	98,22
0, 2, 3, 4, 6, 9, 7	6440	0,61	0,14	5538	0,0577	98,20

нении базовых и нескольких групп дополнительных признаков — в таблице 2. Первая строка таблиц — одна и та же — относится к обучению только на базовых признаках.

Значения чисел в столбцах таблиц:

- Разреженность модели — это доля нулевых элементов матрицы параметров $\vec{w}$.
- Разреженность признаков — это доля полностью нулевых столбцов матрицы параметров $\vec{w}$.
- Осталось признаков — среднее количество признаков, используемых моделью для распознавания после разреживания.
- Кросс-энтропия — это минус логарифм правдоподобия модели на тестовом наборе, поделенный на количество его элементов, что равно перекрестной энтропии условных распределений реальных ответов в тестовом наборе и ответов, прогнозируемых моделью (чтобы получить энтропию в битах, нужно поделить на $\ln 2 \approx 0,7$).
- Точность — это доля правильно распознанных моделью цифр из тестового набора в процентах.

Для наглядности разреженность модели и разреженность признаков в некоторых случаях рассчитаны отдельно для базовых признаков и отдельно для дополнительных, а полученные значения указаны в соответствующих столбцах таблиц через дробь (баз/доп).

Эксперименты с максимизацией обоснованности показывают, что из-за досрочной остановки по критерию переобучения параметры λ и μ не достигают неподвижных точек преобразования (17–18). Так как конечные значения параметров λ и μ зависят от обучения эластичной сети на каждом шаге преобразования, они зависят и от их начальных значений. Итерации по λ_k и μ_k, начинающиеся с небольших значений, позволяют достичь максимальной точности при слабой разреженности модели.

3.2. Максимизация разреженности

Если $\lambda_k > \lambda_k^{\max} = \max\limits_{j \in D_k, l=1,..,q} \left| \dfrac{\partial \ln L(0; \mathbf{T}_{\text{train}})}{\partial w_j^l} \right|$, то решением оптимизационной задачи обучения (11) заведомо будет прогноз, не зависящий от признаков: $w = 0$, а распознавание при этом никакое.

Таблица 3

Результат обучения эластичной сети по совокупности базовых и дополнительных признаков с максимальным разреживанием

№	Кол-во призн.	Разреж. модели (баз/доп)	Разреж. призн. (баз/доп)	Ост-сь призн.	Кросс-энтропия	Точность (%)
0	784	0,60	0,16	659	0,2709	92,63
2	784 + 1568	0,94 / 0,83	0,67 / 0,34	1299	0,1189	96,66
4	784 + 168	1,00 / 0,34	1,00 / 0,03	164	0,1784	94,70
5	784 + 49	1,00 / 0,06	1,00 / 0,00	49	0,3446	90,23
6	784 + 784	0,78 / 0,81	0,22 / 0,30	1160	0,1824	94,75
7	784 + 784	0,60 / 1,00	0,16 / 1,00	661	0,2705	92,54
8	784 + 784	1,00 / 0,48	1,00 / 0,00	783	0,3470	89,83
9	784 + 784	1,00 / 0,71	1,00 / 0,05	746	0,2860	91,79

Таблица 4

Результат обучения эластичной сети по совокупности базовых и дополнительных признаков с максимальным разреживанием и детализованным разбиением на группы

№	Кол-во призн.	Разреж. модели (баз/доп)	Разреж. призн. (баз/доп)	Ост-сь призн.	Кросс-энтропия	Точность (%)
0	784	0,79	0,63	292	0,2737	92,18
2	784 + 1568	0,94 / 0,88	0,85 / 0.74	513	0,1127	96,67
3	784 + 1568	0,85 / 0.88	0,68 / 0,72	692	0,1490	95,74
4	784 + 168	0,86 / 0.57	0.82 / 0.48	230	0,1348	95,94
5	784 + 49	0,90 / 0,29	0,79 / 0,10	212	0,2711	92,37
6	784 + 784	0,84 / 0,90	0,78 / 0,87	274	0,1742	94,99
7	784 + 784	0,84 / 0,86	0,72 / 0,68	470	0,2303	93,51
8	784 + 784	0,84 / 0,96	0,76 / 0,95	225	0,2343	93,45
9	784 + 784	0,83 / 0,94	0,76 / 0,92	248	0,1737	95,10

