УДК 621.391

Мухи Алдин Хассан Мохамед
Одесский национальный политехнический университет

МЕТОДОЛОГИЯ ПОСТРОЕНИЯ КОРРЕЛЯЦИОННЫХ АЛГОРИТМОВ КЛАССИФИКАЦИЯ СЕТЕВОГО ТРАФИКА

В рамках данной работы были выявлены недостатки методов классификации сетевого трафика и предложена собственная схема классификации на основе корреляционного анализа. В основу математического аппарата корреляционного анализа сетевого трафика были положены понятия набора пакетов, классов элементов и величины расхождения. Разработан алгоритм построения NN-классификатора, который включал в себя три группы классификации, которые показали высокую точность при обучении на минимальном наборе образцов.

Ключевые слова: *сетевой трафик, корреляционный анализ, методы классификации, величина расхождения, «VG-NN, MIN-NN, MVT-NN.*

Мухі Алдін Хассан Мохамед
Одеський національний політехнічний університет

МЕТОДОЛОГІЯ ПОБУДОВИ КОРЕЛЯЦІЙНИХ АЛГОРИТМІВ КЛАСИФІКАЦІЯ МЕРЕЖЕВОГО ТРАФІКУ

В рамках даної роботи були виявлені недоліки методів класифікації мережевого трафіку і запропонована власна схема класифікації на основі кореляційного аналізу. В основу математичного апарату кореляційного аналізу мережевого трафіку були покладені поняття набору пакетів, класів елементів і величини розбіжності. Розроблено алгоритм побудови NN-класифікатора, який включав в себе три групи класифікації, які показали високу точність при навчанні на мінімальному наборі зразків.

Ключові слова: *мережевий трафік, кореляційний аналіз, методи класифікації, величина розбіжності, «VG-NN, MINNN, MVT-NN.*

Hassan Mohamed Muhi-Aldeen
Odessa National Polytechnic University

METHODOLOGY OF BUILDING CORRELATION ALGORITHMS OF THE NETWORK TRAFFIC CLASSIFICATION

The problems of modern classification methods are identified and the methodology for constructing of algorithms for classifying network traffic based on a correlation analysis of network traffic flows is proposed. The proposed scheme for classifying network taffic based on correlation analysis includes IP packet preprocessing unit, data flow characteristics extraction unit, and data flow correlation analysis unit. The mathematical apparatus of the correlation analysis of network traffic was based on the concepts of bag of packets, query, classes of elements, prior probability, class-conditional probability density and the distance divergence. An algorithm for constructing the NN classifier was developed. It included such classification groups as "AVG-NN", "MIN-NN" and "MVT-NN". These groups showed high accuracy of classification in terms of training on a minimum set of training samples.

Keywords: *network traffic, correlation analysis, classification methods, magnitude of discrepancy, "VG-NN, MIN NN, MVT-NN.*

Введение

Анализ научных публикаций последних лет показывает, что методология классификации сетевого трафика[1-7]играет важную роль в управления компьютерными сетями, в частности позволяет прогнозировать уровень контроля качества обслуживания (QoS:Quality of Service),а также дает возможность построить эффективную стратегию защиты сетевых сервисов от внешних угроз и обеспечить сохранность конфиденциальных данных (DLP: Data Leak Prevention). Исследования показывают, что стандартные методы классификации сетевого трафика [1-3, 7], такие как методы прогнозирования на основе использования портов (port-based prediction) и методы глубокой проверки на основе полезной нагрузки (payload-based deep inspection) утратили свою эффективность в условиях применения активных портов и шифрования программного кода приложений. На сегодняшний день предпочтение отдается методам машинного обучения для классификации трафика на основе статистических характеристик потока данных [2-4, 8], что базируется на автоматическом поиске характерных паттернов, однако не обладает достаточной точностью классификации.

