УДК 004.056

Имамвердиев Я. Н.[1], Сухостат Л. В.[2]

[1]Канд. техн. наук, доцент, зав. отделом, Институт информационных технологий Национальной Академии Наук Азербайджана, Баку, Азербайджан
[2]Канд. техн. наук, старший научный сотрудник, Институт информационных технологий Национальной Академии Наук Азербайджана, Баку, Азербайджан

ОБНАРУЖЕНИЕ АНОМАЛИЙ В СЕТЕВОМ ТРАФИКЕ НА ОСНОВЕ ИНФОРМАТИВНЫХ ПРИЗНАКОВ

Актуальность. Решена актуальная задача оценки информативности признаков данных большой размерности. Объектом исследования являлся сетевой трафик.

Цель работы – анализ данных сетевого трафика на предмет информативности для выявления аномалий в сетевом трафике с целью сокращения пространства признаков.

Метод. Предложен подход для оценки информативности признаков данных большой размерности, обеспечивающий повышение точности выявления аномалий в сетевом трафике и существенно увеличивающий скорость работы алгоритмов классификации. Проанализированы особенности алгоритмов случайного леса и Firefly. В работе для отбора признаков предложен подход на основе интеграции данных алгоритмов. Признаки сортируются в порядке убывания оценки их важности, наименее информативные не рассматриваются. В качестве классификаторов были рассмотрены деревья решений, наивный Байес, Байесовский классификатор, аддитивная логистическая регрессия и метод к-ближайших соседей. Результаты классификации были оценены с использованием пяти метрик: вероятности истинно-положительных и ложно-положительных результатов, F-меры, мер точности и полноты.

Результаты. Эксперименты были проведены в среде Matlab 2016a, где был реализован предложенный алгоритм на наборе данных NSL-KDD. Наилучшие результаты классификации для отобранных признаков были получены методом к-ближайших соседей.

Выводы. Проведенные эксперименты подтвердили работоспособность предложенного подхода, что позволяет рекомендовать его для применения на практике при оценке информативности с целью сокращения пространства признаков и повышения скорости работы алгоритмов классификации. Кроме того, в целях дальнейшего изучения эффективности обнаружения аномалий в сетевом трафике, будет использован набор реальных данных.

Ключевые слова: сетевые атаки, информативность признаков, случайный лес, алгоритм Firefly, NSL-KDD.

НОМЕНКЛАТУРА

RF – Random Forest;
KDD – Knowledge Discovery and Data Mining;
OOB – Out-of-Bag Error;
DoS – Denial of Service Attack;
U2R – Users to Root Attack;
R2L – Remote to Local Attack;
Weka – Waikato Environment for Knowledge Analysis;
Probe – Probing Attack;
TCP – Transmission Control Protocol;
UDP – User Datagram Protocol;
ICMP – Internet Control Message Protocol;
TPR – True Positive Rate;
FPR – False Positive Rate;
AUC – Area under ROC Curve;
BayesNet – Байесовский классификатор;
J48 – деревья решений;
LogitBoost – аддитивная логистическая регрессия;
IBk – метод к-ближайших соседей;
x_{ij} – точка из набора данных;
$\xi(x_{ij})$ – номер класса, которому соответствует значение x_{ij};
c – количество распознаваемых классов;
Θ – область допустимых значений;
I – информативность признака;
P – популяция светлячков;
β – привлекательность светлячка;
γ – коэффициент поглощения света;
r – расстояние;
α – параметр рандомизации;
T_i – ансамбль деревьев решений;
$\hat{\omega}_i(x)$ – класс новых наблюдений;
Q_t – распределение Гаусса случайных чисел.

ВВЕДЕНИЕ

В последнее время с развитием сетевых технологий угрозы безопасности значительно возросли [1, 2]. Таким образом, повышение уровня сетевой безопасности является одним из актуальных вопросов для исследователей [3].

При анализе данных сетевого трафика проблема размерности стоит остро. Размерность имеющихся данных, характеризующаяся различным числом признаков, достигает большого числа показателей. В силу этого необходимо снизить размерность признакового пространства и выделить из них наиболее важные.

Отбор признаков помогает улучшить производительность классификации и ее способности к обобщению. Другим мотивом для отбора признаков является то, что меньшее количество признаков приводит к более интерпретируемым классификаторам, что важно во многих областях (например, биомедицине).

Кроме того, измерение некоторых переменных признаков может быть довольно дорогостоящим с точки зрения денег, требований к хранению и передаче данных или времени на обучение. Данные с меньшей размерностью также могут быть более легко визуализированы. Таким образом, отбор признаков является важной задачей во многих системах классификации образов.

DOI 10.15588/1607-3274-2017-3-13

Во многих исследованиях был сделан вывод о том, что различные алгоритмы отбора признаков имеют различное поведение на различных наборах данных, и, следовательно, опасно использовать только один алгоритм [4].

Методы машинного обучения широко применяются для отбора признаков с целью анализа сетевого трафика на предмет наличия атак. Теоретически алгоритмы машинного обучения могут получить высокую производительность, т.е. могут минимизировать уровень ложных тревог и максимизировать точность обнаружения. Однако обычно требуется бесконечное число обучающих образцов. На практике это условие невозможно в силу ограничения вычислительной мощности и требования ответа в режиме реального времени.

Существует множество алгоритмов работающих на основе имитации поведения природных агентов, таких как рыбы, птицы, насекомые и т.д. Среди них алгоритм Firefly (алгоритм «светлячков») является одним из тех, который может приводить к эффективным решениям большого числа задач [5]. Целью данного исследования является разработка нового подхода для отбора признаков путем интеграции алгоритмов случайного леса (random forest, RF) [6, 7] и Firefly.

