УДК 004.22 + 004.93’11

Д.А. РАЧКОВСКИЙ

ОЦЕНКА СХОДСТВА ВЕКТОРОВ ПО ИХ РАНДОМИЗИРОВАННЫМ БИНАРНЫМ ПРОЕКЦИЯМ

Аннотация. Исследована оценка угла, скалярного произведения и евклидова расстояния вещественных векторов по бинарным векторам с регулируемой разреженностью. Преобразование проведено проецированием с применением бинарной случайной матрицы с элементами $\{0, 1\}$ и выходного порогового преобразования. Дан сравнительный анализ ошибки оценки мер сходства входных векторов по ряду мер сходства выходных бинарных векторов на основе скалярного произведения

Ключевые слова: бинарные случайные проекции, разреженные бинарные представления, оценка сходства векторов.

ВВЕДЕНИЕ

Векторное представление данных является, пожалуй, наиболее распространенным в информационных технологиях. Многие методы поиска по сходству, классификации, регрессии, кластеризации, а также методы восстановления величин по результатам измерений предназначены для работы именно с векторными данными. В то же время увеличиваются объемы сложноструктурированных данных, например XML-структур или ориентированных ациклических графов баз знаний на языках KIF, CycL, включая представления эпизодов в моделях рассуждений по аналогии [1–11]. Однако методы оперирования сложноструктурированными данными вычислительно сложны. Для повышения эффективности обработки их преобразуют в векторы, позволяющие оценить важные в контексте решаемой задачи характеристики (например, меры сходства) и на основе этих оценок решить исходную задачу (см. [7, 12] и ссылки к ним).

Сложность обработки векторных данных пропорциональна размерности и числу векторов. Для работы с большими массивами векторов большой размерности используют ряд подходов. Наиболее очевидным является применение метода «грубая сила» — мощного вычислительного ресурса. Так, эффективная реализация векторных операций поддерживается векторными (со)процессорами и, кроме того, эти операции естественным образом распараллеливаются. Это позволяет использовать параллельные вычислительные средства — многоядерные процессоры, параллельные компьютеры, а также высокопроизводительные вычислительные кластеры и системы распределенных вычислений (см. [13–15] и ссылки к ним).

Другой подход к повышению эффективности обработки векторных данных базируется на том, что для многих применений не требуется точных результатов операций над исходными векторами (например, вычисления скалярных произведений, расстояний и т.п.), достаточно приближенной, но быстрой оценки. Так, в задачах поиска по сходству часто важнее быстро определить приближенных «ближайших соседей», чем долго искать точных. Разработан ряд методов преобразования векторных массивов, предназначенных для реализации этого подхода.

Методы сокращения размерности векторов [16, 17] адаптируются к особенностям их базы, используемой в качестве обучающей выборки. Применяются методы обучения (адаптации) без учителя и с учителем.

Методы обучения без учителя обеспечивают выявление в данных наиболее информативных (согласно заданным критериям) новых координат и осуществляют сокращение размерности исключением некоторых из них. При этом могут ис-

кажаться меры сходства исходных векторов. Например, метод главных компонент (PCA) выявляет ортогональные направления в данных, минимизирующие искажение евклидовых расстояний между векторами. Это приводит к уменьшению оценки расстояний, причем ошибка увеличивается с ростом числа исключенных измерений.

В ряде методов снижения размерности с учителем сходство итоговых векторов отражает «семантическое сходство» исходных данных [18]. Получение этой информации сопряжено с большой работой по оценке сходства исходных объектов, которая обычно выполняется экспертом.

Общим недостатком многих методов сокращения размерности с обучением является вычислительная сложность решения их оптимизационных задач, например, разложением по сингулярным значениям или процедурами градиентного спуска (с возможными локальными минимумами). Кроме того, при применении ряда методов возникают трудности преобразования новых данных, которые не использовались в обучении. Получение их сокращенных представлений может потребовать повторного обучения или оценки сходства со значительной частью обучающей выборки.

Недостатки методов сокращения размерности с учителем обусловили развитие подхода к преобразованию векторных данных без адаптации — так называемого случайного проецирования (random projection) [19–27]. Здесь для преобразования входных векторов в выходные проводится умножение на случайную матрицу, элементы которой — случайно сгенерированные и затем фиксированные числа из некоторого распределения. В ряде работ (например, [19, 20, 26]) исследована точность оценок некоторых мер сходства–различия при использовании случайных матриц определенного вида. Так, в [26] показано, что преобразование простой (с точки зрения генерации и применения) бинарной случайной матрицей с элементами $\{0, 1\}$ позволяет по выходным векторам оценить евклидово расстояние, норму и скалярное произведение входных векторов.

Некоторые подходы и методы требуют специализированных алгоритмов хранения и обработки векторных данных в определенном формате (представлении). Так, разреженные [28, 29] (с малой долей ненулевых компонентов) бинарные векторы используются, например, в ассоциативно-проективных нейронных сетях [30, 31], а также в эффективной бинарной версии [32–34] распределенной ассоциативной памяти [10, 11, 35–38].

Формирование бинарных выходных векторов с регулируемой разреженностью, по которым можно оценить сходство (величину угла) входных вещественных векторов, рассмотрено в [25] для тернарной случайной матрицы с элементами из $\{-1, 0, +1\}$, а в [27] — для бинарной матрицы. В настоящей статье исследуются оценки угла, евклидова расстояния и скалярного произведения входных векторов по ряду характеристик сходства бинарных выходных векторов при использовании для преобразования случайной бинарной матрицы.

ПРОЕЦИРОВАНИЕ СЛУЧАЙНОЙ БИНАРНОЙ МАТРИЦЕЙ И ОЦЕНКА МЕР СХОДСТВА ПО БИНАРНЫМ ВЕКТОРАМ

Аналогично [26, 27] рассмотрим проецирование векторов случайной бинарной матрицей $\mathbf{R}$ с элементами r_{ij} из множества $\{0, 1\}$. Случайные величины (с.в.), реализацией которых являются элементы $\mathbf{R}$ (единицы и нули), независимы и имеют одинаковое распределение (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 вещественные векторы размерности D, и $\mathbf{u} = \mathbf{R}\mathbf{x}$, $\mathbf{v} = \mathbf{R}\mathbf{y}$ — результаты их проецирования (промежуточные векторы размерности d). Следовательно, размерность $\mathbf{R}$ есть $(d \times D)$.

