УДК 004.93.1

Е.В. ВОЛЧЕНКО, к.т.н., доц. ГУИиИИ, г. Донецк

СЕТОЧНЫЙ ПОДХОД К ПОСТРОЕНИЮ ВЗВЕШЕННЫХ ОБУЧАЮЩИХ ВЫБОРОК W-ОБЪЕКТОВ В АДАПТИВНЫХ СИСТЕМАХ РАСПОЗНАВАНИЯ[1]

В работе рассматривается проблема формирования эффективных обучающих выборок в адаптивных системах распознавания. Предложен метод построения взвешенных выборок w-объектов на основе сеточного подхода. Выполнена оценка предложенного метода, доказана его сходимость и вычислена временная сложность. Приведены результаты экспериментальных исследований, подтверждающие высокое качество получаемых выборок w-объектов. Библиогр.: 14 назв.

Ключевые слова: адаптивная система распознавания, w-объект, обучающая выборка, сеточный подход.

Постановка проблемы и анализ литературы. Постоянно увеличивающийся объем информации, скорость изменения объектов и процессов, происходящих в окружающем мире, требует разработки подходов, позволяющих адаптировать автоматические системы к таким изменениям. Данная проблема является одной из наиболее актуальных в современной теории построения систем распознавания [1].

Современные системы распознавания должны не только обеспечивать эффективность классификации объектов по исходным данным, но и отвечать требованиям адаптивности и работы в режиме реального времени [2, 3]. Под адаптивностью принято понимать [2] способность системы изменять свои свойства (словарь признаков, обучающую выборку, решающие правила классификации и т.д.) в соответствии с изменениями распознаваемых объектов. Работа в реальном времени требует построения решающих правил, позволяющих принять решение о классификации за выделенное время [3].

Информация об изменении распознаваемых объектов поступает в системы распознавания в большинстве случаев в виде новых объектов обучающей выборки, количество которых может достигать десятков тысяч, поэтому для адаптивных систем одной из ключевых проблем является проблема предобработки исходных выборок данных.

[1] Работа выполнена при содействии гранта Президента Украины для поддержи научных исследований молодых ученых №GP/F32/130 "Разработка теоретических основ и методов реализации открытых обучающихся систем автоматического распознавания: способы оптимизации обучающих выборок и методы построения взвешенных решающих правил классификации".

Предобработка данных в системах распознавания является итеративным процессом и включает:

1) очистку данных, которая заключается в удалении шума, пропусков в данных и данных низкого качества [1];

2) сжатие данных, включающее нахождение минимального признакового пространства и репрезентативного множества данных на основе методов редукции и трансформации [2];

3) объединение данных, позволяющее уменьшить объем данных с сохранением исходной информации с помощью эвристических алгоритмов [1, 3].

Задача сжатия данных при условии неизменяющегося словаря признаков может быть решена двумя способами. Первый способ заключается в отборе некоторого множества объектов исходной обучающей выборки, каждый из которых отвечает предъявляемым требованиям. Наиболее известными алгоритмами, реализующими такой способ, являются алгоритмы STOLP [4], FRiS-STOLP [5], NNDE (Nearest Neighbor Density Estimate) и MDCA (Multiscale Data Condensation Algorithm) [1]. Основными отличиями этих алгоритмов друг от друга являются способ отбора объектов, используемое расстояние и критерий оптимальности полученной выборки. Второй способ состоит в построении множества новых объектов, каждый из которых строится по информации о некотором подмножестве объектов исходной обучающей выборки и обобщает его. Основой алгоритмов данного типа является дискретизация пространства признаков и анализ полученных частей пространства независимо друг от друга [6]. К алгоритмам такого типа можно отнести, например, алгоритм ДРЭТ [4], покрывающий признаковое пространство множеством пересекающихся окружностей, определяя тем самым области, принадлежащие каждому из классов.

