НОВЫЕ СРЕДСТВА
КИБЕРНЕТИКИ, ИНФОРМАТИКИ,
ВЫЧИСЛИТЕЛЬНОЙ ТЕХНИКИ
И СИСТЕМНОГО АНАЛИЗА

Д.А. РАЧКОВСКИЙ

УДК 004.22 + 004.93'11

ПРЕОБРАЗОВАНИЕ ВЕКТОРНЫХ ДАННЫХ СЛУЧАЙНЫМИ БИНАРНЫМИ МАТРИЦАМИ

Аннотация. Предложено использование бинарной случайной матрицы с элементами {0,1} для проецирования входных векторов, имеющих формат с плавающей запятой, в выходные векторы того же формата, но сокращенной размерности. Проанализирована точность оценки скалярного произведения, евклидова расстояния, нормы входных векторов по выходным. Аналитически и экспериментально показано, что ошибка оценки для предложенной случайной проекции меньше, чем для тернарной случайной матрицы.

Ключевые слова: бинарные случайные проекции, снижение размерности, оценка сходства векторов.

ВВЕДЕНИЕ

Значительную часть цифровых данных об объектах различной природы можно представить в векторном виде, т.е как набор признаков, полей, компонентов векторов, соответсвующих объектам. Число объектов и размерность векторов-описаний сложных объектов может составлять тысячи, миллионы и миллиарды. Наличие таких объемов данных дает возможность поиска сходной информации, анализа, классификации, выявления закономерностей, решения задач на основе прецедентов и аналогий [1–5]. В то же время большое количество признаков в описаниях многочисленных сложных объектов усложняет хранение, доступ, анализ и понимание данных.

Поиск сходных данных (примеров, прецедентов, аналогов) имеет самостоятельное значение (например, поиск в Интернете), а также является первым этапом как рассуждений человека на основе прецедентов и аналогий, так и при применении этого продуктивного подхода к анализу и пониманию данных в интеллектуальных технологиях и устройствах. Одно из направлений повышения эффективности поиска сходных векторных данных и их обработки заключается в преобразовании исходных представлений в векторы сокращенной размерности, обработка которых давала бы результаты, согласующиеся с результатами для исходных векторов, но при меньших вычислительных затратах.

Адаптируемые к данным методы сокращения размерности, не использующие информацию от учителя (такие, как Principal Component Analysis (PCA) — метод главных компонент и др.) или использующие ее (Linear Discriminant Analysis (LDA) — линейный дискриминантный анализ и др.) [6], являются вычислительно сложными и трансформируют меры сходства–различия исходных векторов при их оценке по векторам сокращенной размерности.

Вычислительно более эффективным подходом, продуцирующим векторы, позволяющие оценить меры сходства–различия исходных векторов, является применение случайных проекций. В этом случае осуществляется преобразование исходных векторов во вторичное векторное пространство путем умножения на

проекционные матрицы, элементы которых — случайно сгенерированные числа из некоторого распределения. (Заметим, что здесь в отличие от традиционного использования математического термина такая случайная проекционная матрица не является идемпотентной и т.п.) Для ряда распределений элементов случайных проекционных матриц по выходным векторам можно оценить некоторые меры сходства и различия исходных векторов. Это показано для проецирования случайными матрицами с гауссовым распределением [7, 8], тернарным распределением с элементами из $\{-1, 0, +1\}$, а также $\{-1, +1\}$ [9–11], устойчивым распределением [12] и др. Случайные проекции используются и в других задачах, например, для эффективного восстановления разреженных сигналов (Compressed Sensing) [13] или для устойчивого решения дискретной некорректной обратной задачи [14].

Очевидно, что самый простой и вычислительно эффективный с точки зрения генерации и использования вариант случайной проекционной матрицы — бинарная случайная матрица с элементами $\{0,1\}$. Однако до сих пор использование таких матриц и свойства полученных с их помощью векторов, насколько известно, не рассматривались. В настоящей статье исследованы свойства сохранения мер сходства–различия векторов при проецировании данными матрицами.

ПРОЕЦИРОВАНИЕ С ПОМОЩЬЮ СЛУЧАЙНОЙ БИНАРНОЙ МАТРИЦЫ

Рассмотрим проецирование векторов случайной бинарной проекционной матрицей $\mathbf{R}$ с элементами r_{ij} из множества $\{0,1\}$. Распределение элементов $\mathbf{R}$ (единиц и нулей) независимо и одинаково (i.i.d.). Значение 1 каждый r_{ij} принимает с вероятностью q, а значение 0 — с вероятностью $1-q$. Обозначим $\mathbf{x}, \mathbf{y}$ входные векторы (размерности D), $\mathbf{u}^{\mathbf{R}} = \mathbf{R}\mathbf{x}$, $\mathbf{v}^{\mathbf{R}} = \mathbf{R}\mathbf{y}$ — результат их проецирования (размерности d). Соответственно $\mathbf{R}$ имеет размерность $D \times d$. Задача — оценить меры сходства–различия $\mathbf{x}, \mathbf{y}$ по $\mathbf{u}$, $\mathbf{v}$. Как и в [10], где рассмотена тернарная проекционная матрица с элементами из $\{-1, 0, +1\}$, будем оценивать величину скалярного произведения $\langle \mathbf{x}, \mathbf{y} \rangle$, квадрата евклидова расстояния $\|\mathbf{x}-\mathbf{y}\|^2$, квадрата евклидовой нормы $\|\mathbf{x}\|^2$, но для проецирования бинарной случайной матрицей.

