**УДК 004.93'11**

***Е.В. ВОЛЧЕНКО***, канд. техн. наук, доц., Институт информатики и искусственного интеллекта ГВУЗ "Донецкий национальный технический университет", Донецк

# О СПОСОБЕ ОПРЕДЕЛЕНИЯ БЛИЗОСТИ ОБЪЕКТОВ ВЗВЕШЕННЫХ ОБУЧАЮЩИХ ВЫБОРОК

В работе предложена метрика для определения расстояния между объектами обучающих выборок, имеющими вес. Выполнено расширение алгоритма *k*-ближайших соседей на взвешенные выборки *w*-объектов с вычислением расстояния на основе предложенной метрики. Проведены экспериментальные исследования, подтвердившие эффективность предложенного подхода. Библиогр.: 9 назв.

**Ключевые слова:** *w*-объект, обучающая выборка, алгоритм *k*-ближайших соседей, метрика.

**Постановка проблемы и анализ литературы.** Классификация объектов в обучающихся системах распознавания заключается в определении их близости к объектам обучающей выборки на основе выбранной метрики (функции расстояния).

Метрикой называют неотрицательную вещественную функцию $d(X_i, X_j)$, удовлетворяющую следующим условиям [1]:

1) $d(X_i, X_j) \geq 0$ для всех объектов $X_i$ и $X_j$ обучающей выборки $X$;

2) $d(X_i, X_j) = 0$ тогда и только тогда, когда $X_i = X_j$ (аксиома тождества);

3) $d(X_i, X_j) = d(X_j, X_i)$ (аксиома симметрии);

4) $d(X_i, X_j) \leq d(X_i, X_s) + d(X_s, X_j)$, где $X_i$, $X_j$ и $X_s$ – три любые объекта выборки $X$ (аксиома треугольника).

Особенности расположения объектов обучающей выборки в признаковом пространстве, недостаточный (избыточный) объем данных, наличие шума и неполных данных существенно повышают важность выбора метрики, позволяющей выполнять классификацию объектов с наибольшей эффективностью [2]. На сегодняшний день разработано значительное количество метрик, перечень которых можно найти, например, в [3], обеспечивающих высокую эффективность классификации, однако для некоторых видов систем распознавания этот вопрос остается открытым.

Анализ большого числа прикладных задач, решающихся путем построения систем распознавания показал, что на сегодняшний день наиболее востребованными являются адаптивные обучающиеся системы распознавания, характеризующиеся способностью изменять свои свойства (словарь признаков, обучающую выборку, решающие правила классификации и т.д.) в соответствии с изменениями распознаваемых объектов [4]. Определяющей особенностью этих систем является возможность пополнения обучающей выборки новыми объектами на всем протяжении времени работы системы, что приводит к неограниченному росту обучающей выборки и необходимости корректировки решающих правил при добавлении новых объектов.

В предыдущих работах автора [4, 5] для сокращения размера обучающих выборок в адаптивных обучающих системах была предложена и реализована алгоритмически идея перехода к взвешенным обучающим выборкам, каждый *w*-объект которой строится по множеству близкорасположенных в пространстве признаков объектов исходной выборки. Значения признаков *w*-объектов являются центрами масс значений признаков объектов найденных множеств. Вес содержит информацию о взаиморасположении, количестве или качестве заменяемых объектов и, исходя из результатов экспериментальных исследований, проведенных в предыдущих работах, позволяет существенно повысить эффективность работы систем.

Введение дополнительной характеристики для описания *w*-объектов не позволяет корректно выполнять классификацию объектов из-за отсутствия метрики, рассчитывающей расстояние между объектами, имеющими разный вес.

**Целью данной работы** является разработка и анализ метрики для оценки степени близости объектов во взвешенных обучающих выборках.

**Постановка задачи.** Пусть имеется некоторая конечная взвешенная обучающая выборка *w*-объектов $X^W = \{X_1^W, X_2^W, \ldots, X_k^W\}$. Каждый *w*-объект $X_i^W$ этой выборки описывается системой признаков $\{x_{i1}, x_{i2}, \ldots, x_{in}\}$ и весом $p_i$ – целым положительным числом, т.е. $X_i^W = \{x_{i1}, x_{i2}, \ldots, x_{in}, p_i\}$ и представляется точкой в линейном пространстве признаков, т.е. $X_i \in R^n$. Для каждого *w*-объекта известна его классификация $y_i \in V$, где $V = \{V_1, \ldots, V_l\}$ – множество всех классов системы.