Одним из перспективных подходов данного направления является наложение на пространство признаков некоторой «решетки», делящей все пространство на прямоугольные области, называемые в дальнейшем клетками. К алгоритмам, реализующим такой подход, относятся алгоритмы LVQ (learning vector quantization) [6], алгоритм четкого разбиения пространства признаков [7], алгоритм GridDC [8]. Основным отличием алгоритма GridDC от двух других является формирование на выходе новой сокращенной обучающей выборки, а не множества классифицированных клеток, являющихся одновременно и решающим правилом классификации. Недостатком формирования классифицированных клеток является то, что получаемое разбиение в большинстве случаев очень грубо аппроксимирует границы классов и может оказаться как чрезмерно избыточным, так и крайне недостаточным

по числу выделенных клеток [8]. Формирование выборки объектов позволяет использовать для построения решающих правил известные алгоритмы, дающие более эффективные решения классификации.

В предыдущих работах автора, например в [9], была предложена идея перехода к взвешенным сокращенным выборкам w-объектов, имеющим кроме значений признаков дополнительный параметр – вес. 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.

Данная работа является обобщением и развитием идей, предложенных автором в работах [8, 10, 11] и посвящена разработке метода построения взвешенных выборок w-объектов на основе сеточного подхода в адаптивных обучающихся системах распознавания.

Цель статьи – разработка метода построения взвешенной обучающей выборки w-объектов на основе сеточного подхода по исходной выборке и при добавлении новых обучающих объектов в процессе работы системы.

Постановка задачи. В качестве исходных данных дано некоторое множество объектов $X = \{X_1, X_2, \ldots, X_k\}$, представленное в виде объединения непересекающихся классов $X = \bigcup_{i=1}^{l} V_i$ и называемое обучающей выборкой. Каждый объект X_i из X описывается системой признаков, т.е. $X_i = \{x_{i1}, x_{i2}, \ldots, x_{in}\}$, и представляется точкой в линейном пространстве признаков, т.е. $X_i \in R^n$. Для каждого объекта X_i известна его классификация $y_i \in [1,l]$.

Определение 1. Качеством классификации $\frac{N(Z, X)}{|Z|}$ назовем частоту неверной классификации объектов контрольной выборки Z решающим правилом, построенным по исходной обучающей выборке X.

Через $|X|$ будем определять мощность множества X, т.е. количество объектов, принадлежащих этому множеству.

Необходимо сформировать классифицированную взвешенную обучающую выборку w-объектов $X^W = \{X_1^W, X_2^W, \ldots, X_m^W\}$, $y_i^W \in [1, l]$, где $X_i^W = \{x_{i1}, x_{i2}, ..., x_{in}, p_i\}$, p_i – вес i-го w-объекта.

Построенная выборка w-объектов должна отвечать следующим требованиям:

а) размер выборки w-объектов должен быть меньше размера исходной обучающей выборки $\left|X\right| < \left|X^W\right|$;

б) каждый объект новой выборки (w-объект) должен относиться к тому же классу, что и объекты, по которым он сформирован;

в) качество классификации $\frac{N(Z, X^W)}{|Z|}$ решающим правилом, построенным по выборке w-объектов, должно удовлетворять неравенству $\frac{N(Z, X^W)}{|Z|} \geq \frac{N(Z, X)}{|Z|}$.

Построение выборки w-объектов по исходной обучающей выборке на основе алгоритма GridDC. Идеей метода GridDC [8] является наложение сетки на признаковое пространство для формирования множества клеток, определение объектов выборки, принадлежащих каждой из клеток, и их замена на w-объекты. Формирование объектов новой выборки выполняется только в случае принадлежности всех объектов клетки одному классу. Вес w-объектов определяется по количеству объектов исходной выборки, принадлежащих клетке.

Далее приведем пошаговое описание метода. Без потери общности получаемых решений применим стандартный для теории распознавания подход, заключающийся в рассмотрении двухклассовых систем.

