ПРОГРАМНО-ТЕХНІЧНІ КОМПЛЕКСИ

И.В. РУБАН, В.А. МАРТОВИЦКИЙ, Н.В. ЛУКОВА-ЧУЙКО

УДК 004.93

ПОДХОД К КЛАССИФИКАЦИИ СОСТОЯНИЯ СЕТИ НА ОСНОВЕ СТАТИСТИЧЕСКИХ ПАРАМЕТРОВ ДЛЯ ОБНАРУЖЕНИЯ АНОМАЛИЙ В ИНФОРМАЦИОННОЙ СТРУКТУРЕ ВЫЧИСЛИТЕЛЬНОЙ СИСТЕМЫ

Аннотация. Исследован подход к классификации состояния сети на основе статистических параметров. Установлены недостатки методов классификации состояния сети, рассмотрена базовая реализация комитета классификаторов. Предложена модификация комитета классификаторов с использованием нейронной сети в качестве метаклассификатора. Проведены эксперименты для классификации состояния сети.

Ключевые слова: стекинг, классификация, машинное обучение, вычислительные системы, метаобучение.

ВВЕДЕНИЕ

В настоящее время сложно представить повседневную жизнь без информационно-вычислительных систем. Устройства, соединенные между собой, генерируют большие объемы данных, которые могут разместить только дата-центры, построенные на основе кластерной архитектуры. Для их нормального функционирования необходимы мощные коммуникационные системы.

Основной задачей телекоммуникационной инфраструктуры дата-центра, который располагается в конкретном центре обработки данных или объединяет территориально-распределенные ресурсы, является обеспечение прозрачной среды для взаимодействия всех вычислительных элементов, включенных в структуру дата-центра. Это в свою очередь приводит к нарушениям безопасности в суперкомпьютерных технологиях. Так, в 2012 г. нефтегазовая компания Саудовской Аравии «Saudi Aramco» признала, что на ее компьютерные системы совершена кибератака: заражению подверглись все рабочие станций ее сотрудников, включая их суперкомпьютер для моделирования процессов распределения нефти [1].

Примером самой сложной вирусной атаки последнего времени является вирус Flame, представляющий собой наибольшую угрозу хищения информации. Этот вирус нацелен на широкий спектр информации на зараженных компьютерах и выполняет свою задачу лучше любого другого ранее выявленного аналога [2].

Современные требования к безопасности и качеству обслуживания телекоммуникационных сетей состоят в том, что необходимо не только иметь данные о текущем состоянии сети, но и уметь его прогнозировать. Таким образом, актуальна разработка новых методов управления телекоммуникационной сетью. Одним из компонентов подобной системы управления является система мониторинга.

Проведенный анализ системы мониторинга вычислительных кластеров [3] показал, что такие системы способны выдавать пользователю статистические данные об ограниченном наборе параметров сети без учета их взаимосвязи. Поэтому для более полного анализа состояния сети в такие системы необходимо включать дополнительные модули для интеллектуального анализа статистической информации, работа которых формально сводится к решению задачи классификации состояния телекоммуникационной сети.

Применение таких модулей позволит спрогнозировать изменение показателей сети с учетом взаимовлияния и доминирования информационных потоков, что предоставит возможность выявить аномальное состояние не только сетевой инфраструктуры, но и всех компонентов дата-центра в целом.

АНАЛИЗ СОВРЕМЕННЫХ ИССЛЕДОВАНИЙ И ПУБЛИКАЦИЙ

Для решения задачи классификации состояния телекоммуникационной сети используется множество подходов и алгоритмов. Классические решения задачи представлены в теории распознавания образов [4, 5].

В работе [6] рассмотрена методика и средства раннего выявления и противодействия угрозам, нарушения информационной безопасности в результате DDOS-атак на основе байесовской сети с применением разделяющих функций и с решением вопросов проверки гипотез. Недостатком данной методики является то, что она используется для устранения лишь одного класса атак и с ее помощью невозможно выявлять другие.