При проецировании каждый компонент $u_i^{\mathbf{r}}$, $i = 1, \ldots, d$, вектора $\mathbf{u}^{\mathbf{R}}$ формируется в результате скалярного произведения строки $\mathbf{r}_i$ матрицы $\mathbf{R}$ на $\mathbf{x}$: $u_i^{\mathbf{r}} = \langle \mathbf{r}_i, \mathbf{x} \rangle = \sum_{j=1}^{D} r_{ij} x_j$. Вычислим математическое ожидание $E\{u_i^{\mathbf{r}}\}$, где усреднение проводится по различным реализациям строк $\mathbf{r}_i$ для одного и того же (постоянного) входного $\mathbf{x}$:

$$E\{u_i^{\mathbf{r}}\} = E\left\{\sum_{j=1}^{D} r_{ij} x_j\right\} = \sum_{j=1}^{D} x_j E\{r_{ij}\} = q \sum_{j=1}^{D} x_j,$$

так как $E\{r_{ij}\} = 1q + 0(1-q) = q$.

Центрированную случайную величину u_i с $Eu_i = 0$ определим как $u_i = u_i^{\mathbf{r}} - E\{u_i^{\mathbf{r}}\} = u_i^{\mathbf{r}} - q\sum_{j=1}^{D} x_j$. Можно представить u_i в виде

$$u_i = \sum_{j=1}^{D} r_{ij} x_j - q \sum_{j=1}^{D} x_j = \sum_{j=1}^{D} (r_{ij} - q) x_j.$$

Таким образом, такой же результат $\mathbf{u}$ ($\mathbf{v}$) даст и проецирование $\mathbf{u} = \mathbf{P}\mathbf{x}$ ($\mathbf{v} = \mathbf{P}\mathbf{y}$) центрированной случайной матрицей $\mathbf{P} = \mathbf{R} - q$ с элементами $\rho_{ij} = r_{ij} - q$, $E\{\rho_{ij}\} = E\{r_{ij} - q\} = (1-q)q + (0-q)(1-q) = 0$. Анализ проецирования с помощью центрированной $\mathbf{P}$ проще, чем с помощью $\mathbf{R}$, а результаты

одинаковы, если $u_i^{\mathbf{r}}$ преобразовать как $u_i^{\mathbf{r}} - q\sum_{j=1}^{D} x_j \equiv u_i$ и далее работать с u_i. Поэтому будем анализировать проекцию центрированной **P**. Так как ρ_{ij} являются i.i.d., где это уместно, используем ρ вместо ρ_{ij} для компактности. Поскольку $u_i = \sum_{j=1}^{D} \rho_{ij} x_j$ также являются i.i.d., иногда будем рассматривать $u = \sum_{j=1}^{D} \rho_j x_j$, где ρ_j — элементы строки **P**, которая умножается на входной вектор. Введем случайные переменные $\xi_j \equiv x_j \rho_j$, $\zeta_j \equiv y_j \rho_j$. Для них $E\{\xi_j\} = x_j E\{\rho_j\} = 0$; аналогично $E\{\zeta_j\} = 0$. Для $j \neq k$ переменные ξ_j, ζ_k независимы.

Скалярное произведение. Найдем математическое ожидание (м.о.) и дисперсию скалярного произведения $\langle \mathbf{u}, \mathbf{v} \rangle$:

$$E\{\langle \mathbf{u}, \mathbf{v} \rangle\} = \sum_{i=1}^{d} E\{u_i v_i\}.$$

Так как слагаемые $u_i v_i$ — независимые случайные величины (с.в.) для различных i, поэтому дисперсия

$$V\{\langle \mathbf{u}, \mathbf{v} \rangle\} = \sum_{i=1}^{d} V\{u_i v_i\}.$$

Рассмотрим $uv \equiv u_i v_i = \sum_{j=1}^{D} \xi_j \zeta_j + \sum_{j \neq k} \xi_j \zeta_k$. Математическое ожидание

$$E\{uv\} = E\{\sum_{j=1}^{D} \xi_j \zeta_j\} = \sum_{j=1}^{D} x_j y_j E\{\rho_j^2\} = E\{\rho^2\}\langle \mathbf{x}, \mathbf{y} \rangle = q(1-q)\langle \mathbf{x}, \mathbf{y} \rangle,$$

так как ввиду независимости $\sum_{j \neq k} E\{\xi_j \zeta_k\} = \sum_{j \neq k} E\{\xi_j\} E\{\zeta_k\} = 0$ и

$$E\{\rho^2\} = (1-q)^2 q + (0-q)^2 (1-q) = q - q^2.$$

Итак,

$$E\{\langle \mathbf{u}, \mathbf{v} \rangle\} = \sum_{i=1}^{d} E\{u_i v_i\} = q(1-q)\langle \mathbf{x}, \mathbf{y} \rangle d. \tag{1}$$

Поэтому оценка $\langle \mathbf{x}, \mathbf{y} \rangle$ находится из оценки $E^* \langle \mathbf{u}, \mathbf{v} \rangle$ как $E^* \langle \mathbf{u}, \mathbf{v} \rangle / (q(1-q)d)$.

