ПОДХОД К КОДИРОВАНИЮ СТРУКТУРЫ АНСАМБЛЯ КЛАССИФИКАТОРОВ ДЛЯ РЕШЕНИЯ ЗАДАЧИ ПОИСКА ОПТИМАЛЬНОЙ СТРУКТУРЫ С ПОМОЩЬЮ ГЕНЕТИЧЕСКОГО АЛГОРИТМА

Кузнецов Владимир Вячеславович
Аспирант, Рязанский государственный радиотехнический университет
vla8774@gmail.com

Аннотация. Построение эффективного многоклассового классификатора для создания двумерных карт местности представляется сложной задачей, одним из подходов к решению которой — это строить ансамбль классификаторов, каждый классификатор которого будет решать небольшую задачу бинарной классификации. Таким образом, задача составления и обучения классификатора сводится к задаче выбора бинарного классификатора и обучения ансамбля бинарных классификаторов. В настоящей статье рассматривается подход к построению ансамбля классификаторов на основе генетического алгоритма, алгоритм обучения классификатора, кодирование структуры ансамбля классификатора. Генетические алгоритмы являются наиболее распространенными и наиболее изученными эволюционными алгоритмами. Результатом работы генетического алгоритма будет выявленная закономерность, представленная в виде вектора. Для задачи выбора структуры классификатора результатом будет — структура иерархического классификатора. Предложенный подход к кодированию структуры ансамбля классификаторов необходим для кодирования в битовую последовательность для решения задачи поиска оптимальной структуры с помощью генетического алгоритма. Проведено сравнительное исследование алгоритмов, использующих генетический алгоритм для построения ансамбля классификаторов с алгоритмами, которые не используют генетический алгоритм. Результатом проведенных экспериментов является существенное уменьшение процента ошибки и времени распознавания классификаторами, построенными с помощью генетического алгоритма.

Ключевые слова: генетический алгоритм, ансамбль классификаторов, кодирование структуры.

APPROACH TO CODING THE ENSEMBLE STRUCTURE OF CLASSIFIERS FOR SOLVING THE PROBLEM OF SEARCHING FOR AN OPTIMAL STRUCTURE BY MEANS OF A GENETIC ALGORITHM

V. Kuznetsov

Summary. Building an effective multi-class classifier for creating two-dimensional maps of the terrain is a difficult task, one approach to which is to build an ensemble of classifiers, each classifier of which will solve a small binary classification problem. Thus, the task of compiling and training a classifier is reduced to the problem of choosing a binary classifier and training an ensemble of binary classifiers. This article discusses an approach to constructing an ensemble of classifiers based on a genetic algorithm, a classifier learning algorithm, coding the structure of the classifier ensemble. Genetic algorithms are the most common and most studied evolutionary algorithms. The result of the genetic algorithm will be the identified pattern, presented in the form of a vector. For the task of choosing a classifier structure, the result will be — the hierarchical classifier structure. The proposed approach to coding the structure of an ensemble of classifiers is necessary for encoding into a bit sequence to solve the problem of finding the optimal structure using a genetic algorithm. A comparative study of algorithms using a genetic algorithm to build an ensemble of classifiers with algorithms that do not use a genetic algorithm has been carried out. The result of the experiments is a significant reduction in the percentage of error and recognition time by classifiers constructed using a genetic algorithm.

Keywords: genetic algorithm, classifier ensemble, structure coding.

Введение

Построение эффективного многоклассового классификатора для создания двумерных карт местности представляется сложной задачей, одним из подходов к решению которой — это строить ансамбль классификаторов, каждый классификатор которого будет решать небольшую задачу бинарной классификации. Таким образом, задача составления и обучения классификатора сводится к задаче выбора бинарного классификатора и обучения ансамбля бинарных классификаторов.

Алгоритм обучения классификатора

Алгоритм обучения ансамбля бинарных классификаторов представлен на рисунке 1. После формирования обучающего множества необходимо определить структуру классификатора. Для формирования структуры ан-

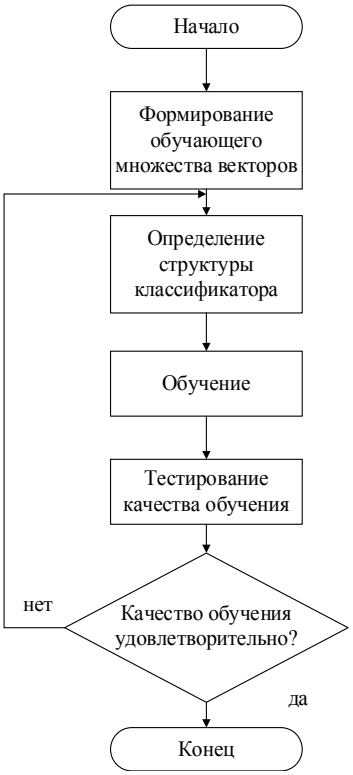


Рис. 1. Алгоритм обучения классификатора

Рис. 2. Алгоритм построения иерархического классификатора

самбля классификаторов воспользуемся генетическим алгоритмом.