В работе [7] рассмотрены классификации, связанные с применением искусственных нейронных сетей. Проведен сравнительный анализ использования нейронных сетей с различными функциями активации и структурой, а также дана оценка скорости работы с указанием количества (в процентах) правильных ответов. Недостатком данных исследований является кодирование категориальных признаков разными целыми числами. Такой вид кодирования недопустим, поскольку к категориальным признакам необходимо применять только операции сравнения.

В работе [8] особое внимание уделено статистической теории распознавания и методу «обобщенный портрет», в [9] в качестве метода классификации предложен метод группового учета аргументов, а в [10] — логические методы распознавания и поиска зависимостей.

Таким образом, можно констатировать, что спектр методов достаточно широк: от классического статистического анализа до аппарата искусственных нейронных сетей. Но ни один из методов не предназначен для работы с большим числом категориальных признаков, а также для определения различных типов кибернетических атак.

Целью настоящей работы является модификация классического комитета различных алгоритмов классификации состояний сети на основе статистических параметров телекоммуникационной инфраструктуры дата-центра. Для этого необходимо:

— рассмотреть модели классификации состояния телекоммуникационной сети;

— построить комитет моделей классификации состояния телекоммуникационной сети;

— сравнить работу комитета с работой базовых классификаторов.

ПОСТАНОВКА ЗАДАЧИ КЛАССИФИКАЦИИ СОСТОЯНИЯ СЕТИ

Классическая задача классификации объектов сформулирована в [11].

Необходимо разработать алгоритм, который сможет по признаковому описанию нового объекта выдать значение его целевого признака.

Набор данных, используемый в настоящей работе для построения модели, смоделирован и получен в сети обмена данными кластерного суперкомпьютера. За основу выбран набор параметров, представленный на соревновании по машинному обучению KDD Cup'99, и добавлены параметры для мониторинга хранилища данных.

В табл. 1–4 приведены примеры исходных данных для задачи классификации состояний сети.

Таблица 1. Параметры TCP

Параметр	Описание
duration	Продолжительность соединения, с
protocol_type	Протокол транспортного уровня
service	Сервис прикладного уровня
number of data bytes source-destination	Входящий поток, байт
number of data bytes destination-source	Исходящий поток, байт
flagstatus	Флаги, установленные в заголовке TCP-пакета
land	Совпадение адресов, иначе 1
wrong_fragment	Количество неправильных фрагментов
urgentpackets	Наличие срочных данных в пакете (флаг URG)

Таблица 2. Характеристики сеанса

Параметр	Описание
hot	Количество hot-индикаторов
num_failed_logins	Количество неудачных попыток входа
logged_in	Успешный вход
num_compromised	Доступ с административными полномочиями
root_shell	Количество попыток доступа с правами администратора
num_root	Количество операций с файлами контроля доступа
num_file_creations	Количество операций создания файла
num_access_files	Количество операций с файлами управления доступом
num_compromised	Количество скомпрометированных статусов
is_hot_login	Принадлежность пользователя к hot-списку
is_guest_login	Признак гостя системы

Таблица 3. Статистика соединения за 2 с

Параметр	Описание
count	Количество соединений с совпадающим хостом
serror_rate	Процент соединений с ошибкой SYN для данного хоста
rerror_rate	Процент соединений с ошибкой REJ для данного хоста
service	Количество соединений с одинаковым исходным портом
same_srv_rate	Процент соединений с совпадающим сервисом
diff_srv_rate	Процент соединений с различным сервисом
srv_count	Количество соединений с совпадающим сервисом
srv_serror_rate	Процент соединений с ошибкой SYN для данной службы источника
srv_rerror_rate	Процент соединений с ошибкой REJ для данной службы источника
srv_diff_host_rate	Процент соединений с различающимися хостами

Таблица 4. Характеристики файловой системы кластера Lustre

Параметр	Описание
num_exports	Количество экспорта на MDT, в том числе другие серверы Lustre
stats	Количество клиентских соединений по NID
lock_count	Количество блокировок
pool.granted	Lustre-распределенный менеджер блокировки (ldlm) передал блокировки
grant_rate	ldlm-блокировка уровня предоставления GR
cancel_rate	ldlm-блокировка уровня отмены CR