1 ПОСТАНОВКА ЗАДАЧИ

Для оценки информативности в работе рассматриваются алгоритмы случайного леса и Firefly, на основе которых отбираются наиболее важные признаки [8].

Обозначим через Θ область допустимых значений. Строки матрицы $X \in R^{m \times n}$ при этом представляют элементы обучающей выборки, $\xi(x_{ij})$ – номер класса, которому соответствует значение x_{ij} j-го признака на i-ом элементе выборки, а c – количество распознаваемых классов. Далее производится оценка информативности $I(x_{ij})$ $(i = 1,...,m)$ j-го признака с областью определения Θ алгоритмом случайного леса. Признаки сортируются в порядке убывания оценки их важности, наименее информативные не рассматриваются.

Далее на основе алгоритма Firefly необходимо сгенерировать популяцию светлячков P, где каждый светлячок соответствует отобранным признакам. При этом необходимо определить изменчивость интенсивности света (variation of light intensity) и формулировку привлекательности (attractiveness formulation). Привлекательность светлячка пропорциональна интенсивности света, которая меняется с расстоянием r и задается в виде,

$$\beta = \beta_0 e^{-\gamma r^2}, \qquad (1)$$

где β_0 превращается в привлекательность при $r = 0$. Движение светлячка k привлекает другого более яркого светлячка l и определяется как

$$y_k^{t+1} = y_k^t + \beta_0 e^{-\gamma r_{kl}^2}(y_l^t - y_k^t) + \alpha t Q_k^t. \qquad (2)$$

Требуется оценить информативность признаков сетевого трафика для повышения скорости работы систем обнаружения вторжений, сохраняя при этом достаточно хорошие результаты.

2 ЛИТЕРАТУРНЫЙ ОБЗОР

При классификации набор данных обычно включает большое количество признаков, которые могут быть релевантными, нерелевантными или избыточными. Избыточные и нерелевантные признаки не пригодны для классификации, и они могут даже снизить эффективность классификатора в отношении большого пространства поиска, которое также известно как «проклятие размерности» [9].

Преимущества отбора признаков включают в себя сокращение вычислительных затрат, экономию дискового пространства, упрощение процедур выбора модели для точного прогнозирования и интерпретации комплексных зависимостей между переменными [10]. Отобранные признаки не только оптимизируют точность классификации, но также уменьшают количество необходимых данных для достижения оптимального уровня производительности процесса обучения [11, 12].

Методы отбора признаков обычно включают в себя стратегию поиска, меру оценки, критерий остановки и валидацию результатов.

Среди двух подходов, используемых для отбора признаков, а именно метода фильтров (filter approach) и метода обертки (wrapper approach), первый работает лучше при анализе данных высокой размерности [11].

Генетический алгоритм является одним из недавних разработок для отбора признаков [13]. В настоящее время он является очень эффективным в научно-технической оптимизации.

Классификация на основе протоколов была предложена с использованием генетического алгоритма с логистической регрессией и применена к набору данных KDD'99 в работах [14, 15].

Гибридный метод для отбора признаков при обнаружении сетевых вторжений представлен в работе [16]. В этой статье, речь идет о новом алгоритме, который сочетает в себе прирост информации и генетический алгоритм.

В [17] представлен современный подход для отбора признаков на основе алгоритма Firefly.

3 МАТЕРИАЛЫ И МЕТОДЫ

В данном разделе приводится описание алгоритмов случайного леса и Firefly.

Случайный лес был предложен Л. Брейманом в статье [6]. Он строится на основе ансамбля деревьев решений, каждый элемент которого получается при помощи бутсрепа (bootstrap) [18, 19]. Называется ансамблем по той причине, что при создании одного дерева используются не все признаки пространства, а только случайно выбранные.

Алгоритм случайного леса заключается в следующем:

Пусть обучающий набор состоит из m образцов, размерность пространства признаков при этом равна n. Строится необходимое число деревьев. С помощью голосования проводится классификация. Объект классификации будет отнесен каждым деревом к одному из классов, и класс, за который проголосовало большее количество деревьев, побеждает.

Для каждого дерева выбирается подвыборка из числа наблюдений и подвыборка из числа переменных [20]. На этой подвыборке обучается дерево.

Далее получается ансамбль деревьев решений $\{T_i\}_{i=1}^{s}$, где s – количество деревьев в ансамбле ($i = 1,2,...,s$).

При предсказании новых наблюдений получается класс $\hat{\omega}_i(x) \in \{\omega_1, \omega_2, ..., \omega_c\}$, предсказанный T_i, т.е. $T_i(x) = \hat{\omega}_i(x)$; где $\hat{\omega}_{rf}^{s}(x)$ – класс, наиболее часто встречающийся в множестве $\{\hat{\omega}_i(x)\}_{i=1}^{s}$ [21].

Для этой задачи можно использовать любые классификаторы, но деревья обладают способностью быстро обучаться. На основе метрики out-of-bag error (OOB) определяется ошибка [22–24].

Преимущества случайных лесов включают:

– значительное повышение точности;

– высокая вычислительная эффективность;

– переподгонка в некоторых случаях решаема (когда количестве признаков больше числа наблюдений в обучающей выборке);

– метод прост в применении.

Их недостатками являются отсутствие наглядного представления процесса принятия решения, а, следовательно, и сложность в интерпретации результатов.

Брейман предложил меры информативности признаков, что позволило строить матрицу близости наблюдений для компенсации перечисленных выше недостатков. Одной из важных задач статистического анализа является нахождение наиболее информативных признаков. А меры информативности дают такую возможность.

