Problems of identification in information systems

УДК 004.8 **doi: 10.20998/2522-9052.2018.4.01**

В. А. Мартовицький, І. В. Рубан, О. В. Сєвєрінов, Н. М. Бологова

Харківський національний університет радіоелектроніки, Харків, Україна

ВІДБІР ПАРАМЕТРІВ МОНІТОРИНГУ МЕРЕЖНОЇ ІНФРАСТРУКТУРИ ДЛЯ КЛАСИФІКАЦІЇ СТАНУ МЕРЕЖІ

Предметом дослідження в статті є етап попередньої обробки даних для алгоритмів машинного навчання і розгляд різних технік попередньої обробки та оцінки інформативності ознак при визначенні параметрів контролю мережевої інфраструктури для більш ефективного інтелектуального аналізу стану мережевої інфраструктури. **Мета роботи** – розгляд різних технік попередньої обробки даних і оцінки інформативності при визначенні параметрів контролю мережної інфраструктури для більш ефективного інтелектуального аналізу. В статті вирішуються наступні **завдання**: розгляд методів відбору параметрів, визначення множини параметрів для оцінки стану мережі. Використовуються методи фільтрації, які організовані на критеріях, що не залежать від методу класифікації; методи обгортки, що грунтуються на інформації про важливість ознак, яка отримана від методів класифікації або регресії, і тому можуть визначити більш глибокі закономірності в даних, ніж фільтри; вбудовані методи, які виконують відбір ознак під час процедури навчання класифікатора, і явно оптимізують набір використовуваних ознак для досягнення кращої точності. Отримано такі результати: проаналізовані різні техніки попередньої обробки та оцінки інформативності ознак при визначенні параметрів контролю мережевої інфраструктури для більш ефективного інтелектуального аналізу стану мережевої інфраструктури. Досліджені результати застосування методів відбору ознак для спрощення різних моделей машинного навчання. Сформовано мінімальний набір параметрів, які потрібні для моніторингу стану мережної інфраструктури. **Висновки**: Застосування методів відбору ознак дозволило зменшити вхідний набір параметрів для методів класифікації стану мережної інфраструктури.

Ключові слова: машинне навчання; відбір ознак; методи фільтрації; методи обгортки; вбудовані методи; мережі.

Вступ

Постановка проблеми. Збільшення кількості інформації, що обробляється обчислювальними кластерами, а також економія на кількості обслуговуючого персоналу потребують використання ефективних засобів моніторингу обчислювальних ресурсів. Результатом цього є зростання кількості параметрів, які повинна відстежувати така система моніторингу. За рахунок великих потоків даних від різних датчиків зростає ймовірність пропуску адміністратором системи негативних змін в контрольованих параметрах мережі обчислювального кластера [1]. Для вирішення даної проблеми в системі моніторингу глобально починають впроваджувати засоби автоматизованого експертного аналізу даних, заснованого на машинному навчанні.

Підготовка даних для використання в машинному навчанні включає декілька етапів. Прийнято вважати, що близько 60-70% часу займає перший етап робочого процесу: очищення, фільтрація і перетворення даних в формат, який підходить для застосування в алгоритмах машинного навчання. На другому етапі виконується попередня обробка і безпосереднє навчання моделей.

Попередня обробка і очищення даних – це важливі етапи, що забезпечують ефективне використання набору даних для машинного навчання. Необроблені дані часто є спотвореними і ненадійними, і в них можуть бути пропущені значення. Використання таких даних при моделюванні може призводити до невірних результатів. Ці завдання є частиною процесу обробки і аналізу даних групи і зазвичай мають на увазі початкове вивчення набору даних, що використовується для визначення і планування необхідної попередньої обробки.

Дана стаття присвячена етапу попередньої обробки даних для алгоритмів машинного навчання і розгляду різних технік попередньої обробки та оцінки інформативності ознак при визначенні параметрів контролю мережевої інфраструктури для більш ефективного інтелектуального аналізу стану мережевої інфраструктури.