ОПИСАНИЕ РАБОТЫ АЛГОРИТМОВ СТЕКИНГА ДЛЯ КЛАССИФИКАЦИИ СОСТОЯНИЯ СЕТИ ДАТА-ЦЕНТРА

Стекинг использует концепцию метаобучения, т.е. пытается обучить каждый классификатор, используя алгоритм, который позволяет обнаружить лучшую комбинацию выходов базовых моделей [12].

Последовательность работы данного алгоритма в упрощенном виде состоит из следующих этапов.

1. Подать на вход алгоритма обучающую выборку $X = \{x_1, \ldots, x_n\}$ и множество базовых алгоритмов классификации $A = \{a_1, \ldots, a_m\}$.

2. Разбить множество X на два непересекающихся подмножества: X^a и X^b.

3. Обучить множество базовых классификаторов A на подмножестве X^a.

4. Тестировать базовые классификаторы A на подмножестве X^b, a_s: $X^b \to Y^b$.

5. Обучить метаалгоритм, используя множество Y^b как входные данные для него, а истинные значения целевой переменной — как выходные значения.

Алгоритм работы классического стекинга представлен на рис. 1.

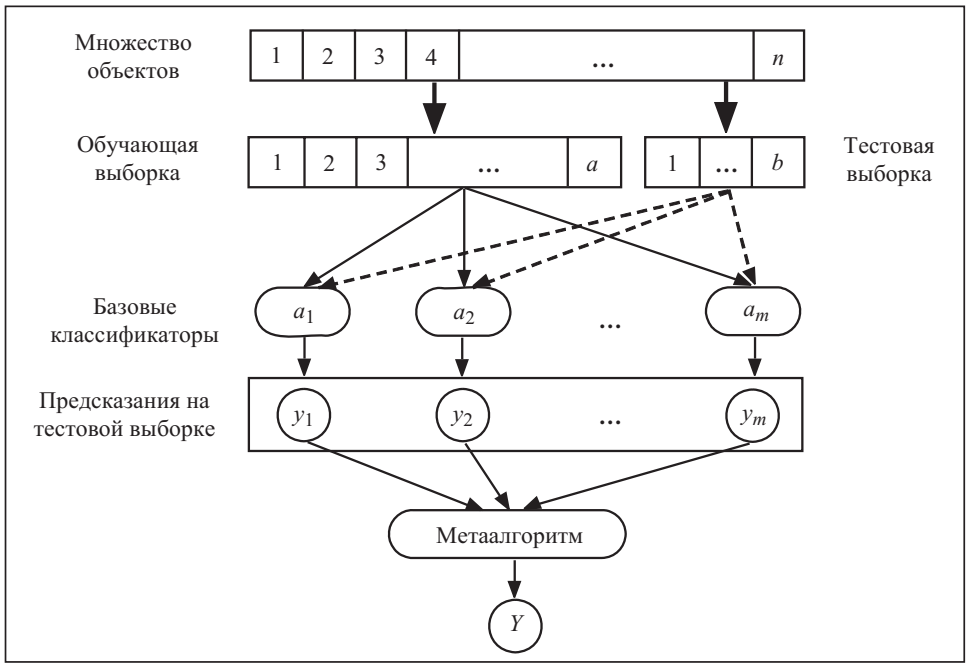

Рис. 1. Схема классического стекинга

Недостатком данного алгоритма является то, что базовые алгоритмы обучаются не на всем множестве объектов. Это в свою очередь приводит к тому, что в обучающее подмножество X^a может не попасть ни одного объекта класса k и в результате получается недообучение множества базовых алгоритмов классификации A и, следовательно, недообучение алгоритма стекинга в целом.

