Поиск эффективных методов снижения размерности при решении задач многоклассовой классификации путем её сведения к решению бинарных задач*

М. Е. Карасиков[1]***, Ю. В. Максимов***[1,2]
karasikov@phystech.edu
[1]МФТИ; [2]ИППИ РАН

Работа посвящена задаче многоклассовой классификации высокой размерности. Рассмотрены способы решения задачи многоклассовой классификации на основе сведения её к задачам бинарной классификации. Исследованы различные подходы к сведению задачи многоклассовой классификации к задачам бинарной классификации и проведено сравнение их эффективностей. Предложены пути повышения производительности классификаторов путем снижения размерности пространства признаков методом случайных проекций. Проведены эксперименты на реальных данных для различных классификаторов, результаты которых отражают характерные зависимости качества классификации и сложности обучения при снижении размерности методом случайных проекций.

Ключевые слова: многоклассовая классификация; One-vs-All; One-vs-One; Error-Correcting Output Codes; лемма Джонсона-Линденштраусса; снижение размерности; случайные проекции.

Dimensionality reduction for multi-class learning problems reduced to multiple binary problems*

M. E. Karasikov[1]***, Y. V. Maximov***[1,2]
[1]MIPT; [2]IITP RAS

Modern machine learning problems, such as image classification, video recognition, text retrieval or engineering diagnostics, leads to the analysis of multi-class learning methods for high-dimensional datasets which can not be solved without data pre-processing. Principal Component Analysis and its randomized versions are some of the most widespread dimensionality reduction methods. We analyze the classification performance of various approaches to multi-class classification (One-vs-One, One-vs-All, Error-Correcting Output Codes) in combination with the dimensionality reduction based on Random Gaussian Projections. Computational efficiency of the Random Projections distinguishes it from other dimensionality reduction methods. With that, low-distortion property of this mapping allows to reduce dimensionality thrice and more with imperceptible quality losses. This leads to an effective and computationally cheap approach for solving multi-class problems in high-dimensional space. Basic theoretical foundations of the approach as well as its computational complexity analysis are discussed. Numerical stability and quality of the method proposed is supported by empirical evaluation of the approach. We provide a number of experiments for different machine learning methods over various real datasets from the open-source machine learning repositories. Experiments show applicability of Random Projections for cheap selection of the most suitable classifier, its parameters optimization and multi-class classification approach selection.

*Исследование выполнено при финансовой поддержке РФФИ в рамках научного проекта № 14-07-31241 мол_а.

Keywords: multi-class classification; One-vs-All; One-vs-One; Error-Correcting Output Codes; Johnson-Lindenstrauss lemma; dimensionality reduction; Random Projections.

Введение

В современных задачах машинного обучения часто возникают задачи многоклассовой классификации, в которых множество некоторых объектов нужно отобразить на множество классов большой мощности. Например, распознавание символов по изображению, классификация речи и текста, медицинская диагностика.

Формально задача классификации заключается в следующем. Пусть $X = \{x_1, \ldots, x_\ell\}$ — множество описаний объектов, $Y = \{y_1, \ldots, y_N\}$ — конечное множество меток классов. Существует целевая функция — отображение $y : X \to Y$, значения которого известны только на объектах обучающей выборки

$$\mathfrak{D} = \left\{(x^1, y^1), \ldots, (x^m, y^m)\right\} \subset X \times Y.$$

Требуется построить алгоритм $a : X \to Y$ — отображение, приближающее целевую функцию y на множестве X. Задачу классификации с $N = 2$ ($N > 2$) классами будем называть бинарной (многоклассовой) задачей. В бинарной задаче положим для удобства $Y = \{-1, +1\}$, а в многоклассовой — $Y = \{1, \ldots, N\}$. Для решения многоклассовой задачи можно использовать два способа. Первый способ состоит в использовании многоклассовых классификаторов, например, решающих деревьев. В нашей работе для решения многоклассовой задачи применяется способ, основанный на использовании классификаторов (линейный дискриминант Фишера [1], SVM [2]), решающих бинарные задачи. При этом многоклассовая задача разбивается на множество бинарных задач, которые решаются независимо с использованием техник бинарной классификации. Процесс разбиения будем называть сведением многоклассовой задачи к бинарным.

Общие сведения о задаче многоклассовой классификации даны в [3, 4]. В [4], также, проведено сравнение различных подходов к сведению многоклассовой задачи к бинарным.

Подходы к сведению многоклассовой задачи к бинарным

— **One-vs-All approach** (OVA) заключается в обучении N классификаторов по следующему принципу

$$f_i(x) = \begin{cases} \geqslant 0, & \text{если } y(x) = i, \\ < 0, & \text{если } y(x) \neq i, \end{cases}$$

которые отделяют каждый класс от остальных. Далее, для каждого $x \in X$ вычисляются все классификаторы и выбирается класс, соответствующий классификатору с большим значением:

$$a(x) = \arg\max_{i=1,\ldots,N} f_i(x).$$

— **One-vs-One approach**. Его так же называют All-vs-All (AVA) approach. В этом случае строятся $N(N-1)$ классификаторов, которые разделяют объекты пар различных классов:

$$f_{ij}(x) = \begin{cases} +1, & \text{если } y(x) = i, \\ -1, & \text{если } y(x) = j. \end{cases}$$

После обучения бинарных классификаторов решение принимается следующим образом:

$$a(x) = \arg\max_{i=1,\ldots,N} \sum_{\substack{j=1,\ldots,N \\ j\neq i}} f_{ij}(x).$$

Сравнение первых двух подходов проведено в [5]. Недостаток подхода OVA состоит во многих случаях отказа от классификации [5]. Однако на распознанных объектах этот способ дает очень хорошие результаты [5].

— **Error-Correcting Output Codes approach** (ECOC) предложен в [6]. ECOC предполагает кодирование меток классов двоичными числами длины F, которое сводит задачу определения неизвестного класса объекта x к определению F неизвестных бит кодового слова класса $y(x)$. Для каждого бита строится бинарный классификатор, отделяющий группу классов со значением $+1$ соответствующего бита от классов со значением -1. Пусть $\mathbf{M} \in \{-1, +1\}^{N\times F}$ — кодовая матрица, в строках которой записаны коды меток классов из Y. Тогда обучаются F классификаторов $f_1, \ldots, f_F$ так, чтобы $f_j(x) = i$ тогда и только тогда, когда $M^i_j = 1$. При классификации нового объекта x вычисляется его кодовое слово $\boldsymbol{f}(x) = [f_1(x), \ldots, f_F(x)]$ и выбирается класс, с ближайшим к $\boldsymbol{f}(x)$ кодовым словом. Для расстояния Хэмминга получим:

$$a(x) = \arg\min_{i=1,\ldots,N} \sum_{j=1}^{F} \left(\frac{1 - \operatorname{sign}(M^i_j f_j(x))}{2} \right).$$

В работе [7] было представлено улучшение ECOC, согласно которому кодовая матрица M допускает нулевые элементы, а классификация происходит по правилу

$$a(x) = \arg\min_{i=1,\ldots,N} \sum_{j=1}^{F} L(M^i_j f_j(x)),$$

где L — некоторая функция потерь. В результате наблюдалось снижение числа ошибок классификации почти на всех представленных тестах.
Реализация изложенных подходов представлена в [8].

Как видно [3, 4, 5, 6, 7], существует множество подходов к сведению многоклассовой задачи к бинарным.

Обратимся теперь к вопросу обучения бинарных классификаторов. Как правило, объекты $x \in X$ задаются векторами в n-мерном евклидовом пространстве $\mathbb{R}^n$, которое мы будем называть пространством признаков. Тогда в качестве бинарных классификаторов, как упоминалось выше, можно использовать линейный дискриминант Фишера [1], SVM [2], а так же строить такие композиции, как AdaBoost [9, 10]. Временная сложность обучения и тестирования названных классификаторов как минимум линейна по числу признаков, и в задачах высокой размерности, т. е. с высокоразмерным пространством признаков (например, в задачах распознавания лиц) все эти методы обладают недостаточно высоким быстродействием. Кроме того, из-за линейной сложности по памяти задачи классификации сверх большой размерности решать без предобработки данных не удается. Таким образом, снижение размерности задачи многоклассовой классификации представляется актуальной задачей.

Часто бывает, что в задаче классификации высокой размерности множество объектов $X \subset \mathbb{R}^n$ лежит на линейном многообразии, размерность которого много меньше размерности исходного пространства признаков. В таких случаях чрезвычайно эффективным оказывается метод главных компонент [11]. Метод главных компонент находит ортогональные направления $\mathbf{w}_1, \ldots, \mathbf{w}_d$ (главные компоненты), вдоль которых выборочная дисперсия максимальна, и проецирует элементы множества X на линейное многообразие

$$\bar{\mathbf{x}} + \mathrm{Lin}(\mathbf{w}_1, \ldots, \mathbf{w}_d).$$

Однако метод главных компонент применим не к любым данным, ведь легко привести пример задачи классификации, для которой классификация значительно усложнится после его применения. Таким образом, одна из проблем метода главных компонент — это зависимость от данных. Еще одним недостатком метода главных компонент является высокая временная сложность $O(\ell n^2 + n^3)$ [12].