Следующие эксперименты были проведены при старте итеративных преобразований для λ_k со значений $\lambda_k^0 = \lambda_k^{\max}$. Эмпирическим путем было установлено, что в качестве начальных значений для μ_k следует выбирать значения, сравнимые по порядку с λ_k^0. В частности, в экспериментах далее полагается $\mu_k^0 = \lambda_k^0$. Результаты такого обучения для некоторых групп дополнительных признаков представлены в таблице 3, из которой видно, что в процессе итеративного преобразования λ_k и μ_k эластичная сеть склонна прореживать одни группы признаков значительно больше других. Так, например, при обучении с использованием дополнительных групп 4, 5, 8 и 9 базовые признаки оказались полностью неиспользуемыми для классификации, а при обучении с группой 7 только базовые признаки и остались.

Эффект неравномерности разреживания по группам признаков связан с тем, что преобразование (17) для параметров λ_k и μ_k зависит от влияния признаков из k-ой группы D_k на обучение эластичной сети. Однако это влияние усредняется по всем признакам $j \in D_k$, поэтому по сути оценивается влияние целой группы признаков.

Чтобы исследовать признаки более детально, необходимо свести это усреднение к минимуму. Для этого было сделано максимально детализованное разбиение признаков на группы (8): каждый признак определен в отдельную группу $D_k, k = 1,...,d$, т. е. $K = d$, где d — размерность пространства признаков. Кроме того, отдельно рассматривается нулевой столбец матрицы $\bar{w}$, для которого $\lambda_0 = 0$, а μ_0 изменяется по формуле (18). При таком разбиении имеется $2d+1$ параметров регуляризации: d параметров λ_k и $d+1$ параметр μ_k. Итерации для λ_k начинались, как и раньше, со значений $\lambda_k^0 = \lambda_k^{\max}$ для достижения максимального разреживания.

Результаты такого обучения на объединении базовых и одной группе дополнительных признаков представлены в таблице 4, а результаты обучения на объединении базовых и нескольких групп дополнительных признаков — в таблице 5.

Таблица 5

Результат обучения эластичной сети при объединении групп признаков 0–9 с максимальным разреживанием и детализованным разбиением на группы

№№	Кол-во призн.	Разреж. модели	Разреж. призн.	Ост-сь призн.	Кросс-энтропия	Точность (%)
0	784	0,79	0,63	292	0,2737	92,18
0,2	2352	0,90	0,78	513	0,1127	96,67
0,2,4	2520	0,93	0,86	354	0,0936	97,10
0,2,3,4	4088	0,92	0,80	810	0,0824	97,52
0,2,3,4,9	4872	0,92	0,81	905	0,0811	97,53
0,2,3,4,6,9	5656	0,95	0,87	747	0,0714	97,82
0,2,3,4,6,7,9	6440	0,95	0,88	773	0,0697	97,90
0,2,3,4,6–9	7224	0,96	0,89	797	0,0707	97,81
0–9	8057	0,96	0,90	805	0,0703	97,86

4. Обсуждение результатов

4.1. Информативность группы признаков

Информативностью группы признаков назовем разность кросс-энтропии модели, обученной только на базовых признаках, и кросс-энтропии модели, обученной на объединении базовых признаков и данной группы. Таким образом, информативность количественно отражает, насколько дополнительная группа признаков оказалась полезной для распознавания. Заметим, что во всех проведенных экспериментах увеличение информативности вело к увеличению точности распознавания.

Из таблицы 1 видно, что в порядке понижения информативности группы признаков разложились в следующем порядке: 2, 4, 3, 9, 6, 7, 8, 1 и 5. Градиенты (2) и проекционные гистограммы (4) отражают структурные характеристики изображения и оказались в начале списка. Зашумленные базовые признаки (1) и зонирование (5) оказались в конце списка. При объединении некоторых групп признаков получается более точное распознавание, однако как добиться максимального распознавания при минимальном наборе признаков?