В настоящей работе предлагается модифицированный алгоритм стекинга. Функционирование данного алгоритма описывается следующим кортежем:

$$(A_{ij}, \mathrm{Out}_{ij}, \mathrm{In}_{ij}, x_n, A_{\mathrm{meta}}),$$

в котором $A_{ij-1}: \mathrm{In}_{ik} \to \mathrm{Out}_{ij}$ способен классифицировать произвольный объект множества In_j; $A_{\mathrm{meta}}: \mathrm{In}_{ik} \to Y$ способен классифицировать произвольный объект множества In_k. Здесь x — множество входных значений; k — количество уровней стекинга; In_k — множество входных объектов i-го алгоритма на j-м уровне; Out_{ij} — множество выходных объектов i-го алгоритма на j-м уровне; i — номер алгоритма на уровне; j — номер уровня стекинга; Y — множество целевых переменных.

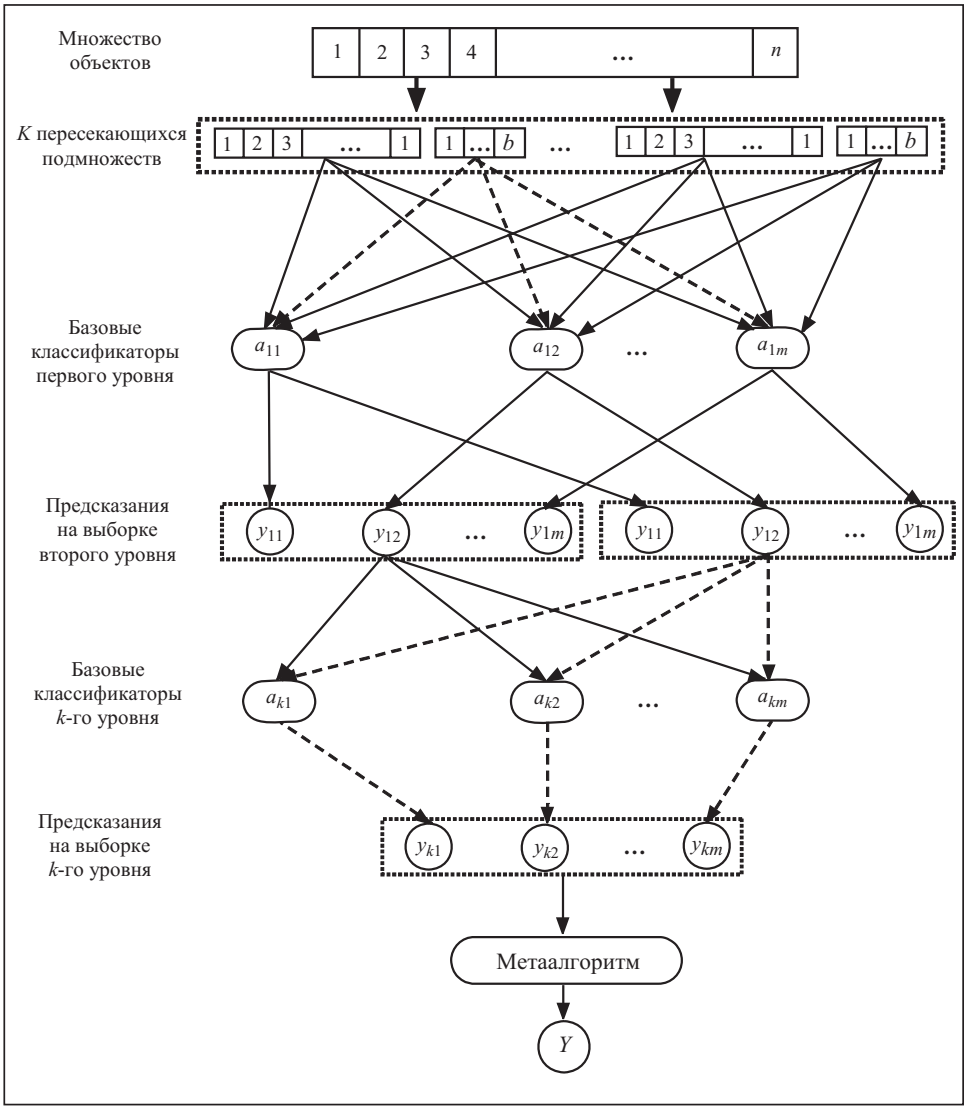

Рис. 2. Схема модифицированного стекинга

Последовательность работы модифицированного алгоритма стекинга состоит из следующих этапов.