Альтернативным методом снижения размерности является метод случайных проекций, заключающийся в случайном проецировании исходного пространства признаков на пространство меньшей размерности. Этот метод будет подробно изложен в главе 1. Сложность генерации проекционной матрицы $O(nd)$, где d — размерность редуцированного пространства признаков. Временная сложность нахождения нового признакового описания объектов — $O(\ell dn)$. Преимуществами последнего метода являются его простота и независимость от исходных данных. При всем этом, метод случайных проекций сравним с методом главных компонент по качеству классификации в редуцированном пространстве [13].

В нашей работе анализируется метод случайных проекций как метод снижения размерности задачи многоклассовой классификации. Приводится сравнение эффективностей основных подходов к сведению многоклассовой задачи к бинарным (One-vs-One, One-vs-All, ECOC) при использовании метода случайных проекций.

Постановка задачи

Дана задача многоклассовой классификации с пространством признаков высокой размерности $X \subset \mathbb{R}^n$. Применяя отображение $A_d : X \to X' \subset \mathbb{R}^d$, $d < n$, будем снижать размерность задачи, тем самым понижая сложность классифицирующего алгоритма. При этом может исказиться метрика, то есть изменятся относительные расстояния между объектами в X', что может привести к потере качества классификации. Чем больше сжатие $\frac{n}{d}$, тем больше возможны потери в качестве классификации. Будем искать то минимальное значение d, при котором ошибка классификации e, определенная некоторым функционалом качества, не превышает некоторой заданной допустимой ошибки e'.

Поставим задачу формально. $X = \{\mathbf{x}_1, \ldots, \mathbf{x}_\ell\} \subset \mathbb{R}^n$ — множество описаний объектов в евклидовом пространстве, $Y = \{1, \ldots, N\}$, $N > 2$ — множество меток классов. Предполагается существование целевой зависимости — отображения $y : X \to Y$, значения которого известны только на объектах обучающей выборки

$$\mathfrak{D} = \big\{(\mathbf{x}^1, y^1), \ldots, (\mathbf{x}^m, y^m)\big\} \subset X \times Y.$$

Пусть задан метод $\mu : \tilde{\mathfrak{D}} \mapsto a$, который $\forall k \geqslant 1$ по произвольной выборке

$$\tilde{\mathfrak{D}} = \big\{(\tilde{\mathbf{x}}^1, y^1), \ldots, (\tilde{\mathbf{x}}^m, y^m)\big\} \subset \mathbb{R}^k \times Y$$

строит алгоритм классификации $a : X \to Y$, решающий задачу многоклассовой классификации с ошибкой $e(a)$, и задана допустимая ошибка классификации e'.

Найти

$$d^* = \min_{e(\mu(A_d(\mathfrak{D})))\leqslant e'} d,$$

где $e(\mu(A_d(\mathfrak{D})))$ — ошибка алгоритма классификации, построенного методом μ по выборке $A_d(\mathfrak{D})$ — образу снижающего размерность отображения A_d, e' — допустимая ошибка классификации.

Многоклассовая классификация

Бинарные классификаторы. Для бинарной классификации, где $Y = \{-1, +1\}$, используются линейные классификаторы $a(\mathbf{x}, \boldsymbol{w}) = \operatorname{sign} f_{\boldsymbol{w}}(\mathbf{x})$, где $f_{\boldsymbol{w}}(\mathbf{x})$ — дискриминантная функция, $\boldsymbol{w}$ — вектор параметров. Обучение классификатора производится путем минимизации эмпирического риска

$$Q(f_{\boldsymbol{w}}, \mathfrak{D}) = \frac{1}{m}\sum_{i=1}^{m} \mathscr{L}\left(y^i f_{\boldsymbol{w}}(\mathbf{x}^i)\right), \tag{1}$$

где функция потерь $\mathscr{L}$ — невозрастающая и неотрицательная. При этом $y^i f_{\boldsymbol{w}}(\mathbf{x}^i)$ называется отступом объекта $\mathbf{x}^i$ относительно алгоритма классификации

$$a(\mathbf{x}, \boldsymbol{w}) = \operatorname{sign} f_{\boldsymbol{w}}(\mathbf{x}). \tag{2}$$

В нашей работе в качестве линейных классификаторов берутся различные модификации SVM и AdaBoost.

В **SVM** [2] за функцию потерь принимается кусочно-линейная функция

$$\mathscr{L}\left(y^i f_{\boldsymbol{w}}(\mathbf{x}^i)\right) = (1 - y^i f_{\boldsymbol{w}}(\mathbf{x}^i))_+,$$

где $(\cdot)_+ = \max\{\cdot, 0\}$. Классификатор $f_{\boldsymbol{w}}$ ищется в виде $f_{\boldsymbol{w}}(\mathbf{x}) = \boldsymbol{w} \cdot \mathbf{x}$, где $\boldsymbol{w}$ — решение задачи безусловной минимизации

$$\frac{1}{2C}\|\boldsymbol{w}\|^2 + \sum_{i=1}^{m}(1 - y^i(\boldsymbol{w} \cdot \mathbf{x}^i))_+ \to \min_{\boldsymbol{w}\in\mathbb{R}^n}.$$

Согласно алгоритму **AdaBoost** [9, 10] дискриминантная функция $f_T(\mathbf{x}) = \sum_{t=1}^{T} \alpha_t h_t(\mathbf{x})$ конструируется из T базовых классификаторов $\{h_t(\mathbf{x})\}_{t=1}^{T} \subset H$, которые обучаются последовательно так, чтобы минимизировать эмпирический риск $Q(f_T, \mathfrak{D})$ (см. формулу 1) с экспоненциальной функцией потерь $\mathscr{L}(M) = e^{-M}$.

В нашей работе базовые классификаторы выбираются из множества

$$H = \left\{h^{\pm}_{j\theta}(\mathbf{x}) = \pm\operatorname{sign}(x_j - \theta) :\ \mathbf{x} = (x_1, \ldots, x_n),\ j = 1, \ldots, n,\ \theta \in \mathbb{R}\right\}. \tag{3}$$

Подходы к сведению многоклассовой задачи к бинарным. Как было сказано выше, в работе предлагается решать задачу многоклассовой классификации путем сведения её к бинарным задачам. При этом удобно использовать конструкцию, предложенную в [7]. Эта конструкция предполагает кодирование меток классов $i \in Y = \{1, \ldots, N\}$ строками $M^i \in \{-1, 0, +1\}^F$ длины F, составляющими кодовую матрицу $[M^i_j]^{N\times F}$. При этом

для каждого столбца M_j, $j = 1, \ldots, F$, матрицы M получаем бинарную задачу, которая заключается в разделении классов

$$Y_j^{+1} = \{i \in Y : \ M_j^i = +1\} \ \text{ и } \ Y_j^{-1} = \{i \in Y : \ M_j^i = -1\}.$$

Формально каждая бинарная задача выглядит следующим образом. $X_j = X$ — множество объектов, $Y_j = \{-1, +1\}$ — множество меток классов. Построить классификатор $a_j : X_j \to Y_j$, аппроксимирующий целевую функцию $y_j : X_j \to Y_j$, значения которой известны только на объектах обучающей выборки

$$\mathfrak{D}_j = \left\{(\mathbf{x}, M_j^y) : \ M_j^y \neq 0, \ (\mathbf{x}, y) \in \mathfrak{D}\right\},$$

где $M \in \{-1, 0, +1\}^{N \times F}$ — матрица, строки которой состоят из кодов меток классов Y.

На этапе распознавания для каждого нового объекта $\mathbf{x}$ вычисляются все F классификаторов $a_1(\mathbf{x}), \ldots, a_F(\mathbf{x})$, и классификация происходит голосованием, то есть объект $\mathbf{x}$ относится к тому классу, который чаще всего встречается в множествах $Y_1^{a_1(\mathbf{x})}, \ldots, Y_F^{a_F(\mathbf{x})}$:

$$a(\mathbf{x}) = \arg\max_{i=1,\ldots,N} \sum_{j=1}^{F} \mathbf{1}\left[i \in Y_j^{a_j(\mathbf{x})}\right] = \arg\min_{i=1,\ldots,N} \sum_{j=1}^{F} \mathbf{1}[a_j(\mathbf{x}) \neq M_j^i]. \tag{4}$$

Этот метод обобщается выбором произвольного расстояния $d(\cdot, \cdot)$ между кодовыми словами M^i классов $i \in Y$ и словом-ответом $\boldsymbol{a}(\mathbf{x}) = [a_1(\mathbf{x}), \ldots, a_F(\mathbf{x})]$:

$$a(\mathbf{x}) = \arg\min_{i=1,\ldots,N} d(M^i, \boldsymbol{a}(\mathbf{x})).$$

Если каждый классификатор $a_j(\mathbf{x})$ задается дискриминантной функцией $f_j(\mathbf{x})$ (см. 2), то расстояние можно определить произвольной функцией потерь L :

$$d_L(M^i, \boldsymbol{f}(\mathbf{x})) = \sum_{j=1}^{F} L(M_j^i f_j(\mathbf{x})).$$