Алгоритм Firefly был предложен Xin She Yang и основан на поведении светлячков [5]. Основной алгоритм Firefly предполагает, что существует P светлячков y_k ($k = 1,..., p$), первоначально произвольно размещенных в пространстве. Интенсивность света I каждого светлячка определяется целевой функцией $f(x)$. В простейшей форме, интенсивность света $I(r)$ изменяется в зависимости от расстояния r монотонно и экспоненциально, как это показано в (3):

$$I = I_0 e^{-\gamma r}, \tag{3}$$

где I_0 – исходная интенсивность света и γ – коэффициент поглощения света. Если $I_i > I_j$, $j \neq i$, то менее яркий светлячок j будет двигаться в направлении более яркого светлячка i.

Привлекательность изменяется в зависимости от расстояния $r_{ij} = d(y_i, y_j)$:

$$r_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}\ . \tag{4}$$

Движение одного светлячка к другому, более привлекательному светлячку определяется как (2). Если яркость j больше, чем i, то передвигаем i к j. Таким образом, сходимость светлячка определяется его движением.

В (2) второе слагаемое обусловлено привлекательностью. В третьем члене α – параметр рандомизации, а Q_t представляет собой распределение Гаусса случайных чисел. Если $\beta_0 = 0$, то движение становится произвольным. Если $\gamma = 0$, то задача сводится к оптимизации роя частиц.

4 ЭКСПЕРИМЕНТЫ

Для проведения экспериментов была рассмотрена база данных сигнатур NSL-KDD [25], построенная на основе базы KDD-99 по инициативе американской Ассоциации перспективных оборонных научных исследований DARPA [26]. Она охватывает широкий спектр различных вторжений, смоделированных в среде, имитирующей сеть Военно-воздушных сил США.

Рассмотренная база NSL-KDD имеет следующие преимущества [27]:

– нет избыточных записей в обучающем наборе, так что классификатор не покажет какой-либо предвзятый результат;

– нет дубликата записей в тестовом наборе. Он содержит некоторые атаки, которые не присутствуют в обучающем наборе;

– количество выбранных записей из каждой группы уровней сложности обратно пропорционально доле записей в исходном наборе данных NSL-KDD.

Обучающий набор данных состоит из 21 различных атак, а тестовый – из 37. Известные виды атак содержатся в обучающем наборе, в то время как новые атаки – это дополнительные атаки в тестовом наборе данных (они отсутствуют в обучающем наборе). Кроме того, количество записей в обучающем наборе составляет 125973 образцов, а в тестовом – 22544. Это преимущество делает его доступным для проведения экспериментов на полных данных без необходимости случайным образом выбирать небольшую часть.

Все атаки в NSL-KDD поделены на четыре группы [28]: DoS (Denial of Service Attack), U2R (Users to Root Attack), R2L (Remote to Local Attack) и Probe (Probing Attack).

Каждая запись имеет 42 атрибута, описывающих различные признаки (табл. 1). Протоколы, которые рассматриваются в NSL-KDD, включают TCP, UDP (User Datagram Protocol) и ICMP (Internet Control Message Protocol).

Метки присваиваются каждой записи либо в качестве типа «атаки», либо как «нормальное» состояние [29].

Для сравнения производительности и эффективности методов обнаружения вторжений в сети используются следующие метрики [30]:

а) Наиболее распространенными метриками для сравнения систем обнаружения вторжений являются вероятности истинно-положительных (True Positive Rate, TPR) и ложно-положительных результатов (False Positive Rate, FPR). FPR является вероятностью получения оповещения, даже если система ведет себя нормально. С другой стороны, вероятность ложно-отрицательных результатов (False Negative Rate, FNR) является вероятностью не дающей сигнала тревоги, даже если поведение системы является вредоносным. Уравнения (5) и (6) представляют FPR и FNR:

$$FPR = \frac{number\ of\ false\ positives}{number\ of\ false\ positives + number\ of\ true\ negatives}, \tag{5}$$

Таблица 1 – Список признаков для каждой записи базы данных NSL-KDD

№	Название признака	№	Название признака	№	Название признака
1	duration	15	su_attempted	29	same_srv_rate
2	protocol_type	16	num_root	30	diff_srv_rate
3	service	17	num_file_creations	31	srv_diff_hast_rate
4	flag	18	num_shells	32	dst_host_count
5	scr_bytes	19	num_access_files	33	dst_host_srv_count
6	dst_bytes	20	num_outbound_cmds	34	dst_host_same_srv_rate
7	land	21	is_host_login	35	dst_host_diff_srv_rate
8	wrong_fragments	22	is_quest_login	36	dst_host_same_src_port_rate
9	urgent	23	count	37	dst_host_srv_diff_host_rate
10	hot	24	srv_count	38	dst_host_serror_rate
11	num_failed_logins	25	serror_rate	39	dst_host_srv_serror_rate
12	logged_in	26	srv_serror_rate	40	dst_host_rerror_rate
13	num_compromised	27	rerror_rate	41	dst_host_srv_rerror_rate
14	root_shell	28	srv_rerror_rate	42	class

$$FNR = \frac{number\ of\ false\ negatives}{number\ of\ false\ negatives + number\ of\ true\ positives}. \quad (6)$$

Следовательно, вероятности TPR и истинно-отрицательных результатов (True Negative Rate, TNR) могут быть определены как:

$$TPR = 1 - FNR \text{ и } TNR = 1 - FPR. \quad (7)$$

По сути, существует компромисс между скоростью ложных срабатываний и частотой ложных отрицательных значений. Если политика обнаружения вторжений становится очень чувствительной, риск FPR будет выше. Таким образом, баланс следует рассматривать между этими двумя рисками (FPR и FNR) в конфигурации системы обнаружения вторжений.