Описание работы генетического алгоритма

Идея генетического алгоритма основана на принципе эволюционного развития биологических существ [1]. Из одного вида выживают особи, которые лучше других приспосабливаются к окружающей среде. Полезные же свойства этих особей закрепляются и накапливаются на генетическом уровне. Возвращаясь к выбору структуры классификатора, отменим, что в качестве популяций используются выборки из обучающего множества.

Генетические алгоритмы являются наиболее распространенными и наиболее изученными эволюционными алгоритмами. Результатом работы генетического алгоритма будет выявленная закономерность, представленная в виде вектора. Для задачи выбора структуры классификатора результатом будет — структура иерархического классификатора.

Дано обучающее множество $V_{teach} = \{v_{teach1}, \cdots, v_{teachi}\}; i = \overline{1,I}$; данные вектора уже классифицированы и каждый из векторов принадлежит определенному классу $C = \{c\}$; Пусть в некотором узле иерархического классификатора сконцентрировано множество V'_{teach}, такое что $V'_{teach} \subset V_{teach}$. В данном случае существуют следующие ситуации:

1. Все элементы множества V'_{teach} относятся к одному классу c_k, тогда данный узел классификатора является стоком и соответствует определенному классу c_k.

2. Множество V'_{teach} содержит примеры, относящиеся к разным классам. Требуется дополнительно уточнить классификацию.

В общем случае, алгоритм построения структуры ансамбля классификатора представлен на рисунке 2.

Кодирование структуры ансамбля классификаторов

Для работы со структурой данные необходимо закодировать в битовую последовательность фиксированной длины так, чтобы, изменение любого ее бита преобразовывало последовательность в непротиворечивую битовую последовательность, так же кодирующую иерархический классификатор с таким же числом узлом. Так же не должно возникать различного рода противоречий, когда происходит скрещивание двух различных

битовых последовательностей. Данное ограничение обуславливается тем, что битовая последовательность должна быть использована в качестве хромосомы генетического алгоритма, который оптимизирует иерархический классификатор с позиции заданного критерия. Длина хромосомы должна быть минимальной, а ее содержание должно быть непротиворечивым для ускорения генетического алгоритма.

Необходимо что бы последовательность битов кодировала бы структуру ансамбля классификатора с числом узлов N. Число классификаторов с N узлами равно N — му числу Каталана [2](1):

$$C_N = \frac{1}{N+1}\binom{2N}{N} = \frac{(2N)!}{(N+1)!\,N!}; \quad (1)$$

Одним из способов построения ансамбля классификаторов является построение дерева бинарных классификаторов, что и будем использовать в данном случае. Рассмотрим ансамбль, каждая структура которого имеет равную вероятность, то энтропия ансамбля будет равна (2):

$$H = -\sum_{i=1}^{C_N} p_i \log_2 p_i = -\sum_{i=1}^{C_N} \frac{1}{C_N}\log_2\left(\frac{1}{C_N}\right) = \log_2 C_N; \quad (2)$$

Из формулы (2.2) следует, что количество битов, необходимое для кодирования структуры ансамбля классификаторов с N узлами, будет равно log_2C_N, с округлением до целого числа в большую сторону. Данная последовательность обладает избыточностью информации по отношению к исходному сообщению из-за округления ее в большую сторону. Величина абсолютной избыточностью будет равна (3):