Найдем дисперсию $V\{uv\} = E\{(uv)^2\} - E^2\{uv\}$:

$$(uv)^2 = \left(\sum_{j=1}^{D} \xi_j \zeta_j + \sum_{j \neq k} \xi_j \zeta_k \right)^2 =$$

$$= \left(\sum_{j=1}^{D} \xi_j \zeta_j \right)^2 + 2\left(\sum_{j=1}^{D} \xi_j \zeta_j \right)\left(\sum_{j \neq k} \xi_j \zeta_k \right) + \left(\sum_{j \neq k} \xi_j \zeta_k \right)^2,$$

$$\left(\sum_{j=1}^{D} \xi_j \zeta_j \right)^2 = \sum_{j=1}^{D} \xi_j^2 \zeta_j^2 + \sum_{j \neq k} \xi_j \zeta_j \xi_k \zeta_k,$$

$$\left(\sum_{j \neq k} \xi_j \zeta_k \right)^2 = \sum_{j \neq k} \xi_j^2 \zeta_k^2 + \sum_{j \neq k} \xi_j \zeta_j \xi_k \zeta_k.$$

Найдем $E\{(uv)^2\}$. Так как $2\left(\sum_{j=1}^{D} \xi_j \zeta_j \right)\left(\sum_{j \neq k} \xi_j \zeta_k \right)$ содержит в каждом слагаемом сомножитель в виде независимой с.в. ζ_k $(k \neq j)$ с $E\{\zeta_k\} = 0$, поэтому $E\left\{ 2\left(\sum_{j=1}^{D} \xi_j \zeta_j \right)\left(\sum_{j \neq k} \xi_j \zeta_k \right) \right\} = 0$. Таким образом,

$$E\{(uv)^2\} = \sum_{j=1}^{D} E\{\xi_j^2\zeta_j^2\} + \sum_{j\neq k} E\{\xi_j^2\zeta_k^2\} + 2\sum_{j\neq k} E\{\xi_j\zeta_j\xi_k\zeta_k\} =$$

$$= E\{\rho^4\}\sum_{j=1}^{D} x_j^2 y_j^2 + E^2\{\rho^2\}\sum_{j\neq k} x_j^2 y_k^2 + 2E^2\{\rho^2\}\sum_{j\neq k} x_j y_j x_k y_k.$$

Получаем

$$E^2\{\rho^2\}\left((E\{\rho^4\}/E^2\{\rho^2\})\sum_{j=1}^{D} x_j^2 y_j^2 + \sum_{j\neq k} x_j^2 y_k^2 + 2\sum_{j\neq k} x_j y_j x_k y_k \right) =$$

$$= E^2\{\rho^2\}\left((E\{\rho^4\}/E^2\{\rho^2\} - 3)\sum_{j=1}^{D} x_j^2 y_j^2 + 2\langle \mathbf{x}, \mathbf{y}\rangle^2 + \|\mathbf{x}\|^2 \|\mathbf{y}\|^2 \right),$$

так как

$$\|\mathbf{x}\|^2 \|\mathbf{y}\|^2 = \sum_{j=1}^{D} x_j^2 \sum_{j=1}^{D} y_j^2 = \sum_{j=1}^{D} x_j^2 y_j^2 + \sum_{j\neq k}^{D} x_j^2 y_k^2 ,$$

$$\langle \mathbf{x}, \mathbf{y}\rangle^2 = \left(\sum_{j=1}^{D} x_j y_j \right)^2 = \sum_{j=1}^{D} x_j^2 y_j^2 + \sum_{j\neq k}^{D} x_j y_j x_k y_k.$$

Итак,

$$V\{uv\} = E\{(uv)^2\} - E^2\{uv\} =$$

$$= E^2\{\rho^2\}\left((E\{\rho^4\}/E^2\{\rho^2\} - 3)\sum_{j=1}^{D} x_j^2 y_j^2 + \langle \mathbf{x}, \mathbf{y}\rangle^2 + \|\mathbf{x}\|^2 \|\mathbf{y}\|^2 \right),$$

$$V\{\langle \mathbf{u}, \mathbf{v}\rangle\} = \sum_{i=1}^{d} V\{u_j v_j\} = V\{uv\}d. \tag{2}$$

Евклидово расстояние. Найдем м.о. и дисперсию квадрата евклидова расстояния $\|\mathbf{u} - \mathbf{v}\|^2$:

$$E\{\|\mathbf{u} - \mathbf{v}\|^2\} = \sum_{i=1}^{d} E\{(u_i - v_i)^2\},\ V\{\|\mathbf{u} - \mathbf{v}\|^2\} = \sum_{i=1}^{d} V\{(u_i - v_i)^2\},$$

$$(u-v)^2 = \left(\sum_{j=1}^{D} (\xi_j - \zeta_j) \right)^2 = \sum_{j=1}^{D} (\xi_j - \zeta_j)^2 + \sum_{j\neq k} (\xi_j - \zeta_j)(\xi_k - \zeta_k),$$

$$E\{(u-v)^2\} = \sum_{j=1}^{D} E\{(\xi_j - \zeta_j)^2\} + \sum_{j\neq k} E\{(\xi_j - \zeta_j)(\xi_k - \zeta_k)\}.$$