При проецировании каждый компонент u_i, $i = 1, \ldots, d$, вектора $\mathbf{u}$ первоначально формируется как скалярное произведение строки $\mathbf{r}_i$ матрицы $\mathbf{R}$ на $\mathbf{x}$:

$$u_i = \langle\, \mathbf{r}_i, \mathbf{x} \,\rangle = \sum_{j=1}^{D} r_{ij} x_j. \tag{1}$$

Компоненты u_i являются i.i.d. с.в. Математическое ожидание (м.о.) u_i:

$$E\{u_i\} = E\left\{\sum_{j=1}^{D} r_{ij} x_j\right\} = q \sum_{j=1}^{D} x_j, \tag{2}$$

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

$$V\{u_i\} = V\left\{\sum_{j=1}^{D} r_{ij} x_j\right\} = \sum_{j=1}^{D} V\{r_{ij} x_j\} = \sum_{j=1}^{D} x_j^2 V\{r_{ij}\} = ||x||_2^2 \,(q - q^2), \tag{3}$$

так как r_{ij} есть i.i.d., $E\{r_{ij}^2\} = 1^2 q + 0^2 (1-q) = q$ и $V(r_{ij}) = E\{r_{ij}^2\} - (E\{r_{ij}\})^2 = q - q^2$.

Стандартизация u_i осуществляется вычитанием м.о. (2) $q \sum_{j=1}^{D} x_j$ (центрирование) и делением на среднеквадратичное отклонение $||x||_2 \sqrt{q - q^2}$ (квадратный корень из (3)). Распределение стандартизованной с.в. u_i сходится к гауссову с нулевым средним и единичной дисперсией. Скорость сходимости u_i исследуется в [27] (обзор проблемы сходимости для общего случая см. в [39, 40]).

Формирование бинарного выходного вектора осуществляется бинаризующим пороговым преобразованием промежуточного вектора $\mathbf{u} \to \mathbf{z}$:

$$z_i = 1 \text{ при } u_i > t_i;\ z_i = 0 \text{ в противном случае, } i = 1, \ldots, d, \tag{4}$$

где t_i — величина порога для i-го компонента выходного вектора. Будем использовать одинаковые величины порогов для всех компонентов, $t_i \equiv t$. Заданная вероятность p единичного компонента z бинарного выходного вектора $\mathbf{z}$ определяется выбором соответствующего порога t_p. Для стандартизованной с.в. u с гауссовым распределением

$$p(z=1) = p(u > t_p) = \frac{1}{\sqrt{2\pi}} \int_{t_p}^{\infty} e^{-\eta^2/2} d\eta = 1 - \Phi(t_p),$$

где Φ — гауссова кумулятивная функция распределения. Величина t_p для обеспечения нужной p определяется как квантиль гауссова распределения, соответствующая $1-p$. Векторы при $p < 0.5$ являются разреженными.

Оценка угла между входными векторами. В [25] для оценки угла θ между входными векторами $\mathbf{x}$, $\mathbf{y}$ предлагается использовать оценки вероятности p_{join} совпадения единичных компонентов $z_{1,i} = 1$ и $z_{2,i} = 1$ в векторах $\mathbf{z}_1$ и $\mathbf{z}_2$ после бинаризации $\mathbf{u} \to \mathbf{z}_1$ с порогом t_1 и $\mathbf{v} \to \mathbf{z}_2$ с порогом t_2. Согласно многомерной центральной предельной теореме совместное распределение с.в. (u_i, v_i) сходится к двумерному гауссову. Для стандартизованных (u_i, v_i) эту вероятность можно определить как результат вычисления интеграла от двумерного гауссова распределения

$$p_{join} \equiv p(z_{1,i} = 1,\ z_{2,i} = 1 \,|\, \theta, t_1, t_2) = p(u_i > t_1,\ v_i > t_2 \,|\, \theta) =$$

$$= \frac{1}{2\pi(1 - \cos^2\theta)} \int_{t_1}^{\infty} \int_{t_2}^{\infty} \exp\left(-\frac{\eta_1^2 - 2\eta_1\eta_2 \cos\theta + \eta_2^2}{2(1 - \cos^2\theta)}\right) d\eta_1 d\eta_2. \tag{5}$$

Таким образом, угол θ связан функциональной зависимостью с p_{join}: $\theta = f(p_{join})$, где f — функция, обратная (5). Поэтому θ можно оценить следующим образом: протабулировать (5); преобразовать входные векторы в выходные $\mathbf{z}_1$ и $\mathbf{z}_2$ по (1), (4); оценить p_{join} как $\langle \mathbf{z}_1, \mathbf{z}_2 \rangle / d$; найти в таблице значение p_{join}, ближайшее к $\langle \mathbf{z}_1, \mathbf{z}_2 \rangle / d$, и использовать соответствующий ему угол θ^* как оценку θ.

Определим м.о. и дисперсию оценки угла, применив линеаризацию функции случайного аргумента [41]. Пусть имеется с.в. ξ с известными м.о. $E\{\xi\} = E_\xi$ и дисперсией $V\{\xi\} = V_\xi$, а также другая с.в. ζ, связанная с ξ как $\zeta = \varphi(\xi)$, причем в окрестности $E\{\xi\} = E_\xi$ функция φ близка к линейной. Поэтому связь ξ с ζ при малых отклонениях от среднего можно представить с использованием одношагового разложения в ряд Тейлора как $\zeta \approx \varphi(E_\xi) + \varphi'(E_\xi)(\xi - E_\xi)$. Тогда м.о. и дисперсию ζ аппроксимируем как

$$E\{\zeta\} \approx \varphi(E_\xi) + \varphi'(E_\xi) E\{(\xi - E_\xi)\} = \varphi(E_\xi), \tag{6}$$

$$V\{\zeta\} \approx V\{\varphi(E_\xi)\} + (\varphi'(E_\xi))^2 V\{(\xi - E_\xi)\} = (\varphi'(E_\xi))^2 V_\xi. \tag{7}$$