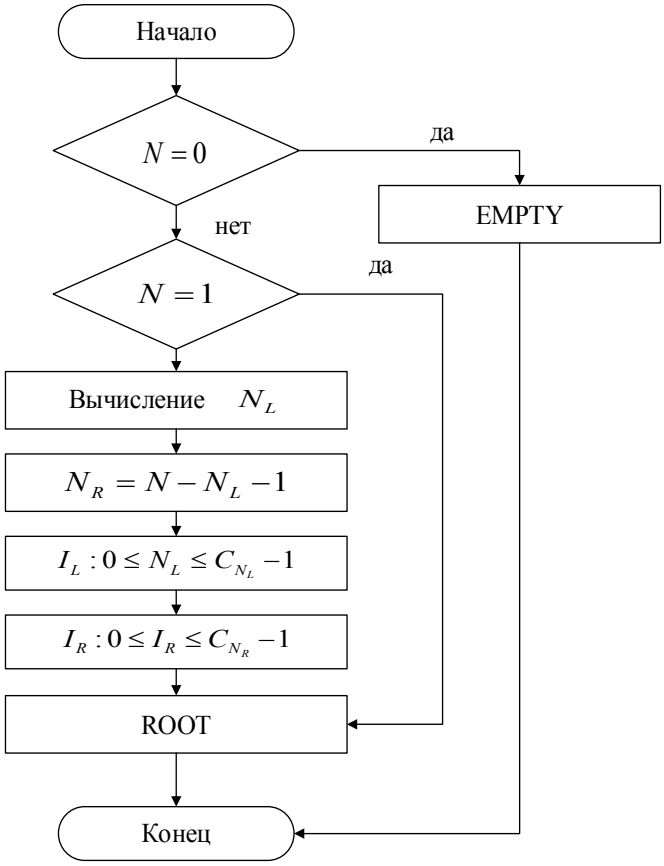
$$R_a = \left[\log_2 C_N\right] - \log_2 C_N; \quad (3)$$

Относительная избыточность (4):

$$R = \frac{[\log_2 C_N] - \log_2 C_N}{[\log_2 C_N]} \quad (4)$$

Так как число всех возможных структур ансамбля равно числу Каталана C_N, то любое число от 0 до C_N-1 возможно сопоставить с конкретной структурой ансамбля. Задача кодирования заключается в нахождении алгоритма однозначного преобразования индекса классификатора в структуру ансамбля классификаторов.

Для решения данной задачи предположим, что индексу 0 соответствует ансамбль с N узлами, ориентированный полностью по левую сторону. Индексу C_N-1 соответствует ансамбль с N узлами, ориентированный полностью по правую сторону. Тогда введем следующие обозначения, Llink — корневой узел левой стороны или Empty, если левая сторона отсутствует, Rlink — корневой узел правой стороны или Empty, если правая сторона отсутствует.



Алгоритм построения иерархического классификатора.

Природа иерархической структуры подсказывает возможность применения рекурсивного алгоритма рисунок 2. На каждом шаге алгоритма по заданному индексу структуры и числу необходимо следующее:

К решению задачи вычисления числа N_L можно подойти исходя из того, что нулевому индексу соответствует классификатор, полностью ориентированное по левую сторону, для $I = 0$ имеем $N_L = N - 1$, $N_R = 0$. Число комбинаций левой стороны будет равно C_{N-1}. Тогда, превые C_{N-1} индексов перечисляют различные комбинации левой стороны классификатора, состоящей из $N - 1$ узлов, при пустой правой стороне.

При этом левая сторона комбинаций $C_{N-1} - 1$ будет полностью ориентированно по правую сторону. На индексе C_{N-1} один узел должен быть принесён слева направо: $I = C_N - 1$; $N_L = N - 2$, $N_R = 1$. Следующие C_{N-2} индексы перечисляют различные комбинации левой стороны $(N_L = N - 2)$, при правой стороне, состоящей из одного узла.

На комбинации $C_{N-1} + C_{N-1}$ еще один узел должен быть перенесен слева направо: $I = C_{N-1} + C_{N-2}$; $N_L =$

Таблица 1. Сравнение алгоритмов.

	Тестовая местность 1		Тестовая местность 2		Тестовая местность 3		Тестовая местность 4	
	Процент ошибки распознавания образов	Время, с	Процент ошибки распознавания образов	Время, с	Процент ошибки распознавания образов	Время, с	Процент ошибки распознавания образов	Время, с
SVM	5	7,4	3,2	6,7	3,9	6,4	4,5	5,3
SVM+	0,15	3,6	0,18	4,1	0,2	4,3	0,2	4,1
DL	3,9	5,3	2,3	6,6	4,5	5,8	0,3	7,8
DL+	0,1	3,5	0,09	3,9	0,25	4,1	0,15	3,7

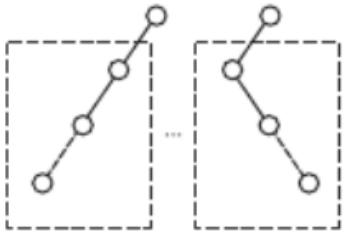


Рис. 2. Первые C_{N-1} комбинаций ансамбля классификаторов.

Рис. 3. Комбинации с C_{N-1} ансамбля классификаторов.