Шаг 1. Формирование сетки. Рассчитывается шаг клетки *s* по формуле:

$$s = \left\lfloor 1 + \frac{\left(\sum_{i=1}^{n} (\max\{x_i\} - \min\{x_i\}) \right)^n * (\lfloor \ln(k) \rfloor - 1)}{n * \prod_{i=1}^{n} (\max\{x_i\} - \min\{x_i\})} \right\rfloor,$$

где $\lfloor \ldots \rfloor$ – оператор округления до ближайшего целого значения; $\max\{x_i\}$ – максимальное значение *i*-го признака среди всех объектов выборки, $\min\{x_i\}$ – минимальное значение.

Выполняется разбиение признакового пространства R^n по каждому из n признаков на интервалы длиной s (наложение прямоугольной сетки), результатом которого является множество клеток G. Далее для каждого объекта выборки X определяется клетка, которой этот объект принадлежит.

Утверждение 1. Объект X_i принадлежит некоторой клетке G_j тогда и только тогда, когда каждое из значений его признаков входит в интервал значений соответствующих признаков данной клетки.

В результате формирования сетки и обработки объектов исходной обучающей выборки будут сформированы непересекающиеся подмножества X_{G_j} объектов, принадлежащих клетке G_j, $j = \overline{1,|G|}$.

Шаг 2. Формирование значений признаков w-объектов.

Возможны следующие варианты обработки содержимого клеток.

1. Если все объекты клетки принадлежат к одному классу, то значения признаков объекта новой выборки рассчитываются как координаты центра масс объектов этой клетки:

$$x_{jt} = \frac{1}{\left|X_{G_j}\right|} \sum_{X_i \in X_{G_j}} x_{it}\ ,\ t = \overline{1,n}\ .$$

2. Если клетка не содержит ни одного объекта, то объект новой выборки не формируется.

3. Если клетка содержит объекты нескольких классов, то она делится на две равные по размеру клетки (поочередно вертикально или горизонтально) до тех пор, пока любая из клеток внутри начальной клетки не будет содержать объекты только одного класса. Далее по каждой из полученных клеток формируются объекты новой выборки (согласно случаям 1 и 2).

Классификация w-объекта определяется по классификации объектов, по которым он сформирован.

Шаг 3. Определение веса w-объектов. Вес w-объекта равен количеству объектов исходной выборки, принадлежащих клетке, т.е.

$$p_j = \left|X_{G_j}\right|.$$

В результате выполнения алгоритма будет получена новая взвешенная обучающая выборка w-объектов X^W.

Пополнение взвешенной обучающей выборки w-объектов. Задачу пополнения выборки w-объектов новыми обучающими объектами целесообразно рассматривать относительно только одного некоторого объекта $X_a = \{x_{a1}, x_{a2}, \ldots, x_{an}\}$, набор признаков которого совпадает с набором признаков объектов исходной выборки, поскольку новые обучающие объекты поступают в систему распознавания последовательно и обрабатываются аналогично. Для решения задачи пополнения обучающей выборки w-объектов новыми классифицированными объектами в процессе работы системы на основе сеточного подхода предлагается следующий алгоритм.

Шаг 1. Определяется клетка, которой принадлежит добавляемый объект X_a.

Шаг 2. Выполняется корректировка существующего или построение нового w-объекта, соответствующего найденной клетке.

1. Если найденный w-объект относится к тому же классу, что и добавляемый объект, то вес w-объекта увеличивается на единицу и значения его признаков пересчитываются:

$$x_{jt} = \frac{x_{jt} + x_{at}}{2}, \quad t = \overline{1,n}, \quad p_j = p_j + 1.$$

2. Если найденной клетке не соответствует ни один w-объект (т.е. ни один объект исходной выборки или добавленный ранее не принадлежал этой клетке), то формируется новый w-объект со значениями признаков, равными значениям признаков добавляемого объекта, и единичным весом:

$$x_{jt} = x_{at}, \quad t = \overline{1,n}, \quad p_j = 1.$$

3. Если классификация найденного w-объекта не совпадает с классификацией добавляемого объекта, то выполняется деление этой клетки на две равные по размеру клетки (поочередно вертикально или горизонтально), как и при построении w-объектов по исходной выборке, до тех пор, пока все построенные клетки не будут содержать объекты только одного класса.