Для достижения максимальной точности предлагается следующий метод отбора признаков:

1. Инициализировать набор признаков для обучения только базовыми признаками.
2. Добавлять к набору группу с наибольшей информативностью из оставшихся в рассмотрении.
3. Проводить обучение на получившемся наборе.
4. Повторять шаги $2-4$, пока точность классификации улучшается.

Следуя данному алгоритму, мы получили результаты в таблице 2. Начиная только с базовых признаков и точности классификации $92,69\%$ и добавляя последовательно группы дополнительных признаков, мы получили набор, состоящий из групп 0, 2, 3, 4, 6 и 9, на котором достигается точность $98,22\%$.

Данный результат значительно ниже уже достигнутого на базе MNIST с помощью сверточных нейронных сетей, глубокого обучения и раздутия обучающей выборки. Однако в сравнении с распознавателями той же сложности (1,2-layer NN, SVM Gaussian Kernel) полученная точность выше [9]. Кроме того, генерируя сложные группы дополнительных признаков, можно получить тем же методом и более высокую точность.

4.2. Разреженность

Эксперименты по обучению на группах дополнительных признаков с максимальным разреживанием (таблица 3) показывают, что добавление признаков улучшило результат только в случае групп 2, 3, 4 и 6. В остальных случаях получилась меньшая точность, чем при обучении только на базовых признаках. При обучении с группами дополнительных признаков 4, 5, 8 и 9 базовые признаки были полностью прорежены моделью, общая разреженность по признакам при этом составила $50-60\%$, а получившаяся точность в среднем на $2,03\%$ меньше, чем в случае обучения, нацеленного на максимальную точность (таблица 1).

Улучшить разреживание позволяет более детализованное разбиение на группы признаков. В предельном случае — каждый признак считается отдельной группой разбиения (8). По таблице 4 видно,

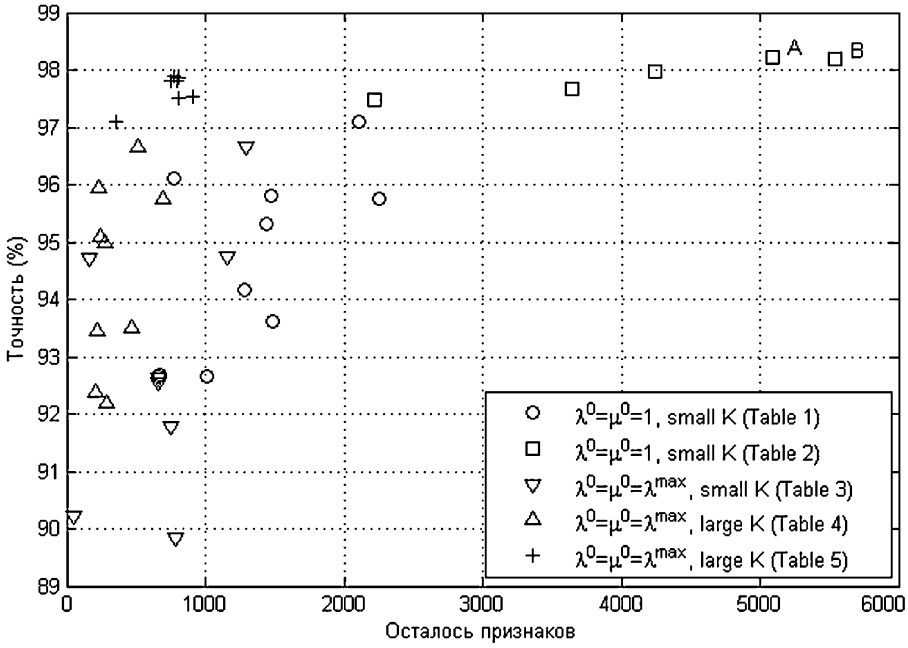

Рис. 1. Достижимые соотношения точности распознавания и количества оставшихся признаков для разных способов обучения. *small K* – разбиение на крупные группы признаков ($K < 10$), *large K* – максимально детализованное разбиение ($K = d$). Точка A – точность: $98,22\%$, осталось приз.: 5089 (таблица 2, строка 6). Точка B – точность: $98,20\%$, осталось приз.: 5538 (таблица 2, строка 7)