Имеется некоторый объект $X_s = \{x_{s1}, x_{s2}, \ldots, x_{sn}\}$, заданный только набором признаков (для единообразия присвоим ему вес равный единице, т.е. $p_s = 1$, тогда $X_s^W = \{x_{s1}, x_{s2}, \ldots, x_{sn}, p_s\}$). Необходимо выполнить классификацию объекта $X_s^W$, для чего требуется построить функцию $d_W(X_s^W, X_i^W)$ оценки расстояния между классифицируемым объектом и объектами взвешенной обучающей выборки.

**Построение метрики на взвешенных обучающих выборках.** Выбор метрик в задачах распознавания ограничивается, в первую очередь, сложностью их вычисления [6] и близостью к реальному топологическому разделению пространства признаков на области, соответствующие классам системы [7].

По результатам анализа особенностей расположения объектов взвешенной выборки в пространстве признаков может быть предложена следующая метрика.

Пусть каждый *w*-объект взвешенной обучающей выборки представляется материальной точкой в признаковом пространстве $R^n$ и имеет массу, равную весу *w*-объекта.

Тогда "близость" двух материальных точек (двух *w*-объектов) $X_i^W = \{x_{i1}, x_{i2}, \ldots, x_{in}, p_i\}$ и $X_j^W = \{x_{j1}, x_{j2}, \ldots, x_{jn}, p_j\}$ в пространстве признаков может быть определена по силе притяжения между ними

$$F_{ij} = \frac{p_i \cdot p_j}{r_{ij}^2} = \frac{p_i \cdot p_j}{\left\| X_i^W - X_j^W \right\|} = \frac{p_i \cdot p_j}{\sqrt{\sum_{o=1}^{n} \left(x_{io} - x_{jo}\right)^2}}. \qquad (1)$$

Два *w*-объекта $X_i^W$ и $X_j^W$ являются ближайшими, если сила притяжения между ними, рассчитанная по формуле (1), максимальна.

Поскольку при вычислении расстояний два объекта являются ближайшими, если расстояние между ними минимально, в качестве метрики для определения расстояния между *w*-объектами будем использовать величину, обратную к (1).

***Теорема 1.*** Функция

$$d_W(X_i^W, X_j^W) = \frac{\sqrt{\sum_{o=1}^{n} (x_{io} - x_{jo})^2}}{p_i \cdot p_j} \qquad (2)$$

является метрикой.

*Доказательство.* Покажем, что формула (2) определяет метрику, т.е. удовлетворяет условиям 1 – 4.

***Свойство 1.*** Для любой пары *w*-объектов $X_i^W$ и $X_j^W$ взвешенной обучающей выборки $X^W$ $d_W(X_i^W, X_j^W) \geq 0$.

В формуле (2) выражение $\sqrt{\sum_{o=1}^{n}(x_{io} - x_{jo})^2}$ является евклидовой метрикой, для которой данное свойство выполняется. Согласно постановке задачи веса *w*-объектов $p_i \geq 1$ и $p_j \geq 1$, поэтому для (2) свойство 1 выполняется для любых $X_i^W$ и $X_j^W$.

***Свойство 2.*** Для любой пары *w*-объектов $X_i^W$ и $X_j^W$ взвешенной обучающей выборки $X^W$ $d_W(X_i^W, X_j^W) = 0$ тогда и только тогда, когда $X_i^W = X_j^W$.

Пусть $x_{io} = x_{jo}$ для всех $o = \overline{1,n}$, тогда $\frac{\sqrt{\sum_{o=1}^{n}(x_{io} - x_{jo})^2}}{p_i \cdot p_j} = \frac{\sqrt{\sum_{o=1}^{n}(x_{io} - x_{io})^2}}{p_i \cdot p_j} = \frac{0}{p_i \cdot p_j} = 0$ (по условию $p_i \geq 1$ и $p_j \geq 1$), т.е. для (2) свойство 2 выполняется для любых $X_i^W$ и $X_j^W$.

***Свойство 3.*** Для любой пары *w*-объектов $X_i^W$ и $X_j^W$ взвешенной обучающей выборки $X^W$ $d_W(X_i^W, X_j^W) = d_W(X_j^W, X_i^W)$.

Поскольку $\sqrt{\sum_{o=1}^{n}(x_{io} - x_{jo})^2} = \sqrt{\sum_{o=1}^{n}(x_{jo} - x_{io})^2}$, то, свойство 3 выполняется для любых $X_i^W$ и $X_j^W$.

***Свойство 4.*** Для любых *w*-объектов $X_i^W$, $X_j^W$ и $X_k^W$ выборки $X^W$ $d_W(X_i^W, X_j^W) \leq d_W(X_i^W, X_k^W) + d_W(X_k^W, X_j^W)$.