Так как $\xi_j - \zeta_j$ и $\xi_k - \zeta_k$ независимы для $j \neq k$ и $E\{(\xi_j - \zeta_j)\} = 0$, то

$$\sum_{j\neq k} E\{(\xi_j - \zeta_j)(\xi_k - \zeta_k)\} = \sum_{j\neq k} E\{(\xi_j - \zeta_j)\} E\{(\xi_k - \zeta_k)\} = 0.$$

Поэтому

$$E\{(u-v)^2\} = \sum_{j=1}^{D} E\{(\xi_j - \zeta_j)^2\} = \sum_{j=1}^{D} E\{\rho_j^2\}(x_j - y_j)^2 = (q - q^2)\|\mathbf{x} - \mathbf{y}\|^2 ,$$

$$E\{\|\mathbf{u} - \mathbf{v}\|^2\} = (q - q^2)\|\mathbf{x} - \mathbf{y}\|^2 \, d. \tag{3}$$

Определим дисперсию $V\{(u-v)^2\} = E\{(u-v)^4\} - E^2\{(u-v)^2\}$:

$$(u-v)^4 = \left(\sum_{j=1}^{D} (\xi_j - \zeta_j)^2 + \sum_{j\neq k} (\xi_j - \zeta_j)(\xi_k - \zeta_k) \right)^2 =$$

$$= \left(\sum_{j=1}^{D} (\xi_j - \zeta_j)^2 \right)^2 + 2\left(\sum_{j=1}^{D} (\xi_j - \zeta_j)^2 \right)\left(\sum_{j\neq k} (\xi_j - \zeta_j)(\xi_k - \zeta_k) \right) +$$

$$+ \left(\sum_{j\neq k} (\xi_j - \zeta_j)(\xi_k - \zeta_k) \right)^2 ,$$

$$\left(\sum_{j=1}^{D} (\xi_j - \zeta_j)^2 \right)^2 = \sum_{j=1}^{D} (\xi_j - \zeta_j)^4 + \sum_{j \neq k} (\xi_j - \zeta_j)^2 (\xi_k - \zeta_k)^2,$$

$$\left(\sum_{j \neq k} (\xi_j - \zeta_j)(\xi_k - \zeta_k) \right)^2 = 2 \sum_{j \neq k} (\xi_j - \zeta_j)^2 (\xi_k - \zeta_k)^2 + \sum \ldots .$$

Найдем $E\{(u-v)^4\}$. При $D \geq 2$ выражение

$$2 \left(\sum_{j=1}^{D} (\xi_j - \zeta_j)^2 \right) \left(\sum_{j \neq k} (\xi_j - \zeta_j)(\xi_k - \zeta_k) \right)$$

содержит в каждом слагаемом сомножитель в виде независимой с.в. $(\xi_k - \zeta_k)$ $(k \neq j)$, поэтому

$$E \left\{ 2 \left(\sum_{j=1}^{D} (\xi_j - \zeta_j)^2 \right) \left(\sum_{j \neq k} (\xi_j - \zeta_j)(\xi_k - \zeta_k) \right) \right\} = 0,$$

так как каждое слагаемое включает сомножитель $E\{(\xi_k - \zeta_k)\} = E\{\xi_k\} - E\{\zeta_k\} = 0.$ Аналогично $E\{\sum \ldots\} = 0$. Таким образом,

$$E\{(u-v)^4\} = \sum_{j=1}^{D} E\{(\xi_j - \zeta_j)^4\} + 3 \sum_{j \neq k} E\{(\xi_j - \zeta_j)^2 (\xi_k - \zeta_k)^2\} =$$

$$= E\{\rho^4\} \sum_{j=1}^{D} (x_j - y_j)^4 + 3E^2\{\rho^2\} \sum_{j \neq k} (x_j - y_j)^2 (x_k - y_k)^2 =$$

$$= E^2\{\rho^2\} \left((E\{\rho^4\} / E^2\{\rho^2\} - 3) \sum_{j=1}^{D} (x_j - y_j)^4 + 3 \| \mathbf{x} - \mathbf{y} \|^4 \right),$$

так как

$$\| \mathbf{x} - \mathbf{y} \|^4 = (\| \mathbf{x} - \mathbf{y} \|^2)^2 = \left(\sum_{j=1}^{D} (x_j - y_j)^2 \right)^2 =$$

$$= \sum_{j=1}^{D} (x_j - y_j)^4 + \sum_{j \neq k} (x_j - y_j)^2 (x_k - y_k)^2.$$

Получаем

$$V\{(u-v)^2\} = E\{(u-v)^4\} - E^2\{(u-v)^2\} =$$

$$= E^2\{\rho^2\} \left((E\{\rho^4\} / E^2\{\rho^2\} - 3) \sum_{j=1}^{D} (x_j - y_j)^4 + 2 \| \mathbf{x} - \mathbf{y} \|^4 \right),$$

$$V\{\| \mathbf{u} - \mathbf{v} \|^2\} = \sum_{i=1}^{d} V\{(u_i - v_i)^2\} = V\{(u-v)^2\} d. \qquad (4)$$

Анализ оценок. Выражения для м.о. и дисперсии квадрата евклидовой нормы вектора получаются из формул (1), (2) или (3), (4) как $\langle \mathbf{u}, \mathbf{u} \rangle$ или $\| \mathbf{u} - 0 \|^2$.