б) Выражение (8) представляет собой меру полноты (recall), которая определяется как доля нормального поведения:

$$\text{recall} = \frac{number\ of\ true\ positives}{number\ of\ true\ positives + number\ of\ false\ negatives}. \quad (8)$$

Тем не менее, мера полноты недостаточно содержательна, так как она может быть получена тривиальным образом путем классификации всех типов поведения, как вредоносных.

в) Существует еще одна метрика, называемая мерой точности (precision), которая решает эту проблему:

$$precision = \frac{number\ of\ true\ negatives}{number\ of\ true\ negatives + number\ of\ false\ positives}. \quad (9)$$

При классификации всего трафика как нормального, мера точности достигается полностью.

г) F-мера является показателем, который сочетает в себе меры точности и полноты:

$$F - measure = \frac{2 \times recall \times precision}{recall + precision}. \quad (10)$$

д) ROC-кривая (Receiver Operator Characteristic – рабочая характеристика (приемника)) сравнивает частоту TPR с FPR [30]. Одно важное ограничение этой метрики состоит в том, что она вычисляет общую производительность системы обнаружения вторжений на всех исходных данных. Чем выше значение площади под ROC-кривой (area under ROC curve, AUC), тем лучше производительность метода.

Разрабатываемый подход был проанализирован с использованием следующих классификаторов:

– наивного Байесовского классификатора (NaiveBayes);

– деревьей решений (J48);

– аддитивной логистической регрессии (Additive Logistic Regression – LogitBoost);

– Байесовского классификатора (BayesNet);

– метода к-ближайших соседей (IBk).

В ходе тестирования были использованы реализации данных алгоритмов в программной системе Weka 3.8.0.

5 РЕЗУЛЬТАТЫ

Эксперименты были проведены в ОС Windows® 10–64 с процессором Core i7 (2,5 ГГц), 8.0 Гб ОЗУ. Оценка информативности признаков проводилась в среде Matlab 2016a на наборе данных NSL-KDD. Параметры алгоритмов Firefly и случайного леса, использованные в эксперименте, приведены в табл. 2.

В результате ошибка алгоритма случайного леса составила 0,08% при количестве деревьев равном 30 и значении OOB равном 0,03%, что показывает хорошую работу подхода.

Пять различных алгоритмов классификации (деревья решений, наивный Байес, Байесовский классификатор, аддитивная логистическая регрессия и метод к-ближайших соседей) сравниваются в программной среде Weka 3.8.0 при отобранных информативных признаках (Таблица 3) и 41 признаке базы данных NSL-KDD.

В таблицах 4 и 5 приводятся результаты работы алгоритмов классификации сетевых атак на основе 41 признака и отобранных информативных признаков из имеющегося набора данных соответственно.

Наилучшие результаты были отмечены жирным шрифтом (табл. 4–5). В качестве метрик были рассмотрены TPR, FPR, Precision, Recall, F-мера и AUC.

Таблица 2 – Параметры алгоритмов оценки информативности признаков

Число светлячков	7
Параметр рандомизации (α)	0,2
Привлекательность	2
Коэффициент поглощения света (γ)	1
Число деревьев	30

Из табл. 4–5 можно сделать заключение, что наилучшие результаты были получены методом IBk. Согласно метрике FPR наименьший процент ошибки классификации для 41 признака был достигнут методом NaiveBayes, а для отобранных 25 признаков – методом IBk.

Сравнение производительности алгоритмов (Таблица 4) показало, что метод BayesNet превосходит остальные по метрике AUC (92,5%). Анализируя полученные данные по метрике F-мера уменьшение размерности вектора признаков согласно информативности привело к улучшению работы методов J48, NaiveBayes, LogitBoost и IBk.

Сравнение значений AUC для рассмотренных классификаторов более наглядно демонстрируется на рис. 1 (красным цветом обозначены результаты для 41 признака, а синим – 25 признаков на основе предложенного подхода).

Таблица 3 – Результаты оценки информативности признаков

Подход	Отобранные признаки
Алгоритм Firefly	22,26,30,5,37,14,7,13,27,21,10,18,24,23,11,12,31,1,20,36
Случайный лес	5,2,23,24,36,10,8,4,34,40,31,32,35,22,6,27,1,33,37,16,14,11,29,13,28
Предлагаемый подход	22,26,5,2,23,24,36,37,14,13,27,21,10,18,11,12,31,1,20,8,29,28,40,35,6

Таблица 4 – Сравнение производительности алгоритмов классификации для 41 признака

Метод	TPR (%)	FPR (%)	Precision (%)	Recall (%)	F-мера (%)	AUC (%)
J48	75,8	13,2	76,7	75,8	**74,0**	81,8
NaiveBayes	70,2	**11,7**	75,8	70,2	70,7	86,5
BayesNet	71,5	19,2	78,6	71,5	67,0	**92,5**
LogitBoost	74,5	15,8	78,0	74,5	73,3	90,6
IBk	**76,8**	16,3	**81,2**	**76,8**	72,6	82,3

Таблица 5 – Сравнение производительности алгоритмов классификации для отобранных признаков

Метод	TPR (%)	FPR (%)	Precision (%)	Recall (%)	F-мера (%)	AUC (%)
J48	76,5	14,7	80,9	76,5	74,3	82,1
NaiveBayes	59,5	7,9	73,4	59,5	64,9	84,7
BayesNet	73,9	16,9	80,2	73,9	69,8	93,6
LogitBoost	78,8	11,5	81,8	78,8	78,7	93,5
IBk	**99,6**	**0,2**	**99,6**	**99,6**	**99,6**	**100**