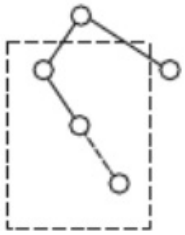


Рис. 4. Комбинации $C_{N-1} + C_{N-2} + 1$ по иерархического классификатора

$N-3$, $N_R = 2$. Начиная с $N_R = 2$ число возможных структур правой стороны ансамбля классификаторов отлично от единицы и равно C_{N_R}. В общем случае это правило можно распространить и на $N_R = \{0,1\}$, так как $C_0 = C_1 = 1$. Поэтому, для каждого N_L от $N-1$ до 0 число различных комбинаций левой и правой стороны ансамбля классификаторов равно $C_{N_L} C_{N-N_L-1}$.

$$\sum_{i=0}^{K} C_i C_{N-i-1} < I$$

$$\sum_{i=0}^{K+1} C_i C_{N-i-1} > I$$

$$N_R = K$$

$$N_L = N - K - 1. \quad (5)$$

Это соответствует формуле рекуррентного соотношения для вычисления чисел Каталана: $C_0 = 1$ и

$$C_N = \sum_{i=0}^{N-1} C_i C_{N-i-1}$$

для $N>0$. С учетом изложенного, для получения чисел N_L и N_R необходимо найти такое K, при котором будут выполняться условия (5).

Для определения параметров I_L и I_R необходимо предварительно вычислить индекс текущей комбинации для заданных значений N_L и

$$N_R: \hat{I} = I - \sum_{i=0}^{K} C_i C_{N-i-1}$$

значение $\hat{I}$ равно разности заданного индекса комбинации и индекса последней комбинации, на которой было выполнено перемещение узла из левого поддерева в правое (т.е. уменьшение числа N_L и увеличение числа N_R).

При перечислении комбинаций ансамбля (с заданными N_L и N_R) правый потомок фиксируется, и последовательно рассматриваются C_{N_L} комбинаций левого потомка. Далее осуществляется переход к следующей комбинации правого потомка, и процедура повторяется. Следовательно,

$$I_R = \left[\frac{\hat{I}}{C_{N_L}}\right] \quad I_L = \hat{I} - I_R C_{N_L},$$

скобки будут обозначать — округление до ближайшего целого в меньшую сторону. В ходе работы алгоритма была получена структура ансамбля классификаторов.

Экспериментальная часть

Был произведен сравнительный анализ алгоритмов, использующих для построения структуры ансамбля классификаторов генетический алгоритм и не использующих. В качестве простых классификаторов в составе ансамбля

классификаторов были использованы машины опорных векторов (SVM) и классификаторы, обучаемые с применением методик глубокого обучения, стохастической машины Больцмана и метода сопряжённого градиента (DL).

Где:

- SVM+ — алгоритм распознавания образов, представляющий из себя ансамбль бинарных классификаторов, основанных на применении машины опорных векторов, построенный с использованием генетического алгоритма;
- DL+ — алгоритм распознавания образов, представляющий из себя ансамбль бинарных классификаторов, обучаемых с применением методик глубокого обучения, стохастической машины Больцмана, метода сопряжённых градиентов, построенный с использованием генетического алгоритма;
- SVM — алгоритм распознавания образов, представляющий из себя ансамбль бинарных классификаторов, основанных на применении машины опорных векторов;
- DL — алгоритм распознавания образов, представляющий из себя ансамбль бинарных классификаторов, обучаемых с применением методик глубокого обучения, стохастической машины Больцмана, метода сопряжённых градиентов;
- тестовая местность 1 — плотная застройка частного сектора на территории города;
- тестовая местность 2 — плотная многоэтажная застройка в черте города;
- тестовая местность 3 — частная застройка на границе с многоэтажной застройкой;
- тестовая местность 4 — неплотная частная застройка за городом.

Результатом проведенных экспериментов является существенное уменьшение процента ошибки и времени распознавания классификаторами, построенными с помощью генетического алгоритма. Результаты точности и времени классификации приведены в таблице 1.

Заключение

Генетические алгоритмы являются надстройкой над существующими алгоритма классификации. Обучение каждого из алгоритмов, входящих в иерархический классификатор, производится с использованием стандартного для соответствующей модели метода. Результаты экспериментов позволяют говорить о применимости и эффективности генетического алгоритма в построение классификаторов для создания двумерных карт местности, т.к. с помощью данного алгоритма была получена структура иерархического классификатора, позволяющая увеличить точность и уменьшить время на классификацию объектов в задачи построения двумерной карты местности.

ЛИТЕРАТУРА

1. David. Genetic Algorithms in Search, Optimization and Machine Learning — Addison-Wesley Longman Publishing Co., Inc. Boston, MA, USA — 1989.
2. Davis T. Catalan numbers // geometer.org: веб-сайт. URL: http://www.geometer.org/ mathcircles/catalan.pdf (дата обращения: 17.08.2016).

Журнал «Современная наука: актуальные проблемы теории и практики»