Тогда классификация будет происходить следующим образом:

$$a(\mathbf{x}) = \arg\min_{i=1,\ldots,N} d_L(M^i, \boldsymbol{f}(\mathbf{x})), \quad \boldsymbol{f}(\mathbf{x}) = [f_1(\mathbf{x}), \ldots, f_F(\mathbf{x})]. \tag{5}$$

Рассмотрим подробно построение кодовой матрицы M. Для подхода One-vs-All число бинарных классификаторов равно числу классов: $F = N$, а каждый бинарный классификатор $a_j(\mathbf{x})$ $(j = 1, \ldots, F)$ обучается так, чтобы отделять объекты класса с меткой j от остальных объектов множества X. Тогда метки классов кодируются следующим образом:

$$M = \begin{pmatrix} +1 & -1 & \ldots & -1 \\ -1 & +1 & \ldots & -1 \\ \vdots & \vdots & \ddots & \vdots \\ -1 & -1 & \ldots & +1 \end{pmatrix}_{N \times N}.$$

Для подхода All-vs-All число классификаторов равно $F = \binom{N}{2}$, и классификаторы разделяют каждую пару классов. Для этого подхода кодовая матрица M записывается

в виде:

$$M = \begin{pmatrix} +1 & +1 & \dots & +1 & +1 & 0 & \dots & \dots & 0 \\ -1 & 0 & \dots & 0 & 0 & +1 & \dots & \dots & 0 \\ 0 & -1 & \dots & 0 & 0 & -1 & \dots & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \dots & \dots & \vdots \\ 0 & 0 & \dots & -1 & 0 & 0 & \dots & \dots & +1 \\ 0 & 0 & \dots & 0 & -1 & 0 & \dots & \dots & -1 \end{pmatrix}_{N \times \binom{N}{2}}.$$

Подход ECOC [6] обобщает подходы One-vs-All и All-vs-All. Его преимуществом является возможность использования кодов, исправляющих ошибки. Пусть минимальное расстояние Хэмминга между строками кодовой матрицы M равно $d_{\min}$. Тогда корректирующая способность кода равна $t = \left[\frac{d_M - 1}{2}\right]$. Свойство кода исправлять ошибки повышает точность классификации, т. к. код гарантированно восстанавливает исходное слово, если произошло не более t ошибок бинарных классификаторов. В этом случае многоклассовая классификация (см. формулу 4) происходит корректно. Таким образом, высокая корректирующая способность кода уменьшает число ошибок многоклассовой классификации. Однако с ростом длины кода вместе с корректирующей способностью растет и число бинарных классификаторов, а, значит, и среднее число ошибок. Так как корректирующая способность t такого кода растет линейно вместе с длиной кода, как и число ошибок бинарных классификаторов, то улучшение качества многоклассовой классификации за счет увеличения длины кода ограничено. В нашей работе для построения кодовой матрицы M в подходе ECOC применяются БЧХ коды и случайное кодирование. При случайном кодировании генерируется множество случайных кодовых матриц и из них выбирается матрица с максимальной корректирующей способностью.

Снижение размерности

В последующих рассуждениях этого раздела будем опираться на работу [14]. Для снижения размерности пространства признаков будем использовать случайное линейное отображение

$$A: \quad \mathbb{R}^n \supset X \to X' \subset \mathbb{R}^d. \tag{6}$$

Пусть $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_\ell]$ — транспонированная матрица объектов-признаков в исходном пространстве признаков, $\mathbf{A} = [\xi_{ij}]^{d \times n}$ — проекционная матрица, соответствующая отображению 6, где ξ_{ij} — независимые центрированные одинаково распределенные случайные величины с мат. ожиданием $\mathsf{E}\xi_{ij} = 0$ и дисперсией $\mathsf{D}\xi_{ij} = \sigma^2$. Транспонированная матрица объектов-признаков в редуцированном пространстве запишется в виде $\mathbf{X}' = \mathbf{A}\mathbf{X}$.

Будем искать отображения 6, сохраняющие попарные расстояния между объектами множества X с точностью $\varepsilon \in (0, 1)$, аналогично условию леммы Джонсона-Линденштрауса [14]:

$$\forall \mathbf{x}', \mathbf{x}'' \in X \quad (1 - \varepsilon)\|\mathbf{x}' - \mathbf{x}''\|_2^2 < \|\mathbf{A}\mathbf{x}' - \mathbf{A}\mathbf{x}''\|_2^2 < (1 + \varepsilon)\|\mathbf{x}' - \mathbf{x}''\|_2^2. \tag{7}$$

$$\|\mathbf{A}\mathbf{x}' - \mathbf{A}\mathbf{x}''\|_2^2 = (\mathbf{x}' - \mathbf{x}'')^\intercal \mathbf{A}^\intercal \mathbf{A}(\mathbf{x}' - \mathbf{x}'') = (\mathbf{x}' - \mathbf{x}'')^\intercal (\mathbf{I} + \underbrace{\mathbf{A}^\intercal \mathbf{A} - \mathbf{I}}_{\mathbf{\Sigma}})(\mathbf{x}' - \mathbf{x}'')$$

Условие 7 можно переписать в виде

$$\forall \mathbf{x}', \mathbf{x}'' \in X \quad -\varepsilon\|\mathbf{x}' - \mathbf{x}''\|_2^2 < (\mathbf{x}' - \mathbf{x}'')^\intercal \mathbf{\Sigma} (\mathbf{x}' - \mathbf{x}'') < \varepsilon\|\mathbf{x}' - \mathbf{x}''\|_2^2.$$

Логично брать такую проекционную матрицу $\mathbf{A}$, чтобы $\mathsf{E}\boldsymbol{\Sigma} = 0$.

$$\mathsf{E}[\boldsymbol{\Sigma}]_{ij} = \mathsf{E}[\mathbf{A}^\intercal\mathbf{A} - \mathbf{I}]_{ij} = \mathsf{E}\sum_{k=1}^{d}\xi_{ki}\xi_{kj} - \delta_{ij} = (\sigma^2 d - 1)\delta_{ij},$$

где $\delta_{ij} = \mathbf{1}[i = j]$ — символ Кронекера. Положим $\sigma^2 = \frac{1}{d}$.

Оценим вероятность того, что случайная матрица $\mathbf{A} = [\xi_{ij}]^{d\times n}$ независимых гауссовских случайных величин $\xi_{ij} \sim \mathcal{N}(0, \frac{1}{d})$ удовлетворяет условию 7. Для этого сначала оценим вероятность того, что отображение A сильно изменит длину некоторого фиксированного вектора $\mathbf{u} \in \mathbb{R}^n$. Пусть ортогональная матрица

$$\mathbf{C}: \quad \mathbf{Cu} = [\underbrace{\|\mathbf{u}\|_2, 0, \ldots, 0}_{n}]^\intercal = \|\mathbf{u}\|_2\boldsymbol{e}_1, \quad \mathbf{C}^\intercal\mathbf{C} = \mathbf{C}\mathbf{C}^\intercal = \mathbf{I}.$$

$$\|\mathbf{Au}\|_2^2 = \mathbf{u}^\intercal\mathbf{A}^\intercal\mathbf{Au} = \mathbf{u}^\intercal\mathbf{C}^\intercal\underbrace{\mathbf{CA}^\intercal}_{\mathbf{B}^\intercal}\underbrace{\mathbf{AC}^\intercal}_{\mathbf{B}}\mathbf{Cu} = \|\mathbf{BCu}\|_2^2 = \|\mathbf{u}\|_2^2\|\mathbf{Be}_1\|_2^2.$$

Поскольку $\mathbf{A} \sim \mathcal{N}^{d\times n}\left(0, \frac{1}{d}\right)$, то $\mathbf{B} = \mathbf{AC}^\intercal \sim \mathcal{N}^{d\times n}\left(0, \frac{1}{d}\right)$.