Для сравнения ошибки оценок скалярного произведения, евклидова расстояния, нормы исходных векторов по векторам после случайной проекции использовалось относительное среднеквадратичное отклонение (коэффициент вариации) $V^{1/2} / E$:

$$V^{1/2}\{\langle \mathbf{u}, \mathbf{v}\rangle\}/E\{\langle \mathbf{u}, \mathbf{v}\rangle\} =$$

$$= \frac{1}{\langle \mathbf{x}, \mathbf{y}\rangle\sqrt{d}}\left((E\{\rho^4\}/E^2\{\rho^2\}-3)\sum_{j=1}^{D} x_j^2 y_j^2 + \langle \mathbf{x}, \mathbf{y}\rangle^2 + \|\mathbf{x}\|^2 \|\mathbf{y}\|^2 \right)^{1/2},$$

$$V^{1/2}\{\|\mathbf{u}-\mathbf{v}\|^2\}/E\{\|\mathbf{u}-\mathbf{v}\|^2\} =$$

$$= \frac{1}{\|\mathbf{x}-\mathbf{y}\|^2\sqrt{d}}\left((E\{\rho^4\}/E^2\{\rho^2\}-3)\sum_{j=1}^{D} (x_j - y_j)^4 + 2\|\mathbf{x}-\mathbf{y}\|^4 \right)^{1/2}. \quad (5)$$

Для рассматриваемой бинарной случайной проекционной матрицы имеем

$$E\{\rho^4\} = (1-q)^4 q + (0-q)^4 (1-q) = (q-q^2)(1-3(q-q^2)),$$

$$E\{\rho^4\}/E^2\{\rho^2\} = (q-q^2)(1-3(q-q^2))/(q-q^2)^2 = 1/(q-q^2)-3.$$

Для тернарной случайной проекционной матрицы с элементами $-1/\sqrt{q}$ (с вероятностью $q/2$), $+1/\sqrt{q}$ (с вероятностью $q/2$), 0 (с вероятностью $1-q$):

$$E\{\rho^2\} = (1/\sqrt{q})^2 q/2 + (-1/\sqrt{q})^2 q/2 = 1,$$

$$E\{\rho^4\} = (1/\sqrt{q})^4 q/2 + (-1/\sqrt{q})^4 q/2 = 1/q,\ E\{\rho^4\}/E^2\{\rho^2\} = 1/q$$

(см. также [10]). При $q < 2/3$ справедливо $1/(q-q^2)-3 < 1/q$, поэтому ошибка оценок для предлагаемых бинарных случайных проекций меньше, чем для тернарных при одинаковой вероятности q ненулевого элемента матрицы.

Для случайной проекционной матрицы с элементами из гауссова распределения [8, 10] имеем

$$V^{1/2}\{\langle \mathbf{u}, \mathbf{v}\rangle\}/E\{\langle \mathbf{u}, \mathbf{v}\rangle\} =$$

$$= \frac{1}{\langle \mathbf{x}, \mathbf{y}\rangle\sqrt{d}}(\langle \mathbf{x}, \mathbf{y}\rangle^2 + \|\mathbf{x}\|^2 \|\mathbf{y}\|^2)^{1/2},\ V^{1/2}\{\|\mathbf{u}-\mathbf{v}\|^2\}/E\{\|\mathbf{u}-\mathbf{v}\|^2\} = \sqrt{\frac{2}{d}}. \quad (6)$$

Сравнивая (5) и (6), видим, что так как $1/(q-q^2)-3-3<0$ при $q \approx [0.2113; 0.7887]$ (т.е. при $1/2-1/(2\sqrt{3}) < q < 1/2+1/(2\sqrt{3})$), в этом диапазоне бинарные случайные проекции обеспечивают точность выше гауссовых (наилучший результат достигается при $q = 0.5$). С другой стороны, ускорение проецирования требует $q << 0.5$, где бинарные проекции не столь эффективны из-за наличия в ошибке слагаемых с положительными коэффициентами при $\sum_{j=1}^{D} x_j^2 y_j^2$ и $\sum_{j=1}^{D} (x_j - y_j)^4$. Однако при $D >> 1$ их вклад мал (для данных с конечным четвертым моментом), поэтому получаем точность, сравнимую с точностью гауссовых случайных проекций, и для случая $q << 0.5$.

ЭКСПЕРИМЕНТАЛЬНОЕ ИССЛЕДОВАНИЕ

Поведение ошибки $V^{1/2}/E$ исследовалось для квадрата евклидовых нормы и расстояния, а также для скалярного произведения векторов при проецировании случайными бинарными (элементы $\{0,1\}$) и тернарными (элементы $\{-1, 0, +1\}$) матрицами с различными параметрами. Экспериментально полученные ошибки (вычисленные по выборочным средним и дисперсиям) сравнивались с аналитическими выражениями, включая ошибку для гауссовых случайных проекций.