Аналіз останніх досліджень та публікацій. Система моніторингу мережі необхідна для контролю стану всієї мережевої інфраструктури, з усіма пристроями і системами. Адміністратори можуть спостерігати за всіма компонентами мережевої інфраструктури, яка використовує певний інтерфейс і обмінюється інформацією про свій стан за стандартним протоколом.

Для виявлення різних видів загроз та порушень система моніторингу повинна відстежувати велику кількість параметрів компонентів мережі, які реалізуються на канальному, мережному, сеансовому і прикладному рівнях моделі OSI.

Спостереження на даних рівнях моделі OSI дає можливість контролювати ступінь використання ресурсів системи, а також знаходити несправності, пов'язані з роботою обладнання, що потрібно для підтримки високої надійності функціонування мережевої інфраструктури.

В [2] описується технологія використання нейронних мереж для вирішення завдання виявлення аномальної мережної активності в мережі. Також розглянута методика збору даних про мережеву активність і виділення параметрів мережних пакетів для подальшого аналізу. У даній статті здійснюється контроль на одному з рівнів мережевої інфраструктури, що призводить до зменшення спроможності виявлення загроз, направлених на різноманітні об'єкти мережевої інфраструктури.

В [3] автор для прогнозування і класифікації аномального трафіку використовує векторну машину (SVM) з урахуванням показників ефективності. Основні параметри включають дисперсію, автокореляцію та самоподібність. Перевагою даного методу є те, що він не використовує дані заголовків пакетів. Це дозволяє впроваджувати цей метод в системах реального часу. Недоліком даного підходу є те, що параметри, які передаються, включають в себе невеликий набір даних, за рахунок чого даний метод може ефективно і точно виявляти аномальний трафік лише протягом короткого періоду часу.

У [4] розглянуто підхід до мережевого моніторингу на основі технології мобільних та інтелектуальних агентів. Недолік даного підходу полягає у відсутності фільтрації даних від шумових і неінформативних параметрів, що значно ускладнює та інколи навіть погіршує модель машинного навчання. Перед ускладненням алгоритму треба переконатися в тому, що вже неможливо підвищити точність його роботи шляхом лише зміни параметрів. Тому одним з основних параметрів, що вимагає ретельної оптимізації в кожному алгоритмі, є складність моделі.

Починати ускладнення треба не з алгоритму, а з додавання параметрів, що мають фізичну інтерпретацію, важливу для прогнозування цільової величини, – їх зазвичай формулюють експерти предметної області. Оскільки часто подібні ознаки є нелінійними перетвореннями початкових даних, то складні алгоритми не здатні їх відтворити самостійно. Разом з тим, оскільки ці параметри мають характерний фізичний зміст, їх врахування в простих моделях нерідко дозволяє такі моделі зробити більш точними, ніж складні моделі, які не враховують цих параметрів.

Метою даної роботи є розгляд різних технік попередньої обробки даних і оцінки інформативності при визначенні параметрів контролю мережевої інфраструктури для більш ефективного інтелектуального аналізу. Для досягнення поставленої мети були сформовані такі завдання:

– розглянути методи відбору параметрів;

– визначити множину параметрів для оцінки стану мережі на прикладі даних, представлених в [5].

Виклад основного матеріалу

Методи відбору ознак. Інтелектуальний аналіз даних великої розмірності в даний час є дуже важливим. Рішення такої задачі нерідко ускладнюється завдяки не дуже великим вибіркам або наявності некорелюючих ознак по відношенню до цільової змінної, а також надмірних ознак.

Якісні дані – це необхідна умова для створення якісних моделей прогнозування. Щоб уникнути появи ситуації «сміття на вході, сміття на виході» і підвищити якість даних і, як наслідок, ефективність моделі, необхідно провести моніторинг працездатності даних, як можна раніше виявити проблеми і вирішити, які дії щодо попередньої обробки і очищення даних необхідні [6]. Тому виникає задача відбору інформаційних ознак.