Таким образом, в рамках данной работы были ***рассмотрены современные научные публикации*** посвященные классификация сетевого трафика на основе статистических характеристик потока [2-4, 6-8] с использованием алгоритмов классификации с обучением (supervised classification algorithms) и без обучения (unsupervised classification algorithms). Был

проведен анализ работ посвященных классификации сетевого трафика с использованием методов кластеризации, которые не включают предварительно подготовленный набор данных о классах трафика [9, 10]. Среди классификаторов, которые работают с предварительно подготовленным набором данных, что используется для обучения, были выделены две категории: параметрические классификаторы [3, 11, 12] и непараметрические классификаторы [13]. Из параметрических классификаторов были выделены:

- алгоритм C4.5 для построения деревьев решений (C4.5 decision tree algorithm);
- метод опорных векторов (SVM: Support Vector Machines);
- байесовская сеть;
- нейронные сети.

В свою очередь, из параметрических классификаторов был выделен метод k-ближайших соседей (k-NN: k-nearest neighbors). Показано, что использование параметрических классификаторов включает в себя этап интенсивного обучения необходимый для определения параметров классификатора, в то время как непараметрические классификаторы (в т.ч. NN-классификатор для $k=1$) принимают классификационное решение на основе доступных паттернов из множества, предоставленного сетевым трафиком [3, 14].

В результате проведенного анализа актуальных исследований в данной области были определены недостатки, которые характеризуют отдельные методы. Так, было отмечено, что эффективность непараметрических классификатора в значительной степени зависит от объема данных, которые используются при его обучении, и в случае недостаточного объема в условиях ограниченного времени подготовки системы и ее ресурсоемкости не представляется возможным определить на основе данного классификатора все классы сетевого трафика. В то же время методы классификации с предварительным обучением в значительной степени зависят от уровня учебного набора данных, что приводит к необходимости подготовки высококвалифицированного персонала, который мог бы вручную выделить пригодные паттерны и маркировать их. Таким образом, ***остается нерешенной задача*** построения алгоритмов классификация сетевого трафика, которые способны эффективно обучаться на небольшом очень наборе тренировочных паттернов, подготовленных вручную.

Целью данной работы является разработка методологии классификации сетевого трафика с использованием корреляционных методов, которая обеспечивает эффективное обучение классификатора на минимальном наборе учебных образцов. ***Для решения поставленной задачи*** было предложено использовать непараметрический подход, который включает корреляцию потоков сетевого трафика, а также провести анализ предложенного подхода на основе математической модели и рассчитать параметры, характеризующие его эффективность.

Классификация сетевого трафика на основе корреляционного анализа

Схема классификации сетевого трафика на основе корреляционного анализа включает в себя такие функциональные элементы как (рис. 1):

- первичная обработка IP-пакета;
- выделение и сопоставление характеристик потока данных;
- корреляционный анализ потока данных;
- набор для обучения.

На этапе первичной обработки система перехватывает IP-пакеты в среде сетевого ресурса информационной системы (СРИС)и строит на основе их заголовков потоки данных (рис. 1).