Число $k = \langle \mathbf{z}_1, \mathbf{z}_2 \rangle$ совпадающих единичных компонентов при вероятности их совпадения p_{join} для различных реализаций бинарных векторов размерности d, а также $p^*_{join} \equiv \langle \mathbf{z}_1, \mathbf{z}_2 \rangle / d$ и оценка угла $\theta^*_{join} = f(p^*_{join})$ являются с.в. Применяя (6), (7) к $\theta^*_{join} = f(p^*_{join})$, получаем

$$E\{\theta^*_{join}\} = f(E\{p^*_{join}\}),\ V\{\theta^*_{join}\} = (f'(E\{p^*_{join}\}))^2 V\{p^*_{join}\}. \tag{8}$$

Случайная величина $k = \langle \mathbf{z}_1, \mathbf{z}_2 \rangle$ имеет биномиальное распределение [25]: $P(k) = \beta(k, d, p_{join})$ с м.о. $E\{k\} = dp_{join}$ и дисперсией $V\{k\} = dp_{join}(1 - p_{join})$. Поэтому

$$E\{p^*_{join}\} = E\{k/d\} = p_{join},$$

$$V\{p^*_{join}\} = V\{k/d\} = V\{k\}/d^2 = p_{join}(1 - p_{join})/d. \tag{9}$$

С учетом (9) запишем (8) в виде

$$E\{\theta^*_{join}\} = f_{join}(p_{join}),\ V\{\theta^*_{join}\} = (f'_{join}(p_{join}))^2\, p_{join}(1 - p_{join})/d. \tag{10}$$

По известному значению вероятности p_{join} определить оценку значения производной $f'_{join}(p_{join}) \equiv \theta'(p_{join})$ можно по табулированной функции (5), найдя по p_{join} ближайшее значение в таблице и соответствующий ему угол θ, а также вычислив $\Delta\theta / \Delta p$, где $\Delta\theta$ и Δp — разность значений углов и вероятностей соответственно между значением в найденной и соседней ячейках таблицы. Следовательно, если известен θ, для определения «аналитического» м.о. и дисперсии его оценки по p^*_{join} найдем в табулированной (5) по θ соответствующую вероятность p_{join} и $f'(p_{join}) \equiv \theta'(p_{join})$ и воспользуемся (10).

Кроме оценки угла по эмпирической вероятности p_{join}, исследуем его оценку по эмпирической условной вероятности p_{cond} совпадения единичных компонентов выходных бинарных векторов, а также по эмпирической вероятности p_{equ} совпадения значений их компонентов (и единичных, и нулевых).

Выразим p_{cond} через p_{join}. Для бинарных векторов с одинаковой вероятностью p единичного компонента имеем

$$p_{cond} \equiv p(z_{1,i} = 1 \mid z_{2,i} = 1) = p(z_{1,i} = 1,\ z_{2,i} = 1) / p(z_{2,i} = 1) \equiv p_{join} / p. \tag{11}$$

Зависимость p_{cond} от угла найдем делением (5) на p.

Определив $p^*_{cond} \equiv \langle \mathbf{z}_1, \mathbf{z}_2 \rangle / |\mathbf{z}_2|$, где $|\mathbf{z}_2|$ — число единичных компонентов $\mathbf{z}_2$, можно вычислить

$$E\{p^*_{cond}\} = E\{k/(pd)\} = p_{cond},$$

$$V\{p^*_{cond}\} = V\{k/(pd)\} = V\{k\}/(pd^2) = p_{cond}(1-p_{cond})/(pd), \qquad (12)$$

так как максимально могут совпасть pd единичных компонентов. Значения $f'(p_{cond}) \equiv \theta'(p_{cond})$ определим по табулированной p_{cond} аналогично p_{join} и получим

$$E\{\theta^*_{cond}\} = f_{cond}(p_{cond}),$$

$$V\{\theta^*_{cond}\} = (f'_{cond}(p_{cond}))^2 p_{cond}(1-p_{cond})/(pd). \qquad (13)$$

Для p_{equ} запишем

$$p_{equ} \equiv p(z_{1,i} = z_{2,i}) = 1 - p(z_{1,i} \neq z_{2,i}) = 1 - (p(z_{1,i}=1) +$$
$$+ p(z_{2,i}=1) - 2p(z_{2,i}=1, z_{2,i}=1)) = 1 - (2p - 2p_{join}). \qquad (14)$$

Зависимость p_{equ} от угла найдем по p_{join} (5) и p.

Определив $p^*_{equ} \equiv 1 - \mathbf{z}_1 \oplus \mathbf{z}_2 / d$, где $\oplus$ — покомпонентная операция «исключающее ИЛИ», аналогично (9), (10), (12), (13) получим

$$E\{p^*_{equ}\} = p_{cond},\ V\{p^*_{equ}\} = p_{equ}(1-p_{equ})/d, \qquad (15)$$

$$E\{\theta^*_{equ}\} = f_{equ}(p_{equ}),\ V\{\theta^*_{equ}\} = (f'_{equ}(p_{equ}))^2 p_{equ}(1-p_{equ})/d. \qquad (16)$$

Как показано в [25], вероятность p_{join} совпадения единичных компонентов выходных бинарных векторов монотонно уменьшается с увеличением угла θ. Согласно (11), (14) это справедливо и для p_{cond}, p_{equ}. Следовательно, оценки этих вероятностей p^*_{join}, p^*_{cond}, p^*_{equ} могут быть полезными как меры сходства входных векторов (без вычисления по ним оценок угла).

Оценка скалярного произведения и евклидова расстояния. Как отмечалось ранее, стандартизация u_i для получения бинарных выходных векторов требует знания евклидовых норм исходных векторов $||\mathbf{x}||$, $||\mathbf{y}||$. Вычисление нормы проводится однократно для одного вектора и полученное число запоминается. Эту информацию совместно с рассмотренной ранее оценкой угла θ^* по бинарным выходным векторам можно использовать для оценки скалярного произведения $\langle \mathbf{x}, \mathbf{y} \rangle^*$ и (квадрата) евклидова расстояния $||\mathbf{x} - \mathbf{y}||^{2*}$:

$$\langle \mathbf{x}, \mathbf{y} \rangle^* = ||\mathbf{x}||\,||\mathbf{y}|| \cos\theta^*, \qquad (17)$$

$$||\mathbf{x} - \mathbf{y}||^{2*} = ||\mathbf{x}||^2 + ||\mathbf{y}||^2 - 2||\mathbf{x}||\,||\mathbf{y}|| \cos\theta^*. \qquad (18)$$

Для скалярного произведения с использованием линеаризации (7) имеем

$$V\{\langle \mathbf{x}, \mathbf{y} \rangle^*\} = ||\mathbf{x}||^2 ||\mathbf{y}||^2 V\{\cos\theta^*\} =$$

$$= ||\mathbf{x}||^2 ||\mathbf{y}||^2 V\{\theta^*\} \left(\frac{d\cos\theta}{d\theta} \right)^2 = ||\mathbf{x}||^2 ||\mathbf{y}||^2 V\{\theta^*\} \sin^2\theta. \qquad (19)$$

Здесь $V\{\theta^*\}$ зависит от способа оценки угла (10), (13), (15), (16).