Виходячи з природи ознак можна виділити дві основні причини відбору ознак:

1. *Велика кількість ознак, що істотно збільшує час роботи класифікаторів*. На сьогодні розвиваються ансамблеві методи машинного навчання, тому час, необхідний на обчислення, може стати просто величезним через велику кількість ознак. Також це може призвести до відмови в обслуговуванні за рахунок переповнення оперативної пам'яті. Це тягне необхідність модифікації алгоритмів класифікації для кожної платформи окремо.

2. *Зі збільшенням кількості ознак часто знижується точність прогнозування*. Особливо, якщо в даних багато шумових ознак (обмаль корелюючих з цільовою змінною). Також це призводить до появи дубльованих інформаційних ознак, і, як наслідок, до перенавчання (overfitting).

Методи відбору ознак діляться на 3 категорії.

Методи фільтрації організовані на критеріях, які не залежить від методу класифікації. Наприклад, такі, як кореляція ознак з цільовим вектором, критерії інформативності. Даний метод використовується до застосування алгоритмів класифікації. Перевагою методів фільтрації є те, що їх можна застосовувати в якості попередньої обробки для зниження розмірності множини ознак і подолання перенавчання [7].

Фільтри використовуються для відбору ознак в кластеризації, для побудови початкового наближення. Недолік таких методів – неможливість виявлення складних зв'язків між ознаками.

Прикладом фільтрації ознак є метод взаємної інформації. В основі цього методу лежить поняття ентропії інформації

$$H(X) = -\sum_{x_i \in X} p(x_i) \cdot \log_2(p(x_i)), \qquad (1)$$

де $p(x_i)$ – ймовірність того, що змінна X приймає значення x_i. У даному прикладі ця ймовірність розраховується як кількість прикладів, в яких $X = x_i$, поділене на всі приклади.

Для розрахунку кореляції між змінними використовуються ще дві величини:

$$H(Y \mid X = x_i) \; - \qquad (2)$$

частинна умовна ентропія – ентропія $H(Y)$, розрахована тільки для тих записів, для яких $X=x_i$;

$$H(Y \mid X) = \sum_{x_i \in X} p(x_i) \cdot H(Y \mid X = x_i) \; - \qquad (3)$$

умовна ентропія – щільність розподілу безперервної випадкової величини X.

Різниця між цими двома величинами визначає ступінь кореляції (взаємну інформацію) між значеннями X і Y, і наскільки вона велика:

$$IG(Y \mid X) = H(Y) - H(Y \mid X). \quad (4)$$

Методи обгортки грунтуються на інформації про важливість ознак, яка отримана від методів класифікації або регресії, і тому можуть визначити більш глибокі закономірності в даних, ніж фільтри [8]. Обгортки можуть використовувати будь-який класифікатор, який визначає ступінь важливості ознак.

Розрізняють два підходи в реалізації цих методів: методи включення (forward selection) і виключення (backwards selection) ознак. Перші починаються з пустої підмножини, до якої поступово додаються різні ознаки [9]. У другому випадку метод починається з підмножини, яка дорівнює вихідній множині ознак, і з нього поступово видаляються ознаки. При цьому кожен раз здійснюється перерахунок класифікатора.

Один із прикладів таких методів – рекурсивне видалення ознак (recursive feature elimination). Як випливає з назви, він відноситься до алгоритмів поступового виключення ознак із загального пулу.

З огляду на зовнішню оцінку вагових характеристик (наприклад, коефіцієнти лінійної моделі), метою якої є рекурсивне усунення ознак відмітимо таке. По-перше, алгоритм оцінки навчається за першим набором ознак і визначає важливість кожної ознаки. Потім найменш важливі ознаки видаляються з їх поточного набору. Ця процедура рекурсивно повторюється доти, поки в кінцевому результаті не буде досягнута бажана кількість ознак.