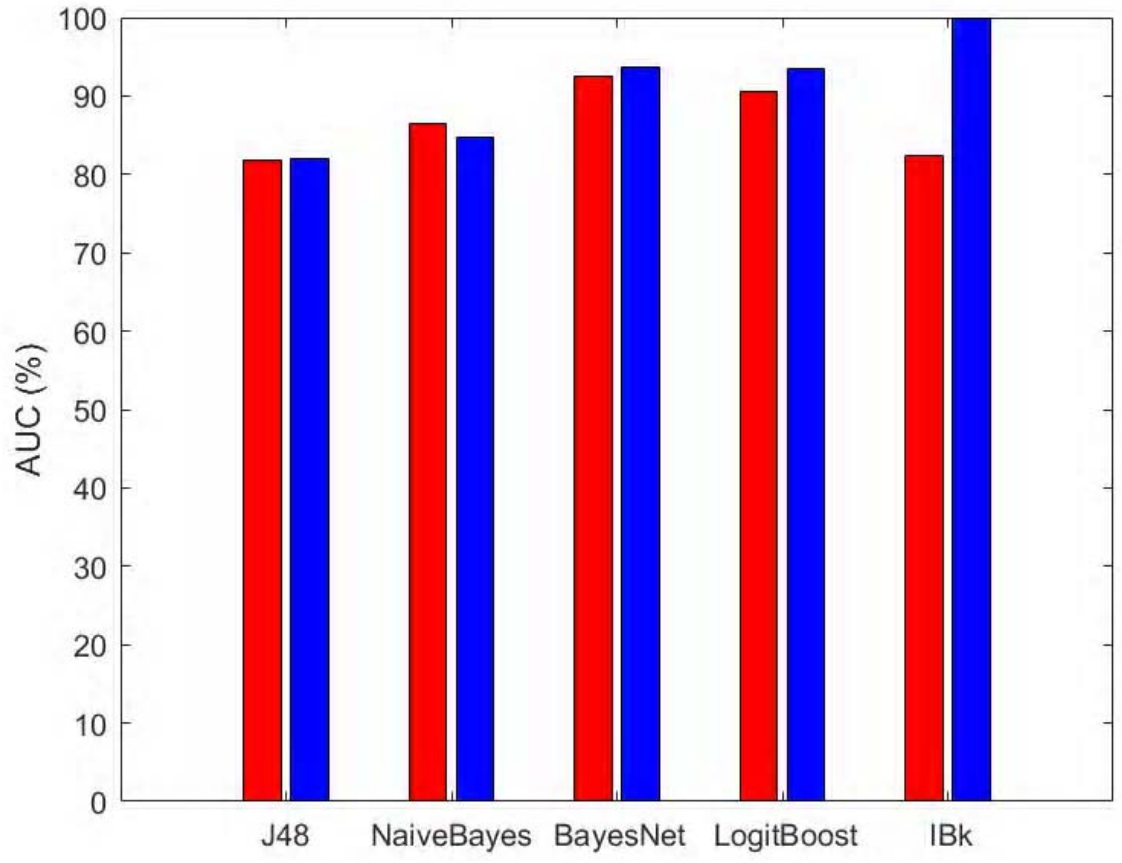

Рисунок 1 – Сравнение производительности методов классификации для базы данных NSL-KDD

6 ОБСУЖДЕНИЕ

Предложенный подход оценки информативности признаков данных большой размерности обеспечивает повышение точности выявления аномалий в сетевом трафике. Соответственно, сокращение пространства признаков существенно увеличивает скорость работы алгоритмов классификации.

Тестирование подхода проводилось на основе метрик TPR, FPR, Precision, Recall, F-мера и AUC. Наилучшие результаты были получены для метода к-ближайших соседей, однако он требует больших временных затрат. В силу этого предпочтение можно было бы отдать методам BayesNet или LogitBoost.

ВЫВОДЫ

Цель текущего исследования состояла в анализе данных высокой размерности на предмет информативности для выявления аномалий в сетевом трафике. В работе для выявления информативных признаков, используемых для обнаружения атак в сетевом трафике, были рассмотрены алгоритмы случайного леса и Firefly. В качестве классификаторов были рассмотрены деревья решений, наивный Байес, Байесовский классификатор, аддитивная логистическая регрессия и метод к-ближайших соседей.

Таким образом, экспериментальные результаты показывают, что предлагаемый подход достигает перспективной производительности при обнаружении сетевых атак на основе информативных признаков.

Хотя предложенный алгоритм отбора информативных признаков имеет обнадеживающую производительность, она может быть дополнительно повышена за счет оптимизации стратегии поиска. Кроме того, в целях дальнейшего изучения эффективности обнаружения аномалий в сетевом трафике, будет использован набор реальных данных.