Далее по каждой из полученных клеток формируются w-объекты по следующим правилам.

Если клетка содержит только добавляемый объект X_a, то построение w-объекта аналогично случаю 2 данного алгоритма.

Для остальных полученных клеток формируются w-объекты со значениями признаков, равными центрам этих клеток и весом пропорционально количеству построенных клеток, т.е.

$$p'_j = \left\lfloor p_j \cdot \frac{1}{q} \right\rfloor,$$

где q – количество клеток, полученных после разбиения исходной клетки.

Отметим, что разбиение клетки на две равные клетки вне зависимости от количества объектов в этой клетке, принадлежащих разным классам не влияет на эффективность работы систем распознавания, поскольку при построении решающих правил учитывается вес w-объектов.

Оценка метода построения выборки w-объектов на основе сеточного подхода. Анализ предложенного метода и обучающей выборки w-объектов позволяет сформулировать следующие утверждения.

Утверждение 2. Выборка w-объектов формируется по всем объектам исходной выборки.

Утверждение 3. Никакие два и более w-объектов не строятся по одному и тому же объекту исходной выборки.

Утверждение 4. W-объект принадлежит тому же классу, что и объекты исходной выборки, по которым он сформирован. Количество классов выборки w-объектов равно количеству классов объектов исходной выборки.

Утверждение 5. Вес w-объекта является целым числом и принимает значения от 1 до количества объектов $\left|V_j\right|$ некоторого класса j.

Доказательство корректности данных утверждений непосредственно следует из имеющихся исходных данных и предложенных алгоритмов.

Теорема 1. Алгоритм построения выборки w-объектов сходится и его временная сложность равна $O(k \log k)$.

Доказательство. Основными элементами, обрабатываемыми в предложенных алгоритмах, являются множества обучающих объектов, принадлежащих клеткам. Количество клеток, на которые разбивается n-мерное признаковое пространство равно

$$N_{\mathrm{G}} = \left\lfloor \prod_{i=1}^{n} \left(\frac{\max\{x_i\} - \min\{x_i\}}{s} \right) \right\rfloor.$$

Обработка клеток проводится до тех пор, пока каждая клетка, полученная путем разбиения исходной, не будет содержать объекты только одного класса. Поскольку, согласно постановке задачи, никакие два объекта обучающей выборки не могут иметь одинаковые значения всех признаков и принадлежать разным классам, количество разбиений клеток конечно. Следовательно, алгоритм сходится за конечное число шагов.

Рассчитаем временную сложность алгоритма построения выборки w-объектов по исходной выборке. За единицу примем время обработки одного объекта исходной выборки.

На первом шаге алгоритма для расчета шага клетки выполняется поиск минимальных и максимальных значений признаков объектов, что требует nk итераций.

На втором шаге определяется принадлежность объектов клеткам, что требует k итераций.

На третьем шаге выполняется построение w-объектов по клеткам. При этом разбиение клеток для обеспечения однозначной классификации требует выполнения не более чем $2^{|V_j|}$ итераций.

Таким образом, временная сложность алгоритма построения выборки w-объектов по исходной выборке составляет $nk + k + N_G 2^{|V_j|}$ итераций. Если предположить, что каждая клетка содержит строго по одному объекту выборки, то $N_G = k$. Если предположить, что некоторая клетка содержит все объекты некоторого класса j, количество объектов которого близко размеру всей выборки, то $2^{|V_j|} \approx 2^k$. Следовательно, временная сложность алгоритма равна $O(k \log k)$.

Отметим, что временная сложность алгоритма пополнения взвешенной обучающей выборки w-объектов рассчитывается аналогично и также составляет $O(k \log k)$.