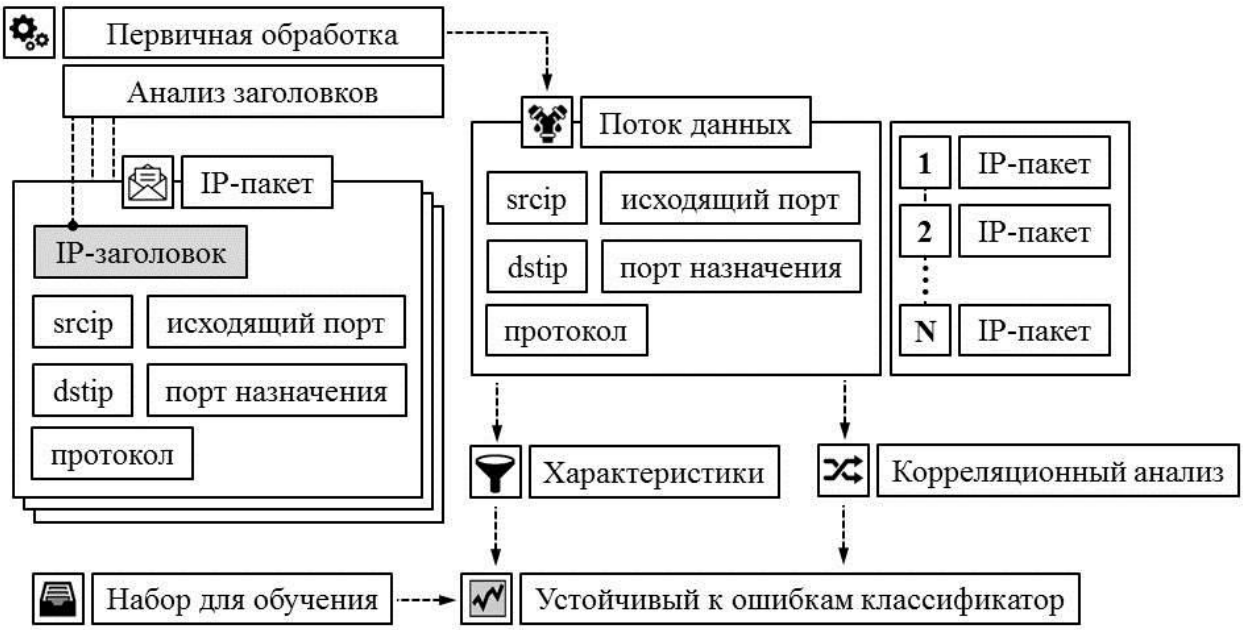

Рис. 1. **Схема классификация сетевого трафика на основе корреляционного анализа**

Потоки данных при этом состоят из последовательностей IP-пакетов, включающих такие пять элементов как:

- srcip;
- srcport (исходящий порт);
- dstip;
- dstport (порт назначения);
- протокол.

Для представления каждого отдельного потока данных извлекаются соответствующие наборы статистических характеристик. Данный этап необходим для формирования подмножества характеристик IP-пакетов, которые могут быть использованы для построения надежных моделей классификации. В свою очередь корреляционный анализ потока данных основывается на применение в отношении информационных элементов IP-пакетов корреляционной функции. Предлагается анализ корреляции потока для корреляции информации в потоках трафика. В качестве последнего этапа данные узла извлечения наборов статистических характеристик и данные корреляционного анализа потоков данных объединяются, и на их основе строится устойчивый к ошибкам классификатор, который способен эффективно обучаться при помощи минимальных наборов для обучения.

Новизна данного методологического подхода, таким образом, заключается в выделении данных о корреляции элементов потоков данных и их дальнейшем использовании в процессе классификации. В рамках стандартного подхода потоки данных рассматриваются как отдельные и независимые группы элементов, в то время как корреляционная информация может значительно улучшить показатели классификации, особенно когда размер набора для обучения в значительной мере ограничен. В предложенной модели анализ корреляции потока данных можно рассматривать как как принципиально новый компонент классификации сетевого трафика, который может быть использован для увеличения эффективности современных и наиболее эффективных алгоритмов в данной области. Тем не менее, следует отдельно отметить, что математический аппарат, который будет использоваться в рамках такого подхода, может быть чрезмерно ресурсоемким для аппаратно-программной платформы классификации и, следовательно, данную проблему необходимо рассмотреть отдельно.

Математическая модель корреляционного анализа сетевого трафика

В основе математического аппарата корреляционного анализа сетевого трафика лежит понятие набора пакетов $x_1, x_2 .... x_n$, состоящего из n элементов, который определяется через

запрос $Q\,x_1, x_2 .... x_n$. Каждый набор пакетов включает в себяIP-пакеты сгенерированные одним приложением, коэффициент корреляции которых должен быть отличным от нуля.

Согласно теореме Байеса классификатор построенный в соответствии с принципом вычисления апостериорного максимума (MAP: maximum-a-posteriori)позволяет наиболее эффективно минимизировать среднюю ошибку классификации [14]. Таким образом, для построения оптимального классификатора максимального подобия (ML:Maximum-Likelihood) элемента запроса *Q* можно использовать следующее выражение:

$$w_{opt} = \arg\max_w (P(w|Q)) \tag{1}$$

что эквивалентно:

$$w_{opt} = \arg\max_w (p(Q|w)) \tag{2}$$

где через *w* определяются классы элементов набора пакетов, *P* — априорная вероятность, *p* — условное распределение случайной величины. Величину *p(Q|w),* переходя от запроса *Q* к набору пакетов $x_1, x_2 .... x_n$, с учетом наивного байесовского допущения (Naive-Bayes assumption) можно записать как:

$$p(Q|w) = \prod_{x \in Q} p(x|w), \tag{2}$$

что позволяет упростить выражение (2) до вида:

$$w_{opt} = \arg\max_w \left(\frac{\sum_{x \in Q} p(x|w)}{\|Q\|}\right). \tag{4}$$