Тогда получим

$$\begin{aligned}\mathsf{P}\left\{\left|\|\mathbf{Au}\|_2^2 - \|\mathbf{u}\|_2^2\right| \geqslant \varepsilon\|\mathbf{u}\|_2^2\right\} &= \mathsf{P}\left\{\left|\|\mathbf{Be}_1\|_2^2 - 1\right| \geqslant \varepsilon\right\} = \\ &= \mathsf{P}\left\{\left|\sum_{i=1}^{d}\xi_{i1}^2 - 1\right| \geqslant \varepsilon\right\} = \\ &= \mathsf{P}\left\{\sum_{i=1}^{d}\xi_{i1}^2 \leqslant 1-\varepsilon\right\} + \mathsf{P}\left\{\sum_{i=1}^{d}\xi_{i1}^2 \geqslant 1+\varepsilon\right\}.\end{aligned}$$

Оценим вероятности $\mathsf{P}\left\{\sum\limits_{i=1}^{d}\xi_{i1}^2 \leqslant 1-\varepsilon\right\}$ и $\mathsf{P}\left\{\sum\limits_{i=1}^{d}\xi_{i1}^2 \geqslant 1+\varepsilon\right\}$ отдельно.

$$\begin{aligned}&\mathsf{P}\left\{\sum_{i=1}^{d}\xi_{i1}^2 \leqslant 1-\varepsilon\right\} = \\ &= \sup_{t<0}\mathsf{P}\left\{\exp\left(t\sum_{i=1}^{d}\xi_{i1}^2\right) \geqslant \exp\left(t\left(1-\varepsilon\right)\right)\right\} \overset{\text{(неравенство Маркова)}}{\leqslant} \sup_{t<0}\frac{\mathsf{E}\left[\exp\left(t\sum_{i=1}^{d}\xi_{i1}^2\right)\right]}{\exp\left(t\left(1-\varepsilon\right)\right)} = \\ &= \sup_{t<0}\frac{\left(\mathsf{E}\left[\exp\left(t\xi^2\right)\right]\right)^d}{\exp\left(t\left(1-\varepsilon\right)\right)} = \\ &= \sup_{t<0}\left[\left(\int_{-\infty}^{+\infty}e^{tx^2}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{x^2}{2\sigma^2}}dx\right)^d\exp\left(-t\left(1-\varepsilon\right)\right)\right] = \\ &= \sup_{t<0}\left[\left(\int_{-\infty}^{+\infty}e^{t\frac{x^2}{d}}\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}dx\right)^d\exp\left(-t\left(1-\varepsilon\right)\right)\right] = \\ &= \sup_{t<0}\left[\left(1-\frac{2t}{d}\right)^{-\frac{d}{2}}\exp\left(-t\left(1-\varepsilon\right)\right)\right].\end{aligned}$$

$$\frac{d}{dt}\left[\left(1-\frac{2t}{d}\right)^{-\frac{d}{2}}\exp\left(-t\left(1-\varepsilon\right)\right)\right]=0$$
$$\Leftrightarrow 1-\left(1-\frac{2t}{d}\right)\left(1-\varepsilon\right)=0$$
$$\Leftrightarrow t=-\frac{d}{2}\frac{\varepsilon}{1-\varepsilon}.$$

Итак,

$$\mathsf{P}\left\{\sum_{i=1}^{d}\xi_{i1}^2\leqslant 1-\varepsilon\right\}\leqslant\sup_{t<0}\left[\left(1-\frac{2t}{d}\right)^{-\frac{d}{2}}\exp\left(-t\left(1-\varepsilon\right)\right)\right]=$$
$$=\left(1+\frac{\varepsilon}{1-\varepsilon}\right)^{-\frac{d}{2}}\exp\left(\frac{d}{2}\varepsilon\right)=\exp\left(\frac{d}{2}\left(\varepsilon+\ln(1-\varepsilon)\right)\right)\leqslant\exp\left(-\frac{d}{2}\frac{\varepsilon^2}{2}\right).$$

Аналогично доказывается

$$\mathsf{P}\left\{\sum_{i=1}^{d}\xi_{i1}^2\geqslant(1+\varepsilon)\right\}\leqslant\exp\left(-\frac{d}{2}\left(\frac{\varepsilon^2}{2}-\frac{\varepsilon^3}{3}\right)\right).$$

Возьмем

$$d(\gamma,\varepsilon)=\left\lceil\frac{4\gamma\ln(\ell)}{\frac{\varepsilon^2}{2}-\frac{\varepsilon^3}{3}}\right\rceil,\ \gamma\geqslant 1.$$

Получим оценку для вероятности сильного изменения длины вектора $\mathbf{u}\in\mathbb{R}^n$:

$$\mathsf{P}\left\{\left|\|\mathbf{A}\mathbf{u}\|_2^2-\|\mathbf{u}\|_2^2\right|>\varepsilon\|\mathbf{u}\|_2^2\right\}\leqslant 2\exp\left(-\frac{d(\gamma,\varepsilon)}{2}\left(\frac{\varepsilon^2}{2}-\frac{\varepsilon^3}{3}\right)\right)\leqslant\frac{2}{\ell^{2\gamma}}.$$

Далее можно получить оценку вероятности того, что матрица $\mathbf{A}$ удовлетворяет условию 7:

$$\mathsf{P}\left\{(1-\varepsilon)\|\mathbf{x}'-\mathbf{x}''\|_2^2<\|\mathbf{A}\mathbf{x}'-\mathbf{A}\mathbf{x}''\|_2^2<(1+\varepsilon)\|\mathbf{x}'-\mathbf{x}''\|_2^2\quad\forall\mathbf{x}',\mathbf{x}''\in X\right\}\geqslant$$
$$\geqslant 1-\binom{\ell}{2}\frac{2}{\ell^{2\gamma}}=1-\frac{\ell-1}{\ell^{2\gamma-1}}=1-\ell^{2-2\gamma}+\ell^{1-2\gamma}.$$

Положив $\gamma=1$, получим, что проекционная матрица

$$\mathbf{A}=[\xi_{ij}]^{d\times n},\ \text{где}\ \xi_{ij}\sim\mathcal{N}\left(0,\frac{1}{d(1,\varepsilon)}\right),\ d(1,\varepsilon)=\left\lceil\frac{4\ln(|X|)}{\frac{\varepsilon^2}{2}-\frac{\varepsilon^3}{3}}\right\rceil,$$

удовлетворяет условию 7 с вероятностью не меньшей

$$p\geqslant\frac{1}{|X|}.\tag{8}$$

Заметим, что для генерации случайной матрицы $\mathbf{A}=[\xi_{ij}]^{d\times n}$ можно использовать и другие распределения случайных величин ξ_{ij} с мат. ожиданием $\mathsf{E}\xi_{ij}=0$ и дисперсией $\mathsf{D}\xi_{ij}=\frac{1}{d}$.

Например, в работе [15] рассматривается случай следующего распределения:

$$\xi_{ij} = \begin{cases} +\sqrt{\frac{s}{d}}, & p = \frac{1}{2s}, \\ 0, & p = 1 - \frac{1}{s}, \\ -\sqrt{\frac{s}{d}}, & p = \frac{1}{2s}. \end{cases}$$

Частный случай для $s = 3$ ранее был предложен в работе [16]. С таким распределением ξ_{ij} матрица $\mathbf{A}$ в среднем имеет разреженность $1 - \frac{1}{s}$, что уменьшает сложность метода случайных проекций в s раз.

Заметим также, что многие современные алгоритмы классификации эффективны при работе с разреженными матрицами, поэтому использование метода случайных проекций с такими классификаторами оправдано лишь в тех случаях, когда матрица объектов-признаков $\mathbf{X}$ не является сильно разреженной. Иначе классификация в задаче высокой размерности с разреженной матрицей объектов-признаков может оказаться вычислительно эффективнее классификации в низкоразмерной задаче, где матрица объектов-признаков не является разреженной, т. к. метод случайных проекций не сохраняет разреженность.

Алгоритм

Предлагается следующий алгоритм многоклассовой классификации:

Input: матрица объектов-признаков $\mathbf{X}^{\ell \times n}$

1. Сгенерировать проекционную матрицу $\mathbf{A} \sim \mathcal{N}^{n \times d}\left(0, \frac{1}{d}\right)$.
2. Найти новые описания объектов: $\mathbf{X}' = \mathbf{X}\mathbf{A}$.
3. Решить задачу многоклассовой классификации с новыми описаниями объектов $\mathbf{X}'$.

В следующей главе представлены эксперименты многоклассовой классификации для различных бинарных классификаторов и подходов к сведению многоклассовой задачи к бинарным.

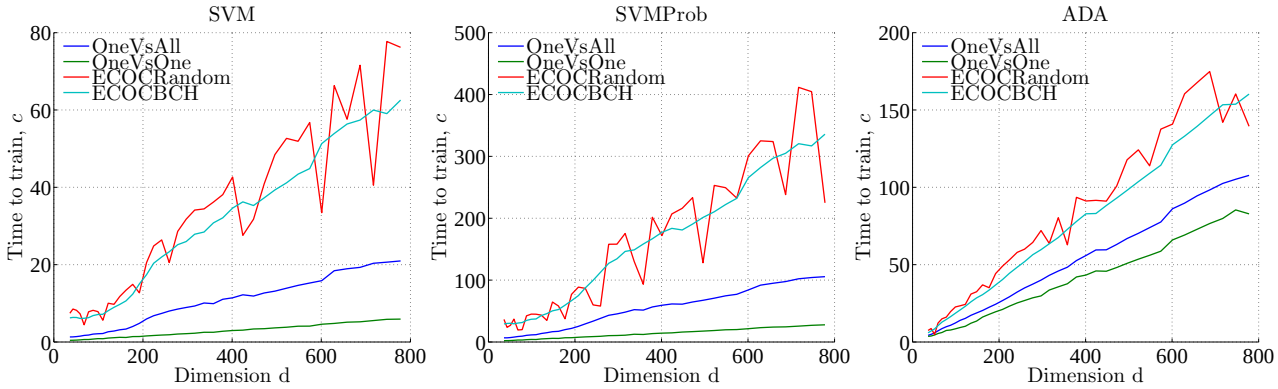
Вычислительный эксперимент

Вычислительный эксперимент проводился с целью демонстрации увеличения производительности бинарных классификаторов SVM и AdaBoost с применением метода случайных проекций как метода снижения размерности задачи многоклассовой классификации. Для сведения исходной многоклассовой задачи к бинарным использовались подходы All-vs-All, One-vs-All и ECOC (BCH и Random). Далее под качеством многоклассовой классификации тестовой выборки X, $|X| = \ell$, будем понимать долю правильных ответов $\frac{1}{\ell}\sum_{i=1}^{\ell} \mathbf{1}[a(x_i) = y(x_i)]$. В данном разделе для краткости метод случайных проекций будем называть RP методом.