Для квадрата евклидова расстояния с использованием (7) получаем

$$V\{||\mathbf{x} - \mathbf{y}||^{2*}\} = 4||\mathbf{x}||^2 ||\mathbf{y}||^2 V\{\cos\theta^*\} =$$

$$= 4||\mathbf{x}||^2 ||\mathbf{y}||^2 V\{\theta^*\} \sin^2\theta = 4V\{\langle \mathbf{x}, \mathbf{y} \rangle^*\}. \qquad (20)$$

ЭКСПЕРИМЕНТАЛЬНОЕ ИССЛЕДОВАНИЕ

Изучалось поведение ошибки оценки угла между входными векторами, скалярного произведения и евклидова расстояния между ними. Оценки определялись по выходным бинарным векторам различной разреженности. В качестве ошибки оценки угла использовалась дисперсия V; ошибка оценки скалярного произведения и расстояния измерялась коэффициентом вариации $V^{1/2}/E$. Результаты зависимости значения ошибки оценки угла от размерности входных и выходных векторов приведены в [27], где исследовалась зависимость значения ошибки от величины угла между входными векторами.

Для преобразования входных вещественных векторов **x**, **y** в промежуточные **u**, **v** использовались случайные матрицы с бинарными $\{0, 1\}$ и тернарными элементами $\{-1, 0, +1\}$, вероятность ненулевого элемента матриц $q = \{0.5, 0.1\}$. Величина порога t для стандартизованных значений компонентов промежуточных векторов u_i выбиралась для поддержания вероятности единичного компонента $p = \{0.5, 0.1\}$ в выходных бинарных векторах, т.е. $t_p = \{0.0, 1.282\}$.

Использовались входные векторы с $D = 1000$ и выходные бинарные векторы с $d = 1000$. Сходство варьировалось путем конкатенации векторов **a**, **b**, **c**: $\mathbf{x} = (\mathbf{a}\ \mathbf{b})$, $\mathbf{y} = (\mathbf{c}\ \mathbf{b})$ разной размерности. Например, если размерности **a** и **c** равны нулю, получаем одинаковые векторы **b** размерности d, а если размерность **b** равна нулю, получаем разные векторы **a**, **c** размерности d. Компоненты **a**, **b**, **c** генерировались случайно из равномерного распределения в $[-D, +D]$. Так как размерности входных векторов ($D = 1000$) велики, их нормы близки. Типичное значение угла между случайными векторами — приблизительно 90°, т.е. случайные векторы почти ортогональны. Для случайных векторов с равномерно распределенными положительными компонентами угол обычно составляет 40°–45°. Для таких векторов достичь ортогональности можно при непересекающихся частях векторов $\mathbf{x} = (\mathbf{a}\ 0)$, $\mathbf{y} = (0\ \mathbf{b})$.

При оценках совместной вероятности p^*_{join} величина $\langle \mathbf{z}_1, \mathbf{z}_2 \rangle / d$ может превысить максимально возможное значение $p^{\max}_{join} = p$. Чтобы обеспечить соответствие экспериментальных результатов формулам (9)–(11), для случая $p^*_{join} > p^{\max}_{join}$ использовалось «зеркальное» преобразование $p - (p^*_{join} - p)$, а полученное значение угла считалось отрицательным.

Усреднение результатов проводилось по 10000 реализациям случайной матрицы.

Эксперименты показали, что значения ошибок при исследованных параметрах близки для бинарной и тернарной случайных матриц. Поэтому приведем результаты для бинарной матрицы.

На рис. 1 показаны зависимости ошибки V оценки угла θ между входными векторами от его величины. Здесь оценки определены по p^*_{join}, p^*_{cond}, p^*_{equ}; обозначения T соответствуют значениям, полученным из (10), (13), (16), а E — экспериментальным результатам.

Для $p = 0.5$ аналитические значения ошибок оценки угла близки к соответствующим им экспериментальным. Оценки угла по p^*_{join} имеют наибольшую ошибку (ее значение убывает с увеличением угла), по p^*_{equ} — наименьшую ошибку, а по p^*_{cond} ошибка имеет промежуточное значение. Нулевую ошибку при нулевом угле дают p^*_{equ} и p^*_{cond}. Значение ошибки растет с увеличением угла. (В приведенных экспериментах нулевой угол не использовался).

Для $p = 0.1$ ошибки, как и следовало ожидать, превышают соответствующие значения для $p = 0.5$. Аналитические значения ошибок оценки угла близки к со-