На практике большую эффективность показывает классификатор на основе логарифмической вероятности, что определяется подобным выражению (4) образом:

$$w_{opt}^{\log} = \arg\max_w \left(\frac{\sum_{x \in Q} \log(p(x|w))}{\|Q\|}\right). \tag{5}$$

В итоге, NN-классификатор $p_{NN}(x|w)$ может быть построен через определения подобия на основе функции ядерной оценки плотности распределения, в качестве которой можно использовать функцию Гаусса:

$$\begin{cases} p_{NN}(x|w) = \dfrac{\sum_{x_t \in w} K(\Delta x)}{\|w\|} \\ K(\Delta x) = \exp\left(-\dfrac{\|\Delta x\|}{2\sigma^2}\right) \\ \Delta x = x - x_t \end{cases} \tag{6}$$

где x_t — тренировочный образец из набора «обучения с учителем». Аналогично (5) логарифмическая вероятность для выражения (6) может быть определена как:

$$\log(p_{NN}(x|w)) = \frac{\min_{xt \in w} \|\Delta x\|}{2\sigma^2 \|w\|}. \tag{7}$$

Поскольку $2\sigma^2\|w\|$ является константой для любых x_i набора пакетов, эта величина никак не влияет на процесс квалификации и классификатор может быть определен как:

$$w_{opt}^{\log} = \arg\min_w \left(\frac{\sum_{xt\in w} \min_{xt\in w} \|\Delta x\|}{\|Q\|}\right). \quad (8)$$

Наиболее просто показать применение данного подхода на примере бинарной классификации, т.е. различения двух классов w_1 и w_2 (рис. 2).

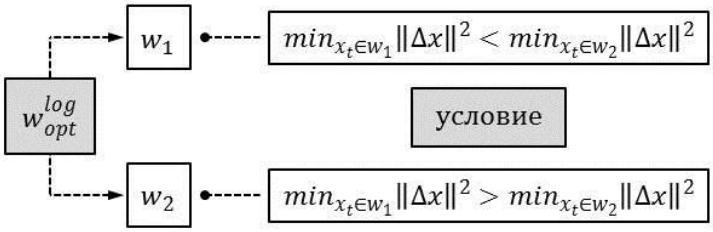

Рис. 2. **Базовая схема различения классов при классификации сетевого трафика**

Таким образом, через расстояния $\min_{xt\in wi}\|\Delta x\|^2$ можно определить величину расхождения (distance divergence) между классами w_1 и w_2:

$$\delta_x = \min_{xt\in w1}\|\Delta x\|^2 - \min_{xt\in w2}\|\Delta x\|^2 \quad (9)$$

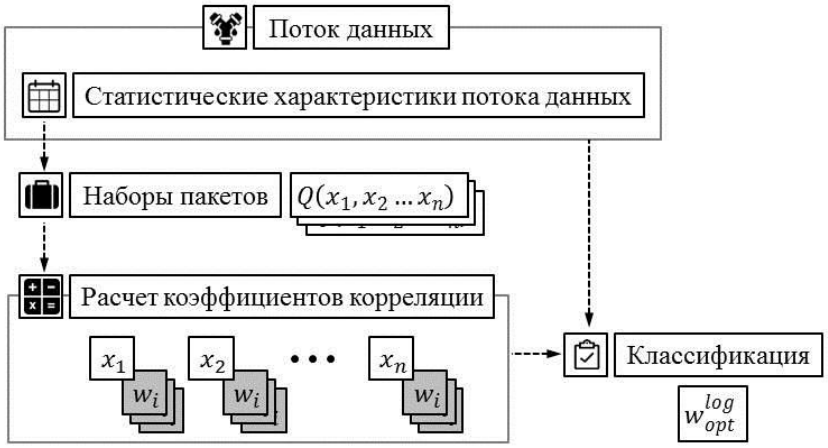

Рис. 3. **Базовый алгоритм построения NN-классификатора на основе корреляционного анализа**

