УДК 004.8 doi: 10.26102/2310-6018/2019.24.1.031

[1, 2]С.В. Пальмов, [1]А.А. Диязитдинова, [1]О.Ю. Губарева

ИССЛЕДОВАНИЕ ТЕЛЕКОММУНИКАЦИОННОГО ТРАФИКА СРЕДСТВАМИ АНАЛИТИЧЕСКОЙ СИСТЕМЫ ORANGE

[1]*ФГБОУ ВО «Поволжский государственный университет телекоммуникаций и информатики», Самара, Россия*
[2] *ФГБОУ ВО «Самарский государственный технический университет», Самара, Россия*

Упростить задачу обеспечения информационной безопасности можно за счет применения технологии интеллектуального анализа данных. Данная технология может быть использована для прогнозирования атак на информационную систему. Дерево решений – один из эффективных инструментов построения прогностических моделей. Orange – аналитическая система, в которой содержится большое число алгоритмов интеллектуального анализа данных, включая дерево решений. Средствами указанной системы выполнен анализ реальных данных о сетевых атаках, полученных в ходе экспериментального исследования, с целью прогнозирования DDoS-атак. Для оценки качества работы использовались пять метрик: правильность, специфичность, точность, полнота и F-мера. Итоги проведенного анализа представлены в табличном виде. Полученные результаты были сравнены с прогнозами, созданными iWizard-E – интеллектуальной системой поддержки принятия решений, использующей модифицированный алгоритм дерева решений. iWizard-E превосходит Orange по первым трем метрикам, но уступает по последним двум. Реализации указанного алгоритма в системах Orange и iWizard-E не могут быть применены для проведения анализа данных вышеприведенного вида, поскольку формируют прогнозы, обладающие низкой достоверностью. Необходимо провести усовершенствование дерева решений, направленное на повышение качества генерируемых прогностических моделей в разрезе увеличения значений метрики «полнота».

Ключевые слова: искусственный интеллект, интеллектуальный анализ данных, система Orange, принятие решений, трафик, *F*-мера.

Введение

Задача обеспечения информационной безопасности информационных телекоммуникационных систем может реализовываться различными методами и средствами [1]. Одним из подходов является применение искусственного интеллекта, в частности, технологии интеллектуального анализа данных (*Data Mining*) [2, 3].

В [4] авторы продемонстрировали возможности указанной технологии, на примере интеллектуальной системы поддержки принятия решений (ИСППР) «*iWizard-E*» [5], в разрезе прогнозирования *DDoS*-атак [6]. Вторая задача упомянутого исследования состояла в оценке аналитических способностей ИСППР. В итоге, был сделан вывод о невысокой эффективности данного программного обеспечения и необходимости его

модернизации. Однако авторы решили прежде изучить возможности сторонних реализаций, построенных на принципах, схожих с *iWizard-E*, а именно использующих для формирования прогностических моделей алгоритм дерева решений [7]. Это позволит упростить формулировку стратегии, направленной на совершенствование ИСППР.

В качестве стороннего продукта задействована аналитическая система *Orange* [8]. В ней присутствуют три критерия оценки качества разбиения (*Gini Index, Information Gain* и *Gain Ratio*) из четырех, содержащихся в *iWizard-E*, что и повлияло на выбор авторов статьи; в свободном доступе программное обеспечение с реализованным критерием *Entropy* найти не удалось [9].

Материалы и методы

С целью получения реальной статистики, касающейся сетевых атак, был выполнен эксперимент, в ходе которого с помощью системы мониторинга Zabbix был осуществлен захват и фиксация дюжины трафиков при нормальном устойчивом состоянии и двух трафиков во время проведения сетевой атаки. Исследование выполнено на сегменте сети оператора связи. Анализировались данные с Интернет-ресурса *https://top.mail.ru/Rating/* (Рейтинг@*mail.ru*). Трафик детально смоделирован посредством сервера оператора, на который в дальнейшем была произведена *DDoS*-атака, с помощью генерации паразитного трафика с нескольких устройств.

