УДК 004.85

НЕОДНОРОДНЫЙ АНСАМБЛЕВЫЙ АЛГОРИТМ КЛАССИФИКАЦИИ РАЗНОТИПНЫХ ДАННЫХ

© 2017 О.К. Альсова, И.М. Стубарев

Новосибирский государственный технический университет

Статья поступила в редакцию 11.12.2017

В статье предложен неоднородный ансамблевый алгоритм, предназначенный для классификации разнотипных данных. Алгоритм основан на итерационном применении одиночных (базовых) классификаторов на исходной обучающей выборке и включении в ансамбль только тех классификаторов, относительная ошибка которых не превосходит заданный порог. С использованием алгоритма выполнено построение нескольких ансамблей классификаторов на выборке из архива данных по машинному обучению и на реальных медицинских данных. Сравнительное тестирование показало преимущества использования предложенного ансамблевого алгоритма по сравнению с одиночными классификаторами (повышение точности классификации, уменьшение дисперсии ошибки классификатора).

Ключевые слова: одиночный (базовый) алгоритм классификации, неоднородный ансамблевый алгоритм, bagging, bootstrap – выборка, дерево решений, логистическая регрессия, нейронная сеть.

1. ВВЕДЕНИЕ

Под ансамблевым алгоритмом классификации понимается совокупность одиночных классификаторов, решения которых объединяются (агрегируются) определенным образом для получения окончательного классификационного решения. Согласно литературным источникам использование ансамбля классификаторов позволяет повысить точность классификации при решении прикладных задач [1,2]. К настоящему моменту разработаны различные методы построения ансамбля классификаторов. Среди них наиболее распространены методы bagging, boosting, основанные на манипуляции с исходной обучающей выборкой данных с целью построения нескольких классификаторов и последующей комбинацией (агрегацией) полученных решений [3,4].

Большинство работ в этой области посвящено построению однородного ансамбля, состоящего из моделей классификаторов одного типа (например, деревьев решений). Однако, теоретические и экспериментальные исследования показывают, что необходимым и достаточным условием точности ансамбля является различность и независимость составляющих его классификаторов (ошибочная классификация разных объектов из обучающей выборки) [5,6]. Это условие достигается, во-первых, при обучении классификаторов на различных подмножествах выборки исходных данных, во-вторых, при использовании в ансамбле различных моделей классификаторов в качестве базовых (например, деревьев решений, логистической регрессии, нейронной сети и т.д.). Поэтому, перспективным направлением представляется разработка и исследование неоднородных ансамблевых алгоритмов.

В данной работе предложен неоднородный ансамблевый алгоритм, использующий процедуру бэггинга (bagging) для формирования случайных подвыборок из исходного обучающего множества с последующим построением базовых классификаторов разных типов на различных его подмножествах. В качестве базовых выбраны алгоритмы, которые позволяют классифицировать объекты, описанные разнотипными признаками, измеренными как в количественной, так и в качественной шкалах.

Для тестирования алгоритма использовались данные из репозитория данных по машинному обучению (*http://archive.ics.uci.edu/ml/datasets*) и реальные данные о хирургическом лечении больных с патологией аорты. Данные предоставлены Сибирским федеральным биомедицинским исследовательским центром имени академика Е.Н. Мешалкина.

2. НЕОДНОРОДНЫЙ АНСАМБЛЕВЫЙ АЛГОРИТМ КЛАССИФИКАЦИИ

Идея неоднородного ансамблевого алгоритма классификации заключается в итерационном применении одиночных классификаторов на обучающей выборке и учете в итоговом решении при построении ансамбля вклада только тех классификаторов, ошибка классификации которых не превосходит заданный порог.

Исходная выборка данных разбивается на две части: обучающую (используется для обучения классификатора) и тестовую.