Данный математический аппарат позволяет определить базовый алгоритм построения NN-классификатора на основе корреляционного анализа (рис. 3), а также выделить основные группы классификации, которые можно разработать на основе данного алгоритма и определить их эффективность. Основные группы NN-классификатора при этом включают в себя [15]:

- группа «AVG-NN»,которая для принятия решения по набору пакетов комбинирует все значения расстояний потока данных;
- группа «MIN-NN», которая для принятия решения по набору пакетов выбирает минимальное расстояние потока данных;
- группа «MVT-NN», которая для принятия решения по набору пакетов комбинирует все решения о потоках.

На основе полученных математических выражений можно точно описать каждую из групп и определить зависимость ее эффективности от размера набора для обучения.

Результаты применения корреляционного анализа сетевого трафика

Как было показано в предыдущем разделе в рамках рассмотренного подхода NN-классификатор агрегирует предсказанные значения потоков данных. При этом значение элемента *x* пакета, определяется через использование его минимального расстояния до обучающих выборок класса *w:*

$$d_x = \min_{xt \in w} \|\Delta x\|^2 \quad (10)$$

На основе этого можно построить выражения для расстояний для трех групп принятия решения по классификации набора пакетов, рассмотренных в предыдущем разделе. Так для группы «AVG-NN»рассматривается минимальное значение по *Q*:

$$d_Q^{AVG} = \frac{\sum_{x \in Q} \min_{xt \in w} \|\Delta x\|^2}{\|Q\|} \quad (11)$$

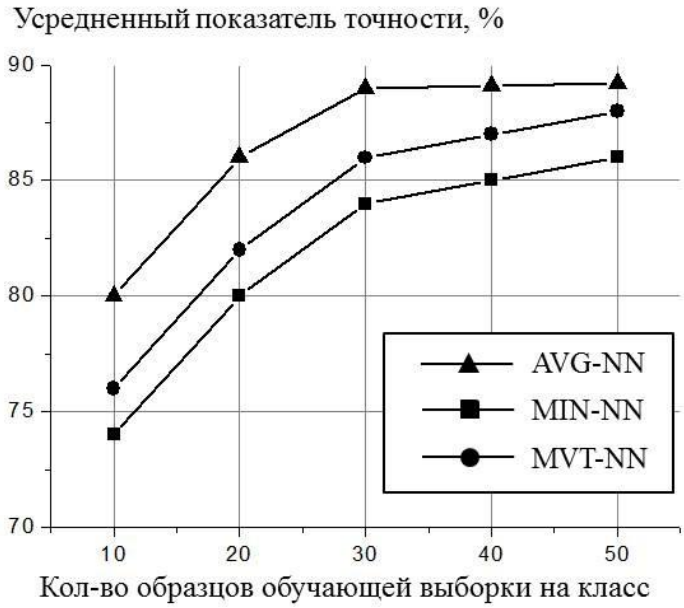

Рис. 4. **Зависимость точности корреляционного анализа от размера набора для обучения**

В свою очередь группа «MIN-NN» описывается через математическое выражение:

$$d_Q^{MIN} = \min_{x \in Q} (in_{xt \in w} \|\Delta x\|^2) \quad (12)$$

т.е. через минимально расстояние между *Q* и *w*. Существенно более сложный подход существует для определения группа «MVT-NN» [15]:

$$\begin{cases} \{w_{opt} = \arg\max_w (\sum_{x \in Q} v_w(x)) \\ v_w(x) = \begin{cases} 1 для w = w_{opt}^x \\ 0 для w \neq w_{opt}^x \end{cases} \\ w_{opt}^x = \arg\max_w (\min_{xt \in w} \|\Delta x\|^2) \end{cases} \quad (13)$$

Для каждой из групп был проведен анализ точности в зависимости от размера набора для обучения (10, 20…50 образцов). Результаты эксперимента приведены на рис. 4.

Все три группы показали высокую эффективность работы в условиях обучения на минимальном наборе. Точность разработанных алгоритмов превышает точность безкорреляционного анализа в 1,5-2 раза. Наилучшие показатели точности были получены для группа «AVG-NN».