В итоге все данные (83347 записи) были приведены к следующему виду (генеральная совокупность):

Таблица 1 – Структура файла, содержащего генеральную совокупность

Название параметра	Описание параметра
time	время фиксации значений *visitors*, *input* и *output*
visitors	число пользователей в сети в заданный момент времени, шт.
input	входящий трафик, Мбайт
output	исходящий трафик, Мбайт
alarm	наличие «скачка» входящего трафика (да/нет)

Свидетельством наличия *DDoS*-атаки является скачок входящего трафика. Скачок есть резкое увеличение (> 50%) входящего трафика.

Цель исследования сформулирована следующим образом:

1) Средствами системы *Orange* на основе значений атрибутов *visitors, input* и *output* в произвольно выбранный момент времени *t*, сформировать прогноз относительно вероятности скачка входящего трафика, который определяется целевым атрибутом «*alarm*» (рассчитан авторами).

2) Провести сравнительный анализ результатов, полученных при помощи *iWizard-E* и *Orange*.

Результаты

Для более корректной реализации второй задачи, была проведена серия экспериментов в формате, описанном в [4, С.117], однако, под воздействием причин объективного характера были внесены следующие изменения:

- используется три критерия разбиения вместо четырех (см. выше);
- эксперимент №6 не проводился, поскольку в Orange нет возможности задавать пороговое значение критерия разбиения, при превышении которого будет выполнено разбиение узла.

Также с целью упрощения сравнительного анализа, Таблица, содержащая значения метрик, которые использовались для оценки эффективности работы аналитической системы, имеет структуру, аналогичную той, что применялась в случае *iWizard-E*. Вышеупомянутые изменения, внесенные в структуру экспериментов, явно отражены в Таблице (см. Таблицы 1 и 2).

Таким образом, было проведено 8 экспериментов, каждый из которых состоял из трех частей, за исключением эксперимента №2. Для всех экспериментов средствами *Orange* были сгенерированы случайные стратифицированные выборки на основе генеральной совокупности. Для эксперимента №2 в *Orange* была произведена дискретизация всех атрибутов из Таблицы 1, кроме *time* и *alarm*. Использовался следующий тип дискретизации: *Equal-Frequency Discretization* (число интервалов - 10).

На основании выборок выполнялось построение прогностических моделей, а их тестирование осуществлялось с использованием генеральной совокупности (83347 записи).

Таблица 2 – Эксперименты

Номер эксперимента	Настройки эксперимента
1	Р = 100, О = 2500
2	О = 2500, дискретизация атрибутов = да
3	Р = 75, О = 2500
4	Р = 50, О = 2500
5	Р = 25, О = 2500
6	Р = 100, О = 2500, КР> 0.001 **[не реализовано в *Orange*]**
7	О = 2000
8	О = 1500
9	О = 1000

В Таблице 2: Р (разбиение) – это максимальное количество разбиений узлов дерева решений, при достижении которого его построение завершается; О (объем) – число записей в выборке; «КР» – значение критерия разбиения, при котором узел может быть разбит.

Качество прогностических моделей оценивалось посредством использования нижеприведенных метрик:

Правильность (*accuracy, classification accuracy*) *Пр = П / О* – описывает способность модели правильно распознавать записи в итоговом файле; *П* – число правильно распознанных записей; *О* – общее количество записей в генеральной совокупности.

Специфичность (*specificity*) *Сп = ИО / ЛП* характеризует способность прогностической модели правильно распознавать ситуации, когда скачок будет отсутствовать; *ИО* – истинно-отрицательный результат, *ЛП* – ложноположительный результат.

Точность (*precision*) *Тн = ИП / (ИП + ЛП)* - это доля записей, действительно характеризующихся наличием скачка относительно всех записей, которые были отнесены моделью к этому классу; *ИП* – истинно-положительный результат.