1. Подать на вход алгоритма обучающую выборку $X = \{x_1, \ldots, x_n\}$ и множество базовых алгоритмов классификации $A = \{a_1, \ldots, a_m\}$.

2. Разбить множество X на K пересекающихся подмножеств путем равномерной выборки L объектов с возвращением. При этом каждое подмножество строить с использованием различных объектов исходной выборки X. (Примерно 37 % объектов остаются вне подмножества и не используются при построении K-го подмножества.)

3. Обучить множество базовых классификаторов A k-го уровня стекинга на подмножестве K.

4. Тестировать базовые классификаторы k-го уровня на множестве объектов, которые не вошли в K-е подмножество.

5. Обучить метаалгоритм, используя множество объектов, которые не вошли в K-е подмножество как входные данные для метаалгоритма, а истинные значения целевой переменной — как выходные значения.

Схема модифицированного алгоритма стекинга представлена на рис. 2.

Данный алгоритм позволяет избавиться от недообучения и его можно применять на обучающих выборках небольших размеров.

ОБСУЖДЕНИЕ РЕЗУЛЬТАТОВ ИССЛЕДОВАНИЯ МЕТОДОВ КЛАССИФИКАЦИИ СОСТОЯНИЯ СЕТИ ДАТА-ЦЕНТРА

Модифицируемый алгоритм стекинга позволяет использовать меньшее число объектов обучающей выборки, а также приводит к поэтапному уменьшению признакового пространства для метаклассификатора с меньшей степенью корреляции.

Дана оценка эффективности предложенного алгоритма с работой базовых классификаторов и классического стекинга. Для проведения сравнительного анализа использовались данные, представленные на чемпионате по машинному обучению KDD Cup'99, и данные, полученные при мониторинге сетевой инфраструктуры учебного дата-центра, развернутого на основе сетевой файловой системы Lustre.

Рассмотрим классы атак на сети в обучающей и тестовой выборках KDD Cup'99:

- обучающая выборка: back, buffer_overflow, ftp_write, guess_passwd, imap, ipsweep, land, loadmodule, multihop, neptune, nmap, phf, pod, portsweep, rootkit, satan, smurf, spy, teardrop, warezclient, warezmaster, normal;
- тестовая выборка: guess_passwd, imap, ipsweep, land, loadmodule, multihop, neptune, nmap, phf, pod, portsweep, rootkit, satan, smurf, spy, teardrop, warezclient, warezmaster apache2, httptunnel, mailbomb, mscan, named, perl, processtable, ps, saint, sendmail, snmpgetattack, snmpguess, sqlattack, udpstorm, worm, xlock, xsnoop, xtermbuffer_verflow, Neptune, warezmaster, smurf, normal.

Представим классы DDOS-атак, которые проводились на сеть учебного дата-центра:

- обучающая выборка: back, neptune, pod, spy, normal;
- тестовая выборка: back, neptune, pod, smurf, spy, teardrop, normal.

Для проведения эксперимента использовались такие базовые классификаторы: kNN, наивный классификатор Байеса, деревья классификации, SVM. В качестве метаклассификатора использовался многослойный персептрон.

Первый этап эксперимента включает подготовку данных для обучения базовых классификаторов. Поскольку в эксперименте используются такие алгоритмы машинного обучения, как kNN и SVM, которые чувствительны к масштабированию данных, для численных признаков будем применять нормализацию по методу минимакса.

Для категориальных признаков использовалась кодировка, рассмотренная в [13], суть которой заключается в следующем.