Альсова Ольга Константиновна, кандидат технических наук, доцент кафедры вычислительной техники.
E-mail: alsova@corp.nstu.ru
Стубарев Игорь Михайлович, магистрант.
E-mail: igorekiks@gmail.com

На входе ансамблевого алгоритма задаются следующие исходные данные:

- $X^{train} = \{x_{ij}^{train}\}, i = \overline{1, n_{train}}; j = \overline{1, p},$

где x_{ij}^{train} – значение *j*-го признака, измеренного на *i*-ом объекте обучающей выборки; *p* – число признаков; n_{train} – число объектов в обучающей выборке;

- $X^{test} = \{x_{ij}^{test}\}, i = \overline{1, n_{test}}; j = \overline{1, p},$

где x_{ij}^{test} – значение *j*-го признака, измеренного на *i*-ом объекте тестовой выборки; *p*– число признаков; n_{test} – число объектов в тестовой выборке;

- $Y^{train} = \{y_i^{train}\}, i = \overline{1, n_{train}}$,

где $y_i^{train} \in Y = \{1,..., l\}$ – номер класса *i*-го объекта обучающей выборки; *l* – количество классов;

и входные параметры:

- $C = \{c_i\}, i = \overline{1, k}$,

где c_i – *i*-ый базовый классификатор, *k* – количество базовых классификаторов;

- *T* – число итераций;

- $\varepsilon_{\min}, \varepsilon_{\max}$ – соответственно минимальная и максимальная пороговые ошибки для включения базового классификатора в ансамбль. В алгоритме реализована возможность задания минимальной пороговой ошибки классификации для того, чтобы исключить ситуацию переобучения классификатора.

Разработанный алгоритм реализован в программной системе классификации разнотипных данных на основе ансамбля алгоритмов [7,8], среда разработки – *IDENetBeans 8.0.2.*, язык программирования – *Java*, с использованием библиотеки анализа данных *weka*.

В качестве базовых классификаторов могут выступать различные методы классификации. На настоящий момент в программной системе реализовано семь базовых классификаторов, а именно: деревья решений (алгоритмы *CART*, *CHAID*, *C4.5*, *ID3*), нейронная сеть (многослойный персептрон), логистическая регрессия (алгоритм мультиномиальной логистической регрессии с возможностью классификации разнотипных данных) [9], алгоритм *k*-взвешенных ближайших соседей.

Необходимо построить ансамбль классификаторов:

$$A = \{a_t\}, t = \overline{1, N},$$

где *N* – количество классификаторов в ансамбле и выполнить классификацию объектов из тестовой выборки:

$$Y^{test} = \{y_i^{test}\}, i = \overline{1, n_{test}},$$

где y_i^{test} – номер класса *i*-го объекта, на основе использования ансамбля.

Для построения ансамбля классификаторов используется итерационный алгоритм, на каждой итерации $t = 1,..., T$ выполняется следующая последовательность действий:

Формируется бутстреп-выборка:

$$X_t^{b-train} = \{x_{ij}^{b-train}\}, i = \overline{1, n_{train}}; j = \overline{1, p}$$

из исходной обучающей выборки X^{train}.

Обучающая выборка формируется с помощью процедуры бэггинга (*bagging*), которая основана на формировании случайных подвыборок (бутстреп-выборок) из исходного обучающего множества. Количество объектов в этих подвыборках такое же, как и у исходного обучающего множества. Причем одни объекты могут отбираться по несколько раз, а другие ни разу.

Выбирается случайным образом (равновероятно) базовый классификатор c_i из множества C:

$$a_t = \underset{1 \le i \le k}{rnd}\{c_i\}.$$

Обучается классификатор a_t на выборке $X_t^{b-train}$.

Вычисляется ошибка ε (относительное количество неверно классифицированных объектов), построенной модели классификатора a_t.

Включается в ансамбль классификатор в случае, если ошибка классификации удовлетворяет условиям: $\varepsilon_{\min} \le \varepsilon \le \varepsilon_{\max}$.

В результате применения алгоритма формируется ансамбль классификаторов.

Далее выполняется итоговая классификация объектов из тестовой выборки на основе использования построенного ансамбля с применением метода большинства голосов (либо метода взвешенного голосования) для формирования общего классификационного решения. Псевдокод разработанного ансамблевого алгоритма приведен в листинге 1.

