**УДК 004.93.1**

***Е.В. ВОЛЧЕНКО***, канд. техн. наук, доц. ГУИиИИ (г. Донецк)

# РАСШИРЕНИЕ МЕТОДА ГРУППОВОГО УЧЕТА АРГУМЕНТОВ НА ВЗВЕШЕННЫЕ ОБУЧАЮЩИЕ ВЫБОРКИ W-ОБЪЕКТОВ

В работе рассматривается задача построения решающих правил классификации в адаптивных системах распознавания. Рассмотрена возможность построения решающих правил методом группового учета аргументов по взвешенной выборке w-объектов. Предложены способы использования веса w-объектов в алгоритмах метода группового учета аргументов. Приведены результаты экспериментальных исследований, подтверждающие высокое качество получаемых решающих правил. Табл.: 2. Библиогр.: 13 назв.

**Ключевые слова**: решающее правило, метод группового учета аргументов, взвешенная выборка w-объектов.

**Постановка проблемы и анализ литературы.** Исследования в области построения систем автоматического распознавания в последние годы связаны, в первую очередь, с расширением области их практического использования [1, 2]. Это связано с разработкой большого количества разнообразных устройств (роботов, систем технической и медицинской диагностики, персональных, мобильных и карманных компьютеров), автоматическая работа которых невозможна без распознавания текущего состояния объектов, процессов, явлений и состояний, с которыми эти устройства работают. Не меньшее влияние на развитие систем распознавания оказывают новые информационные технологии, в том числе и технологии Интернет. В этом направлении на основе построения систем распознавания решаются задачи построения новостных порталов и электронных библиотек с автоматической рубрикацией документов, почтовых серверов с возможностью "спам"-фильтрации электронной корреспонденции.

Большинство современных прикладных задач, решаемых путем построения систем распознавания, характеризуется большим объемом исходных данных и возможностью добавления новых данных уже в процессе работы систем. Именно поэтому основными требованиями, предъявляемыми к современным системам распознавания, являются:

– адаптивность, состоящая в возможности системы в процессе работы изменять свои характеристики (для обучающихся систем распознавания – корректировать решающие правила классификации) при изменении окружающей среды (добавлении новых объектов обучающей выборки);

– работа в реальном времени, предполагающая наличие возможности формирования решений о классификации за ограниченное время;

– высокая эффективность классификации для линейно разделимых и пересекающихся в признаковом пространстве классов [3].

Системы распознавания, отвечающие таким требования, принято называть адаптивными системами распознавания.

Добавление новых объектов, являющееся для большинства прикладных задач достаточно интенсивным, приводит к существенному увеличению размера обучающей выборки, что, в свою очередь, приводит к:

1) увеличению времени корректировки решающих правил (особенно, если используемый метод обучения требует построения нового решающего правила, а не его частичной корректировки);

2) ухудшению качества получаемых решений из-за проблемы переобучения на выборках большого объема [4, 5].

Эти особенности адаптивных обучающихся систем распознавания делают необходимой предобработку обучающих выборок путем сокращения их размера. Известные алгоритмы сокращения размера выборок STOLP, ДРЭТ [6] FRiS-STOLP [7], NNDE (Nearest Neighbor Density Estimate) и MDCA (Multiscale Data Condensation Algorithm) [3] основаны на выборе некоторого подмножества объектов обучающей выборки и удалении остальных. Такой подход, на наш взгляд, не позволяет учитывать плотность распределения объектов выборки в пространстве признаков и динамику изменения значений признаков добавляемых объектов. В работе [8] нами был предложен метод сокращения обучающей выборки путем построения взвешенной выборки w-объектов, позволяющий выполнять сокращение обучающей выборки при не ухудшении качества классификации и решить проблему добавления новых объектов в выборку при минимальном увеличении её размера.