Вбудовані методи виконують відбір ознак під час процедури навчання класифікатора, і саме вони явно оптимізують набір використовуваних ознак для досягнення кращої точності [10]. Основним методом з цієї категорії є регуляризація. Є різні її реалізації, але основний принцип є загальним. Якщо розглянути роботу класифікатора без регуляризації, то вона полягає в побудові такої моделі, яка найкращим чином налаштувалася б на передбачення всіх точок тренувального сету. Наприклад, якщо алгоритмом класифікації є лінійна регресія, то підлаштовуються коефіцієнти полінома, який апроксимує залежність між ознаками і цільовою змінною.

Ідея регуляризації полягає в тому, щоб побудувати алгоритм, який мінімізує не тільки помилку, але і кількість використовуваних змінних [11].

Перевагою вбудованих алгоритмів є те, що вони, як правило, знаходять рішення швидше, уникаючи перепідготовки даних з нуля, при цьому зникає необхідність розділяти дані на навчальну і тестову множину. Разом з тим на даний час невідомі будь-які вбудовані методи, що дозволяють вирішити всі існуючі задачі.

Прикладом таких методів є метод регуляризації Тихонова (ridge regression). Розглянемо його так само на прикладі лінійної регресії. Якщо в тестовому наборі дана матриця ознак A і вектор цільової змінної b, то рішення буде мати вигляд $Ax = b$.

У процесі роботи алгоритму мінімізується вираз

$$\|Ax - y\|^2 + \alpha \|x\|^2, \quad (5)$$

де перший доданок є середньоквадратичною помилкою, а другий – регуляризуючим оператором (сума квадратів всіх коефіцієнтів, помножена на альфа). У процесі роботи алгоритму розміри коефіцієнтів будуть пропорційні важливості відповідних змінних, а перед тими змінними, які дають найменший внесок в усунення помилки, будуть наближатися до нуля.

Параметр α дозволяє налаштовувати внесок регуляризуючого оператора в загальну суму. З його допомогою можна вказати пріоритет – точність моделі або мінімальну кількість використовуваних змінних.

Для проведення порівняльного аналізу використовувалися дані з чемпіонату по машинному навчанню KDD 1999 і дані, отримані під час моніторингу мережевої інфраструктури навчального дата-центру, розгорнутого на основі мережної файлової системи Lustre, докладний опис яких представлено в табл. 1. Дані мають 38 ознак, з них: 27 числових ознак і 11 категоріальних.

Таблиця 1 – ***Параметри мережі***

№	Параметри
1	Тривалість з'єднання, с
2	Протокол транспортного рівня
3	Сервіс прикладного рівня
4	Вхідний потік, байт
5	Вихідний потік, байт
6	Прапори, встановлені в заголовку TCP-пакету
9	Наявність термінових даних в пакеті (прапор URG)
10	Кількість гарячих індикаторів
11	Кількість невдалих спроб входу
12	Успішний вхід
13	Доступ з правами адміністратора
14	Кількість спроб доступу з правами адміністратора
15	Кількість операцій з файлами контролю доступу
16	Кількість операцій створення файлу
17	Кількість операцій з файлами управління доступом
20	Ознака гість системи
21	Кількість з'єднань з співпадаючим хостом
22	Відсоток з'єднань з помилкою SYN
23	Кількість з'єднань з одним каналом вихідного порту
24	Відсоток з'єднань з помилкою REJ
25	Відсоток з'єднання з сервісом, що збігається
26	Відсоток з'єднань з різними послугами
27	Кількість зв'язків з сервісом, що збігається
28	Відсоток з'єднань з помилкою SYN джерела
31	Відсоток з'єднань з помилкою REJ джерела
32	Кількість експорту на MDT, в тому числі інші сервери Luster
33	Кількість клієнтських з'єднань по NID
34	Кількість блокувань
35	Luster-розподілений менеджер блокування (ldlm) передавав блокування
36	ldlm-блокування рівня надання GR
37	ldlm-блокування рівня відміни CP
38	Кількість вихідних команд у ftp-сеансі

У табл. 2 розставлені в порядку убування значення кореляції ознаки щодо цільової змінної.