Листинг 1. Псевдокод неоднородного ансамблевого алгоритма

Вход: $X^{train}, Y^{train}, X^{test}$ – обучающая и тестовая выборки;

C – множество базовых классификаторов;

$\varepsilon_{\min}, \varepsilon_{\max}$ – минимальная и максимальная пороговые ошибки классификации; *BootstrapSample(X)* – функция формирования бутстреп-выборки из обучающей выборки *X*.

Выход: *A* – ансамбль классификаторов;

Y^{test} – номера классов объектов тестовой выборки.

Алгоритм: $A=\varnothing$, $\varepsilon = 0$ (ошибка классификации)

for *t=1* ***to*** *T*

do $X_t^{b-train}$ *=BootstrapSample(* X^{train} *)*

$a_t = \underset{1 \le i \le k}{rnd}\{c_i\}$ – выбор классификатора из множества C

обучение классификатора a_t на выборке $X_t^{b-train}$

for *i=1* ***to*** n_{train}

do $pred_i$ = предсказанный класс для *i*-го объекта классификатором a_t

if ($pred_i \neq y_i^{train}$) ***then*** $\varepsilon = \varepsilon + \varepsilon_i / n_{train}$

end for

if $(\varepsilon_{\min} \leq \varepsilon \leq \varepsilon_{\max})$ ***then*** $A = A \cup \{a_t\}$ – включение классификатора в ансамбль

end for

//классификация объектов из тестовой выборки

for *i=1* ***to*** n_{test}

do ***for*** *t=1* ***to*** *N*

do y_{it}^{test} = предсказание класса *i*-го объекта классификатором a_t

end for

$$y_i^{test} = \arg\max_{1 \leq t \leq N}(count\_freq(y_t^{test})$$

end for

return $A,\ y_i^{test}, i = \overline{1, n_{test}}$.

3. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТАЛЬНЫХ ИССЛЕДОВАНИЙ

Разработанный неоднородный ансамблевый алгоритм тестировался на двух выборках исходных данных, а именно, использовались данные по характеристикам стекла Glass из репозитория данных машинного обучения и реальные медицинские данные о хирургическом лечении больных с патологией аорты. Решалась задача классификации объектов. В первом случае определялся тип (класс) стекла на основе анализа его характеристик, во втором случае – тип нарушения мозгового кровообращения (НМК) для каждого пациента выборки на основе анализа исходных признаков о его состоянии. В табл. 1. представлены количественные характеристики данных.

В проведенном исследовании в качестве базовых использовались семь классификаторов, а именно, четыре алгоритма построения деревьев решений (ID3, CART, C4.5, CHAID), мультиномиальная логистическая регрессия, нейронная сеть (многослойный персептрон) и метод k-взвешенных ближайших соседей.

Точность базовых классификаторов и неоднородного ансамблевого алгоритма оценивалась, во-первых, на всей исходной выборке, которая использовалась как обучающая, во-вторых, на тестовой выборке. Для формирования и оценки точности на тестовой выборке применялся алгоритм 10-ти кратной 10-ти блочной кросс-проверки (10×10-fold cross validation), который используется в качестве стандарта в машинном обучении для сравнения и тестирования классификаторов. Алгоритм заключается в разбиении исходной выборки случайным образом на 10 непересекающихся частей (подвыборок) одинаковой (или почти одинаковой) длины со стратификацией классов. Каждая часть по очереди становится контрольной (тестовой выборкой), при этом обучение классификатора производится по остальным 9 частям. Точность классификатора определяется как средняя ошибка классификации на контрольных подвыборках. Этапы тестирования по 10-ти контрольным подвыборкам повторяются 10 раз. На последнем шаге алгоритма полученные результаты усредняются для итогового расчета точности классификации на тестовой выборке.

Для каждой выборки исходных данных были построены базовые классификаторы и два варианта ансамблевых алгоритмов (с включением в исходное множество С только четырех алгоритмов формирования деревьев решений и с включением в исходное множество С всех базовых классификаторов). Итоговая классификация объектов выполнялась на основе применения метода взвешенного голосования.