СПИСОК ЛИТЕРАТУРЫ

1. Dua S. Data mining and machine learning in cybersecurity / S. Dua, X. Du. – Boca Raton, FL: CRC Press, 2011. – 256 p. DOI: 10.1201/b10867
2. Saxe J. Why security data science matters and how its different: pitfalls and promises of data science based breach detection and threat intelligence [Electronic resource]. – 2015. – Access mode: https://www.blackhat.com/us-15/speakers/Joshua-Saxe.html
3. Gates C. Challenging the anomaly detection paradigm: a provocative discussion / C. Gates, C.Taylor // Proceedings of the Workshop on New Security Paradigms. – 2007. – P. 21–29. DOI: 10.1145/505202.505211
4. Molina L. C. Feature selection algorithms: a survey and experimental evaluation / L. C. Molina, L. Belanche, A. Nebot // Proceedings of IEEE International Conference on Data Mining. – 2002. – P. 306–313. DOI: 10.1109/ICDM.2002.1183917
5. Yang X.-S. Firefly algorithms for multimodal optimization / X.-S. Yang // Stochastic Algorithms: Foundations and Applications. – 2009. – Vol. 5792. – P. 169–178. DOI: 10.1007/978-3-642-04944-6_14
6. Breiman, L. Random forests / L. Breiman // Machine Learning. – 2001. – № 1. – P. 5–32. DOI: 10.1023/A:1010933404324
7. Random forests – Classification manual [Electronic resource]. – 2017. Access mode: http://www.math.usu.edu/adele/Forests/
8. Strobl, C. Danger: High power! – exploring the statistical properties of a test for random forest variable importance / C. Strobl, A. Zeileis // Proceedings in Computational Statistics. – 2008. – P. 59–66.
9. Xue B. Particle swarm optimization for feature selection in classification: Novel initialization and updating mechanisms / B. Xue, M. Zhang, W. N. Browne // Applied Soft Computing. – 2014. – Vol. 18. – P. 261–276. DOI: 10.1109/TSMCB.2012.2227469
10. Feng D. Supervised feature subset selection with ordinal optimization / D. Feng, F. Chen, W. Xu // Knowledge-Based Systems. – 2014. – Vol. 56. – P. 123–140. DOI: 10.1016/j.knosys.2013.11.004
11. Bouaguel W. A fusion approach based on wrapper and filter feature selection methods using majority vote and feature weighting / W. Bouaguel, G. B. Mufti, M. Limam // Proceedings of the International Conference on Computer Applications Technology. – 2013. – P. 1–6. DOI: 10.1109/ICCAT.2013.6522003
12. Wang G. An improved boosting based on feature selection for corporate bankruptcy prediction / G. Wang, J. Ma, S. Yang // Expert Systems with Applications. – 2014. – Vol. 41, № 5. – P. 2353–2361. DOI: 10.1016/j.eswa.2013.09.033
13. Srinivasa K. G. Application of Genetic Algorithms for Detecting Anomaly in Network Intrusion Detection Systems / K. G. Srinivasa // Advances in Computer Science and Information Technology. Networks and Communications. – 2012. – Vol. 84. – P. 582–591. DOI: 10.1007/978-3-642-27299-8_61
14. Yu K. M. Protocol-based classification for intrusion detection / K. M. Yu, M. F. Wu, W. T. Wong // Applied Computer and Applied Computational Science. – 2008. – Vol. 3, № 3. – P. 135–141.
15. Akbar S. Intrusion detection system methodologies based on data analysis / S. Akbar, R. K. Nageswara, J. A. Chandulal // International Journal of Computer Applications. – 2010. – Vol. 5, № 2. – P. 10–20. DOI: 10.5120/892-1266
16. Sethuramalingam S. Hybrid feature selection for network intrusion detection / S. Sethuramalingam, E. R. Naganathan // International Journal of Computer Science and Engineering. – 2011. – Vol. 3, № 5. – P. 1773–1780. DOI: 10.4225/75/57a84d4fbefbb
17. Banati H. Fire Fly based feature selection approach / H. Banati, M. Bajaj // IJCSI International Journal of Computer Science Issues. – 2011. – Vol. 8, № 4. – P. 473–80.
18. Hothorn T. Unbiased recursive partitioning: a conditional inference framework / T. Hothorn, K. Hornik, A. Zeileis // Journal of Computational and Graphical Statistics. – 2006. – Vol. 15, № 3. – P. 651–674. DOI: 10.1198/106186006X133933
19. Breiman L. Stacked Regressions / L. Breiman // Machine Learning. – 1996. – Vol. 24. – P. 49–64. DOI: 10.1007/BF00117832
20. Strobl C. Conditional variable importance for random forests / C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, A. Zeileis // BMC Bioinformatics. – 2008. – Vol. 9, № 1. – P. 25. DOI: 10.1186/1471-2105-9-307
21. Siroky D. Navigating Random Forests and related advances in algorithmic modeling / D. Siroky // Statistics Surveys. – 2009. – Vol. 3. – P. 147–163. DOI: 10.1214/07-SS033
22. Archer K. J. Empirical characterization of random forest variable importance measures / K. J. Archer, R. V. Kimes // Computational Statistics & Data Analysis. – 2008. – № 4. – P. 2249–2260. DOI: 10.1016/j.csda.2007.08.015
23. Strobl C. Bias in random forest variable importance measures: illustrations, sources and a solution / C. Strobl, A.-L. Boulesteix, A. Zeileis, T. Hothorn // BMC Bioinformatics. – 2007. – Vol. 8, № 1. – P. 1471–2105. DOI: 10.1186/1471-2105-8-25
24. Liaw A. Classiыcation and Regression by randomForest / A. Liaw, M. Wiener // R News. – 2002. – Vol. 2, № 3. – P. 18–22.
25. Aggarwal P. Analysis of KDD dataset attributes-class wise for intrusion detection / P. Aggarwal, S. K. Sharma // Procedia Computer Science. – 2015. – Vol. 57. – P. 842–851. DOI: 10.1016/j.procs.2015.07.490
26. McHugh J. Testing intrusion detection systems: a critique of the 1998 and 1999 DARPA intrusion detection system evaluations

as performed by lincoln laboratory / J. McHugh // ACM Transactions on Information and System Security. – 2000. – Vol. 3, № 4. – P. 262–294. DOI: 10.1145/382912.382923
27. Tavallaee M. A detailed analysis of the KDD CUP 99 Data Set / M. Tavallaee, E. Bagheri, W. Lu, A. Ghorbani // Proceedings of the second IEEE Symposium on Computational Intelligence for Security and Defense Applications. – 2009. – P. – 53–58. DOI: 10.1109/CISDA.2009.5356528
28. NSL-KDD data set for network-based intrusion detection systems [Electronic resource]. – 2017. – Access mode: http://nsl.cs.unb.ca/NSL-KDD/
29. Davis J. J. Data preprocessing for anomaly based network intrusion detection: A review / J. J. Davis, A. J. Clark // Computers & Security. – 2011. – Vol. 30, № 6–7. – P. 353–375. DOI: 10.1016/j.cose.2011.05.008
30. Holz T. 13 security measurements and metrics for networks / T. Holz // Dependability Metrics. – 2008. – P. 157–165. DOI: 10.1007/978-3-540-68947-8_13

Статья поступила в редакцию 07.04.2017.
После доработки 26.06.2017.