Таблиця 2 – **Список параметрів мережі**

Методи	Номера ознак
Info gain	5, 3, 6, 4, 30, 29, 33, 34, 35, 38
Chi-squared	5, 3, 6, 4, 29, 30, 33, 34, 35, 12
ReliefF	3, 29, 4, 32, 38, 33, 30, 12, 36, 6
Variance threshold	3, 6, 4, 32, 29, 33, 30, 12, 36, 38

Різна впорядкованість ознак викликана тим, що у кожного методу застосовуються свої алгоритми ранжирування.

Для тестування якості класифікації данні були розбиті на три частини для використання кроссвалідації за трьома фолдами. В якості класифікатора використовувався XGBClassifier, Random Forest, AdaBoost, а помилка алгоритмів оцінювалась по метриці MSE.

На рис. 1 представлена средня помилка за трьома фолдами для кожного з методів фільтрації. Як видно з графіку, після фільтрації ознак якість алгоритму класифікації практично не змінилася. А час, необхідний для побудови моделі класифікації, який є тривалістю процесу навчання класифікатора після застосування кожного методу, істотно зменшився (табл. 3).

Наступним етапом експерименту було використання методу обгортки, а саме використання Random Forest. Число дерев в Random Forest варіювалося від 100 до 500. На рис. 2 наведено графік, на якому для перших 12 ознак показана їх важливість для трьох частин вибірки за допомогою вбудованого алгоритму в Random Forest.

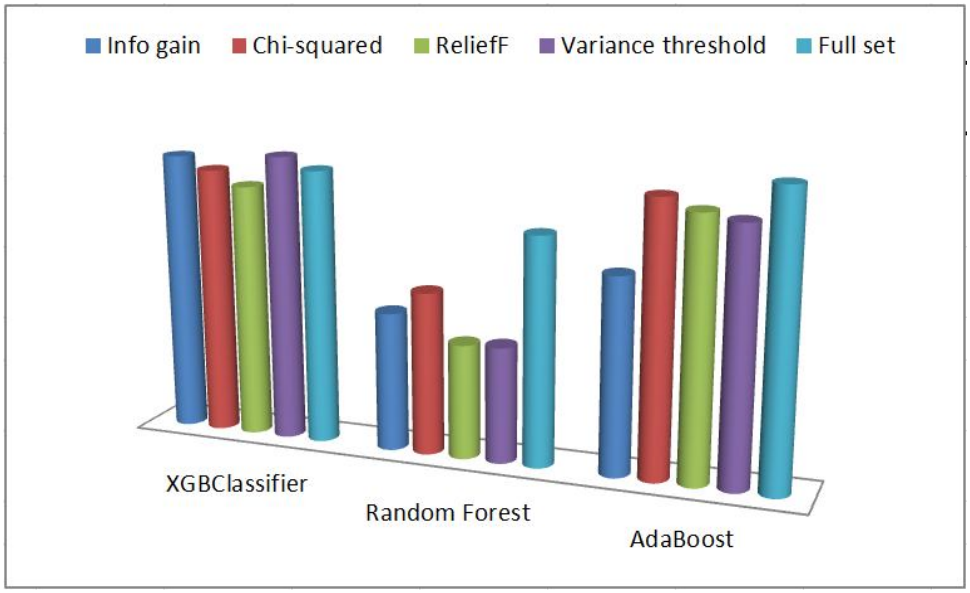

Рис 1. Точність класифікації для методів фільтрації

Таблиця 3 *Показник ефективності*

Метод	Алгоритм	MSE(%)	Час (с)
Info gain	XGBClassifier	88,74	182
	Random Forest	84,34	167
	AdaBoost	86,12	170
Chi-squared	XGBClassifier	88,36	183
	Random Forest	85,08	163
	AdaBoost	88,49	172
ReliefF	XGBClassifier	87,92	183
	Random Forest	83,58	165
	AdaBoost	88,13	172
Variance threshold	XGBClassifier	88,92	182
	Random Forest	83,62	165
	AdaBoost	87,93	170
Full set	XGBClassifier	88,56	263
	Random Forest	87,12	197
	AdaBoost	89,07	243