В табл. 2-3. представлены результаты экспериментов, полученные для двух исходных выборок данных. Для каждого классификатора приведены следующие характеристики: точность классификации на обучающей и на тестовой выборках ($f_{tr.}$ и f_{test}), в %; средняя частота ошибки классификации на тестовой выборке – Aver. error (отношение неверно классифицированных объектов к общему количеству объектов в долях единицы); дисперсия классификатора (D) и 95% доверительный интервал для средней частоты ошибки классификации (95% range). Отметим, что ранее также решалась задача кластеризации медицинских данных, результаты решения которой описаны в [10].

В результате применения неоднородного ансамблевого алгоритма на данных Glass в итоговый ансамбль классификаторов вошли методы построения деревьев решений ID3, CART (для первого варианта) и ID3, CART, нейронная

Таблица 1. Описание исходных данных

Данные	*Объем выборки*	*Количество признаков*	*Количество классов*
Glass	*214*	*9*	*6*
НМК	*124*	*19*	*4*

сеть (для второго варианта). Анализ результатов табл. 2 позволяет сделать вывод об улучшении качества классификации данных с помощью неоднородного ансамблевого алгоритма по сравнению с базовыми классификаторами. Точность классификации на обучающей выборке составила 97,61%, на тестовой – 78,1%, дисперсия классификатора – 0,006. При этом, лучший результат по совокупности характеристик, полученный для базового классификатора, соответствует алгоритму ID3 и составляет на обучающей выборке – 94,86%, на тестовой выборке – 70,19%, дисперсия классификатора – 0,011. Таким образом, применение ансамблевого алгоритма позволило увеличить точность классификации на тестовой выборке на 7,91%, уменьшить дисперсию ошибки классификатора и уменьшить соответственно 95% доверительный интервал для средней частоты ошибки классификации.

На рис. 1. для каждого класса представлены ROC-кривые, построенные на тестовой выборке, в результате применения неоднородного ансамблевого алгоритма. Показатель AUC (Area under ROC curve - площадь под ROC-кривой) для всех классов выше 0,5 и составляет от 0,76 до 0,97, что свидетельствует о высокой точности классификатора.

В результате применения неоднородного ансамблевого алгоритма на выборке медицинских данных в итоговый ансамбль классификаторов вошли методы построения деревьев решений ID3, CART, нейронная сеть и логистическая регрессия. Согласно табл. 3. неоднородный ансамблевый алгоритм позволяет существенно улучшить результаты классификации на тестовой выборке по сравнению с базовыми классификаторами. Точность классификации с помощью неоднородного ансамблевого алгоритма составила 75,21%, тогда как лучший результат классификации с помощью метода ID3 составил только 67,74%. Также ансамблевому алгоритму соответствует меньшая дисперсия классификатора относительно базовых классификаторов и меньший 95% доверительный интервал для средней частоты ошибки классификации.

Таблица 2. Характеристики точности алгоритмов классификации для данных Glass

Алгоритм	$f_{tr.}$, в %	f_{test}, в %	*D*	*Aver. error*	95% *range*
ID3	94,9	70,2	0,011	0,29	0,28; 0,31
C4.5	94,9	69,8	0,008	0,30	0,29; 0,31
CART	90,2	69,5	0,009	0,30	0,29; 0,32
CHAID	89,1	60,6	0,009	0,39	0,38; 0,41
Логистическая регрессия	73,3	63,1	0,01	0,37	0,36; 0,38
Нейронная сеть	85,7	69,0	0,008	0,31	0,29; 0,32
k-ближайших соседей	69,7	62,4	0,008	0,38	0,36; 0,39
Ансамбль (вар. 1)	96,7	75,6	0,007	0,24	0,23; 0,26
Ансамбль (вар. 2)	97,6	78,1	0,006	0,22	0,21; 0,22

Таблица 3. Характеристики точности алгоритмов классификации для медицинских данных