Рассмотрим естественный случай, когда выполняется расчет расстояния от двух *w*-объектов $X_i^W$, $X_j^W$ (обычно принадлежащих

разным классам) до распознаваемого объекта $X_s^W$, т.е. покажем, что $d_W(X_i^W, X_j^W) \le d_W(X_i^W, X_s^W) + d_W(X_s^W, X_j^W)$.

Пусть $\frac{\sqrt{\sum_{o=1}^{n}(x_{io}-x_{jo})^2}}{p_i \cdot p_j} > \frac{\sqrt{\sum_{o=1}^{n}(x_{io}-x_{so})^2}}{p_i \cdot p_s} + \frac{\sqrt{\sum_{o=1}^{n}(x_{so}-x_{jo})^2}}{p_s \cdot p_j}$, т.е.

$$p_s\sqrt{\sum_{o=1}^{n}(x_{io}-x_{jo})^2} > p_j\sqrt{\sum_{o=1}^{n}(x_{io}-x_{so})^2} + p_i\sqrt{\sum_{o=1}^{n}(x_{so}-x_{jo})^2}.$$

Поскольку в худшем случае для евклидовой метрики $\sqrt{\sum_{o=1}^{n}(x_{io}-x_{jo})^2} = \sqrt{\sum_{o=1}^{n}(x_{io}-x_{so})^2} + \sqrt{\sum_{o=1}^{n}(x_{so}-x_{jo})^2}$, то

$$p_s\left(\sqrt{\sum_{o=1}^{n}(x_{io}-x_{so})^2} + \sqrt{\sum_{o=1}^{n}(x_{so}-x_{jo})^2}\right) > p_j\sqrt{\sum_{o=1}^{n}(x_{io}-x_{so})^2} +$$

$$+ p_i\sqrt{\sum_{o=1}^{n}(x_{so}-x_{jo})^2}.$$

В результате преобразований получим выражение

$$(p_s - p_j)\sqrt{\sum_{o=1}^{n}(x_{io}-x_{so})^2} > (p_i - p_s)\sqrt{\sum_{o=1}^{n}(x_{so}-x_{jo})^2}. \qquad (3)$$

Так как $p_s = 1$, а $p_i \ge 1$ и $p_j \ge 1$, то неравенство (3) является неверным и свойство 4 выполняется для любых $X_i^W$, $X_j^W$ и классифицируемого объекта $X_s^W$.

Поскольку для (2) все свойства метрики выполняются, то функция $d_W(X_i^W, X_j^W) = \frac{\sqrt{\sum_{o=1}^{n}(x_{io}-x_{jo})^2}}{p_i \cdot p_j}$ – является метрикой.

Отметим, что рассмотрение случая, когда классифицируемый объект $X_s^W$ имеет неединичный вес, будет рассмотрено в следующих работах.

**Классификация объектов на основе взвешенных обучающих выборок.** Для классификации объектов на основе взвешенных обучающих выборок *w*-объектов будем использовать алгоритм *k*-ближайших соседей [2], широко применяющийся при решении задач классификации в условиях неполных априорных данных. Выбор данного метода для классификации на основе взвешенной обучающей выборки основывается на результатах исследований [5], согласно которым он будет показывать высокую эффективность классификации при использовании сокращенной обучающей выборки. Для классификации объекта $X_s^W$ с помощью метрики (2) найдем *k* ближайших к нему *w*-объектов каждого из классов и отнесем к тому классу, суммарное расстояние до объектов которого минимально.

**Результаты экспериментальных исследований.** Для оценки эффективности предложенного подхода был проведен ряд экспериментальных исследований. В качестве исходных данных были использованы выборки объектов двух классов размером 1000 – 5000 объектов при 20% пересечении областей классов в пространстве признаков, содержащих 5 – 20 признаков распознавания. Для генерации значений признаков использовался нормальный и равномерный законы распределения.

Оценка эффективности классификации объектов на основе взвешенной обучающей выборки по предложенной модификации алгоритма *k*-ближайших соседей выполнялась на тестовых выборках размером 100 объектов, полученных с помощью тех же генераторов, что и обучающие выборки. В качестве критерия оценки эффективности классификации использовалась частота неверных классификаций. Количество "ближайших соседей" равно 10% размера обучающей выборки *w*-объектов.

Анализ полученных результатов позволяет сделать следующие выводы:

1) размер взвешенной выборки *w*-объектов составил в среднем 2,3% размера исходной обучающей выборки;

2) частота неверной классификации объектов тестовой выборки модифицированным методом *k*-ближайших соседей по выборке *w*-объектов уменьшилась в среднем на 7,4% по сравнению с частотой неверной классификации методом *k*-ближайших соседей по исходной выборке.

Отметим, что близкие результаты были получены при 10 – 40% пересечении областей классов в пространстве признаков и изменении количества признаков распознавания.