Результаты экспериментальных исследований. Для оценки эффективности предложенного метода был проведен ряд экспериментальных исследований. В качестве исходных данных были использованы сформированные по нормальному закону распределения выборки объектов двух классов размером 1000 – 5000 объектов, содержащих два признака распознавания. Степень пересечения классов в пространстве признаков изменялась от полной обособленности до пересечения на 40%. В качестве решающего правила классификации по выборке w-объектов использовались модифицированный метод

k-ближайших соседей [12] и модифицированный метод потенциальных функций [13]. Классификация объектов в модифицированном методе *k*-ближайших соседей определяется по k ближайшим w-объектам к распознаваемому объекту $X'=\{x_1, x_2, \ldots, x_n\}$ по следующей метрике:

$$F(X_j^W, X') = \frac{p_j \cdot p_a}{r_{ja}^2} = \frac{p_j \cdot p_a}{\left\| X_j^W - X' \right\|} = \frac{p_j \cdot p_a}{\sum_{t=1}^{n} \left(x_{jt} - x_{at} \right)^2},$$

где $p_a = 1$ – вес распознаваемого объекта, который принимается равным единице.

Два объекта являются ближайшими, если значение $F(X_j^W, X')$ максимально. Объект X'_i относится к тому классу, объектов которого среди k ближайших больше.

Классификация объектов в модифицированном методе потенциальных функций выполняется с помощью решающего правила вида:

$$U(X^W, X') = \sum_{j=1}^{m} U(X_j^W, X'),\ U(X_j^W, X') = \exp(-\alpha \cdot \frac{1}{F(X_j^W, X')}) \cdot$$

Анализ полученных результатов позволяет сделать следующие выводы:

1) частота неверной классификации объектов тестовой выборки методом k-ближайших соседей по выборке w-объектов на 7,4% меньше частоты неверной классификации методом *k*-ближайших соседей по исходной выборке и на 1,6% меньше частоты неверной классификации методом потенциальных функций по исходной выборке;

2) время выполнения классификации предложенным методом на 3,8% меньше времени классификации методом *k*-ближайших соседей по исходной выборке и на 37,9% меньше времени классификации методом потенциальных функций по исходной выборке.

Анализ эффективности использования выборок w-объектов, построенных по предложенному методу, для классификации объектов класса тестовых задач ADS 1 репозитория ISEC (International Statistical Education Centre) [14] показал уменьшение неверных классификаций по сравнению с классическими методами (методом *k*-ближайших соседей и методом потенциальных функций) в среднем на 3,4%.

Также было получено, что из 100 добавляемых новых объектов только для 7 объектов потребовалось построение новых w-объектов, а для 4 – разделение существующих клеток.

Выводы. В работе предложен новый подход к построению взвешенных обучающих выборок w-объектов на основе сеточных алгоритмов в адаптивных системах распознавания. Описан метод формирования значений признаков w-объектов и их весов по исходной обучающей выборке и при добавлении в выборку новых обучающих объектов. Анализ предложенного метода показал его сходимость, низкую временную сложность, корректность обработки объектов исходной выборки. Результаты экспериментальных исследований показали уменьшение частоты неверных классификаций и времени её выполнения решающим правилом, построенным по выборке w-объектов, по сравнению с классическими методами.