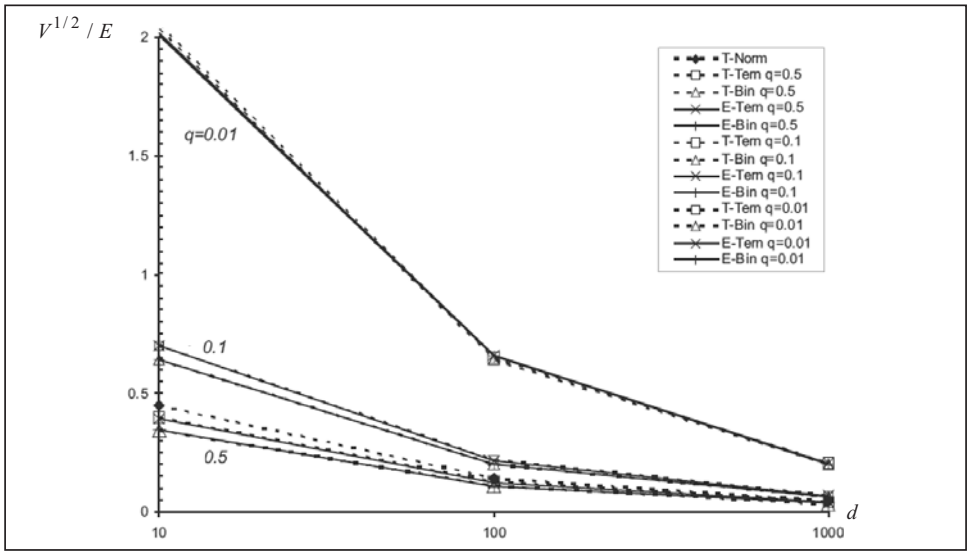

Рис. 1. Зависимость ошибки $V^{1/2}/E$ оценки квадрата евклидова расстояния между входными векторами от размерности d векторов после проекции. Размерность входных векторов $D = 10$

Исследовались матрицы при $q = \{0.5, 0.1, 0.01\}$ (для бинарных матриц q — вероятность 1; для тернарных матриц вероятность 1 и -1 одинакова и равна $q/2$). Использовались входные векторы размерности $D = \{10, 100, 1000, 10000\}$. Их компоненты генерировались из равномерного распределения в $[-D, +D]$, сходство варьировалось путем конкатенации разных долей одинаковых и различных векторов. Заметим, что малые $D = \{10, 100\}$ обычно не интересны для практических приложений и исследовались для иллюстрации особенностей поведения ошибки и подтверждения правильности аналитических выражений. Выходные векторы получали с размерностью $d = \{10, 100, 1000\}$. Для каждого исследованного набора параметров проецирования экспериментальные V, E получены усреднением по 10000 реализациям выходных векторов, соответствующих 10000 реализациям случайных матриц (по 1000 реализаций для $D = 10000, d = 1000$).

Результаты всех экспериментов при всех значениях параметров (и для различных величин сходства между входными векторами) близки к теоретическим. На рис. 1 приведены зависимости ошибки оценки евклидова расстояния между (двумя фиксированными) входными векторами размерности $D = 10$ от размерности векторов после проекции, по которым оценивалось это расстояние. Здесь линии, обозначения которых начинаются с символа T, соответствуют аналитическим выражениям, а с символа E — экспериментально полученным результатам, Norm — гауссова проекционная матрица, Tern — тернарная, Bin — предложенная бинарная; q — вероятность ненулевых элементов в матрице.

Таблица 1

Виды случайных проекций	Значение $V^{1/2}/E$		
	d =10	d =100	d =1000
T-Norm	0.447214	0.141421	0.044721
T-Tern q =0.5	0.447173	0.141409	0.044717
T-Bin q =0.5	0.447133	0.141396	0.044713
E-Tern q =0.5	0.442526	0.140829	0.04454
E-Bin q =0.5	0.448362	0.142213	0.04466
T-Tern q =0.1	0.447495	0.14151	0.044749
T-Bin q =0.1	0.447419	0.141486	0.044742
E-Tern q =0.1	0.441405	0.141205	0.044144
E-Bin q =0.1	0.446487	0.14008	0.044195
T-Tern q =0.01	0.451096	0.142649	0.04511
T-Bin q =0.01	0.451017	0.142624	0.045102
E-Tern q =0.01	0.452649	0.142631	0.044639
E-Bin q =0.01	0.44236	0.141008	0.045312

Ошибки значительно отличаются для различных значений q при одинаковом d. Для $q = 0.5$ ошибка для гауссовой проекционной матрицы выше, чем для тернарной и бинарной, для $q = 0.01$ — заметно ниже. Для $D = 100$ отличия между значениями ошибок уменьшаются, а для $D = 1000$ становятся незначительными. Для $D = 10000$ отличия в значениях ошибок для гауссовой, тернарной и бинарной случайных проекционных матриц очень малы (табл.1). Для всех исследованных q ошибка для бинарных случайных проекций меньше, чем для тернарных, что соответствует аналитическим оценкам (для исследованных параметров отличия незначительны). Аналогичные результаты получены для скалярного произведения и квадрата нормы векторов.