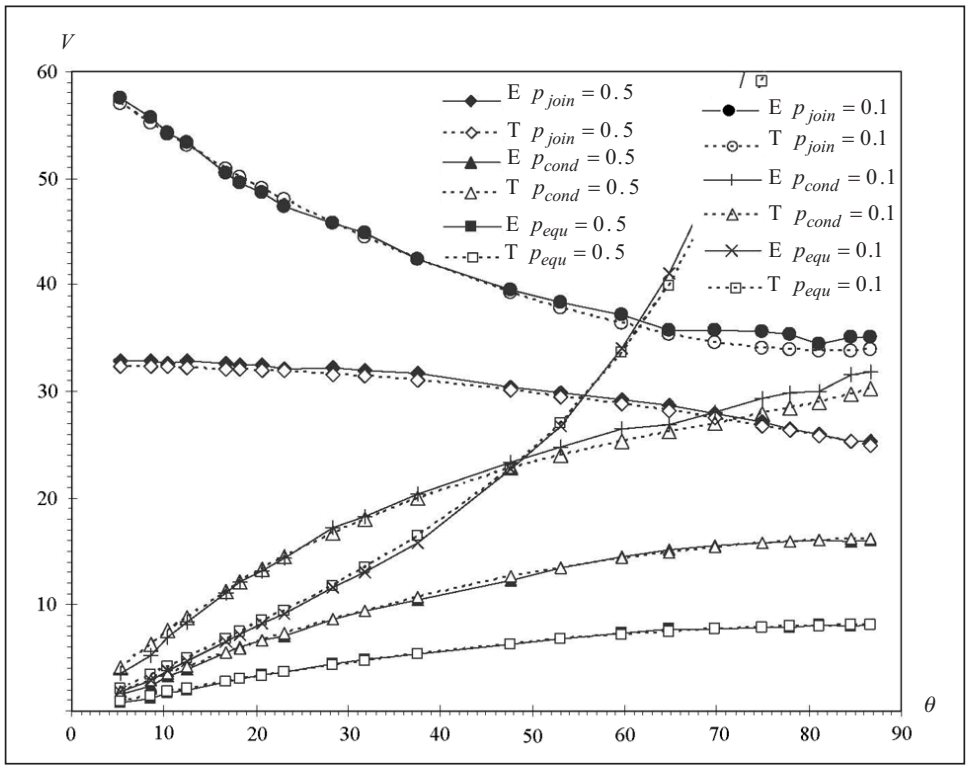

Рис. 1. Зависимости ошибки оценки угла θ между выходными векторами от его величины с $D = 1000$, $d = 1000$, $q = 0.1$

ответствующим им экспериментальным для p^*_{cond} и p^*_{join} во всем диапазоне углов. Ошибка оценки угла по p^*_{cond} по-прежнему меньше ошибки по p^*_{join} с наибольшим разрывом при 0° и с примерно равными значениями при 90°. Значения ошибки оценки угла по p^*_{equ} сильно отличаются от случая $p = 0.5$: если угол больше 50°, ее значение превышает ошибку по p^*_{cond}, а если угол больше 60° — ошибку по p^*_{join}; кроме того, аналитические значения ошибки становятся меньше экспериментальных, т.е. приближение (7) не выполняется. Таким образом, предпочтительна оценка угла по p^*_{cond}, а диапазон применимости оценок по p^*_{equ} ограничен малыми углами и уменьшается при уменьшении доли единичных компонентов в выходных векторах.

На рис. 2 приведены зависимости ошибки $V^{1/2} / E$ (коэффициента вариации) оценки скалярного произведения входных векторов по (17) от величины угла θ для оценок угла, полученных по p^*_{join}, p^*_{cond}, p^*_{equ}. Обозначения T соответствуют значениям ошибки, полученным из (19) и по оценкам угла из (10), (13), (16), а E — экспериментальным результатам.

Для всех исследованных комбинаций параметров значения ошибок оценки скалярного произведения при $p = 0.1$ превышает соответствующие значения при $p = 0.5$; ошибка растет с увеличением угла между входными векторами особенно быстро, если угол больше 50°. Аналогично оценке угла оценка по p^*_{equ} также имеет наименьшую ошибку при $p = 0.5$, но при $p = 0.1$ ошибка становится самой большой, если угол больше 60°. При бóльших значениях углов заметно различие между аналитическими и экспериментальными значениями ошибок для p^*_{equ}, для остальных случаев эти результаты близки. Таким образом, наибольшая точность

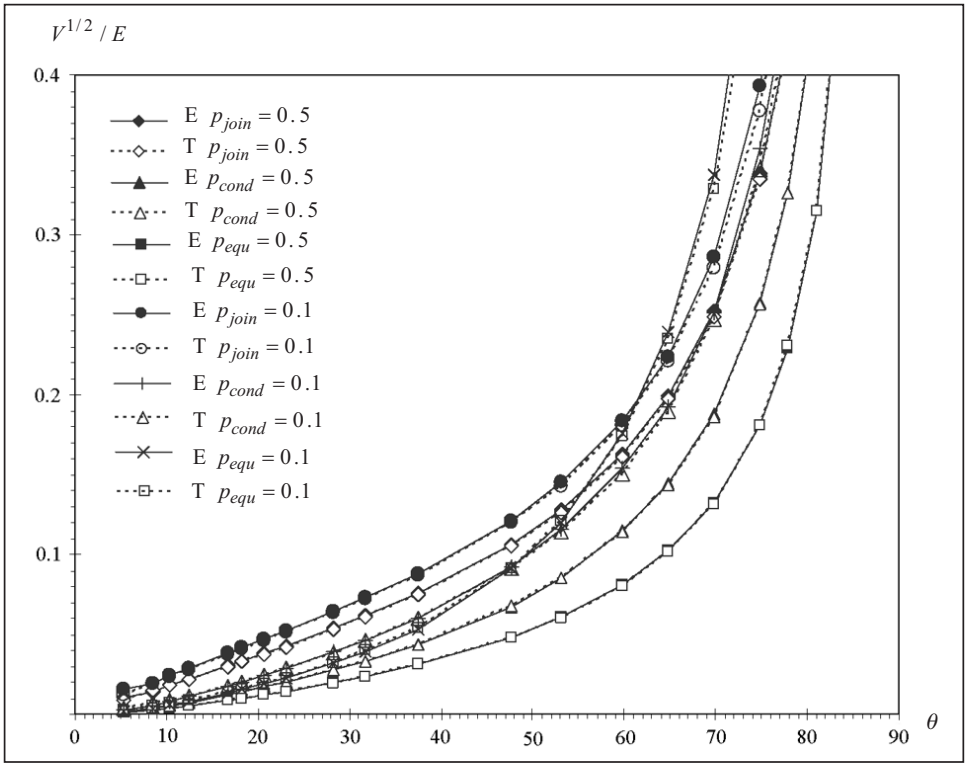

Рис. 2. Зависимость ошибки оценки скалярного произведения от угла θ между векторами с $D = 1000$, $d = 1000$, $q = 0.1$

оценки величины скалярного произведения достигается при малых значениях углов между входными векторами: для $p = 0.5$ следует использовать оценку по p^*_{equ}, а для разреженных выходных бинарных векторов наилучшие результаты достигаются при оценке по p^*_{cond}.