MNIST dataset

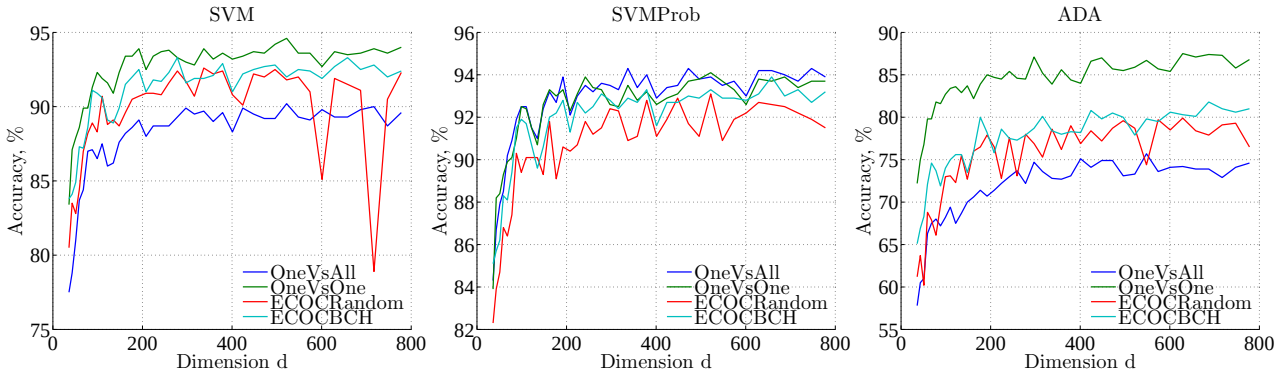
Для вычислительного эксперимента использовалась случайная выборка 2000 объектов для обучения и 1000 объектов для тестирования из базы MNIST [17]:

— of classes: 10
— of data: $60,000$ / $10,000$ (testing)
— of features: 780 / 778 (testing)



(а) Binary classifier: Vote SVM (б) Binary classifier: Probabilistic SVM (в) Binary classifier: AdaBoost

Рис. 1. Зависимость времени обучения бинарных классификаторов Vote SVM, Probabilistic SVM и AdaBoost от размерности редуцированного пространства при одной реализации RP. Data set: MNIST.

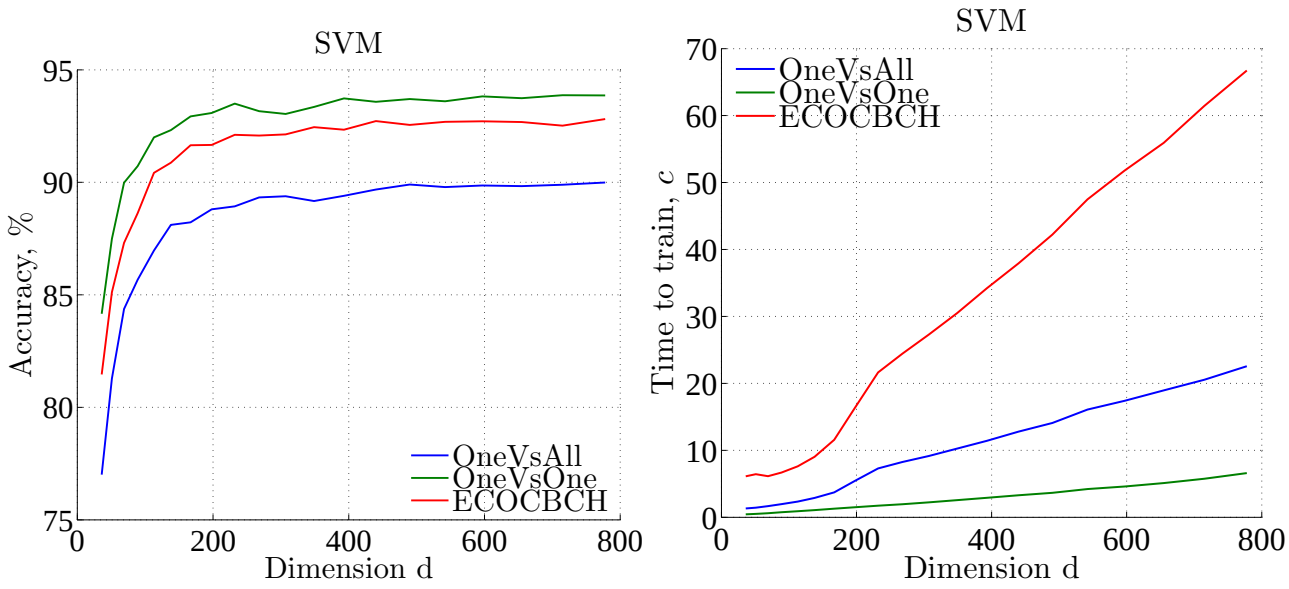


(а) Binary classifier: Vote SVM (б) Binary classifier: Probabilistic SVM (в) Binary classifier: AdaBoost

Рис. 2. Зависимость качества многоклассовой классификации от размерности редуцированного пространства при одной реализации RP. Data set: MNIST.

Задача многоклассовой классификации сводилась к бинарным при помощи подходов OVA, AVA, ECOC-Random (18 столбцов, разреженность 50%) и ECOC-BCH (длина БЧХ кода 15). Задача бинарной классификации решалась классификаторами SVM [18] (Vote SVM), SVMProb [18] (Probabilistic SVM) и AdaBoost [8]. Настройка классификатора SVM производилась с квадратичным ядром и параметрами регуляризации $C = 1$, $\gamma = 1$. В алгоритме AdaBoost использовались 50 базовых классификаторов 3.

Из рисунков 1 и 2 видно, что подход ECOC-BCH оказался предпочтительнее подхода ECOC-Random, так как ECOC-BCH стабильнее по времени обучения и качеству классификации. Можно видеть, что для всех рассмотренных случаев решения задачи многоклассовой классификации зависимость времени обучения от размерности пространства признаков не сильно отличается от линейной. Снизив размерность в 3 раза при помощи RP метода, мы получили тройной прирост в скорости обучения, при этом ошибка возросла на величину порядка $2-3\%$. Рисунок 2 наглядно отображает разброс точности, зависящий от конкретного проецирования.



(а) Binary classifier: Vote SVM (б) Binary classifier: Vote SVM

Рис. 3. Результаты классификации, усредненные по 10 реализациям RP. Data set: MNIST.

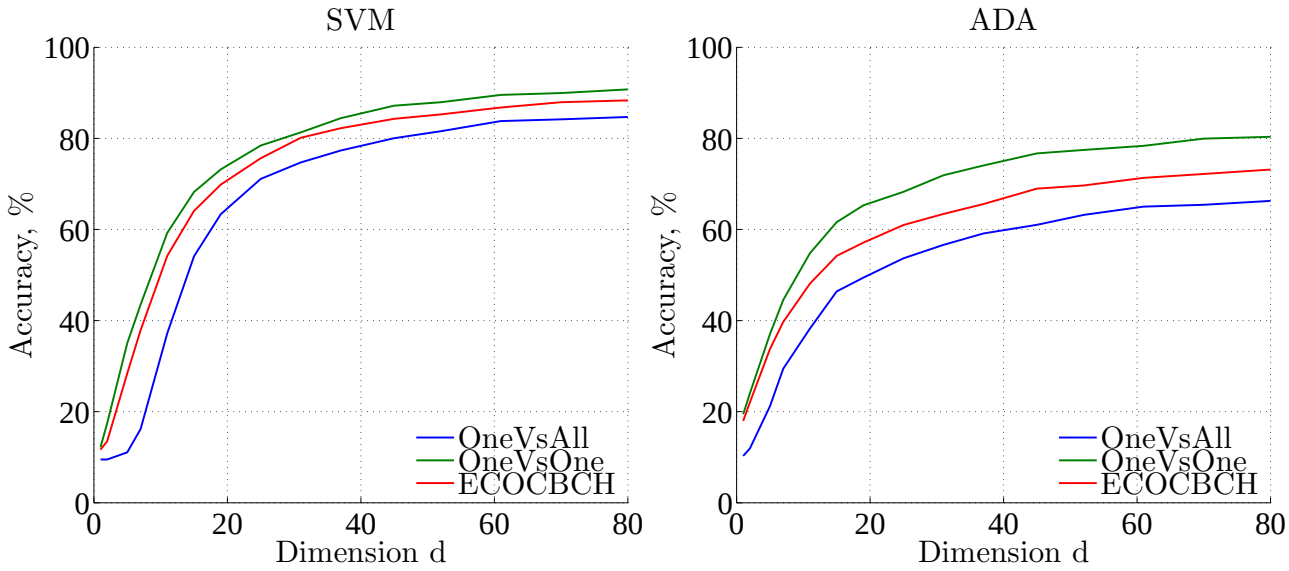
Таблица 1. Качество многоклассовой классификации. Data set: MNIST, Binary classifier: Vote SVM.