Пусть F — некоторое множество вещественных функций, в котором для произвольного натурального числа q существует только одна функция от q переменных. При этом все функции симметричные, т.е. для любой функции $\varphi : R \to R$ — некоторая функция от q переменных из F для любой перестановки σ. Примером таких множеств может являться множество сумм средних арифметических, максимумов и т.д.

Для кодирования значения f_b b-го категориального признака выбираем множество объектов из обучающей выборки со значением $I = \{t \in \{1, 2, \ldots, l\} \mid f_{tb} - f_b\}$, а также выбираем вещественный признак, относительно которого кодируем, на-

Таблица 5. Результаты эксперимента на дата-сете KDD Cup'99

Классификаторы	Количество правильных решений		Количество неправильных решений			
			Ошибки I рода		Ошибки II рода	
	Обучающая выборка	Тестовая выборка	Обучающая выборка	Тестовая выборка	Обучающая выборка	Тестовая выборка
Наивный классификатор Байеса	80.59	64.90	15.68	23.40	3.73	11.70
kNN	85.40	70.40	10.60	19.97	4.00	9.63
SVM	84.30	66.47	10.60	19.93	5.10	13.60
Деревья классификации	85.70	72.80	8.10	15.60	6.20	11.60
Классический стекинг	86.64	71.38	3.09	15.82	10.27	12.80
Модифицированный стекинг	92.01	84.19	4.70	10.49	3.29	5.32

Таблица 6. Результаты эксперимента на дата-сете, полученные при мониторинге вычислительной сети кластерного суперкомпьютера

Классификаторы	Количество правильных решений		Количество неправильных решений			
			Ошибки I рода		Ошибки II рода	
	Обучающая выборка	Тестовая выборка	Обучающая выборка	Тестовая выборка	Обучающая выборка	Тестовая выборка
Наивный классификатор Байеса	79.45	62.45	16.80	25.60	3.75	11.95
kNN	82.80	67.74	12.80	20.20	4.40	12.06
Деревья классификации	80.13	66.47	12.02	19.93	7.85	13.60
Классический стекинг	81.23	69.30	11.07	17.23	7.70	13.47
Модифицированный стекинг	82.44	68.80	5.16	17.82	12.40	13.38

пример, s-й, и кодируем значение f_b значением подходящей функции из F (т.е. функции от $|L|$ переменных) от значений f_{ls}, $l \in L$. Например, кодирование протокола осуществляем его заменой на среднее арифметическое значение продолжительности сеанса для данной категории.

Второй этап — это обучение базовых алгоритмов, классического и модифицированного стекинга. После чего проводим экспериментальное исследование эффективности работы базовых классификаторов, классического и модифицированного стекинга, результаты которого приведены в табл. 5, 6.

Таким образом, предложенная модификация алгоритма стекинга является действенной для предотвращения недообучения базовых алгоритмов.

ЗАКЛЮЧЕНИЕ

В результате проведенных исследований проанализирована работа базовых алгоритмов классификации и комитетов этих алгоритмов для определения состояния телекоммуникационной сети учебного дата-центра.

Исследованы результаты работы базовых классификаторов и классического алгоритма стекинга, что позволило выявить недостатки этого комитета и модифицировать его. Таким образом, доказана возможность применения алгоритма модифицированного стекинга для малых объемов данных, а также показано, что за счет использования разнородных классификаторов стекинг более устойчив к шумам в данных, что позволяет более точно классифицировать состояния телекоммуникационной сети.