На рис. 3 приведены зависимости ошибки $V^{1/2}/E$ (коэффициента вариации) оценки квадрата евклидова расстояния между входными векторами по (18) от величины угла θ, аналитические значения ошибки получены из (20); обозначения в легенде аналогичны рис. 1, 2.

Как и для оценок угла и скалярного произведения, значения ошибок оценки квадрата расстояния при $p = 0.1$ превышают соответствующие значения при $p = 0.5$. Однако ошибка не растет, а уменьшается с увеличением угла. Так как при увеличении угла между векторами растет расстояние между ними, значение исследуемой в настоящей статье относительной ошибки уменьшается, и это компенсирует рост дисперсии угла при увеличении угла для p^*_{cond} и p^*_{equ}. В результате ошибка оценки квадрата расстояния по p^*_{cond} (и даже по p^*_{equ}) не слишком сильно изменяется во всем исследованном диапазоне изменения углов. Самую большую ошибку (и расхождение между аналитическими и экспериментальными результатами) можно получить для p^*_{join} при малых углах, так как для этого случая дисперсия оценки угла максимальна. Отметим, что коэффициент вариации неустойчив в окрестности нуля, и, следовательно, его нельзя применять для адекватной оценки ошибки при близких к нулю значениях угла. Как отмечено ранее, ошибка оценки угла по p^*_{equ} очень быстро растет при приближении величины угла к 90° для разреженных бинарных векторов.

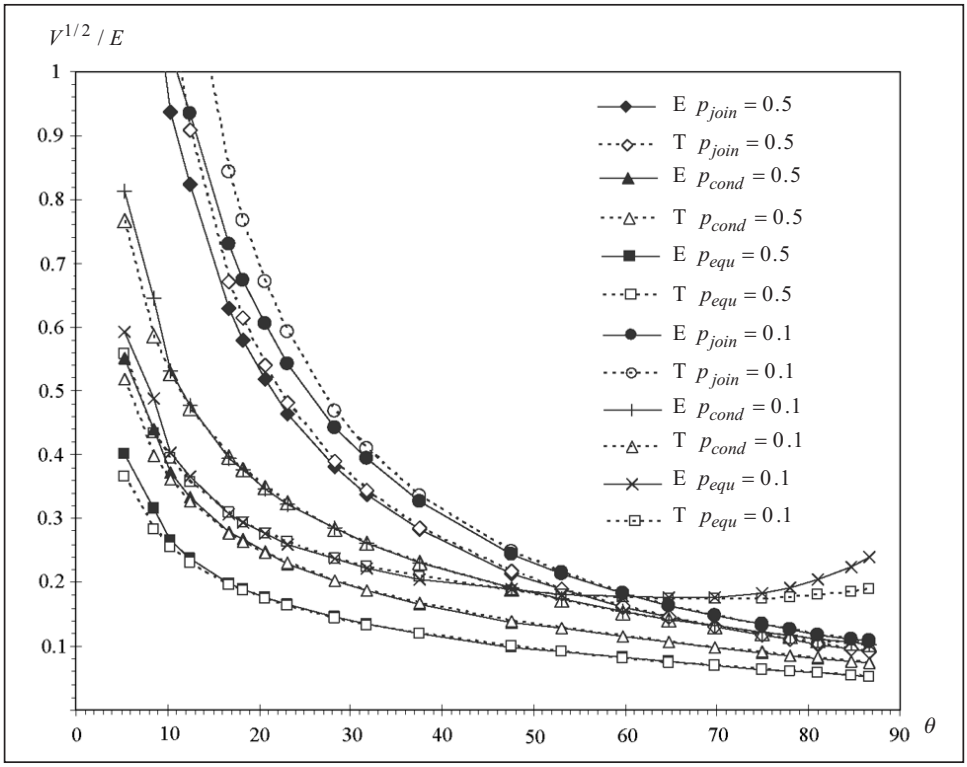

Рис. 3. Зависимость ошибки оценки квадрата евклидова расстояния от угла θ между векторами с $D = 1000$, $d = 1000$, $q = 0.1$

Таким образом, для $p = 0.5$ наименьшую ошибку имеют оценки расстояния по p^*_{equ}; для разреженных выходных бинарных векторов наилучшие результаты достигаются при оценках по p^*_{cond} и по p^*_{equ} (для p^*_{equ}, если величина углов не слишком близка к 90°); при малых значениях углов ошибка оценки расстояния быстро растет.

Для оценок угла между входными векторами по выходным бинарным векторам, а также скалярного произведения и евклидова расстояния между входными векторами по полученной оценке угла и известным нормам входных векторов при их применении в задачах поиска по сходству интересно исследовать гибридный подход, когда для малых углов сходство оценивается по углу, а для больших — по расстоянию.

ЗАКЛЮЧЕНИЕ

Исследованы оценки мер сходства вещественных векторов по бинарным, полученным проецированием случайной бинарной матрицей с элементами $\{0, +1\}$ и выходным пороговым преобразованием, которое позволяет регулировать степень разреженности (долю единичных компонентов) бинарных векторов. Сходство последних оценивалось по мерам на основе (нормированного к их размерности d) скалярного произведения $\langle \mathbf{z}_1, \mathbf{z}_2 \rangle / d \equiv |\mathbf{z}_1 \wedge \mathbf{z}_2| / d$ и связанных с ним $\langle \mathbf{z}_1, \mathbf{z}_2 \rangle / |\mathbf{z}_2|$ и $1 - \mathbf{z}_1 \oplus \mathbf{z}_2 / d = 1 - (|\mathbf{z}_1| + |\mathbf{z}_2| - 2|\mathbf{z}_1 \wedge \mathbf{z}_2|)/d$, где $\oplus$, $\wedge$ — операции покомпонентного XOR и AND соответственно, $|\mathbf{z}|$ — число единичных компонентов в бинарном векторе $\mathbf{z}$. Эти меры являются естественными оценками сходства исходных вещественных векторов $\mathbf{x}, \mathbf{y}$: их значения монотонно убывают с увеличением угла θ между векторами и дают возможность оценить его величину. Оценка угла θ при известных значениях евклидовых норм $||\mathbf{x}||$, $||\mathbf{y}||$ позволила также оценить скалярное произведение $\langle \mathbf{x}, \mathbf{y} \rangle^* = ||\mathbf{x}|| \, ||\mathbf{y}|| \cos\theta^*$ и евклидово расстояние $||\mathbf{x} - \mathbf{y}||^* = (||\mathbf{x}||^2 + ||\mathbf{y}||^2 - 2||\mathbf{x}|| \, ||\mathbf{y}|| \cos\theta^*)^{1/2}$.