Таблица 3 – Результаты экспериментов

Эксперимент	*Пр*	*Сп*	*Тн*	*Пл*	*F-мера*
Э1Ч1	0,8940	0,9438	0,1355	0,1343	0,1349
Э1Ч2	не реализовано в *Orange*				
Э1Ч3	0,8956	0,9462	0,1318	0,1246	0,1281
Э1Ч4	0,9096	0,9643	0,1230	0,0764	0,0943
Э2Ч1	0,9385	1	0	0	0
Э3Ч1	0,8887	0,9399	0,1051	0,1076	0,1063
Э3Ч2	не реализовано в *Orange*				
Э3Ч3	0,8863	0,9369	0,1064	0,1146	0,1104
Э3Ч4	0,9071	0,9618	0,1104	0,0723	0,0874
Э4Ч1	0,8887	0,9399	0,1051	0,1076	0,1063
Э4Ч2	не реализовано в *Orange*				
Э4Ч3	0,8886	0,9380	0,1259	0,1363	0,1309
Э4Ч4	0,9162	0,9715	0,1455	0,0741	0,0982
Э5Ч1	0,8912	0,9403	0,1355	0,1427	0,1390
Э5Ч2	не реализовано в *Orange*				
Э5Ч3	0,8886	0,9380	0,1259	0,1363	0,1309
Э5Ч4	0,9262	0,9834	0,1766	0,0542	0,0829
Э6Ч1	не реализовано в *Orange*				
Э6Ч2					
Э6Ч3					
Э6Ч4					
Э7Ч1	0,8916	0,9423	0,1191	0,1189	0,1190
Э7Ч2	не реализовано в Orange				
Э7Ч3	0,8937	0,9446	0,1222	0,1175	0,1198
Э7Ч4	0,8963	0,9483	0,1159	0,1033	0,1092
Э8Ч1	0,8927	0,9442	0,1125	0,1078	0,1101
Э8Ч2	не реализовано в *Orange*				
Э8Ч3	0,8936	0,9460	0,1021	0,0936	0,0976
Э8Ч4	0,8919	0,9416	0,1313	0,1347	0,1330
Э9Ч1	0,8832	0,9347	0,0896	0,0981	0,0936
Э9Ч2	не реализовано в *Orange*				
Э9Ч3	0,8879	0,9399	0,0947	0,0959	0,0953
Э9Ч4	0,8859	0,9376	0,0927	0,0973	0,0949

Полнота (*Recall*) $Пл = ИП / (ИП + ЛО)$ - характеризует способность классификатора определять как можно большее число положительных ответов из ожидаемых; *ЛО* – ложноотрицательный результат.

F-мера $= 2 * ((Тн * Пл) / (Тн + Пл))$ - является гармоническим средним между точностью и полнотой.

Обсуждение

Наиболее качественные результаты, сформированные интеллектуальной системой поддержки принятия решений «iWizard-E», представлены в Таблице 4.

Результат (прогноз) будет обладать максимальным качеством, если все его метрики будут равны единице.

Следует отметить, что лучшие результаты по метрикам Пр и Сп достигнуты в Э2Ч1, однако значения прочих метрик для указанного случая равны нулю и, поэтому, они были исключены из рассмотрения.

Таблица 4 – Лучшие результаты по каждой из метрик

	Метрики				
	Пр	*Сп*	*Тн*	*Пл*	*F-мера*
№№ экспериментов и их части	Э5Ч4	Э5Ч4	Э5Ч4	Э5Ч1	Э5Ч1
Значения метрик	0,9262	0,9834	0,1766	0,1427	0,1390

Согласно [10] F-мера представляется наиболее предпочтительной метрикой при оценке прогностической модели. Поэтому, более качественной будем считать ту модель, которая была построена в Э5Ч1.

Изучив данные, представленные в Таблице 3, можно сделать вывод, что Orange продемонстрировал в целом высокую эффективность при прогнозировании скачков входящего трафика. Тем не менее, процент истинно-положительных результатов по сравнению с истинно-отрицательными очень мал.

Таким образом, алгоритм деревьев решений, реализованный в *Orange*, не может быть использован для анализа телекоммуникационного трафика вышеуказанного формата.

Заключение

Сравним полученные результаты с [4] (см. Таблицу 5).