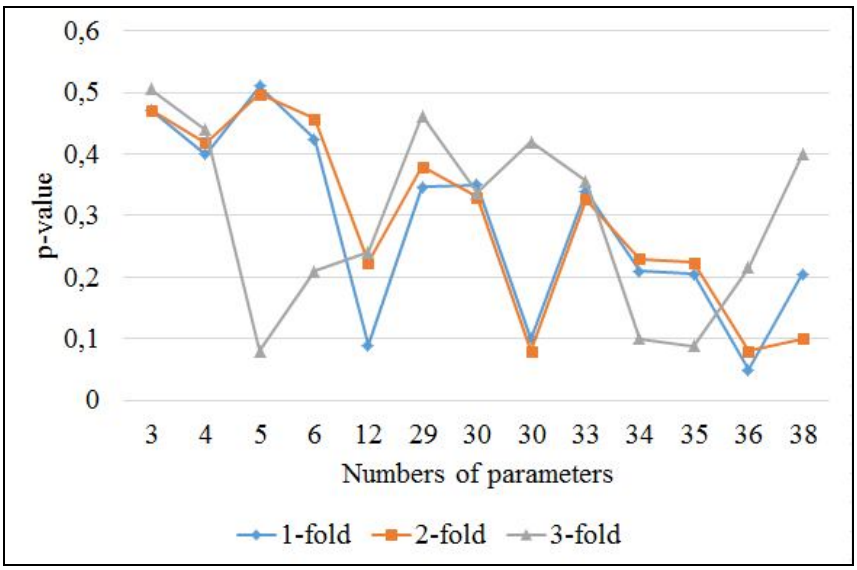

Рис. 2. Важливість ознак

Інші ознаки мали важливість меншу 0,1. Використовуючи лише ті ознаки, які мали максимальну важливість, побудований простий алгоритм, середня помилка якого по метриці MSE мало чим відрізняється від XGBClassifier, але швидкість роботи алгоритму збільшилась.

Висновки

Однією з основних проблем, з якими стикаються існуючі системи виявлення вторгнень і моніторингу, є обробка великих масивів даних, що ускладнює їх інтелектуальний аналіз.

Відбір ознак є важливим етапом побудови алгоритмів машинного навчання. Даний етап необхідний для позбавлення від шумових ознак, завдяки цьому покращується якість і збільшується швидкість роботи машинних алгоритмів. Проведені експерименти підтверджують, що алгоритми відбору ознак за допомогою Random Forest і методи фільтрації ефективно виконують це завдання.

Представлений в статті [5] алгоритм класифікації станів мережі на основі відібраних параметрів дозволить використовувати його в системах реального часу.