Аналитически и экспериментально исследованы зависимости значения ошибки оценки угла, скалярного произведения, евклидова расстояния между входными вещественными векторами от величины угла между ними. Зависимости существенно отличаются для исследованных трех мер сходства бинарных векторов, а также для оценок скалярного произведения и евклидова расстояния. Это дает возможность использовать оценки с наименьшей ошибкой в различных диапазонах значений углов.

Для бинарных и тернарных случайных матриц величины ошибки близки при исследованных значениях параметров. Однако реализация преобразования с помощью бинарной случайной матрицы вычислительно более простая.

Перспективной темой дальнейших работ является исследование влияния на точность оценок сходства модификаций предложенных методов, например: использование вместо вероятности единичного элемента в матрице реальной доли единичных элементов во всей матрице, а также в ее строках и столбцах; использование случайных матриц с фиксированным числом случайно расположенных единичных элементов во всей матрице, а также в ее строках и столбцах; учет реальной доли единичных компонентов в выходных бинарных векторах при оценке угла на основе их скалярного произведения.

Выходные бинарные векторы, в которых важна не семантика отдельных компонентов, а отражение ими сходства представляемых объектов, являются примером рандомизированных распределенных представлений [1–11, 30, 31, 35, 36, 42–45]. Распределенные представления могут также формироваться во внутренних слоях многослойных сетей при обучении [46–48]. Распределенные представления используют для представления семантического сходства [18, 42, 49–51], последовательностей [29–31, 43, 51–54], сложных иерархически структурированных объектов [1, 2, 4–11, 30, 31, 35, 36, 42–45, 55–57], которые требуются для моделей и систем искусственного интеллекта [3, 31, 35, 36, 57–59].

СПИСОК ЛИТЕРАТУРЫ

1. Rachkovskij D.A. Representation and processing of structures with binary sparse distributed codes // IEEE Transactions on Knowledge and Data Engineering. — 2001. — **13**, N 2. — P. 261–276.
2. Rachkovskij D.A. Some approaches to analogical mapping with structure sensitive distributed representations // Journal of Experimental and Theoretical Artificial Intelligence. — 2004. — 16, N 3. — P. 125–145.
3. Stanojevic M., Vranes S. Semantic approach to knowledge processing // WSEAS Transactions on Information Science and Applications. — 2008. — 5(6). — P. 913–922.
4. Slipchenko S.V., Rachkovskij D.A. Analogical mapping using similarity of binary distributed representations // International Journal Information Theories and Applications. — 2009. — **16**, N 3. — P. 269–290.
5. Gayler R.W., Levy S.D. A distributed basis for analogical mapping // Proceedings of the Second International Analogy Conference. NBU Press, Sofia, Bulgaria. — 2009. — P. 165–174.
6. 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.
7. Гриценко В.И.,. Рачковский Д.А, Гольцев А.Д., Лукович В.В., Мисуно И.С., Ревунова Е.Г., Слипченко С.В., Соколов А.М., Талаев С.А. Нейросетевые распределенные представления для интеллектуальных информационных технологий и моделирования мышления // Кибернетика и вычислительная техника. — 2013. — Вып. 173. — С. 7–24.
8. Pickett M, Aha D. Using cortically-inspired algorithms for analogical learning and reasoning // Biologically Inspired Cognitive Architectures. — 2013. — **6.** — P. 76–86.
9. 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), 4–9 Aug. 2013, Dallas, TX, IEEE. — 2013. — P. 1–8.
10. Emruli B., Sandin F. Analogical mapping with sparse distributed memory: A simple model that learns to generalize from examples // Cognitive Computation. — 2014. — **6**, N 1. — P. 74–88.
11. Widdows D., Cohen T. Reasoning with vectors: a continuous model for fast robust inference // Logic Journal of the IGPL. — 2015. — **23**, N 2. — P. 141–173.
12. Indyk P., Matousek J. Low-distortion embeddings of finite metric spaces // Handbook of discrete and computational geometry, discrete mathematics and its applications / J.E. Goodman, J. O'Rourke (Eds). — Boca Raton, FL: Chapman & Hall/CRC, 2004. — P. 177–196.