Для обеспечения работы системы в реальном времени предпочтительным является использование методов построения решающих правил, предполагающих их корректировку, например, путем добавления одного или нескольких слагаемых. Одним из наиболее эффективных методов такого типа является метод группового учета аргументов (МГУА) [9, 10]. В основе данного метода лежат не только принципы обучения с учителем, но и самоорганизация, характерная для самообучающихся систем, что позволяет выполнять направленный адаптивный поиск оптимальных решений. Особенностью данного метода является использование малых выборок из-за накладываемых ограничений на сложность используемых полиномов, что может приводить к ухудшению качества получаемых решений [11]. Использование в данном методе взвешенных обучающих выборок w-объектов позволит, на наш взгляд, увеличить эффективность получаемых решающих правил и расширить данный метод на выборки большого объема.

**Цель статьи** – расширение метода группового учета аргументов для построения решающих правил классификации в обучающихся системах распознавания по взвешенным обучающим выборкам w-объектов.

В качестве исходных данных дано некоторое множество объектов *M*, называемое обучающей выборкой. Каждый объект $X_i$ из $M$ описывается системой из $n$ 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 $y_i \in [1, l]$, где $l$ – число классов системы.

Далее опишем метод построения взвешенной выборки w-объектов, алгоритм построения решающих правил классификации на основе МГУА и особенности его реализации по взвешенной выборке w-объектов.

**Метод построения выборки w-объектов.** В работе [8] нами был предложен метод построения взвешенной обучающей выборки w-объектов для сокращения выборок большого объема в адаптивных системах распознавания. Основой данного метода является выбор множеств близкорасположенных объектов исходной выборки и их замена одним взвешенным объектом новой выборки. Значения признаков каждого объекта новой выборки являются центрами масс значений признаков объектов исходной выборки, которые он заменяет. Введенный дополнительный параметр – вес определяется как количество объектов исходной выборки, которые были заменены одним объектом новой выборки. Предлагаемый метод ориентирован как на сокращение исходной обучающей выборки, так и на анализ необходимости корректировки выборки и быстрое выполнение такой корректировки при пополнении выборки в процессе работы системы.

Построение w-объекта состоит из трех последовательных этапов:

1) построение образующего множества $W_f$, содержащее некоторое количество $d$ объектов исходной выборки, принадлежащих одному классу;

2) формирование вектора $X_i^W = \{x_{i1}, x_{i2}, \ldots, x_{in}\}$ значений признаков w-объекта $X_i^W$ и расчет его веса $p_i$;

3) корректировка исходной обучающей выборки – удаление объектов, включенных в образующее множество $X = X \setminus W_f$.

Построение образующего множества $W_f$ состоит в нахождении начальной точки $X_{f1}$ формирования w-объекта, определении конкурирующей точки $X_{f2}$ и выборе в образующее множество $W_f$ таких объектов исходной выборки, расстояние до каждого из которых от начальной точки меньше, чем расстояние от них до конкурирующей точки. В качестве начальной точки $X_{f1}$ формирования w-объекта используется объект исходной обучающей выборки, наиболее удаленный от всех объектов других классов. Конкурирующая точка $X_{f2}$ выбирается путем нахождения ближайшего к $X_{f1}$ объекта, не принадлежащего тому же классу, что и сам $X_{f1}$, т.е. $y_{f1} \neq y_{f2}$.

Для случая двух классов выбор объектов $\{X_{f1}, X_{f2}, ..., X_d\}$ образующего множества $W_f$ осуществляется по следующему правилу: объект $X_i$ включается в $W_f$, если:

1) он принадлежит тому же классу, что и начальная точка $X_{f1}$;

2) расстояние от рассматриваемого объекта до начальной точки $X_{f1}$ меньше, чем до конкурирующей точки $X_{f2}$;

3) расстояние $R_{i,1}$ от рассматриваемого объекта до начальной точки меньше расстояния $R_{i,2}$ от рассматриваемого объекта до конкурирующей точки и меньше расстояния $R_{1,2}$ между начальной и конкурирующей точками (для случая, когда классы состоят из нескольких отдельных областей признакового пространства).

Таким образом, образующее множество $W_f$ формируется по правилу:

$$W_f = X_{f1} \cup X_{f2} \cup \{X_i \mid R_{i,1} < R_{i,2} < R_{1,2}\}, \quad (1)$$

где $f_1 = \arg\max\limits_{i=1,..,d} \sum\limits_{j=1}^{d} R(X_i, X_j)$,

$f_2 = \arg \min\limits_{\substack{j=1,..,d \\ y_j \neq y_{f1}}} R(X_j, X_{f1})$,

$R_{a,b} = R(X_a, X_b) = \sum\limits_{t=1}^{n} (x_{at} - x_{bt})^2$.

Значения признаков $\{x_{i1}, x_{i2}, ..., x_{in}\}$ нового w-объекта $X_i^W$ формируются по образующему множеству $W_f$ и рассчитываются как координаты центра масс системы из $p_f = |W_f|$ материальных точек (примем, что объекты исходной обучающей выборки, являющиеся в признаковом пространстве материальными точками, имеют массу, равную 1), где $|W_f|$ – мощность множества $W_f$, т.е.

$$x_{it} = \frac{1}{p_f} \sum_{X_j \in W_f} x_{jt}. \quad (2)$$

После формирования очередного w-объекта, все объекты образующего его множества удаляются из исходной обучающей выборки, т.е. $X = X \setminus W_f$. Алгоритм заканчивает свою работу, когда в исходной обучающей выборке не останется ни одного объекта $X = \varnothing$.

**Построение решающих правил классификации с использованием метода группового учета аргументов.** Метод группового учета аргументов основан на принципах теории обучения и самоорганизации, в частности, на принципе массовой "селекции" или самоорганизующемся направленном переборе всевозможных вариантов построения решающего правила классификации с отсечениями [12]. Задача построения решающего правила в МГУА представляется как задача индуктивного построения модели, усложняющейся в процессе работы алгоритма.

Искомой с помощью МГУА моделью (решающим правилом классификации) для рассматриваемой в данной работе задачи будет решающее правило классификации, представляемое в виде полинома Колмогорова-Габбора

$$g(\bar{x}) = \alpha_0 + \sum_{i=1}^{n} \alpha_i \cdot x_i + \sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_{ij} \cdot x_i x_j + \sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{\gamma=1}^{n} \alpha_{ij\gamma} \cdot x_i x_j x_\gamma + \dots ,$$

где $\alpha = \{\alpha_0, \alpha_i, \alpha_{ij}, \alpha_{ij\gamma}, \dots | \, i, j, \gamma = \overline{1,n}\}$ – коэффициенты.

Для определения вектора коэффициентов математической модели по выборке данных применяют метод наименьших квадратов (МНК).

Задача построения решающего правила классификации решается в несколько этапов [11]. Вначале из всех независимых переменных (аргументов) образуются всевозможные группы комбинаций. Традиционно каждая такая комбинация содержит только два аргумента и образует простой полином малого порядка (элементарное решающее правило классификации). Далее выполняется классификация объектов контрольной выборки $S_P$ полученными полиномами и каждый из них оценивается по выбранному критерию селекции. Среди всех полиномов отбирается $R_i$ лучших полиномов, которые являются аргументами новых частных полиномов следующего уровня обобщения и процесс повторяется.

Общая эффективность МГУА обеспечивается следующими принципами:

1) принцип неокончательности промежуточных решений, который состоит в том, что ни одно из полученных на первом уровне решений не принимается за истину, и только часть решений пропускается для дальнейшего усложнения решающего правила;

2) принцип внешнего дополнения, согласно которому для оценки качества полиномов используется контрольная выборка, не участвовавшая в определении коэффициентов полинома;

3) принцип самоотбора, согласно которому на следующий уровень обобщения пропускаются только лучшие полиномы предыдущего уровня;

4) принцип единственности окончательного решения, согласно которому усложнение полиномов происходит до тех пор, пока не прекратится улучшение качества получаемых решающих правил.

Согласно принципу обработки новых объектов обучающей выборки, поступающих в процессе работы системы распознавания, изложенного в [13], данная схема может быть усилена за счет пополнения контрольной выборки новыми объектами и выполнения новых уровней обобщения при неверной классификации добавляемых объектов, что подтверждает эффективность использования МГУА в адаптивных системах распознавания. Для решения проблемы обработки выборок большого размера далее предложим способы учета веса w-объектов при построении полиномов в МГУА.

**Способы использования веса w-объектов при построении решающего правила методом группового учета аргументов.** Для построения оптимального по выбранным параметрам алгоритма МГУА с использованием взвешенной выборки w-объектов предложим возможные способы учета веса w-объектов.

1. Использование веса w-объектов при расчете весовых коэффициентов частных полиномов любого уровня обобщения. При использовании такого способа частный полином, создаваемый, например, на первом уровне алгоритма, будет иметь вид:

$$g(\bar{x}) = \alpha_0 + \alpha_1 k_i x_i + \alpha_2 k_j x_j + \alpha_3 k_i k_j x_i x_j ,$$

где $k_i$ и $k_j$ – вес $i$-го и $j$-го w-объектов соответственно.

Данный способ включения веса w-объектов в частные полиномы позволяет оказывать существенное влияние на принимаемые решения о классификации объектов, имеющих большой вес.

2. Включение веса w-объектов в выбранный критерий селекции. Тогда, например, критерий "минимум смещения плюс регулярность" будет иметь вид

$$\rho_1 = \sqrt{\frac{\sum_{i-1}^{\alpha \cdot n} k_i \cdot (y(s)_{R_1} - y(s)^2_{R_2})}{\alpha \cdot \sum_{i=1}^{n} k_i \cdot y(s)^2_{R_1}} + \frac{\sum_{i=1}^{n_B} k_i \cdot (y(S_i)_{табл} - y(S_i)_M)^2_i}{\sum_{i=1}^{n_B} k_i \cdot y(S_i)^2_{табл}}} \to \min.$$

Отметим, что данный способ учета веса w-объектов позволяет существенно увеличить точность получаемых классификаторов.

3. Использование веса w-объектов для сокращения количества частных полиномов, создаваемых на каждом уровне алгоритма. При реализации такого подхода из всех $Z_N = C^2_{R_{N-1}} = \frac{R_{N-1} \cdot (R_{N-1} - 1)}{2}$ полиномов создается только некоторая их часть, имеющая максимальное значение веса объектов, входящих в полином.

4. Использование веса w-объектов в алгоритмах с базисными переменными. На любом уровне алгоритма в частные полиномы происходит включение объектов исходной выборки. При использовании взвешенной

обучающей выборки определение объектов, включаемых в полиномы, может осуществляться с учетом их веса.

**Результаты экспериментальных исследований.** Для оценки качества решающих правил, получаемых с помощью МГУА по взвешенной обучающей выборке, был проведен ряд экспериментальных исследований. Оцениваемыми параметрами базового и модифицированного алгоритмов МГУА являлись:

– качество классификации объектов контрольной выборки (количество неверных классификаций объектов контрольной выборки построенным решающим правилом – $N(S_k)$);

– количество членов полинома $mk$, от которого зависит скорость выполнения классификации распознаваемых объектов.

Для проведения исследований использовались исходные выборки, созданные по нормальному закону распределения, размером 1500 объектов при различной площади пересечения двух классов, и контрольные выборки размером 100 объектов. Результаты исследований по выбору оптимального способа учета веса w-объектов, описанные выше, усредненные по 50 экспериментам, приведены в табл. 1.

Таблица 1

Зависимость длины решающего правила и количества неверно классифицированных объектов контрольной выборки от способа включения веса w-объектов в искомый полином при изменяющемся размере исходной обучающей выборки

| Площадь пересеч. классов, % | Расчет весовых коэффициентов | | Включение в критерий селекции | | Сокращение кол-ва частных полиномов | | Реализация алгоритма базисных функций | |
|---|---|---|---|---|---|---|---|---|
| | $mk$ | $N(S_k)$ | $mk$ | $N(S_k)$ | $mk$ | $N(S_k)$ | $mk$ | $N(S_k)$ |
| Классы обособ-лены | 7,3 | 0 | 6,4 | 0 | 8,5 | 0 | 10,4 | 0 |
| 0 | 12,6 | 0,02 | 8,1 | 0,06 | 14,7 | 0,04 | 15,5 | 0,02 |
| 10 | 14,1 | 0,16 | 10,6 | 0,4 | 21,1 | 0,30 | 23,7 | 0,21 |
| 20 | 16,5 | 0,34 | 11,9 | 0,7 | 33,0 | 0,56 | 34,2 | 0,45 |
| 30 | 18,8 | 0,52 | 13,4 | 1,2 | 38,9 | 0,84 | 47,1 | 0,68 |
| 40 | 25,8 | 0,72 | 19,4 | 3,7 | 46,3 | 1,16 | 56,3 | 0,94 |

Анализ результатов экспериментальных исследований по выбору способа включения веса w-объектов в решающее правило классификации показал, что наиболее эффективным является способ использования веса объектов взвешенной обучающей выборки при расчете коэффициентов полиномов.

Для оценки эффективности использования взвешенного многорядного алгоритма метода группового учета аргументов по сравнению с базовым многорядным алгоритмом МГУА, использующим исходную обучающую выборку, также проведена серия испытаний, в которых выполнялось

построение решающих правил классификации по исходным выборкам различного размера. Площадь пересечения классов для этих выборок в пространстве признаков составляет 20%. Результаты испытаний, приведенные в табл. 2, являются средними по 50 экспериментам.

Таблица 2

Количество неверно классифицированных объектов контрольных выборок для анализируемых алгоритмов метода группового учета аргументов

| Размер обучающей выборки | $N(S_k)$ базовым алгоритмом МГУА | $N(S_k)$ взвешенным алгоритмом МГУА |
|---|---|---|
| 200 | 0,6 | 0,4 |
| 400 | 1,4 | 1,0 |
| 600 | 2,8 | 1,6 |
| 800 | 4,6 | 2,4 |
| 1000 | 6,8 | 3,0 |

Анализ результатов исследований показал, что независимо от размера обучающей выборки эффективность предложенного в данной работе взвешенного многорядного алгоритма МГУА превышает эффективность базового алгоритма МГУА, построенного по исходной выборке.

Также отметим, что количество уровней построения полиномов во взвешенном алгоритме МГУА в среднем на 20% меньше количества уровней, выполняемых базовым алгоритмом.

**Выводы.** В работе предложен подход к построению решающих правил классификации адаптивных систем распознавания методом группового учета аргументов по взвешенной выборке w-объектов. По результатам анализа особенностей алгоритмов построения решающих правил методом группового учета аргументов предложены способы включения веса w-объектов в решающие правила и дана оценка степени их влияния на получаемые решения. Проведенное экспериментальное исследование эффективности использования взвешенной обучающей выборки w-объектов в качестве исходных данных при построении решающих правил классификации показало уменьшение времени построения решающего правила классификации в среднем на 20%, увеличение скорости классификации распознаваемых объектов среднем на 16,5%, увеличение эффективности классификации объектов контрольных выборок.

Таким образом, на основе предлагаемого автором подхода к решению задачи построения решающих правил классификации методом группового учета аргументов по взвешенным обучающим выборкам w-объектов, была решена близкая к рассмотренным ранее задача, что позволяет говорить об универсальности и эффективности использования взвешенных обучающих выборок w-объектов для решения задач построения обучающихся систем распознавания.

**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.** *Giudici P.* Applied data mining: statistical

methods for business and industry / *P. Giudici.* – Chichester, Wiley & Sons, 2003. – 380 p. **3.** *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. **4.** *Olson D.L.* Advanced Data Mining Techniques / *D.L. Olson, D. Delen* – Springer-Verlag Berlin, 2008. – 180 p. **5.** *Liu H.* Selective Sampling Approach to Active Feature Selection / *H. Liu, H. Motoda, L. Yu* // Artificial Intelligence. – 2004. – V. 159. – № 1–2. – P. 49-74. **6.** *Загоруйко Н.Г.* Прикладные методы анализа знаний и данных / *Н.Г. Загоруйко*. – Новосибирск: Издательство института математики, 1999. – 270 с. **7.** *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. **8.** *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. **9.** *Ивахненко А.Г.* Помехоустойчивость моделирования / *А.Г. Ивахненко, В.С. Степашко*. – К.: Наукова думка, 1985. – 215 с. **10.** *Ивахненко А.Г.* Индуктивный метод самоорганизации моделей сложных систем / *А.Г. Ивахненко* – К.: Наукова думка, 1981. – 296 с. **11.** *Васильев В.И.* Взаимодополняемость метода группового учета аргументов (МГУА) и метода предельных упрощений (МПУ) / *В.И. Васильев* // Искусственный интеллект. – Донецк: ИПИИ, 2001. – № 1. – С. 29-42. **12.** *Белозерский Л.А.* Анализ и обработка априорной информации в конструировании систем автоматического распознавания (САРС) / *Л.А. Белозерский, А.И. Шевченк*о. – Донецк, ИПИИ "Наука і освіта", 2007. – 180 с. **13.** *Волченко Е.В.* Предобработка обучающих выборок в открытых системах автоматического распознавания / *Е.В. Волченко* // Труды Международной научной конференции "Интеллектуальные системы принятия решений и проблемы вычислительного интеллекта (ISDMCI'2009)". – Евпатория. – 2009. – С. 43-51.

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

УДК 004.93'1

**Розширення методу групового врахування аргументів на зважені навчаючі вибірки w-об'єктів / Волченко О.В.** // Вісник НТУ "ХПІ". Тематичний випуск: Інформатика і моделювання. – Харків: НТУ "ХПІ". – 2010. – № 31. – С. 49 – 57.

У роботі розглянуто задачу побудови вирішуючих правил класифікації в адаптивних системах розпізнавання. Розглянуто можливість побудови вирішуючих правил методом групового врахування аргументів по зваженій вибірці w-об'єктів. Запропоновано способи використання ваги w-об'єктів в алгоритмах методу групового врахування аргументів. Наведено результати експериментальних досліджень, що підтверджують високу якість отримуваних вирішуючих правил. Табл.: 2. Бібліогр.: 12 назв.

**Ключові слова:** вирішуюче правило, метод групового врахування аргументів, зважена вибірка w-об'єктів.

UDC 004/93'1

**Expansion of the group method of data handling on the weighted training sample of w-objects / Volchenko E.V.** // Herald of the National Technical University "KhPI". Subject issue: Information Science and Modelling. – Kharkov: NTU "KhPI". – 2010. – №. 31. – P. 49 – 57.

The problem of decision rules construction for classification in adaptive recognition systems is considered. The possibility of decision rules construction on weighted training samples of w-objects with the help of the group method of data handling is considered. The ways of using w-objects weights in the algorithms of the group method of data handling are offered. The effectiveness of the proposed method is shown on the test data. Tables: 2. Refs: 12 titles.

**Key words:** decision rule, group method of data handling, weighted sample of w-objects.

*Поступила в редакцию 13.05.2010*