Список літератури

1. Ruban I. Designing a monitoring model for cluster super–computer / I. Ruban, V. Martovytskyi, N. Lukova-Chuiko // Eastern-European Journal of Enterprise Technologies. – 2016. – №6 (84). – С. 32-37.
2. Катасев А. С. Нейросетевая диагностика аномальной сетевой активности / А. С.Катасев, Д. В. Катасева, А. П. Кирпичников // Вестник Казанского технологического университета. – 2015. – № 18. – С. 163-167.
3. Gavalas Damianos. Advanced network monitoring applications based on mobile/intelligent agent technology / Gavalas Damianos // Computer Communications. – 2000. – № 23. – P. 720-730.
4. Liu L. Anomaly diagnosis based on regression and classification analysis of statistical traffic features / Liu L. // Security and Communication Networks. – 201.4 – № 7. – P. 132-138. – DOI: https://doi.org/10.1016/S0140-3664(99)00232-7.
5. Рубан И.В. Подход к классификации состояния сети на основе статистических параметров для обнаружения аномалий в информационной структуре вычислительной системы / И. В. Рубан, В. А. Мартовицкий, Н. В. Лукова-Чуйко // Кибернетика и системный анализ. – 2018. – № 54 (2). – С. 142-150.
6. Task to prepare data for enhanced machine learning [Electronic resource]. – Access mode: https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/prepare-data.
7. Chandrashekar G. Survey on feature selection methods / G. Chandrashekar, F. Sahin // Computers & Electrical Engineering. – 2014. – № 40 (1). – P. 16-28.
8. Hu Z. Hybrid filter–wrapper feature selection for short-term load forecasting / Z. Hu // Engineering Applications of Artificial Intelligence. – 2015. – № 40 (1). – P. 17-27.
9. Reif M. Efficient feature size reduction via predictive forward selection / M. Reif, F. Shafait // Pattern Recognition. – 2014. – № 47 (4). – P. 1664-1673.
10. Osanaiye O. Ensemble-based multi-filter feature selection method for ddos detection in cloud computing / O. Osanaiye, H. Cai, K.-K. R. Choo, A. Dehghantanha, Z. Xu, M. Dlodlo // EURASIP Journal on Wireless Communicationsand Networking. – 2016. – № 1. – ppP. 130-140.
11. Tibshirani R. J. Exact post-selection inference for sequential regression procedures / R. J. Tibshirani // Journal of the American Statistical Association. – 2016. – № 111(514). – P. 600-620.
12. Дьяконов А. Г. Методы решения задач классификации с категориальными признаками / А. Г. Дьяконов // Прикладная математика и информатика. – 2014. – № 46. – С. 103-127.

References

1. Ruban, I., Martovytskyi, V. and Lukova-Chuiko, N. (2016), “Designing a monitoring model for cluster super–computer“, *Eastern-European Journal of Enterprise Technologies*, Vol. 6, No. 84, pp. 32-37.

2. Katasev, A.S., Kataseva, D.V. and Kyrpychnykov, A.P. (2015), "Neural network diagnosis of abnormal network activity", *Bulletin of Kazan Technological University*, No. 18, pp. 163-167.
3. Liu L. (2014), "Anomaly diagnosis based on regression and classification analysis of statistical traffic features", *Security and Communication Networks*, No. 7, pp. 132-138.
4. Gavalas D. (2000), "Advanced network monitoring applications based on mobile/intelligent agent technology", *Computer Communications*, No. 23, pp 720-730, DOI: https://doi.org/10.1016/S0140-3664(99)00232-7.
5. Ruban, I., Martovytskyi V. and Lukova-Chuiko N. (2016), "An approach to classifying a network state based on statistical parameters for detecting anomalies in the information structure of a computer system ", *Cybernetics and Systems Analysis*, Vol. 54б No. 2, pp. 142-150.
6. Task to prepare data for enhanced machine learning, available at: https://docs.microsoft.com/en-us/azure/machine-learning/team-data-science-process/prepare-data.
7. Chandrashekar, G. and Sahin, F (2014), "Survey on feature selection methods", *Computers & Electrical Engineering*, Vol. 40, No. 1, pp. 16-28.
8. Hu, Z. (2015), "Hybrid filter–wrapper feature selection for short-term load forecasting", *Engineering Applications of Artificial Intelligence*, Vol. 40, No. 1, pp. 17-27.
9. Reif, M. and Shafait, F. (2014), "Efficient feature size reduction via predictive forward selection", *Pattern Recognition*, Vol. 47, No. 4, pp. 1664-1673.
10. Osanaiye, O., Cai, H. Choo, K.-K. R., Dehghantanha, A., Xu, Z. and Dlodlo, M. (2016), "Ensemble-based multi-filter feature selection method for ddos detection in cloud computing", *EURASIP Journal on Wireless Communication sand Networking*, No. 1, pp. 130-140.
11. Tibshirani, R.J. (2016), "Exact post-selection inference for sequential regression procedures", *Journal of the American Statistical Association*, Vol. 111, No. 514, pp. 600-620.
12. Dyakonov A.G. (2014), "Methods for solving classification problems with categorical features", *Applied Mathematics and Computer Science*, Vol. 46, pp. 103-127.

Received (Надійшла) 27.09.2018
Accepted for publication (Прийнята до друку) 07.11.2018

Отбор параметров мониторинга сетевой инфраструктуры для классификации состояния сети

В. А. Мартовицкий, И. В. Рубан, А. В. Северинов, Н. Н. Бологова

Предметом исследования в статье является этап предварительной обработки данных для алгоритмов машинного обучения и рассмотрение различных техник предварительной обработки и оценки информативности признаков при определении параметров контроля сетевой инфраструктуры для более эффективного интеллектуального анализа состояния сетевой инфраструктуры. **Цель** работы - рассмотрение различных техник предварительной обработки данных и оценки информативности при определении параметров контроля сетевой инфраструктуры для более эффективного интеллектуального анализа. В статье решаются следующие **задачи**: рассмотрение методов отбора параметров, определение множества параметров для оценки состояния сети. Используются методы фильтрации, организованные на условиях, независящих от метода классификации, методы обертки, основанные на информации о важности признаков, полученной от методов классификации или регрессии, и поэтому могут определить более глубокие закономерности в данных, чем фильтры, встроенные методы, выполняющие отбор признаков во время процедуры обучения классификатора и явно оптимизирующие набор используемых признаков для повышения точности. Получены следующие результаты: проанализированы различные техники предварительной обработки и оценки информативности признаков при определении параметров контроля сетевой инфраструктуры для более эффективного интеллектуального анализа состояния сетевой инфраструктуры. Исследованы результаты применения методов отбора признаков для упрощения различных моделей машинного обучения. Сформирован минимальный набор параметров, необходимых для мониторинга состояния сетевой инфраструктуры. **Выводы**: Применение методов отбора признаков позволило уменьшить входной набор параметров для методов классификации состояния сетевой инфраструктуры.

Ключевые слова: машинное обучение; отбор признаков; методы фильтрации; методы обертки; встроенные методы; сети.

Selection of network infrastructure monitoring parameters to classify network status

V. Martovytskyi, I. Ruban, O. Sievierinov, N. Bolohova

The subject of research in the article is the stage of preliminary data processing for machine learning algorithms and consideration of various pre-processing techniques and evaluation the informativeness of features-based parameters network infrastructure monitoring for effective intellectual state analysis. **The aim** of the work - to consider various data preprocessing techniques and evaluation of informativeness for determining controls parameters of network infrastructure for more efficient intellectual analysis. The article solves following **tasks**: consideration of methods for selecting parameters; parameter determination for assessing the state of a network filtration methods, based on algorithms that are not related to classification methods; wrapper methods, based on importance features information, obtained from classification or regression methods, which can determine data deeper patterns; embedded methods that perform feature selection during the classifier training procedure and optimize the set of features used to improve accuracy. **Results:** various preliminary processing techniques and evaluation of informativeness of feature were analyzed to determine the parameters of network infrastructure monitoring. The results of feature selection methods were analyzed to simplify the different machine learning models. The minimum parameters set has been formed for monitoring the state of the network infrastructure. **Conclusions:** The use of feature selection methods made it possible to reduce the input parameter set for classifying the state of the network infrastructure methods.

Keywords: machine learning; feature selection; filtration methods; wrapping methods; embedded methods; networks.