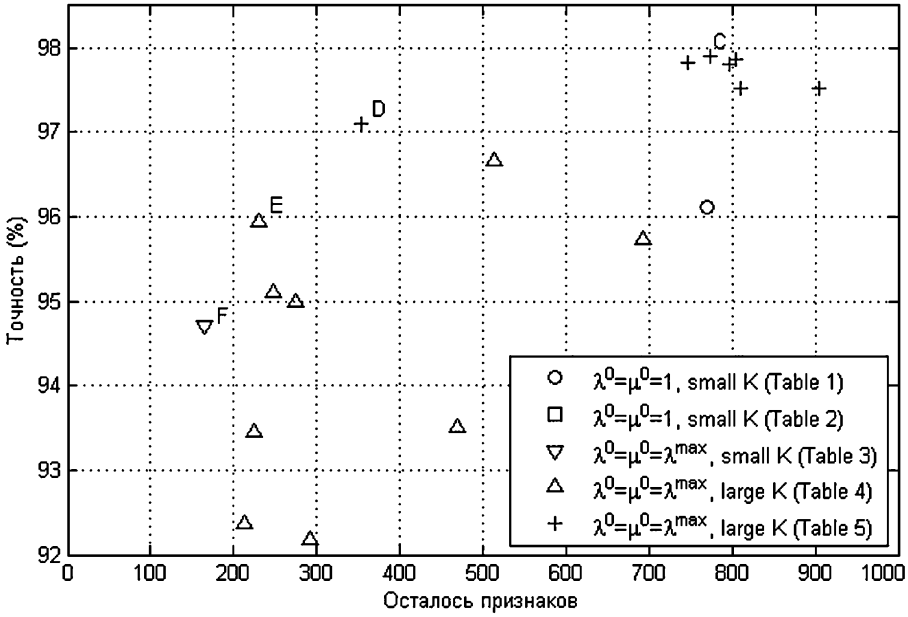

Рис. 2. Достижимые соотношения точности распознавания и количества оставшихся признаков для разных способов обучения (фрагмент). *small K* – разбиение на крупные группы признаков ($K < 10$), *large K* – максимально детализованное разбиение ($K = d$). Точка C – точность: $97,9\%$, осталось признн.: 773 (таблица 5, строка 7). Точка D – точность: $97,1\%$, осталось признн.: 354 (таблица 5, строка 3). Точка E – точность: $95,94\%$, осталось признн.: 230 (таблица 4, строка 4). Точка F – точность: $94,70\%$, осталось признн.: 164 (таблица 3, строка 3)

что разреженность признаков при таком варианте разбиения на группы достигает $80-90\%$, при среднем падении точность только $0,4\%$.

Для обучения с максимальным разреживанием и детализованным разбиением был применен алгоритм отбора признаков на основе информативности, предложенный в п. 4.1 (таблица 5). Итоговый набор признаков содержит группы 0, 2, 3, 4, 6, 7 и 9, на этом наборе достигается точность классификации $97,9\%$ при уровне разреженности по признакам $0,88$, т. е. из всех 6440 признаков набора используются только 773. По сравнению с эластичной сетью, на которой достигается максимальная точность (таблица 2), падение точности составило $0,32\%$ при уменьшении числа эффективных признаков в $6,6$ раз с 5089 до 773.

Для некоторых практических задач классификации приоритетом является большая разреженность модели, которая достигается за счет снижения точности. Предложенный метод позволяет обучать модели с разными соотношениями разреженности и точности. На рисунке 1 отмечены полученные в экспериментах эластичные сети при разных способах обучения (таблицы 1–5) в пространстве «Количество оставшихся признаков — Точность распознавания». Выделенная точка A соответствует эластичной сети с максимальной точностью, она получена при обучении с разбиением на $K=6$ групп признаков и начале итераций для параметров регуляризации со значений $\lambda_k=1,\mu_k=1,k=1,...,K$ (таблица 2, строка 6).

На рисунке 2 изображен фрагмент рисунка 1 для количества оставшихся признаков менее 1000. Выделенная точка C соответствует эластичной сети с точностью $97,9\%$ и количеством оставшихся признаков 773, она получена при обучении с индивидуальным параметром регуляризации для каждого признака, т. е. при разбиении на $K=6440$ групп признаков и начале итераций для параметров регуляризации со значений $\lambda_k=\mu_k=\lambda_k^{\max}$, $k=1,...,K$.