По сравнению с гауссовой проекционной матрицей бинарная матрица требует в 32–64 раза меньше памяти при использовании для представления 1 бита на элемент матрицы вместо 32–64 бит для представления гауссовых случайных величин в формате с плавающей запятой.

ЗАКЛЮЧЕНИЕ

Для преобразования входных векторов в формате с плавающей запятой в выходные в аналогичном формате предложено использовать проецирование случайной матрицей с бинарными элементами $\{0, +1\}$. Выходные векторы позволяют оценивать меры сходства–различия исходных векторов (евклидово расстояние и скалярное произведение, а также евклидову норму); вычислительная эффективность оценки повышается при уменьшении размерности выходных векторов.

Аналитически и экспериментально исследована ошибка оценивания мер сходства–различия. Как и для других типов случайных проекционных матриц, ошибка уменьшается при росте размерности d выходных векторов ($\sim 1/\sqrt{d}$). При фиксированном d для вероятности q в окрестности 0.5 в бинарной проекционной матрице (так называемые «плотные» бинарные матрицы) ошибка оценки мер сходства–различия меньше, чем для гауссовой случайной проекционной матрицы. При этом генерация бинарных случайных величин и операция умножения для реализации проецирования могут выполняться значительно эффективнее, чем для гауссовых случайных величин, за счет отсутствия необходимости умножения чисел с плавающей запятой. Вычислительная эффективность может быть еще выше при уменьшении q (росте «разреженности» бинарной проекционной матрицы). При этом для сохранения точности оценок входные векторы должны иметь достаточно большую размерность, что и предполагается при постановке задачи эффективной оценки сходства многомерных векторов.

По сравнению со случайной матрицей с тернарными элементами бинарная матрица дает меньшую ошибку, обеспечивая при этом более простую генерацию случайной матрицы и реализацию проекции.

Перспективно исследование вопросов эффективной реализации проецирования, а также применимости бинарной проекционной матрицы при получении бинарных выходных векторов (аналогично тому, как исследовано в [11] для тернарной матрицы). Такие отражающие сходство векторы являются примером нейросетевых рандомизированных распределенных представлений, которые могут не только использоваться для эффективного поиска по сходству [15–18], но и запоминаться и обрабатываться в ассоциативных нейронных сетях [19–29], а также служить элементами представлений иерархически структурированных моделей сложных объектов внешнего мира различной природы [30–37].

СПИСОК ЛИТЕРАТУРЫ

1. Гриценко В.И., Рачковский Д.А., Гольцев А.Д., Лукович В.В., Мисуно И.С., Ревунова Е.Г., Слипченко С.В., Соколов А.М. Нейросетевые распределенные представления для интеллектуальных информационных технологий и моделирования мышления // Кибернетика и вычислительная техника. — 2013. — Вып. 173. — С. 7–24.
2. Rachkovskij D.A. Representation and processing of structures with binary sparse distributed codes // IEEE Trans. on Knowledge and Data Engineering. — 2001. — **13**, N 2. — P. 261–276.
3. Rachkovskij D.A. Some approaches to analogical mapping with structure sensitive distributed representations // J. Experimental and Theoretical Artificial Intelligence. — 2004. — **16**, N 3. — P. 125–145.
4. Slipchenko S.V., Rachkovskij D.A. Analogical mapping using similarity of binary distributed representations // International J. Information Theories and Applications. — 2009. — **16**, N 3. — P. 269–290.
5. Rachkovskij D.A., Slipchenko S.V. Similarity-based retrieval with structure-sensitive sparse binary distributed representations // Computational Intelligence. — 2012. — **28**, N 1. — P. 106–129.
6. Burges C.J.C. Dimension Reduction: A Guided Tour // Foundations and Trends in Machine Learning. — 2010. — **2**, N 4. — P. 275–365.
7. Indyk P., Motwani R. Approximate nearest neighbors: towards removing the curse of dimensionality // Proc. of 30th ACM Symposium on Theory of Computing. — 1998. — P. 604–613.
8. Vempala S.S. The Random Projection Method. — Providence, RI: American Mathematical Society, 2004. — 105 p.
9. Achlioptas D. Database-friendly random projections: Johnson–Lindenstrauss with binary coins // J. Computer and System Sciences. — 2003. — **66**, N 4. — P. 671–687.
10. Li P., Hastie T.J., Church K.W. Very sparse random projections // 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — Philadelphia, PA, USA: ACM Press, 2006. — P. 287–296.
11. Rachkovskij D.A., Misuno I.S., Slipchenko S.V. Randomized projective methods for construction of binary sparse upsilon ector representations // Cybernetics and Systems Analysis. — 2012. — **48**, N 1. — P. 146–156.
12. Li P. Very sparse stable random projections for dimension reduction in $l_\alpha (0 < \alpha \le 2)$ norm // 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — San Jose, CA, USA: ACM Press, 2007. — P. 440–449.
13. Donoho D.L. Compressed sensing // IEEE Trans. Informational Theory. — 2006. — **52**, N 4. — P. 1289–1306.
14. Rachkovskij D.A., Revunova E.G. Randomized method for solving discrete ill-posed problems // Cybernetics and Systems Analysis. — 2012. — **48**, N 4. — P. 621–635.
15. Kanerva P., Sjodin G., Kristoferson J., Karlsson R., Levin B., Holst A., Karlgren J. and Sahlgren M. Computing with large random patterns // In: Foundations of Real-World Intelligence. — CSLI Publications, Stanford, California, 2001. — P. 251–311.
16. Мисуно И.С., Рачковский Д.А., Слипченко С.В. Векторные и распределенные представления, отражающие меру семантической связи слов // Математические машины и системы. — 2005. — № 3. — С. 50–67.
17. Sokolov A. LIMSI: learning semantic similarity by selecting random word subsets // Proceedings of the Sixth International Workshop on Semantic Evaluation (SEMEVAL'12). — Association for Computational Linguistics, 2012. — P. 543–546.
18. Sokolov A., Riezler S. Task-driven greedy learning of feature hashing functions // Proceedings of the NIPS'13 Workshop "Big Learning: Advances in Algorithms and Data Management", Lake Tahoe, USA, 2013. — P. 1–5.