Имамвердиев Я. Н.[1], Сухостат Л. В.[2]
[1]Канд. техн. наук, доцент, зав. відділом, Інститут інформаційних технологій Національної Академії Наук Азербайджану, Баку, Азербайджан
[2]Канд. техн. наук, старший науковий співробітник, Інститут інформаційних технологій Національної Академії Наук Азербайджану, Баку, Азербайджан

ВИЯВЛЕННЯ АНОМАЛІЙ У МЕРЕЖЕВОМУ ТРАФІКУ НА ОСНОВІ ІНФОРМАТИВНИХ ОЗНАК

Актуальність. Вирішено актуальне завдання оцінки інформативності ознак даних великої розмірності. Об'єктом дослідження був мережевий трафік.

Мета роботи – аналіз даних мережевого трафіку на предмет інформативності для виявлення аномалій в мережевому трафіку з метою скорочення простору ознак.

Метод. Запропоновано підхід для оцінки інформативності ознак даних великої розмірності, що забезпечує підвищення точності виявлення аномалій в мережевому трафіку і істотно збільшує швидкість роботи алгоритмів класифікації. Проаналізовано особливості алгоритмів випадкового лісу і Firefly. В роботі для відбору ознак запропонований підхід на основі інтеграції даних алгоритмів. Ознаки сортуються в порядку убування оцінки їх важливості, найменш інформативні не розглядаються. Як класифікаторів були розглянуті дерева рішень, наївний Байес, Байєсівський класифікатор, аддитивная логістична регресія і метод до найближчих сусідів. Результати класифікації були оцінені з використанням п'яти метрик: ймовірності істинно-позитивних і хибно-позитивних результатів, F-заходи, заходів точності і повноти.

Результати. Експерименти були проведені в середовищі Matlab 2016a, де був реалізований запропонований алгоритм на наборі даних NSL-KDD. Найкращі результати класифікації для відібраних ознак були отримані методом к-найближчих сусідів.

Висновки. Проведені експерименти підтвердили працездатність запропонованого підходу, що дозволяє рекомендувати його для застосування на практиці при оцінці інформативності з метою скорочення простору ознак і підвищення швидкості роботи алгоритмів класифікації. Крім того, з метою подальшого вивчення ефективності виявлення аномалій в мережевому трафіку, буде використаний набір реальних даних.

Ключові слова: мережеві атаки, інформативність ознак, випадковий ліс, алгоритм Firefly, NSL-KDD.

Imamverdiyev Y. N.[1], Sukhostat L. V.[2]
[1]PhD, Associate Professor, Head of Department, Institute of Information Technology of Azerbaijan National Academy of Sciences, Baku, Azerbaijan
[2]PhD, Senior Researcher, Institute of Information Technology of Azerbaijan National Academy of Sciences, Baku, Azerbaijan

NETWORK TRAFFIC ANOMALIES DETECTION BASED ON INFORMATIVE FEATURES

Context. The urgent task for feature informativeness evaluation of a large amount of data has been solved. The object of the study was a network traffic.

Objective is to analyze the data informativeness for network traffic anomalies detection in order to reduce the feature space.

Method. The approach for feature informativeness evaluation of a large amount of data is proposed to increase the accuracy of the anomaly detection in network traffic. It also substantially increases the computation speed of the classification algorithms. The characteristics of a random forest and Firefly algorithms are considered. In the paper, an algorithm for feature selection based on the integration of these algorithms is proposed. Features are sorted in descending order according to their importance, the least informative ones are not considered. The decision trees, naive Bayes, Bayesian classifier, additive logistic regression and k-nearest neighbors method are considered as classifiers. The quality of the classification results is estimated using six evaluation metrics: true positive rate, false positive rate, precision, recall, F-measure and AUC.

Results. The experiments have been performed in the Matlab environment (2016a) on the NSL-KDD data set, using the proposed algorithm. The best classification results for the selected features have been obtained using k-nearest neighbors method.

Conclusions. The conducted experiments have confirmed the efficiency of the proposed approach and allow recommending it for practical use in feature informativeness evaluation in order to reduce the feature space and increase the computation speed of the classification algorithms. In addition, in order to further study the effectiveness of anomaly detection in network traffic, a real data set will be used.

Keywords: network attacks, feature informativeness, random forest, Firefly algorithm, NSL-KDD.