Как упоминалось ранее, итерации по преобразованию (17–18) при максимизации обоснованности не гарантируют ее сходимость к глобальному максимуму. Конечные значения параметров регуляризации, а следовательно и решение минимизационной задачи эластичной сети (11) зависят от начальных точек λ и μ. Варьируя начальные значения параметров λ и μ для итеративного преобразования и разбиение на группы признаков, можно балансировать между разреженностью и точностью классификации.

Заключение

В статье представлен метод обучения эластичной сети с автоматическим подбором параметров регуляризации посредством итеративного преобразования последних с целью максимизации обоснованности. Предложенный метод не гарантирует сходимость к глобальному максимуму обоснованности, а результат зависит от начальных приближений для параметров регуляризации и разбиения на группы. Однако эффективность метода подтверждается экспериментами на примере распознавания рукописных символов. Самая высокая точность классификации ($97,9\%$) при большой разреженности (более $0,8$) была достигнута при максимально детализованном разбиении признаков на группы и старте итеративного преобразования параметров регуляризации со значений, гарантирующих максимальное разреживание (таблица 5). Данный метод обучения эластичных сетей позволяет создать механизм отбора признаков с возможностью балансирования между разреженностью модели и точностью классификации.

В следующих работах планируется провести тестирование предложенного метода на других открытых наборах данных, включая данные, содержащие большое количество признаков, для которых одним из важнейших свойств обучаемых моделей является высокая разреженность, а также сравнение предложенного метода автоматического подбора параметров с существующими.

Литература

1. *Zou H., Hastie T.* Regularization and Variable Selection via the Elastic Net // Journal of the Royal Statistical Society B. 2005. № 67. P. 301–320.
2. *James G., Witten D., Hastie T., Tibshirani R.* An Introduction to Statistical Learning. M.: Springer, 2013. 204 p.
3. *Guyon I., Elisseeff A.* An Introduction to Variable and Feature Selection // Journal of Machine Learning Research. 2003. № 3. P. 1157–1182.
4. *Nesterov Yu.* Gradient Methods for Minimizing Composite Objective Function. 2007.
5. *Richtarik P., Schmidt M.* Modern Convex Optimization Methods for Large-Scale Empirical Risk Minimization // International Conference on Machine Learning. July 2015.
6. *Нестеров Ю. Е.* Введение в выпуклую оптимизацию. М.: МЦНМО, 2010. 262 с.
7. *Bishop C. M.* Pattern Recognition and Machine Learning. M.: Springer, 2006. 740 p.
8. *Прилепко А. И., Калиниченко Д. Ф.* Асимптотические методы и специальные функции. М.: МИФИ, 1980. 107 с.

9. *Le Cun Y., Bottou L., Bengio Y., Haffner P.* Gradient-based learning applied to document recognition // Proceedings of the IEEE. 1998. № 86 (11). P. 2278–2324. http://yann.lecun.com/exdb/mnist/
10. *Trier O. D., Jain A. K., Taxt T.* Features Extraction Methods for Character Recognition // Pattern Recognition. 1996. № 29 (11). P. 641–662.
11. *The MathWorks Inc.* MATLAB Image Processing Toolbox documentation. http://www.mathworks.com/help/images/
12. *Harris C., Stephens M.* A combined corner and edge detector // Proceedings of the 4th Alvey Vision Conference. 1988. P. 147–151.
13. *Canny J.* A Computational Approach To Edge Detection // IEEE Trans. Pattern Analysis and Machine Intelligence. 1986. № 8 (6). P. 679–698.

Дубнов Игорь Андреевич. Аспирант МФТИ. Окончил в 2013 году Московский физико-технический институт (государственный университет). Область научных интересов: классификация, распознавание образов, нейронные сети. E-mail: dubnov@phystech.edu

Мерков Александр Борисович. Кандидат физико-математических наук, старший научный сотрудник ИСА РАН. Окончил в 1979 году Московский государственный университет имени М. В. Ломоносова. Автор около 20 научных работ и 2 монографий. Область научных интересов: теория особенностей, топология, комбинаторика, статистическая теория обучения. E-mail: alexander.merkov@gmail.com