Таблица 5 - Лучшие результаты по каждой из метрик для *Orange* и *iWizard-E*

Название системы	Метрики и их значения				
	Пр	*Сп*	*Тн*	*Пл*	*F-мера*
Orange	Э5Ч4	Э5Ч4	Э5Ч4	Э5Ч1	Э5Ч1
	0,9262	0,9834	0,1766	0,1427	0,1390
iWizard-E	Э6Ч1	Э6Ч1	Э7Ч1, Э7Ч2, Э7Ч3	Э7Ч4	Э7Ч4
	0,9381	0,9996	0,1845	0,0561	0,0866

Как видно из представленных данных, *iWizard-E* превосходит *Orange* по значениям метрик *Пр*, *Сп* и *Тн*, но заметно уступает по *Пл* и, как следствие, по *F*-мера. Следовательно, первая из упомянутых систем способна корректно распознать большее число ситуаций, чем вторая. Однако, в основном, это будут ситуации, характеризующиеся отсутствием скачка трафика (высокое значение метрики *Сп*).

Подводя итог, можно сказать, что *Orange* превосходит *iWizard-E*, но, тем не менее, обе системы формируют малодостоверные прогнозы, что исключает их использование для решения вышеуказанной задачи. Дальнейшие же усилия авторов будут направлены на повышение эффективности работы *iWizard-E* в разрезе увеличения значения метрики *Пл*.

По нашему мнению, также представляется интересным протестировать обе системы на этом же наборе данных, но с удаленными записями-дубликатами. Результаты такого исследования будут представлены в следующей статье.

ЛИТЕРАТУРА

1. Информационная безопасность предприятия: ключевые угрозы и средства защиты [Электронный ресурс]. – Режим доступа : https://www.kp.ru/guide/informatsionnaja-bezopasnost-predprijatija.html. (дата обращения: 01.02.2019)
2. Mutyala, Nikhil Kumar & Koushik, K.V.s & Sundar, K. John. (2018). Data Mining and Machine Learning Techniques for Cyber Security Intrusion Detection // International Journal of Scientific Research in Computer Science, Engineering and Information Technology. 2018. Vol. 3. No. 3. – Pp. 162 – 167. DOI 10.13140/RG.2.2.35197.26085.
3. Gunderman, D. Security Analysts Becoming ‘Data-Mining Gurus’? Q&A With Bay Dynamics’ Ryan Stolte [Электронный ресурс]. – Режим доступа: https://www.cshub.com/attacks/interviews/security-analysts-becoming-data-mining-gurus-qa-with-bay-dynamics-ryan-stolte. (дата обращения: 01.02.2019)
4. Пальмов, С.В. Анализ телекоммуникационного трафика с помощью интеллектуальной системы поддержки принятия решений / С.В. Пальмов, А.А. Мифтахова, О.Ю. Губарева // Наука и бизнес: пути развития. – М.: ТМБпринт. – 2018. – №8(86). – С. 116–122.
5. Мифтахова, А.А. Использование методов искусственного интеллекта для повышения успеваемости студентов вузов / А.А. Мифтахова // Наука и бизнес: пути развития. – М. : ТМБпринт. – 2017. – №5(71). – С. 7–12.

6. DoS и DDoS-атаки: значение и различия [Электронный ресурс]. – Режим доступа: https://ddos-guard.net/ru/info/blog-detail/dos-i-ddos-ataki-znachenie-i-razlichiya. (дата обращения: 01.02.2019)
7. Sharma, Himani & Kumar, Sunil. (2016). A Survey on Decision Tree Algorithms of Classification in Data Mining. International Journal of Science and Research // International Journal of Science and Research. 2016. Vol. 5. No. 4. – Pp. 2094 – 2097.
8. Data Mining Fruitful and Fun [Электронный ресурс]. – Режим доступа: https://orange.biolab.si/.(дата обращения: 01.02.2019)
9. Информационная энтропия [Электронный ресурс]. − Режим доступа: http://ru.math.wikia.com/wiki/Информационная_энтропия. (дата обращения: 01.02.2019)
10. Баженов, Д. Оценка классификатора (точность, полнота, F-мера) [Электронный ресурс]. – Режим доступа: http://bazhenov.me/blog/2012/07/21/classification-performance-evaluation.html. (дата обращения: 01.02.2019)