REFERENCES

1. Dua S., Du X. Data mining and machine learning in cybersecurity. Boca Raton, FL, CRC Press, 2011, 256 p. DOI: 10.1201/b10867
2. Saxe J. Why security data science matters and how its different: pitfalls and promises of data science based breach detection and threat intelligence [Electronic resource], 2015, Access mode: https://www.blackhat.com/us-15/speakers/Joshua-Saxe.html
3. Gates C., Taylor C. Challenging the anomaly detection paradigm: a provocative discussion, *Proceedings of the Workshop on New Security Paradigms,* 2007, pp. 21–29. DOI: 10.1145/505202.505211
4. Molina L. C., Belanche L., Nebot A. Feature selection algorithms: a survey and experimental evaluation, *Proceedings of IEEE International Conference on Data Mining*, 2002, pp. 306–313. DOI: 10.1109/ICDM.2002.1183917
5. Yang X.-S. Firefly algorithms for multimodal optimization, *Stochastic Algorithms: Foundations and Applications*, 2009, Vol. 5792, pp. 169–178. DOI: 10.1007/978-3-642-04944-6_14
6. Breiman, L. Random forests, *Machine Learning*, 2001, No. 1, pp. 5–32. DOI: 10.1023/A:1010933404324
7. Random forests – Classification manual [Electronic resource], 2017, Access mode: http://www.math.usu.edu/adele/Forests/
8. Strobl C., Zeileis A. Danger: High power! – exploring the statistical properties of a test for random forest variable importance, *Proceedings in Computational Statistics*, 2008, pp. 59–66.
9. Xue B., Zhang M., Browne W. N. Particle swarm optimization for feature selection in classification: Novel initialization and updating mechanisms, *Applied Soft Computing,* 2014, Vol. 18, pp. 261–276. DOI: 10.1109/TSMCB.2012.2227469
10. Feng D., Chen F., Xu W. Supervised feature subset selection with ordinal optimization, *Knowledge-Based Systems*, 2014, Vol. 56, pp. 123–140. DOI: 10.1016/j.knosys.2013.11.004
11. Bouaguel W., Mufti G. B., Limam M. A fusion approach based on wrapper and filter feature selection methods using majority vote and feature weighting, *Proceedings of the International Conference on Computer Applications Technology*, 2013, pp. 1–6. DOI: 10.1109/ICCAT.2013.6522003
12. Wang G., Ma J., Yang S. An improved boosting based on feature selection for corporate bankruptcy prediction, *Expert Systems with Applications*, 2014, Vol. 41, No. 5, pp. 2353–2361. DOI: 10.1016/j.eswa.2013.09.033
13. Srinivasa K. G. Application of Genetic Algorithms for Detecting Anomaly in Network Intrusion Detection Systems, *Advances in Computer Science and Information Technology. Networks and Communications*, 2012, Vol. 84, pp. 582–591. DOI: 10.1007/978-3-642-27299-8_61
14. Yu K. M., Wu M. F., Wong W. T. Protocol-based classification for intrusion detection, *Applied Computer and Applied Computational Science*, 2008, Vol. 3, No. 3, pp. 135–141.
15. Akbar S., Nageswara R. K., Chandulal J. A. Intrusion detection system methodologies based on data analysis, *International Journal of Computer Applications*, 2010, Vol. 5, No. 2, pp. 10–20. DOI: 10.5120/892-1266
16. Sethuramalingam S., Naganathan E. R. Hybrid feature selection for network intrusion detection, *International Journal of Computer Science and Engineering*, 2011, Vol. 3, No. 5, pp. 1773–1780. DOI: 10.4225/75/57a84d4fbefbb
17. Banati H., Bajaj M. Fire Fly based feature selection approach, *IJCSI International Journal of Computer Science Issues*, 2011, Vol. 8, № 4, pp. 473–80.
18. Hothorn T., Hornik K., Zeileis A. Unbiased recursive partitioning: a conditional inference framework, *Journal of Computational and Graphical Statistics*, 2006, Vol. 15, No. 3, pp. 651–674. DOI: 10.1198/106186006X133933
19. Breiman L. Stacked Regressions, *Machine Learning*, 1996, Vol. 24, pp. 49–64. DOI: 10.1007/BF00117832
20. Strobl C., Boulesteix A.-L., Kneib T., Augustin T., Zeileis A. Conditional variable importance for random forests, *BMC Bioinformatics*, 2008, Vol. 9, No. 1, pp. 25. DOI: 10.1186/1471-2105-9-307
21. Siroky D. Navigating Random Forests and related advances in algorithmic modeling, *Statistics Surveys*, 2009, Vol. 3, pp. 147–163. DOI: 10.1214/07-SS033
22. Archer K. J., Kimes R. V. Empirical characterization of random forest variable importance measures, *Computational Statistics & Data Analysis,* 2008, No. 4, pp. 2249–2260. DOI: 10.1016/j.csda.2007.08.015
23. Strobl C., Boulesteix A.-L., Zeileis A., Hothorn T. Bias in random forest variable importance measures: illustrations, sources and a solution, *BMC Bioinformatics*, 2007, Vol. 8, No. 1, pp. 1471–2105. DOI: 10.1186/1471-2105-8-25
24. Liaw A., Wiener M. Classiыcation and Regression by randomForest. R News, 2002, Vol. 2, No. 3, pp. 18–22.
25. Aggarwal P., Sharma S. K. Analysis of KDD dataset attributes-class wise for intrusion detection, *Procedia Computer Science*, 2015, Vol. 57, pp. 842–851. DOI: 10.1016/j.procs.2015.07.490
26. McHugh J. Testing intrusion detection systems: a critique of the 1998 and 1999 DARPA intrusion detection system evaluations as performed by lincoln laboratory, *ACM Transactions on Information and System Security*, 2000, Vol. 3, No. 4, pp. 262–294. DOI: 10.1145/382912.382923
27. Tavallaee M., Bagheri E., Lu W., Ghorbani A. A detailed analysis of the KDD CUP 99 Data Set, *Proceedings of the second IEEE Symposium on Computational Intelligence for Security and Defense Applications*, 2009, pp. 53–58. DOI: 10.1109/CISDA.2009.5356528
28. NSL-KDD data set for network-based intrusion detection systems [Electronic resource], 2017, Access mode: http://nsl.cs.unb.ca/NSL-KDD/
29. Davis J. J., Clark A. J. Data preprocessing for anomaly based network intrusion detection: A review, *Computers & Security,* 2011, Vol. 30, No. 6–7, pp. 353–375. DOI: 10.1016/j.cose.2011.05.008
30. Holz T. 13 security measurements and metrics for networks, *Dependability Metrics*, 2008, pp. 157–165. DOI: 10.1007/978-3-540-68947-8_13