Список литературы: 1. *Larose D.T.* Discovering knowledge in Data: An Introduction to Data Mining / *D.T. Larose.* – New Jersey, Wiley & Sons, 2005. – 224 p. **2.** *Pal S.K.* Pattern Recognition Algorithms for Data Mining: Scalability, Knowledge Discovery and Soft Granular Computing / *S.K. Pal, P. Mitra.* – Chapman and Hall/CRC, 2004. – 280 p. **3.** *Olson D.L.* Advanced Data Mining Techniques / *D.L. Olson, D. Delen.* – Springer-Verlag Berlin, 2008. – 180 p. **4.** *Загоруйко Н.Г.* Прикладные методы анализа знаний и данных / *Н.Г. Загоруйко.* – Новосибирск: Издательство института математики, 1999. – 270 с. **5.** *Zagoruiko N.G.* Methods of Recognition Based on the Function of Rival Similarity */ N.G. Zagoruiko, I.A. Borisova, V.V. Dyubanov, and O.A. Kutnenko* // Pattern Recognition and Image Analysis. – 2008. – Vol. 18. – №. 1. – P. 1–6. **6.** *Kohonen T.* Self-Organizing Maps */ T. Kohonen.* – Springer-Verlag, 1995. – 501 p. **7.** *Субботин С.А.* Метод обучения нейро-нечеткой сети распознаванию образов на основе прямоугольного разбиения пространства признаков / *С.А. Субботин* // Складні системи і процеси. – 2009. – № 1. – С. 111–111. **8.** *Волченко Е.В.* Метод сокращения обучающих выборок GridDC / *Е.В. Волченко, И.В. Дрозд* // Искусственный интеллект. – 2010. – № 4. – С. 185–190. **9.** *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. **10.** *Волченко Е.В.* Пополнение взвешенных обучающих выборок w-объектов в адаптивных системах распознавания, построенных на основе сеточного подхода / *Е.В. Волченко* // Труды Международной научной конференции "Интеллектуальные системы принятия решений и проблемы вычислительного интеллекта (ISDMCI'2011)". – Херсон: ХНТУ, 2011 – Том 2. – С. 205-209. **11.** *Волченко Е.В.* Построение взвешенных обучающих выборок w-объектов на основе сеточного подхода / *Е.В. Волченко* // Доклады 15-й Всероссийской конференции "Математические методы распознавания образов (ММРО-15)", Петрозаводск, 2011. – М.: Макс-Пресс, 2011. **12.** *Волченко Е.В.* Адаптация решающих правил в открытых системах распознавания / *Е.В. Волченко* // Тезисы докладов VII международной научно-практической конференции "Математическое и программное обеспечение интеллектуальных систем" (MPZIS-2009). – Днепропетровск, 2009. – С. 62–64. **13.** *Волченко Е.В.* Модифицированный метод потенциальных функций / *Е.В. Волченко* // Бионика интеллекта. – 2006. – № 1 (64). – С. 86–92. **14.** http://www.isical.ac.in/~miu

Статья представлена д.ф.-м.н. проф., проректором по науч.-педагог. и учеб. работе ГУИиИИ МОН Украины Миненко А.С.

УДК 004.93.1

Сітковий підхід побудови зважених навчаючи вибірок w-об'єктів у адаптивних системах розпізнавання / Волченко О.В. // Вісник НТУ "ХПІ". Тематичний випуск: Інформатика і моделювання. – Харків: НТУ "ХПІ". – 2011. – № 36. – С. 12 – 22.

У роботі розглянуто проблему формування ефективних навчаючих вибірок у адаптивних системах розпізнавання. Запропоновано метод побудови зважених вибірок w-об'єктів на основі сіткового підходу. Виконано оцінку запропонованого методу, доведено його збіжність та підраховано часову складність. Наведено результати експериментальних досліджень, що підтверджують високу якість отримуваних вибірок w-об'єктів. Бібліогр.: 14 назв.

Ключові слова: адаптивна система розпізнавання, w-об'єкт, навчаюча вибірка, сітковий підхід.

UDC 004.93.1

Grid approach to the construction of weighted training samples of w-objects in adaptive recognition systems / Volchenko E.V. // Herald of the National Technical University "KhPI". Subject issue: Information Science and Modelling. – Kharkov: NTU "KhPI". – 2011. – № 36. – P. 12 – 22.

Problem of the effective training samples formation in adaptive recognition systems is considered in the work. Method of constructing a weighted samples of w-objects based on the grid approach is proposed. Estimation of the proposed method has been implemented, and proved its convergence and time complexity. Experimental results confirming the high quality of w-objects samples are presented. Refs.: 14 titles.

Keywords: adaptive recognition system, w-object, training samples, grid-base approach.

Поступила в редакцию 15.07.2011