S.V. Palmov[1,2], A.A. Diyazitdinova[1], O.Y. Gubareva[1]

TELECOMMUNICATION TRAFFIC ANALYSIS USING ORANGE ANALYTICAL SYSTEM

[1]Povolzhskiy State University of Telecommunications and Informatics, Samara, Russia
[2]Samara State Technical University, Samara, Russia

To simplify the task of ensuring information security is possible through data mining usage. This technology can be used to predict attacks on the information systems. Decision tree is one of the effective tools for predictive models building. Orange is an analytical system that contains a large number of data mining algorithms, including a decision tree. With help of the system made an analysis of real data on network attacks obtained during the experimental study, with the aim of predicting DDoS attacks. Five metrics were used to assess the quality of work: accuracy, specificity, precision, recall and F-measure. The results of the analysis are presented in tabular form. The results were compared with the forecasts created by iWizard-E, an intelligent decision support system using a modified decision tree algorithm. iWizard-E surpasses Orange in the first three metrics, but inferior in the last two. The implementation of this algorithm in the Orange and iWizard-E systems cannot be applied to analyze the data of the above type, since they form forecasts with low reliability. It is necessary to improve the decision tree aimed at improving the quality of the generated prognostic models in the context of increasing the values of the "completeness" metric.

Keywords: artificial intelligence, data mining, Orange system, decision making, traffic, F-measure.

REFERENCES

1. Enterprise Information Security: Key Threats and Remedies. Available at: http://www.iccwbo.ru/blog/2016/obespechenie-informatsionnoy-bezopasnosti/ (accessed 01.02.2019). (In Russ.)
2. Mutyala, Nikhil Kumar & Koushik, K.V.s & Sundar, K. John. (2018). Data Mining and Machine Learning Techniques for Cyber Security Intrusion Detection // International Journal of Scientific Research in Computer Science, Engineering and Information Technology. 2018. Vol. 3. No. 3. – Pp. 162 – 167. DOI 10.13140/RG.2.2.35197.26085.
3. Gunderman, D. Security Analysts Becoming 'Data-Mining Gurus'? Q&A With Bay Dynamics' Ryan Stolte. Available at: https://www.cshub.com/attacks/interviews/security-analysts-becoming-data-mining-gurus-qa-with-bay-dynamics-ryan-stolte (accessed 01.02.2019)
4. Palmov, S.V., Miftakhova A.A., Gubareva O. Yu. The Analysis of Telecommunication Traffic by Intelligent Decision Support System // Nauka i biznes: puti razvitiya = Science and Business: Development Ways. 2018. №8 (86). – Pp. 116–122. (In Russ.)
5. Miftakhova, A.A. [Artificial Intelligence for Improving Students' Performance] // Nauka i biznes: puti razvitiya = Science and Business: Development Ways. 2017. №5 (71). – Pp. 7–12. (In Russ.)
6. DoS and DDoS attacks: meaning and differences. Available at: https://ddos-guard.net/ru/info/blog-detail/dos-i-ddos-ataki-znachenie-i-razlichiya (accessed 01.02.2019) (In Russ.)
7. Sharma, Himani & Kumar, Sunil. (2016). A Survey on Decision Tree Algorithms of Classification in Data Mining. International Journal of Science and Research // International Journal of Science and Research. 2016. Vol. 5. No. 4. – Pp. 2094 – 2097.
8. Data Mining Fruitful and Fun. Available at: https://orange.biolab.si/ (accessed 01.02.2019).
9. Informational Entropy. Available at: http://ru.math.wikia.com/wiki/Informatsionnaya_entropiya (accessed 01.02.2019). (In Russ.)
10. Bazhenov, D [Classifier rating (Precision, Recall, F-measure)]. Available at: http://bazhenov.me/blog/2012/07/21/classification-performance-evaluation.html (accessed 01.02.2019). (In Russ.)