Сжатие $\frac{d}{n}$	Подходы											
	One-vs-All			One-vs-One			ECOC-Random			ECOC-BCH		
	min	*avg*	*max*	*min*	*avg*	*max*	*min*	*avg*	*max*	*min*	*avg*	*max*
0.05	76.6	77.3	78.2	82.2	83.3	84.4	75.5	79.6	81.3	81.2	82.1	82.7
0.07	81.2	82.5	85.4	88.3	88.8	89.4	82.9	85.0	86.2	84.6	86.1	87.6
0.11	84.4	85.3	86.4	89.4	90.6	92.5	85.1	87.3	89.9	86.4	88.6	90.1
0.15	86.0	87.0	88.1	91.5	92.1	92.8	85.7	89.2	90.5	89.5	90.1	91.3
0.19	87.5	88.0	88.7	91.1	92.5	93.4	88.8	90.1	91.0	89.7	91.1	92.0
0.25	87.8	88.6	89.2	92.2	92.8	93.5	83.5	89.7	91.3	90.7	91.9	92.8
0.30	88.0	88.8	89.6	92.3	93.2	93.8	89.7	90.6	92.1	90.7	91.8	92.5
0.37	**89.0**	**89.6**	**90.2**	**93.0**	**93.5**	**94.1**	**91.0**	**91.5**	**92.7**	91.3	92.3	93.0
0.44	88.8	89.3	90.1	92.6	93.2	94.0	89.5	**91.3**	92.3	**92.0**	92.3	93.1
0.52	88.9	89.6	90.2	93.1	93.7	94.1	88.8	90.9	92.8	91.8	92.4	**93.8**
0.60	**89.1**	89.8	90.5	93.2	93.7	94.3	**90.5**	**91.5**	92.8	91.4	92.4	93.3
0.69	88.7	89.8	90.5	93.0	**93.8**	94.4	79.4	90.3	92.7	91.9	92.7	93.4
0.79	89.4	89.7	90.4	93.2	**93.8**	94.4	90.4	91.7	92.4	92.3	92.7	**93.7**
0.89	**89.5**	90.0	90.5	**93.4**	93.7	94.1	89.2	91.6	92.7	92.5	92.9	93.5
1.00	89.2	90.0	90.5	93.2	93.7	94.2	91.4	92.2	92.8	92.5	92.9	93.3

На рисунке 3 показаны средние для качества многоклассовой классификации и времени обучения классификаторов Vote SVM.

Детальные результаты с серии 10 реализаций RP представлены в таблице 1.

На рисунке 4 показана усредненная по 20 случайным проекциям зависимость качества классификации для алгоритмов SVM и AdaBoost в сильно редуцированных пространствах

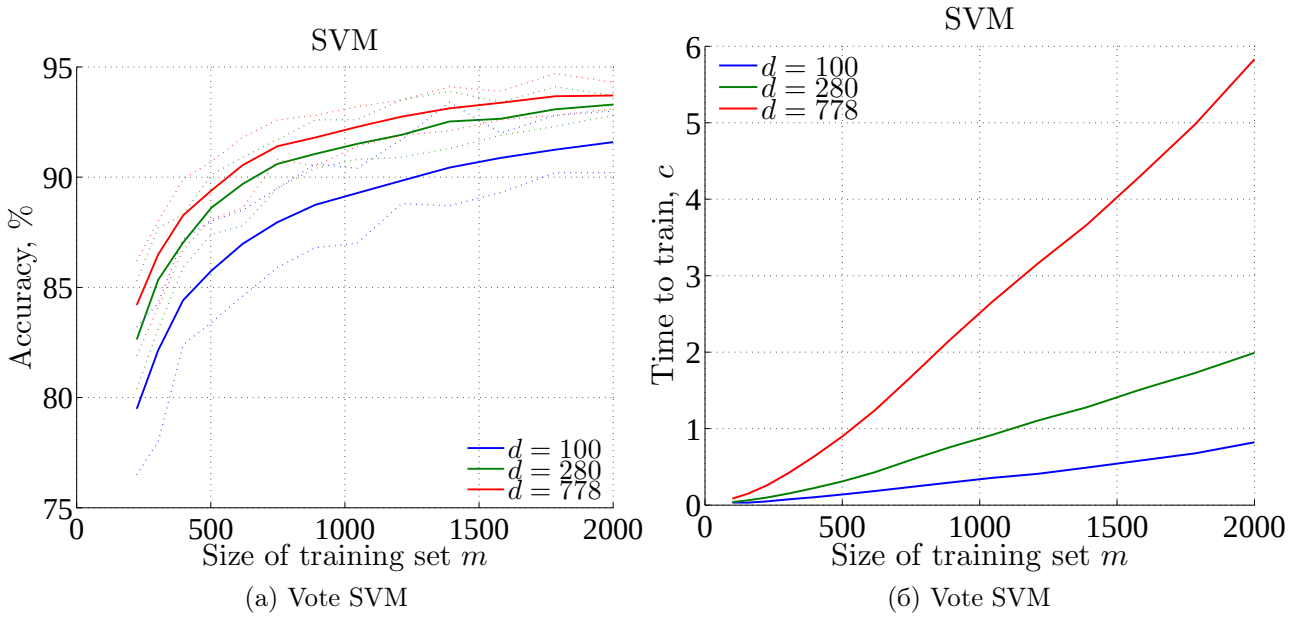


(а) Binary classifier: Vote SVM (б) Binary classifier: AdaBoost

Рис. 4. Качество классификации в сильно редуцированных пространствах, усредненное по 12 реализациям RP. Data set: MNIST.

признаков. На нем видно, что существенные потери в качестве начинаются после сжатия до размерности $d = 30$.

Рисунок 5 показывает устойчивость RP метода относительно мощности обучающей выборки. Тонкими пунктирами обозначены максимальные отклонения от среднего за 20 реализаций RP.



(а) Vote SVM (б) Vote SVM

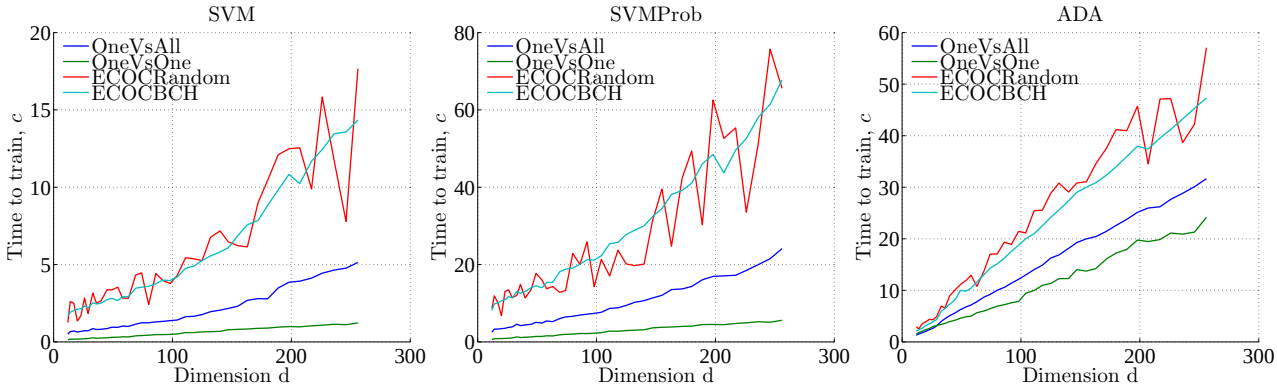
Рис. 5. Зависимость качества многоклассовой классификации от размера обучающей выборки. Усреднение по 20 реализациям RP. Data set: MNIST.

USPS dataset

Использовалась случайная выборка 2000 объектов для обучения и 1000 объектов для тестирования из базы USPS:

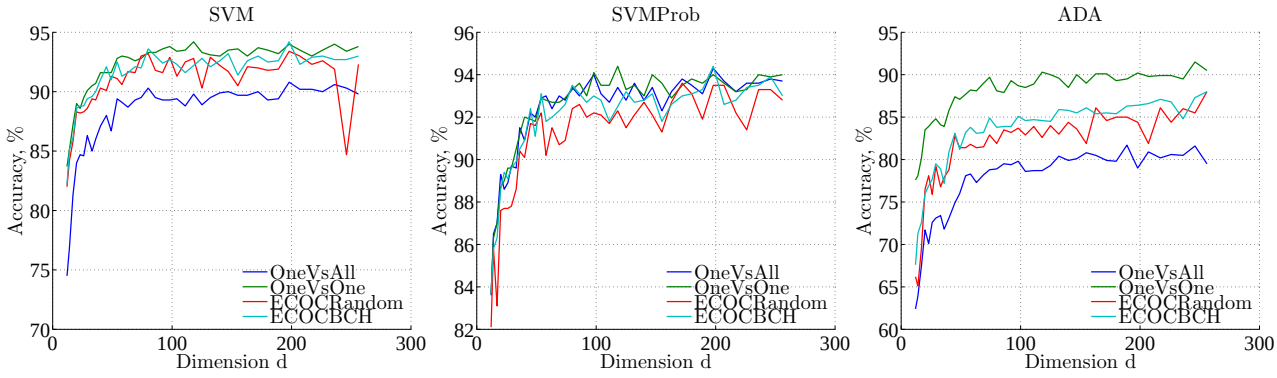
— of classes: 10
— of data: 7,291 / 2,007 (testing)
— of features: 256

В отличие от данных MNIST, матрица объектов-признаков USPS не является разреженной. Многоклассовая задача классификации решалась при тех же условиях, что в эксперименте с данными 1 за исключением того, что ядро бинарного классификатора SVM выбиралось кубическое.