19. Frolov A., Kartashov A., Goltsev A., Folk R. Quality and efficiency of retrieval for Willshaw-like autoassociative networks. I. Correction // Network: Computation in Neural Systems. — 1995. — **6**, N 4. — P. 513–534

20. Frolov A., Kartashov A., Goltsev A., Folk R. Quality and efficiency of retrieval for Willshaw-like autoassociative networks. II. Recognition // Network: Computation in Neural Systems. — 1995. — **6**, N 4. — P. 535–549.

21. Frolov A.A., Husek D., Muraviev I.P. Informational capacity and recall quality in sparsely encoded Hopfield-like neural network: Analytical approaches and computer simulation // Neural Networks. — 1997. — **10**, N 5. — P. 845–855.

22. Frolov A.A., Husek D., Polyakov P.Yu. Recurrent-neural-network-based boolean factor analysis and its application to word clustering // IEEE Transactions on Neural Networks. — 2009. — **20**, N 7. — P. 1073–1086.

23. Frolov A.A., Rachkovskij D.A., Husek D. On informational characteristics of sparsely encoded binary auto-associative memory // 9-th International Conference on Neural Information Processing ICONIP'02. — Orchid Country Club, Singapore. — 2002. — P. 235–238.

24. Frolov A.A., Rachkovskij D.A., Husek D. On information characteristics of Willshaw-like auto-associative aemory // Neural Network World. — 2002. — **12**, N 2. — P. 141–158.

25. Frolov A. A., Husek D., Rachkovskij D.A. Time of searching for similar binary vectors in associative memory // Cybernetics and Systems Analysis. — 2006. — **42**, N 5. — P. 615–623.

26. Nowicki D.W., Dekhtyarenko O.K. Averaging on Riemannian manifolds and unsupervised learning using neural associative memory // Proc. ESANN 2005. — Bruges, Belgium, April, 27–29, 2005. — P. 181–189.

27. Nowicki D., Siegelmann H. Flexible kernel memory // PLoS one. — **5**, N 6. — e10955. doi:10.1371/journal.pone.0010955

28. Nowicki D., Verga P., Siegelmann H. Modeling reconsolidation in kernel associative memory // PloS one. — 2013. — **8**, N 8. — e68189. doi:10.1371/journal.pone.0068189

29. Emruli B., Gayler R.W., Sandin F. Analogical mapping and inference with binary spatter codes and sparse distributed memory // International Joint Conference on Neural Networks (IJCNN), August, 4–9, 2013. — 2013. — P. 1–8.

30. Kussul E.M., Rachkovskij D.A. Multilevel assembly neural architecture and processing of sequences // In A.V. Holden & V. I. Kryukov (Eds.), Neurocomputers and Attention: Vol. II. Connectionism and neurocomputers. Manchester and New York: Manchester University Press, 1991. — P. 577–590.

31. Амосов Н.М., Байдык Т.Н., Гольцев А.Д., Касаткин А.М., Касаткина Л.М., Куссуль Э.М., Рачковский Д.А. Нейрокомпьютеры и интеллектуальные роботы. — Киев: Наук. думка, 1991. — 269 с.

32. Rachkovskij D.A., Slipchenko S.V., Kussul E.M., Baidyk T.N. Binding procedure for distributed binary data representations // Cybernetics and Systems Analysis. — 2005. — **41**, N3. — P. 319–331.

33. Letichevsky A., Godlevsky A., Letichevsky A. Jr, Potienko S., Peschanenko V. The properties of predicate transformer in VRS system // Cybernetics and Systems Analysis. — 2010. — **46**, N 4. — P. 521–532.

34. Letichevsky A., Letychevsky A. Jr, Peschanenko V. Insertion modeling system // Lecture Notes in Comput. Sci. — 2011. — **7162**. — P. 262–274.

35. Gallant S.I., Okaywe T.W. Representing objects, relations, and sequences // Neural computation. — 2013. — **25**, N 8. — P. 2038–2078.

36. Rachkovskij D.A., Kussul E.M., Baidyk T.N. Building a world model with structure-sensitive sparse binary distributed representations // Biologically Inspired Cognitive Architectures. — 2013. — **3**. — P. 64–86.

37. Letichevsky A.A. Theory of interaction, insertion modeling, and cognitive architectures // Biologically Inspired Cognitive Architectures. — 2014. — **8**. — P. 19–32.

Поступила 24.12.2013