Также были проведены экспериментальные исследования эффективности предложенного подхода на выборках репозитория ISEC (International Statistical Education Centre) [9], для которых частота неверной классификации объектов модифицированным методом *k*-ближайших соседей по выборке *w*-объектов уменьшилась в среднем на 5,3%.

**Выводы.** В работе предложен способ оценки расстояния между объектами взвешенной обучающей выборки и классифицируемыми объектами с помощью новой метрики. Выполнено расширение алгоритма *k*-ближайших соседей на взвешенные выборки *w*-объектов с вычислением расстояния на основе предложенной метрики.

Результаты экспериментальных исследований показали устойчивое уменьшение частоты неверных классификаций в среднем на 7,4%, что позволяет сделать вывод об эффективности использования взвешенных выборок *w*-объектов в адаптивных обучающихся системах.

Автор благодарит к.ф.-м.н., с.н.с. И.С. Грунского за ряд ценных замечаний и внимание к данной работе.

**Список литературы:** **1.** *Дюран Б.* Кластерный анализ / *Б. Дюран, П. Оделл*. – М.: Статистика, 1977. – 130 с. **2.** *Theodoridis S.* Pattern Recognition / *S. Theodoridis, K. Koutroumbas*. – San Diego: Academic Press, 2008. – 823 p. **3.** *Воронин Ю.А.* Теория классифицирования и её приложения / *Ю.А. Воронин*. – Новосибирск: Наука, 1985. – 232 с. **4.** Розробка теоретичних засад і методів реалізації відкритих систем автоматичного розпізнавання, що навчаються: способи оптимізації навчаючих вибірок і методи побудови зважених вирішуючих правил класифікації [Текст]: звіт з НДР (заключний): Тема GP/F32/130, Грант Президента України для підтримки наукових досліджень молодих учених на 2011 рік / керівник роботи О.В. Волченко. – Донецк, ГВУЗ "ДонНТУ", 2011. – 67 с. **5.** *Волченко Е.В.* Метод построения взвешенных обучающих выборок в открытых системах распознавания / *Е.В. Волченко* // Доклады 14-й Всероссийской конференции "Математические методы распознавания образов (ММРО-14)", Суздаль, 2009. – М.: Макс-Пресс, 2009. – С. 100 – 104. **6.** *Гороховатский В.А.* Метрики на множествах ключевых точек изображений / *В.А. Гороховатский* // Бионика интеллекта. – 2008. – № 2 (69). – С. 45 – 50. **7.** *Рудаков К.В.* О структуре метрических технологий Data Mining / *К.В. Рудаков, Г.В. Никитов* // Искусственный интеллект. – 2002. – № 2. – С. 218 – 220. **8.** *Павлов Д.А.* Модифицированный алгоритм классификации типа *k*-ближайших соседей / *Д.А. Павлов, А.П. Серых* // Фундаментальная и прикладная математика. – 2000. – Том 6. – № 2. – С. 533-548. **9.** http://www.isical.ac.in/~miu

*Статью представил д.ф.-м.н. доц., зав. кафедрой системного анализа и моделирования Института информатики и искусственного интеллекта ГВУЗ "ДонНТУ" Миненко А.С.*

УДК 004.93'11

**Спосіб визначення близькості об’єктів зважених навчаючих вибірок / Волченко О.В.** // Вісник НТУ "ХПІ". Серія: Інформатика та моделювання. – Харків: НТУ "ХПІ". – 2012. – № 38. – С. 38 – 45.

У роботі запропоновано метрику для визначення відстані між об'єктами навчаючих вибірок, що мають вагу. Виконано розширення алгоритму *k*-найближчих сусідів на зважені вибірки *w*-об’єктів з обчисленням відстані на основі запропонованої метрики. Проведено експериментальні дослідження, що підтвердили ефективність запропонованого підходу. Бібліогр.: 9 назв.

**Ключові слова:** *w*-об'єкт, алгоритм *k*-найближчих сусідів, метрика.

UDC 004.93’1

**Method for determining the proximity of objects of weighted training samples / Volchenko E.V.** // Herald of the National Technical University "KhPI". Subject issue: Information Science and Modeling. – Kharkov: NTU "KhPI". – 2012. – №. 38. – P. 38 – 45.

In article the metrics for distance determination between the objects of weighted training samples is proposed. Expansion of k-nearest neighbors algorithm on the weighed samples of w-objects with distance calculation on the basis of the proposed metrics is done. Experimental results are confirmed the efficiency of the offered approach. Refs: 9 titles.

**Key words:** *w*-object, *k*-nearest neighbors algorithm, metric.

*Поступила в редакцию 29.04.2012*