(а) Binary classifier: Vote SVM (б) Binary classifier: Probabilistic SVM (в) Binary classifier: AdaBoost

Рис. 6. Зависимость времени обучения бинарных классификаторов Vote SVM, Probabilistic SVM и AdaBoost от размерности редуцированного пространства при одной реализации RP. Data set: USPS.

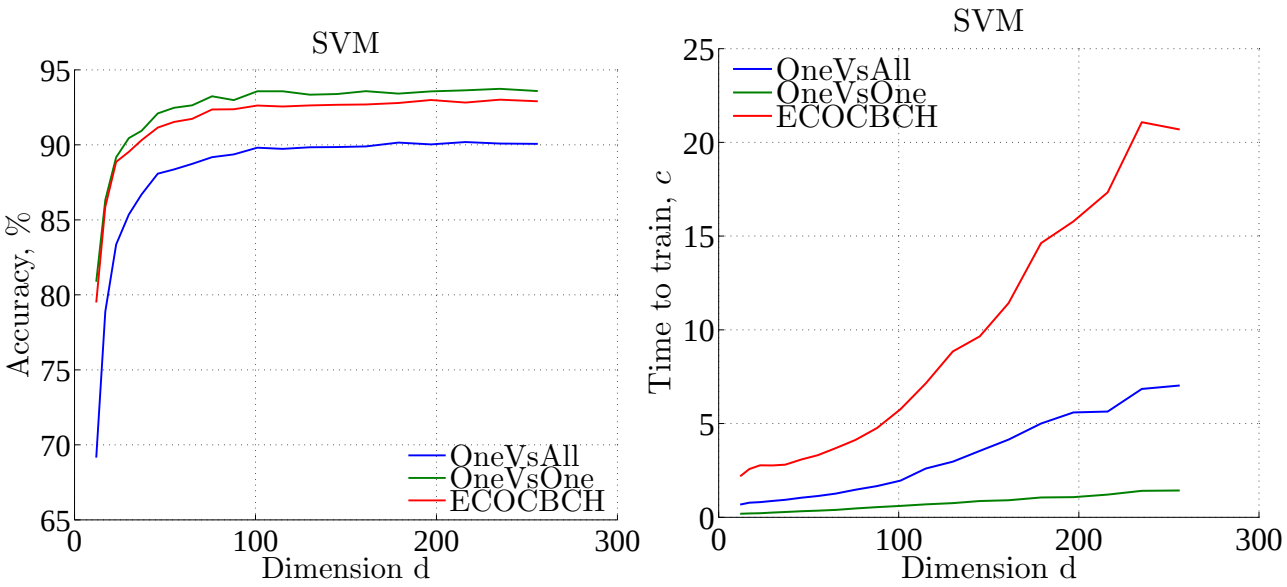


(а) Binary classifier: Vote SVM (б) Binary classifier: Probabilistic SVM (в) Binary classifier: AdaBoost

Рис. 7. Зависимость качества многоклассовой классификации от размерности редуцированного пространства при одной реализации RP. Data set: USPS.

Рисунки 6 и 7, в целом, аналогичны рисункам 1 и 2 из прошлого параграфа.

На рисунке 8 показана зависимость средних для качества многоклассовой классификации и времени обучения классификаторов Vote SVM от размерности редуцированного пространства d.



(а) Binary classifier: Vote SVM (б) Binary classifier: Vote SVM

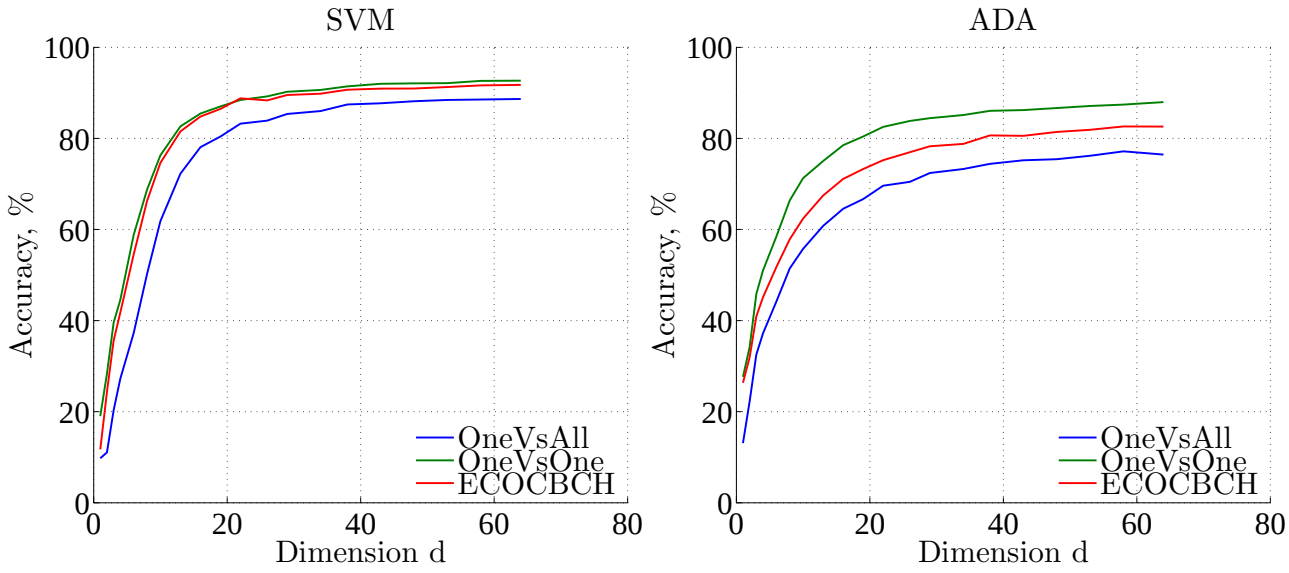
Рис. 8. Результаты классификации, усредненные по 15 реализациям RP. Data set: USPS.

Таблица 2. Качество многоклассовой классификации. Data set: USPS, Binary classifier: Vote SVM.

Сжатие $\frac{d}{n}$	Подходы											
	One-vs-All			One-vs-One			ECOC-Random			ECOC-BCH		
	min	*avg*	*max*	*min*	*avg*	*max*	*min*	*avg*	*max*	*min*	*avg*	*max*
0.05	65.4	69.0	73.8	76.8	80.8	85.6	74.5	78.7	81.0	76.7	79.6	84.7
0.07	78.4	81.4	83.7	84.9	87.1	88.5	83.2	85.5	87.6	85.3	87.0	88.2
0.11	83.7	85.2	87.1	88.9	90.0	91.5	87.1	88.6	90.3	87.9	88.9	90.1
0.15	86.0	87.1	88.7	90.5	91.6	92.5	87.0	89.5	90.8	89.7	90.9	**92.8**
0.20	86.2	87.8	88.9	91.1	92.2	93.1	86.4	90.3	91.9	90.5	91.4	92.6
0.25	87.3	88.6	89.9	91.7	92.6	93.4	**90.1**	91.2	92.5	90.7	91.6	92.5
0.30	88.4	89.2	90.4	92.0	92.9	93.9	85.7	91.1	92.5	**91.1**	92.1	93.0
0.37	88.2	89.3	90.4	92.7	93.2	93.8	**90.9**	91.8	92.9	90.7	92.1	93.0
0.44	88.1	89.6	90.3	92.8	93.4	94.0	**90.3**	91.8	92.7	91.8	92.5	93.1
0.52	**89.2**	89.8	**91.0**	**93.0**	93.6	94.3	86.8	92.2	**93.5**	**92.1**	**92.9**	**93.7**
0.60	88.2	89.7	90.4	92.8	93.5	94.2	**91.0**	92.2	**93.6**	91.8	92.7	**93.6**
0.69	88.7	89.7	90.8	92.8	93.5	94.0	**91.8**	92.3	92.9	91.8	92.6	93.3
0.79	**89.3**	**90.0**	90.8	92.8	93.6	94.4	86.4	91.9	93.1	91.8	92.7	93.3
0.89	88.6	89.9	91.1	93.3	93.7	94.5	**91.6**	**92.4**	93.1	**92.4**	**93.**0	**93.7**
1.00	89.3	90.2	91.2	93.2	93.6	93.9	84.9	91.6	93.3	92.0	92.7	93.3

Детальные результаты с серии 15 реализаций RP представлены в таблице 2.