Алгоритм	$f_{tr.}$, в %	f_{test}, в %	*D*	*Aver. error*	95% *range*
ID3	82,3	67,7	0,013	0,32	0,31; 0,34
C4.5	83,1	66,1	0,014	0,34	0,32; 0,36
CART	85,5	60,5	0,012	0,40	0,38; 0,41
CHAID	79,0	66,1	0,009	0,34	0,33; 0,35
Логистическая регрессия	85,5	62,4	0,013	0,38	0,36; 0,40
Нейронная сеть	90,3	66,9	0,013	0,33	0,32; 0,35
k-ближайших соседей	73,4	60,1	0,010	0,40	0,39; 0,41
Ансамбль	87,6	75,2	0,001	0,25	0,24; 0,26

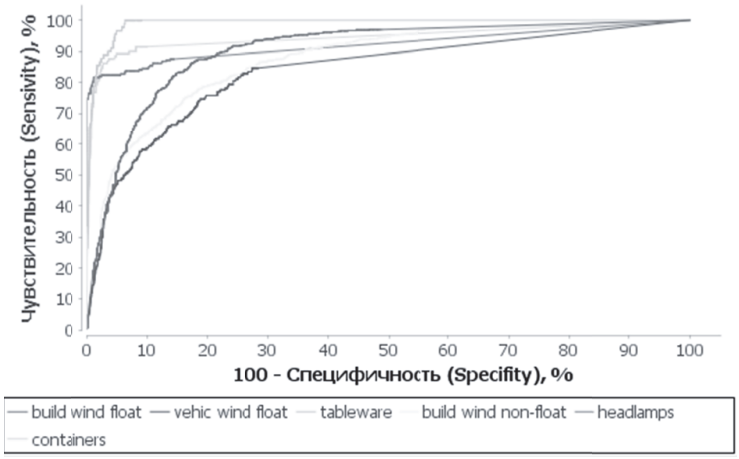

Рис. 1. ROC-кривая по данным Glass

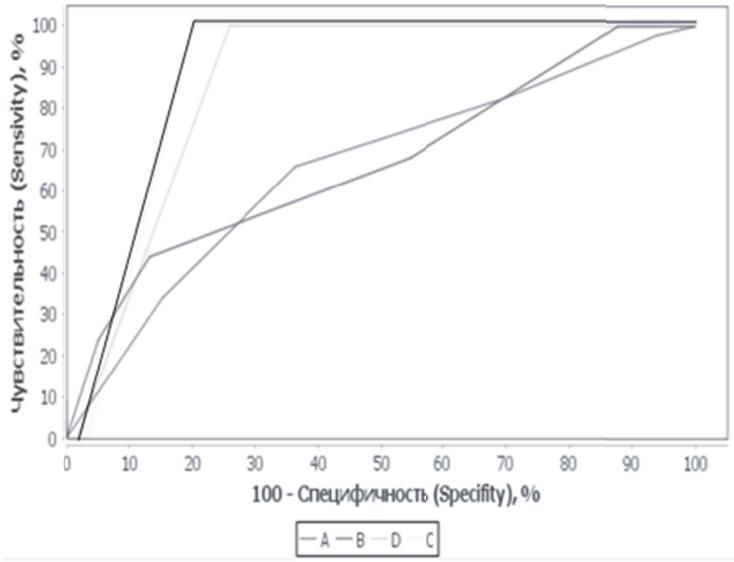

Рис. 2. ROC-кривая по медицинским данным (НМК)

На рис. 2. для каждого класса представлены ROC-кривые, построенные по медицинским данным на тестовой выборке в результате применения неоднородного ансамблевого алгоритма. Показатель AUC составляет от 0,66 до 0,86 в зависимости от класса и для всех классов выше 0,5.

Таким образом, из проведенных результатов вычислительных экспериментов на разных исходных выборках данных, можно сделать вывод о том, что предложенный неоднородный ансамблевый алгоритм позволяет достичь более высокой точности классификации объектов по сравнению с базовыми алгоритмами и обеспечить более стабильные результаты (уменьшение дисперсии классификатора, уменьшение доверительного интервала для средней частоты ошибки классификации).