Выводы

Проведенный в рамках данной работы анализ позволил выявить недостаткисовременных методы классификации и предложить методологию построения алгоритмов классификация

сетевого трафика на основе корреляционного анализа потоков сетевого трафика. В основу математического аппарата корреляционного анализа сетевого трафика были положены понятия набора пакетов, запроса, классов элементов, вероятности определения классов и величины расхождения между классами. Был разработан алгоритм построения NN классификатора, который включал в себя такие группы классификации как «AVG-NN», «MIN-NN» и «MVT-NN». Данные группы показали высокую точность классификации в условиях обучения на минимальном наборе учебных образцов.

Список использованных источников:

1. Alizadeh, H., &Zúquete, A. (2016). Traffic classification for managing Applications' networking profiles. Security and Communication Networks, 9(14), 2557-2575.
2. T.T. Nguyen and G. Armitage, "A Survey Of Techniques for Internet Traffic Classification Using Machine Learning," IEEE Comm. Surveys Tutorials, vol. 10, no. 4, pp. 56-76, Oct.-Dec. 2008.
3. H. Kim, K. Claffy, M. Fomenkov, D. Barman, M. Faloutsos, and K. Lee, "Internet Traffic Classification Demystified: Myths, Caveats, and the Best Practices,"Proc. ACM CoNEXTConf.,pp. 1-12, 2008.
4. Sharifi, H., Akbari, M. K., &Javadi, B. (2014). Performance modelling of adaptive routing communication networks in multi-cluster systems under bit-reversal traffic. International Journal of Communication Networks and Distributed Systems, 12(4), 442.
5. Goldman, A. (n.d.). Scalable Algorithms for Complete Exchange on Multi-Cluster Networks. 2nd IEEE/ACM International Symposium on Cluster Computing and the Grid (CCGRID02).
6. Y. Xiang, W. Zhou, and M. Guo, "Flexible Deterministic Packet Marking: An IP Traceback System to Find the Real Source of Attacks,"IEEE Trans. Parallel Distributed Systems, vol. 20, no. 4, pp. 567-580, Apr. 2009.
7. Ghofrani, F., Jamshidi, A., &Keshavarz-Haddad, A. (2015). Internet traffic classification using Hidden Naive Bayes model. 2015 23rd Iranian Conference on Electrical Engineering.
8. Tae, P. H., & Lee, B. (2016). Automated Construction Cost Estimation System Using DB Modeling of a TBM Construction Classification System. International Journal of Database Theory and Application, 9(7), 107-120.
9. L. Bernaille, R. Teixeira, I. Akodkenou, A. Soule, and K. Salamatian, "Traffic Classification on the Fly," Proc ACM SIGCOMM, vol. 36, pp. 23-26, Apr. 2006.
10. J. Erman, A. Mahanti, M. Arlitt, I. Cohen, and C. Williamson, "Offline/Realtime Traffic Classification Using Semi-Supervised Learning,"PerformanceEvaluation,vol. 64, nos. 9-12, pp. 1194-1213, Oct. 2007.
11. N. Williams, S. Zander, and G. Armitage, "A Preliminary Performance Comparison of Five Machine Learning Algorithms for Practical IP Traffic Flow Classification,"Proc ACM SIGCOMM, vol. 36, pp. 5-16, Oct. 2006.
12. T. Auld, A.W. Moore, and S.F. Gull, "Bayesian Neural Networks for Internet Traffic Classification," IEEE Trans. Neural Networks, vol. 18, no. 1, pp. 223-239, Jan. 2007.
13. Zhao, S., Zhang, Y., & Chang, P. (2017). Network Traffic Classification Using Tri-training Based on Statistical Flow Characteristics. 2017 IEEE Trustcom/BigDataSE/ICESS.
14. Casas, P., &Fiadino, P. (2013). Mini-IPC: A minimalist approach for HTTP traffic classification using IP addresses. 2013 9th International Wireless Communications and Mobile Computing Conference (IWCMC)..
15. Wang, W., Zhu, M., Zeng, X., Ye, X., & Sheng, Y. (2017). Malware traffic classification using convolutional neural network for representation learning. 2017 International Conference on Information Networking (ICOIN).

Стаття надійшла до редакції 11.10.2018