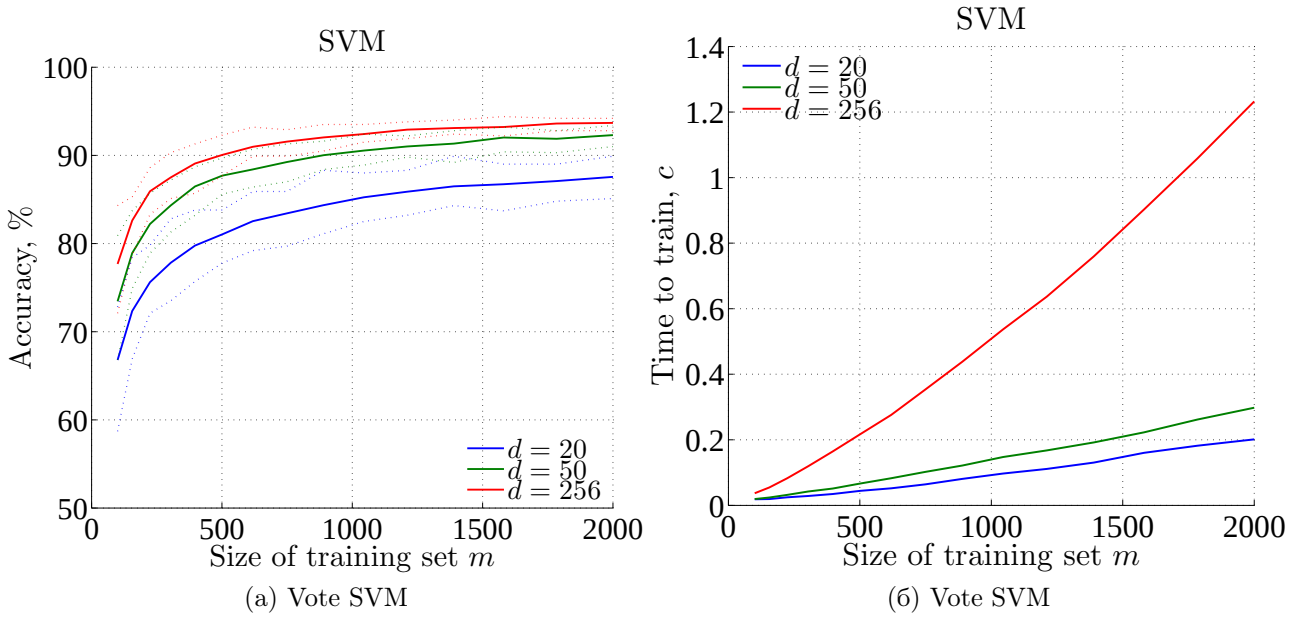
На рисунке 9 показана усредненная по 10 случайным проекциям зависимость качества классификации для алгоритмов SVM и AdaBoost в сильно редуцированных пространствах признаков. На нем видно, что существенные потери в качестве начинаются после сжатия до размерности $d = 20$.



(а) Binary classifier: Vote SVM (б) Binary classifier: AdaBoost

Рис. 9. Качество классификации в сильно редуцированных пространствах, усредненное по 10 реализациям RP. Data set: USPS.

Как и в предыдущем параграфе, рисунок 10 показывает устойчивость RP метода. Тонкими пунктирами обозначены максимальные отклонения от среднего за 40 реализаций RP.



(а) Vote SVM (б) Vote SVM

Рис. 10. Зависимость качества многоклассовой классификации от размера обучающей выборки. Усреднение по 40 реализациям RP. Data set: USPS.

Выводы

В проведенных экспериментах было выяснено, что RP метод снижения размерности наиболее устойчив при использовании подходов One-vs-One и ECOC-BCH. Подход ECOC-Random является наименее устойчивым, что видно на рисунках 2 и 7. Однако

средняя зависимость качества многоклассовой классификации для всех рассмотренных подходов One-vs-All, One-vs-One, ECOC-Random и ECOC-BCH приблизительно одинаковая с точностью до сдвига, продиктованного выбором функции потерь в формуле 5 и геометрией задачи, избирающей приоритетный подход.

Заключение

В работе было показано, что предложенный метод случайных проекций снижения размерности задачи многоклассовой классификации устойчив по отношению к обучающей выборке для всех рассмотренных подходов к сведению многоклассовой задачи классификации к множеству бинарных задач. Метод случайных проекций, как правило, позволяет снизить размерность в два – четыре раза с потерей качества решения многоклассовой задачи порядка 5%. Метод слабо зависит от данных, вычислительно эффективен и достаточно прост в применении, что позволяет использовать его в частности при прототипировании алгоритмов анализа данных.

Результаты экспериментов свидетельствуют о том, что наиболее эффективна работа метода случайных проекций в задачах с полными данными. Существенно, что для всех рассмотренных наборов данных предпочтительный подход к сведению многоклассовой задачи к бинарным можно определить при редуцированной размерности пространства признаков. Таким образом, выбор оптимального подхода и функции потерь для конкретной задачи может проводиться при сильном сжатии пространства признаков, что значительно снижает вычислительные затраты.

Литература

[1] *Fisher R. A.* The use of multiple measurements in taxonomic problems // *Annals of Eugenics*, 1936. Vol. 7, No. 7. P. 179–188.

[2] *Cortes C., Vapnik V.* Support-vector networks // *Machine Learning*, 1995. Vol. 20, No. 3. P. 273–297. Available at: `http://dx.doi.org/10.1023/A:1022627411411`.

[3] *Xia F.* Advanced statistical methods in nlp: Multi-class classification. 2012. Available at: `http://courses.washington.edu/ling572/winter2012/slides/ling572_class13_multiclass.pdf`.

[4] *Rifkin R.* Lecture on multiclass classification. 2008. Available at: `http://www.mit.edu/~9.520/spring08/Classes/multiclass.pdf`.

[5] *Tax D. M. J., Duin R. P. W.* Using two-class classifiers for multiclass classification // ICPR. Vol. 2. 2002. P. 124–127. Available at: `http://dx.doi.org/10.1109/ICPR.2002.1048253`.

[6] *Dietterich T. G., Bakiri G.* Solving multiclass learning problems via error-correcting output codes // *Journal of Artificial Intelligence Research*, 1995. Vol. 2. P. 263–286.

[7] *Allwein E. L., Schapire R. E., Singer Y.* Reducing multiclass to binary: A unifying approach for margin classifiers // *Journal of Machine Learning Research*, 2000. Vol. 1. P. 113–141.

[8] *Escalera S., Pujol O., Radeva P.* Error-correcting ouput codes library // *Journal of Machine Learning Research*, 2010. Vol. 11. P. 661–664. Available at: `http://doi.acm.org/10.1145/1756006.1756026`.

[9] *Freund Y., Schapire R.* A short introduction to boosting // *Journal of Japanese Society for Artificial Intelligence*, 1999. Vol. 14, No. 5. P. 771–780. Available at: `http://citeseer.nj.nec.com/freund99short.html`.

[10] *Freund Y., Schapire R. E.* A decision-theoretic generalization of on-line learning and an application to boosting // *Journal of Computer and System Sciences*, 1997. Vol. 55, No. 1. P. 119–139. Available at: http://www.sciencedirect.com/science/article/pii/S002200009791504X.

[11] *Pearson K.* On lines and planes of closest fit to systems of points in space // *Philosophical Magazine*, 1901. Vol. 2. P. 559–572.

[12] *Golub G. H., Van Loan C. F.* Matrix Computations. 2nd edition. Baltimore: Johns Hopkins University Press, 1989.

[13] *Goel N., Bebis G., Nefian A.* Face recognition experiments with random projection // Proceedings of Society of Photo-Optical Instrumentation Engineers (SPIE) Conference / Ed. by A. K. Jain, N. K. Ratha. Vol. 5779. 2005. P. 426–437.

[14] *Dasgupta S., Gupta A.* An elementary proof of a theorem of johnson and lindenstrauss // *Random Struct. Algorithms*, 2003. Vol. 22, No. 1. P. 60–65. Available at: http://dx.doi.org/10.1002/rsa.10073.

[15] *Li P., Hastie T. J., Church K. W.* Very sparse random projections // Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. KDD '06. New York, NY, USA: ACM, 2006. P. 287–296. Available at: http://doi.acm.org/10.1145/1150402.1150436.

[16] *Achlioptas D.* Database-friendly random projections: Johnson-lindenstrauss with binary coins // *J. Comput. Syst. Sci.*, 2003. Vol. 66, No. 4. P. 671–687.

[17] *LeCun Y., Bottou L., Bengio Y., Haffner P.* Gradient-based learning applied to document recognition // *Proceedings of the IEEE*, 1998. Vol. 86, No. 11. P. 2278–2324. MNIST database available at http://yann.lecun.com/exdb/mnist/.

[18] *Chang C.-C., Lin C.-J.* LIBSVM: A library for support vector machines // *ACM Transactions on Intelligent Systems and Technology*, 2011. Vol. 2, No. 1. P. 27:1–27:27. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm.