УДК 004.93’11

Е.В. ВОЛЧЕНКО,канд.техн.наук, доц., Институт информатики и искусственного интеллекта ГВУЗ «ДонНТУ», Донецк,
В.С. СТЕПАНОВ, ассис., Институт информатики и искусственного интеллекта ГВУЗ «ДонНТУ», Донецк,

МЕТОД w-MIEF ПОСТРОЕНИЯ РАБОЧЕГО СЛОВАРЯ ПРИЗНАКОВ НА ОСНОВЕ ВЗВЕШЕННЫХ ОБУЧАЮЩИХ ВЫБОРОК

Роботу присвячено вирішенню задачі обробки початкових даних у адаптивних системах розпізнавання, що навчаються. Запропоновано метод w-MIEF побудови робочого словника ознак по зважених навчаючих вибірках w-об’єктів, що базується на відборі ознак із максимальною індивідуальною ефективністю класифікації. Наведено результати експериментальних досліджень, що підтверджують ефективність запропонованого методу.
Ключові слова: адаптивна система розпізнавання, навчаюча вибірка, w-об’єкт, словник ознак.

Работа посвящена решению задачи предобработки исходных данных в адаптивных обучающихся системах распознавания. Предложен метод w-MIEF построения рабочего словаря признаков по взвешенным обучающим выборкам w-объектов, основанный на отборе признаков с максимальной индивидуальной эффективностью классификации. Приведены результаты экспериментальных исследований, подтверждающие эффективность предложенного метода.
Ключевые слова: адаптивная система распознавания, обучающая выборка, w-объект, словарь признаков.

Problem of the effective training samples processing in adaptive recognition systems is considered in the work. Method w-MIEF of feature subset’s construction based on the selection of feature with maximum individual efficiency of recognition is proposed. Experimental results confirming the high quality of proposed method are presented.
Key words: adaptive recognition system, training samples, w-object, feature set.

1. Введение

Появление большого количества технических устройств, работа которых напрямую зависит от распознавания текущего состояния объектов, процессов, явлений и состояний, с которыми эти устройства работают, является одной из основных причин активного расширения области практического применения систем распознавания [1]. Немаловажную роль в расширении области применения систем распознавания играет развитие информационных технологий, в том числе и сети Интернет. В этой области на основе систем распознавания образов решаются задачи построения электронных библиотек с автоматической рубрикацией текстов, новостных порталов, почтовых серверов с возможностью фильтрации нежелательной электронной корреспонденции. Именно поэтому наряду с необходимостью обеспечения эффективности классификации, современные системы распознавания должны отвечать требованиям работы в режиме реального времени, адаптации к изменениям в распознаваемых объектах и окружающем мире, предоставлять возможность работы с неполными и

зашумленными данными [1, 2]. Появление этих требований обусловлено, в первую очередь, большим объемом данных о распознаваемых объектах, как известных при создании систем, так и поступающих в процессе их функционирования.

Необходимость обработки постоянно увеличивающегося объема данных в режиме реального времени обусловило появление нового класса систем распознавания, называемых адаптивными. При построении адаптивных систем к задачам, традиционно решаемым при построении систем распознавания, добавляются задачи предобработки исходных и поступающих в процессе работы данных, корректировки решающих правил и словаря признаков при добавлении новых данных, заполнения пробелов в данных и удаления выбросов.

Задача предобработки исходных данных, представленных в виде выборки значений признаков объектов, может решаться в двух направлениях [3]:

1) сокращение количества объектов в выборке путем их частичного удаления или замены выбранного подмножества одним объектом;

2) выбором подмножества признаков из словаря для сокращенного описания каждого объекта выборки.

В предыдущих работах авторов, например в [4], для решения задачи сокращения количества объектов в выборке был предложен переход к взвешенным выборкам w-объектов, показавший свою эффективность при решении многих прикладных задач. Построение каждого w-объекта осуществляется путем замены множества близкорасположенных объектов исходной выборки, а вес w-объекта характеризует заменяемое множество.

Сокращение пространства признаков (построение рабочего словаря) путем удаления неинформативных или неопределенных для многих объектов признаков является одной из наиболее актуальных и сложных задач, решаемых при построении систем распознавания. Выбор системы информативных признаков позволяет существенно сократить временные затраты на измерение значений признаков распознаваемых объектов и выполнение классификации, уменьшить объем хранимых данных и упростить описание объектов [1, 3]. Выборки w-объектов также могут содержать малоинформативные признаки, удаление которых не ухудшит, а иногда и улучшит эффективность классификации.

Целью данной работы является объединение двух подходов к решению задачи предобработки исходных данных, состоящее в разработке метода сокращения пространства признаков на основе взвешенных обучающих выборок w-объектов.

2. Анализ существующих методов построения рабочего словаря признаков

На сегодняшний день выделяется три основные парадигмы выбора признаков: фильтрующие, оберточные и встраиваемые методы [2]. Фильтрующие методы [2, 5, 6] оценивают каждый признак независимо от остальных, с учетом принадлежности к классам в обучающей выборке и определяют рейтинг всех признаков. В словарь будут включены признаки, имеющие максимальный рейтинг. Оберточные методы [2, 7, 8] используют классические алгоритмы поиска из теории искусственного интеллекта.

Алгоритмы такого типа оценивают каждое изменение лучшего подмножества признаков с помощью некоторого алгоритма обучения. Встраиваемые методы [2, 9, 10] создают линейную прогнозирующую модель, с помощью которой производится попытка одновременно максимизировать точность аппроксимации и минимизировать количество используемых признаков. Примером такой модели, является алгоритм построения дерева решений, в котором при достижении очередного узла ветвления должен быть выбран новый признак. В данных алгоритмах наравне с проблемой выбора оптимального подмножества признаков решается задача определения оптимальной последовательности отбора признаков.

Существующие методы отбора признаков по способу реализации можно разделить на 3 группы: методы полного перебора, эвристические методы и методы произвольного поиска. Методы, основанные на оценке всех возможных подмножеств признаков распознавания, являются эффективными только для словарей и выборок малого размера из-за низкой скорости поиска решений, поэтому для адаптивных систем, характеризующихся большим объемом постоянно добавляемых данных, использоваться не могут.

Эвристические методы основаны на некоторых предположениях о свойствах оптимальных решений, оцениваемых по выбранному критерию. В качестве критерия оценки качества получаемых словарей наиболее часто используется критерий информативности, показывающий долю правильно классифицированных объектов по выбранному признаку [11]. По способу определения информативности признаков выделяют следующие группы методов:

1) методы, постепенно дополняющие (уменьшающие) набор информативных признаков, пока последующее изменение не повысит эффективность классификации [2, 3, 7];

2) методы, осуществляющие выбор подмножества признаков по выбранному комплексному критерию, оценивающему признаки по выбранному набору параметров [2, 5, 6];

3) методы, использующие в процессе выбора подмножества информативных признаков произвольную составляющую (например, для выбора начального подмножества или для изменения выбранного подмножества в процессе работы метода и т.д.) [3, 8-10].

Проведенный анализ показывает, что наиболее эффективными на сегодняшний день являются методы, основанные на оценке как каждого из признаков в отдельности, так и выбранного подмножества в целом на основе эвристических критериев, соответствующих особенностям исходных данных. При этом стоит отметить, что проблема сокращения времени определения словаря признаков при условии обеспечения высокой эффективности классификации является одной из основных нерешенных проблем в области распознавания.

Далее предложим метод построения словаря признаков на основе взвешенных выборок w-объектов, позволяющих выполнять существенное сокращение размера обучающих выборок без уменьшения эффективности классификации.

3. 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 априорным словарем признаков и весом – неотрицательным 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. Для каждого класса задана априорная вероятность P_i.

Необходимо построить сокращенный рабочий словарь из m $(m < n)$ признаков при условии сохранения начального уровня эффективности классификации.

4. Метод построения рабочего словаря признаков на основе взвешенных обучающих выборок

Предлагаемый в данной работе метод w-MIEF основан на индивидуальной оценке информативности признаков по обобщенному критерию, включающему в себя отношение дискриминанта Фишера и коэффициент индивидуальной эффективности распознавания. Основной особенностью предлагаемого метода является использование в качестве исходных данных взвешенной выборки w-объектов.

Алгоритм w-MIEF состоит из следующих этапов.

1. Инициализация алгоритма (значений дискриминанта Фишера $fisher_i = 0, i = \overline{1,n}$ и среднего значения $fisherAvg = 0$).

2. Вычисление отношения дискриминанта Фишера для всех признаков априорного словаря

$$fisher_i = \frac{\sum_{j=1}^{l} P_j \cdot D_{ji}}{\sum_{j=1}^{l} \sum_{\substack{a=1 \\ a \neq j}}^{l} \left(P_j \cdot P_a \cdot (m_{ji} - m_{ai})^2 \right)},$$

где D_{ji} - дисперсия значений i-ого признака по j-ому классу,

m_{ji} - среднее значение i-ого признака по j-ому классу.

3. Вычисление степени покрытия классов признаками априорного словаря

$$\mathrm{cov}erage_{ji} = \frac{\sum_{a=1}^{S_i} featValItems_{ja}}{m_{ji} \cdot P_j}$$

где S_i- количество уникальных найденных значений признака x_i,

$featValItems_{ja}$ - количество объектов класса j с заданным значением i-ого признака.

4. Вычисление взвешенной эффективности признаков

$$weightedEff_i = \frac{\sum_{j=1}^{l} featValWeight_j}{\sum_{a=1}^{k} p_a}$$

где $featValWeight_j$ - суммарный вес объектов класса j с заданным значением i-ого признака,

p_a - вес w-объекта.

5. Выбор наилучших признаков:

1) с максимальной эффективностью по степени перекрытия классов

$$BestCoverage = \max_{i=1,n} \sum_{j=1}^{l} \text{cov}erage_{ji},$$

2) с максимальной эффективностью по информативности:

$$BestFi = \max_{i=1,n} weightedEff_i.$$

6. Формирование рабочего словаря признаков по параметрам максимальной эффективности по степени перекрытия классов и информативности признаков по следующим правилам:

1) эффективность текущего признака совпадает с эффективностью лучшего признака;

2) эффективность признака больше 0 и данный признак улучшает распознавание хотя бы одного класса;

3) эффективность признака больше 0 и распознавание хотя бы одного из классов лежит в пределах порога t, заданного пользователем.

Отметим, что наличие пользовательского порога t, связно с тем, что возможны ситуации, когда признаки не улучшают распознавание хотя бы одного из классов, но, тем не менее, имеют схожий с лучшим уровень распознавания классов. Таким образом, если для рассматриваемого признака разница между лучшим и текущим уровнем распознавания любого из классов находится в пределах $(0; t]$, то данный признак также помечается, как информативный.

Данный алгоритм получил название weighted Maximum Individual Efficiency Filter (w-MIEF, фильтр признаков по максимальной индивидуальной эффективности), т.к. он позволяет выбирать информативные признаки на основе их максимальной индивидуальной эффективности признаков.

Асимптотическая временная сложность алгоритма составляет $O(MN)$.

5. Экспериментальные исследования

В качестве тестовых наборов данных для оценки эффективности предложенного метода использовались выборки репозитория UCI [12]. В таблице 1 представлено описание используемых исходных наборов данных и результаты построения по ним выборок w-объектов по алгоритму, представленному в работе [4] (количество признаков и классов при переходе к выборкам w-объектов не изменяется).

Таблица 1. Наборы данных из репозитория UCI

Набор данных	Исходная выборка			Кол-во w-объектов после сокращения
	Кол-во объектов	Кол-во признаков	Кол-во классов	
Multifeature	2000	649	10	617
LIBRAS	360	90	15	193
Vehicle	846	18	4	433
Waveform	5000	21	3	2541
Wisconsin Cancer	569	32	2	116

Для оценки эффективности классификации по полученным словарям признаков использовался алгоритм построения дерева решений C4.5 [1]. Пользовательский порог t метода w-MIEF выбирался экспериментально для обеспечения максимальной эффективности классификации. Тестирование проводилось путем выполнения десятикратной кросс-проверки с выбором 10% объектов в обучающую выборку.

Для сравнительной оценки эффективности работы алгоритма w-MIEF был использован один из наиболее эффективных алгоритмов построения словарей признаков Relief [6]. Результаты тестирования представлены в таблицах 2-3.

Таблица 2. Результаты классификации (размер словаря признаков)

Набор данных	Исходный словарь признаков	Сокращение словаря признаков алгоритмом Relief	Сокращение словаря признаков алгоритмом w-MIEF
Multifeature	649	135	117
LIBRAS	90	47	42
Vehicle	18	12	9
Waveform	21	11	10
Wisc. Can.	32	12	15

Таблица 3. Результаты классификации
(точность и время выполнения классификации)

Набор данных	Исходный словарь признаков		Сокращение словаря признаков алгоритмом Relief		Сокращение словаря признаков алгоритмом w-MIEF	
	Точность, %	Время, мс	Точность, %	Время, мс	Точность, %	Время, мс
Multifeature	96,84	89761,4	95,56	4281	94,39	8043,7
LIBRAS	77,88	5548,6	75,72	2850	72,98	1514,8
Vehicle	82,48	1001,6	79,38	684,6	79,17	336,7
Waveform	88,46	32368,2	85,04	14282,2	84,18	4649,9
Wisc. Can.	96,50	571,6	96,32	235,6	94,09	131,8
Среднее значение:	88,43	25850,2	86,40	4466,68	85,16	2935,38

На основе представленных результатов экспериментальных исследований можно сделать выводы о том, что:

1) по сравнению с исходной выборкой предложенный метод позволяет сократить количество признаков и время выполнения распознавания на 45% - 80% при сохранении высокой эффективности классификации;

2) по сравнению с алгоритмом Relief предложенный метод позволяет сократить количество признаков и время выполнения распознавания на 25% - 40% при сохранении высокой эффективности классификации.

Необходимо отметить, что на выборках большего размера метод w-MIEF показывал лучшие результаты по эффективности классификации.

Этот результат подтверждает сделанный в предыдущих работах авторов вывод о необходимости перехода к взвешенным выборкам w-объектов при наличии в исходных выборках более 2500 объектов.

6. Выводы

В данной работе предложено комплексное решение задачи обработки обучающих выборок в адаптивных системах распознавания, заключающееся в одновременном сокращении количества объектов выборок и признаков априорного словаря. Разработан новый метод w-MIEF построения рабочего словаря признаков, позволяющий анализировать признаки объектов обучающих выборок, содержащих весовые коэффициенты. Предложенный метод формирует рабочий словарь, содержащий не дискриминирующие признаки с достаточно высокой индивидуальной эффективностью. Пользовательский параметр разработанного метода позволяет выбирать не только признаки, имеющие лучшую эффективность распознавания, но и признаки, имеющие меньшую эффективность распознавания в пределах задаваемого порога и при этом обладающие хорошей способностью к разделению классов.

Предложенный метод был протестирован с использованием реальных тестовых данных, традиционно используемых для проверки качества классификаторов. Тестирование проводилось путем кросс-проверки с использованием классификатора C4.5. Результаты исследований показали, что метод w-MIEF позволяет выполнить сокращение словаря признаков при условии сохранения эффективности классификации, полученной при использовании исходных выборок.

Результаты данной работы подтверждают эффективность использования взвешенных выборок w-объектов при построении адаптивных 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**. *Liu H.* Computational methods of feature selection / H. Liu, H. Motoda.– Chapman & Hall/CRC data mining and knowledge discovery, 2007. – 440 p.**3**. *Загоруйко Н.Г.* Прикладные методы анализа данных и знаний / Н.Г. Загоруйко. – Новосибирск: ИМ СО РАН, 1999. – 270 с.**4**. *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**.5.** Yu Lei Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution / Lei Yu, Huan Liu // Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003), Washington DC, 2003. – In ICML. – P. 856-863.**6.** *Kononenko I.* Estimating attributes: Analysis and extensions of RELIEF/ I. Kononenko // European Conference on Machine Learning. – Catania, Italy, Springer Verlag, New York, 1994. – P. 171-182.**7**. *Pudil P*. Floating search methods in feature selection / P. Pudil, Novovi˘cov´a, J. Kittler // Pattern Recognition Letters 15. – 1994. – P. 1119-1125.**8**. Stracuzzi D.J. Randomized variable elimination / D.J. Stracuzzi, P.E. Utgoff // Journal of Machine Learning Research 5. – 2004. – P. 1331-1364.**9**. *Skalak D.B.* Prototype and feature selection by sampling and random mutation hill climbing / D.B. Skalak [In W.W. Cohen and H. Hirsh, editors] // Machine Learning: Proceedings of the Eleventh International Conference. – New Brunswick, NJ, Morgan Kaufmann, 1994. – P. 293-301.**10.** *Liu H. A* probabilistic approach to feature selection. / H. Liu, R. Setino [In L. Saitta, editor] // Machine Learning: Proceedings of the Thirteenth International Conference on Machine Learning. – Bari, Italy, Morgan Kaufmann, 1996. – P. 319–327.**11**. *Файнзильберг Л.С.* Оценка полезности признаков при решении задач диагностики в

статистической постановке / Л.С. Файнзильберг // Математические машины и системы. – № 1. – 1998. – С. 57 – 64.**12.** *Merz C.J.* UCI Repository of machine learning datasets / C.J. Merz, P.M. Murphy // Information and Computer Science University of California, Irvine, CA, 1998. – Режим доступа: http://www.ics.uci.edu/~mlearn/databases

Поступила в редколлегию 01.03.2012

УДК 656.13+612.821

Н.У. ГЮЛЕВ, канд.техн.наук, доц., ХГАГХ, Харьков,
В.К. ДОЛЯ, докт. техн.наук, проф.,зав.каф., ХГАГХ, Харьков

МОДЕЛЬ ИЗМЕНЕНИЯ ФУНКЦИОНАЛЬНОГО СОСТОЯНИЯ ВОДИТЕЛЯ-ФЛЕГМАТИКА В ТРАНСПОРТНОМ ЗАТОРЕ

Исследовано влияние различных факторов на функциональное состояние водителя-флегматика. Представлена математическая модель влияния транспортного затора на функциональное состояние водителя-флегматика.
Ключевые слова: функциональное состояние, математическая модель, транспортный затор, показатель активности регуляторных систем.

Досліджено вплив різних факторів на функціональний стан водія-флегматика. Представлена математична модель впливу транспортного затору на функціональний стан водія-флегматика.
Ключові слова: функціональний стан, математична модель, транспортний затор, показник активності регуляторних систем.

The effect of various factors on the functional state of the driver-phlegmatic. A mathematical model of the impact of traffic congestion on the functional state of the driver-phlegmatic.
Key words: functional status, the mathematical model, the transport route, the index of activity of regulatory systems.

1. Введение

Пребывание в транспортном заторе приводит к ухудшению функционального состояния водителя. Происходит временное нарушение некоторых психофизиологических функций водителя, от которых во многом зависит безопасность работы транспортной системы [1,2]. При этом транспортные заторы влияют на водителей по разному в зависимости от темперамента.

2. Постановка проблемы

Транспортные заторы возникают вследствие превышения интенсивности транспортного потока над пропускной способностью улиц и дорог. Это приводит к снижению скорости транспортных средств и увеличению времени передвижения.

Транспортными средствами управляют водители с различной квалификацией и различными психофизиологическими характеристиками [3-6]. От психофизиологии водителя и его функционального состояния зависит время реакции водителя и динамический габарит автомобиля, который влияет на характеристики транспортного потока [7].