ЗАКЛЮЧЕНИЕ

В представленной работе описан неоднородный ансамблевый алгоритм, основанный на применении комплекса базовых классификаторов, относящихся к разным классам. Проведенные исследования показали, что применение алгоритма в качестве инструмента решения задачи классификации позволяет найти наилучшую комбинацию базовых классификаторов, которая обеспечит максимальную точность классификации объектов с помощью ансамбля. Выполнено тестирование предложенного алгоритма на разных выборках исходных данных, которое позволяет сделать вывод о более высокой точности классификации данных с помощью предложенного алгоритма по сравнению с базовыми классификаторами. Дальнейшим направлением исследований является решение задачи определения оптимальных параметров ансамблевого алгоритма (метод формирования подвыборок из исходного множества данных, метод выбора классификатора на каждой итерации, метод голосования для формирования общего классификационного решения) в автоматическом режиме.

СПИСОК ЛИТЕРАТУРЫ

1. Multiple Classifier Systems / *J. Kittler & F. Roli (editors)* // Proc. of 2nd International Workshop, MCS2001, (Cambridge, UK, 2-4 July 2001) / Lecture Notes in Computer Science. V. 2096. Springer-Verlag, Berlin.
2. *Vishwath P., Murty M.N., Bhatnagar* C. Fusion of multiple approximate nearest neighbor classifier for fast and efficient classification // Information fusion. 2004. V. 5. Pp. 239-250.
1. *Quinlan J.R.* Bagging, boosting and C4.5 // Proceedings of AAA/IAAI. 1996. V. 1. Pp. 725-730.
1. *Breiman L.* Bagging predictors // Machine Learning.1996.V. 24, No. 2. Pp. 123–140.
1. *Tumer K., N.C. Oza* Decimated input ensembles for improved generalization // Proceedingsof the International Joint Conference on Neural Networks. Washington, DC. 1999.
1. *Чистяков С.П.* Случайные Леса: Обзор // Труды Карельского научного центра РАН. 2013. № 1. С. 117 – 136.
1. Батыгин *Р.И.,* Альсова *О.К.* Программная система классификации разнотипных данных на основе ансамбля алгоритмов (ECA - Ensemble Classification Algorithms): Свидетельство о государственной регистрации программы для ЭВМ № 2017610788. 2017.
1. *Batygin R.I., Alsova O.K.* Software system for different types of data classification based on the ensemble algorithms // Actual problems of electronic instrument engineering (APEIE–2016) : proceedings. Novosibirsk, 2016. V. 1. Part 2. Pp. 506-509.
1. *Альсова О.К., Альсов С.А.* Алгоритм мультиномиальной классификации разнотипных медицинских данных // Естественные и технические науки. 2015. № 11. С.386-389.
10. *Альсова О.К.* Алгоритмы кластеризации разнотипных медицинских данных на примере решения медицинской задачи // Труды СПИИРАН, 2014. № 6. С. 156-169.

HETEROGENEOUS ENSEMBLE ALGORITHM FOR CLASSIFICATION OF DIFFERENT TYPES OF DATA

© 2017 O.K. Alsova, I.M. Stubarev

Novosibirsk State Technical University

In this article developed heterogeneous ensemble algorithm for classification of different types of data is proposed. The algorithm is based on the iterative use of single (basic) classifiers on the initial training sample and inclusion in the ensemble only those classifiers whose relative error does not exceed a predetermined threshold. With the algorithm a few ensembles were designed for data from machine learning database and for real medical data. The comparative testing shows the advantages of the proposed ensemble algorithm compared with the single classifiers (the increase of classification accuracy, the decrease of the variance of the classifier).

Keywords: single (basic) classification algorithm, heterogeneous ensemble algorithm, bagging, bootstrap – sample, decision tree, logistic regression, neural network.

Olga Alsova, Candidate of Technics, Associate Professor at the Computer Engineering Department.
E-mail: alsova@corp.nstu.ru
Igor Stubarev, Master's Student.
E-mail: igorekiks@gmail.com