13. Upadhyaya S.R. Parallel approaches to machine learning — A comprehensive survey // Journal of Parallel and Distributed Computing. — 2013. — **73**, N3. — P. 284–292.
14. Kussul N., Shelestov A., Skakun S., Kravchenko O. High-performance intelligent computations for environmental and disaster monitoring // International Journal Information Technologies and Knowledge. — 2009. — **3**, N 2. — P. 135–156.
15. Kussul N., Shelestov A., Skakun S. Grid technologies for satellite data processing and management within international disaster monitoring projects // Grid and Cloud Database Management / S Fiore, G. Aloisio (Eds). — Berlin; Heidelberg: Springer-Verlag, 2011. — P. 279–306.
16. Van der Maaten L.J.P., Postma E.O., van den Herik H.J. Dimensionality reduction: A comparative review // Tech. Rep. TiCC-TR 2009-005, Tilburg Centre Creative Comput., Tilburg Univ., Tilburg, The Netherlands, 2009. — 35 p.
17. Burges C.J.C. Dimension reduction: A guided tour // Foundations and Trends in Machine Learning. — 2010. — **2**, N 4. — P. 275–365.
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. Vempala S.S. The Random Projection Method. — Providence, RI: American Mathematical Society, 2004. — 105 p.
20. 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.
21. Revunova E.G., Rachkovskij D.A. Using randomized algorithms for solving discrete ill-posed problems // Information Theories and Applications. — 2009. — **16**, N 2. — P. 176–192.
22. Revunova E.G. Study of error components for solution of the inverse problem using random projections // Mathematical Machines and Systems — 2010. — N 4. — P. 33–42.
23. Revunova E.G., Rachkovskij D.A. Stable transformation of a linear system output to the output of system with a given basis by random projections // The 5th International Workshop on Inductive Modelling (IWIM-2012). — 2012. — P. 37–41.
24. 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.
25. Rachkovskij D.A., Misuno I.S., Slipchenko S.V. Randomized projective methods for construction of binary sparse vector representations // Cybernetics and Systems Analysis. — 2012. — **48**, N 1. — P. 146–156.
26. Rachkovskij D.A. Vector data transformation with random binary matrices // Cybernetics and Systems Analysis. — 2014. — **50**, N 6. — P. 960–968.
27. Rachkovskij D.A. Formation of similarity-reflecting binary vectors with random binary projections // Cybernetics and Systems Analysis. — 2015. — **51**, N 2. — P. 313–323.
28. Rinkus G. Quantum computation via sparse distributed representation // NeuroQuantology. — 2012. — **10**, N 2. — P. 311-315.
29. Rinkus G.J. Sparsey™: Event recognition via deep hierarchical sparse distributed codes // Frontiers in Computational Neuroscience. — 2014. — 8, Article 160. — P. 1–44.
30. Kussul E.M., Rachkovskij D.A. Multilevel assembly neural architecture and processing of sequences // Neurocomputers and Attention: Vol. II. Connectionism and neurocomputers / A.V. Holden, V.I. Kryukov (Eds.). — Manchester; New York: Manchester University Press, 1991. — P. 577–590.
31. 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.
32. Kartashov A., Frolov A., Goltsev A., Folk R. Quality and efficiency of retrieval for Willshaw-like autoassociative networks: III. Willshaw-Potts model // Network: Computation in Neural Systems. — 1997. — **8**, N 1. — P. 71–86.
33. Frolov A.A., Rachkovskij D.A., Husek D. On information characteristics of Willshaw-like auto-associative memory // Neural Network World. — 2002 — **12**, N 2. — P. 141–158.
34. 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.
35. Kleyko D., Osipov E., Senior A., Khan A.I., Sekercioglu Y.A. Holographic graph neuron: a bio-inspired architecture for pattern processing. — 2015. — http://arxiv.org/pdf/1501.03784v1.pdf.
36. Emruli B., Sandin F., Delsing J. Vector space architecture for emergent interoperability of systems by learning from demonstration // Biologically Inspired Cognitive Architectures. — 2015. — **11.** — P. 53–64.

37. 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.

38. Knoblauch A., Palm G., Sommer F.T. Memory capacities for synaptic and structural plasticity // Neural Computation. — 2010. — **22**, N2. — P. 289–341.

39. Korolev V., Shevtsova I. An improvement of the Berry–Esseen inequality with applications to Poisson and mixed Poisson random sums // Scandinavian Actuarial Journal. — 2012. — **2012**, N 2. — P. 81–105.

40. Shevtsova I.G. On the absolute constants in the Berry–Esseen-type inequalities // Doklady Mathematics. — 2014. — **89**, N 3. — P. 378–381.

41. Венцель Е.С. Теория вероятностей. — М.: Наука, 1969. — 576 с.

42. Widdows D., Cohen T. Real, complex, and binary semantic vectors // Lecture Notes in Computer Science. — 2012. — **7620**. — P. 24–35.

43. Омельченко Р.С. Программа проверки орфографии (Spellchecker) на основе распределенных представлений // Проблемы программирования. — 2013. — № 4. — С. 35–42.

44. Cohen T., Widdows D., Wahle M., Schvaneveldt R. Orthogonality and orthography: introducing measured distance into semantic space // Lecture Notes in Computer Science. — 2014. — **8369.** — P. 34–46.

45. Kanerva P., Sjodin G., Kristoferson J., Karlsson R., Levin B., Holst A., Karlgren J., Sahlgren M. Computing with large random patterns // Foundations of Real-World Intelligence. — Stanford (California): CSLI Publications, 2001. — P. 251–311.

46. Reznik A.M., Galinskaya A.A., Dekhtyarenko O.K., Nowicki D.W. Preprocessing of matrix QCM sensors data for the classification by means of neural network // Sensors and Actuators B. — 2005. — **106.** — P. 158–163.

47. Chernodub A.N. Direct method for training feed-forward neural networks using batch extended Kalman filter for multi-step-ahead predictions // Lecture Notes in Computer Science. — 2013. — **8131.** — P. 138–145.

48. Chernodub A.N. Training Neural Networks for classification using the Extended Kalman Filter: A comparative study // Optical Memory and Neural Networks. — 2014. — **23**, N 2. — P. 96–103.

49. 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.

50. 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.

51. Sokolov A. Vector representations for efficient comparison and search for similar strings // Cybernetics and Systems Analysis. — 2007. — **43**, N 4. — P. 484–498.

52. Kleyko D., Osipov E. On bidirectional transitions between localist and distributed representations: The case of common substrings search using vector symbolic architecture // Procedia Computer Science. — 2014. — **41.** — P. 104–113.

53. Rasanen O., Kakouros S. Modeling dependencies in multiple parallel data streams with hyperdimensional computing // Signal Processing Letters, IEEE. — 2014. — **21**, N 7. — P. 899–903.

54. Recchia G.L., Sahlgren M., Kanerva P., Jones M.N. Encoding sequential information in semantic space models: Comparing holographic reduced representation and random permutation // Computational Intelligence and Neuroscience. — Article ID 986574. — 2015. — 18 p.

55. Kvasnicka V., Pospichal J. Deductive rules in holographic reduced representation // Neurocomputing. — 2006. — **69.** — P. 2127–2139.

56. Gallant S.I., Okaywe T.W. Representing objects, relations, and sequences // Neural computation. — 2013. — **25**, N8. — P. 2038–2078.

57. Sandin F., Khan A.I., Dyer A.G., Amin A.H.M., Indiveri G., Chicca E., Osipov E. Concept learning in neuromorphic vision systems: What can we learn from insects? // Journal of Software Engineering and Applications. — 2014. — **7**, N 5. — P. 387–395.

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

59. Letichevsky A.A., Letychevskyi O.O., Peschanenko V.S., Huba A.A. Generating symbolic traces in the insertion modeling system // Cybernetics and Systems Analysis. — 2015. — **51**, N 1. — P. 5–15.

Поступила 17.09.2014