СПИСОК ЛИТЕРАТУРЫ

1. Bronk C., Tikk-Ringas E. The cyber attack on Saudi Aramco. *Survival*. 2013. Vol. 55, N 2. P. 81–96.
2. Flamer — самая сложная вирусная угроза последнего времени. URL: https://xakep.ru/2013/10/19/flamer-samaya-slozhnaya-virusnaya-ugroza-poslednego-vremeni (дата обращения 04.07.2017).
3. Рубан И., Мартовицкий В., Лукова-Чуйко Н. Разработка модели мониторинга кластерных суперкомпьютеров. *Восточно-Европейский журнал передовых технологий*. 2016. Т. 6, № 2. С. 32–37.
4. Воронцов К.В. Машинное обучение (курс лекций). 2009. URL: http://bit. ly/ml_course (дата обращения 04.07.2017).
5. Смеляков К.С., Построение эффективной стратегии сегментации с использованием дерева решений. *Системи озброєння і військова техніка*. 2013. № 4. С. 125–127.
6. Oke G., Loukas G., Gelenbe E. Detecting denial of service attacks with bayesian classifiers and the random neural network. *Fuzzy Systems Conference, 2007 (FUZZ-IEEE 2007)*. IEEE International. IEEE, 2007. P. 1–6.
7. Высочина О.С., Шматков С.И., Салман Л.М. Анализ систем мониторинга телекоммуникационных сетей. *Радіоелектроніка, інформатика, управління*. 2010. Т. 23, № 2. С. 139–142.
8. Вапник В.Н., Червоненкис А.Я. Теория распознавания образов. Москва: Наука, 1974. 416 с.
9. Ивахненко А.Г. Системы эвристической самоорганизации в технической кибернетике. Киев: Техніка, 1971. 372 с.
10. Лбов Г.С. Методы обработки разнотипных экспериментальных данных. Отв. ред. Растригин Л.А. Новосибирск: Наука, 1981. 160 с.
11. Воронцов К.В. Математические методы обучения по прецедентам (теория обучения машин). URL: http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf (дата обращения 04.07.2017).
12. Wolpert D.H. Stacked generalization. *Neural Networks*. 1992. Vol. 5, N 2. P. 241–259.
13. Дьяконов А.Г. Методы решения задач классификации с категориальными признаками. *Прикладная математика и информатик*а. 2014. Т. 46. С. 103–127.

Надійшла до редакції 16.11.2017

І.В. Рубан, В.О. Мартовицький, Н.В. Лукова-Чуйко

ПІДХІД ДО КЛАСИФІКАЦІЇ СТАНУ МЕРЕЖІ НА ОСНОВІ СТАТИСТИЧНИХ ПАРАМЕТРІВ ДЛЯ ВИЯВЛЕННЯ АНОМАЛІЙ В ІНФОРМАЦІЙНІЙ СТРУКТУРІ ОБЧИСЛЮВАЛЬНОЇ СИСТЕМИ

Анотація. Досліджено підхід до класифікації стану мережі на основі статистичних параметрів. Встановлено недоліки методів класифікації стану мережі, розглянуто базову реалізацію комітету класифікаторів. Запропоновано модифікацію комітету класифікаторів з використанням нейронної мережі як метакласифікатора. Наведено експерименти для класифікації стану мережі.

Ключові слова: стекінг, класифікація, машинне навчання, обчислювальні системи, метанавчання.

I. Ruban, V. Martovytskyi, N. Lukova-Chuiko

APPROACH TO CLASSIFICATION OF NETWORK CONDITION ON THE BASIS OF STATISTICAL PARAMETERS FOR DETECTION OF ANOMALIES IN THE INFORMATION STRUCTURE OF THE COMPUTING SYSTEM

Abstract. The authors investigate the approach to classification of network condition on the basis of statistical parameters. The shortcomings of the methods of classification of network condition are revealed and basic implementation of the committee of qualifiers is considered. A modification of the committee of qualifiers with use of a neural network as the meta qualifier is proposed. Experiments are carried out for classification of network condition.

Keywords: stacking, classification, machine learning, computing systems, meta-learning.

Рубан Игорь Викторович,
доктор техн. наук, профессор, проректор по научно-методической работе Харьковского национального университета радиоэлектроники, e-mail: ruban_i@ukr.net.

Мартовицкий Виталий Александрович,
аспирант Харьковского национального университета радиоэлектроники, e-mail: martovytskyi@gmail.com.

Лукова-Чуйко Наталья Викторовна,
кандидат физ.-мат. наук, доцент кафедры Киевского национального университета имени Тараса Шевченко, e-mail: lukova@ukr.net.