TRAFFIC INFORMATION ANALYSIS USING DEEP LEARNING ARTIFICIAL NEURAL NETWORKS

Ivelina Stefanova Balabanova, Teodora Valentinova Zhorova

Technical University of Gabrovo,
Department of Communications Equipment and Technologies,
Gabrovo, Bulgaria,
teddy.tedun@gmail.com

Abstract

The paper systematizes the results of a study of multilayer artificial neural networks for the analysis and identification of areas of Internet traffic data content. The object of the research is traffic data of corporate company customers in segmented time zones and geographical areas. The systematic procedures with deep learning techniques on backpropagation neural networks have been carried out in MATLAB software. Selected quality indicators from learning processes of multilayer structures at different applied ratios between neurons in hidden layers are analyzed. The network performance, correlation metrics and classification quality have were evaluated. High levels of accuracy in a correct recognition of the data inthe information sets have been established about the defined classification groups with maximum limitation of the Mean-Square Error, the Mean Absolute Error and etc.

Keywords: Internet traffic; multilayer neural networks; deep learning; accuracy; mean-squared error.

1. ВЪВЕДЕНИЕ

Анализът и параметричната оценка на Интернет трафика са важна задача, свързана с подобряване на качеството на обслужване на частни и корпоративни потребители по отношение на различни функции като:

- класификация и прогнозиране на мрежовия трафик;
- трафична маршрутизация и мрежова сигурност;
- контрол на грешките и претоварването на информационните трасета;
- QoS (Quality of Service) и QoE (Quality of Experience) управление [1].

Един от най-често използваните изчислителни апарати за целта се явяват изкуствените невронни мрежи. Налице е голямо разнообразие от научни изследвания, засягащи този аспект на управление в ИКТ системите.

Широко използвани за различни сценарии са комбинирани подходи с включване на конволюционни и рекурентни невронни мрежи [2, 3]. Други инструменти, използващи извлечени статистически трафични характеристики, са Naïve Bayes алгоритъм, метод на опорните вектори (Support Vector Machine), линейни и квадратични дискриминантни класификатори в [4, 5]. В огромната си част съществуващите проучвания засягат детектиране и установяване на типа на аномалии и нарушения в целостта на предаваната информация [6-8].

В настоящия доклад е предложен подход за идентификация на зони на потребление на Интернет съдържание чрез многослойни невронни мрежи с обратно разпространение на грешката с дълбоко обучение на базата на синтез на невронни структури при различно съотношение между изчислителните единици в скритите слоеве.

2. ПОСТАНОВКА НА ЗАДАЧАТА ЗА ИДЕНТИФИКАЦИЯ НА ЗОНИ НА ИНТЕРНЕТ ПОТРЕБЛЕНИЕ

Поставена е задачата за диагностика и идентификация на трафика чрез многослойни изкуствени невронни мрежи относно ин-

формационната извадка, обхващаща 78 WEB корпоративни компании клиенти за гр. Чикаго, САЩ. Обучаващото множество включва следните входни трафични променливи:

- Променлива №1: Flows, flows/s (kilo-);
- Променлива №2: IPv4 - Packet size, Mean;
- Променлива №3: IPv6 - Packet size, Mean;
- Променлива №4: Mean Transition Rate, pkts/s (kilo-), спрямо дефинирани класификационни групи:
- Клас №1: Chicago city area "1";
- Клас №2 или Class №2: Chicago city area "2".

В хода на проведен предварителен синтез бяха използвани класически изкуствени невронни мрежи с право разпространение на сигналите и обратно разпространения на грешката (Feed-Forward Neural Networks - FFNNs) при използване на Scaled Conjugate Gradient алгоритъм на обучение. Бяха изследвани невронни архитектури с приложена тангенс-сигмоидална и softmax функции на активация в междинния и изходния слой при три случая на разделяне на информационния набор за процесите на обучение, валидиране и тестване – 70:15:15, 60:20:20 и 50:25:25. Като най-подходящо беше установено третото зададено съотношение между данните в информационния масив, за което бяха наблюдавани удовлетворяващи нива на точността над 90.0 %.

С оглед на търсене на по-ефективни невронни архитектури се премина към процедури по увеличаване на броя на скритите слоеве и прилагане на принципите на „дълбокото обучение". За конкретния случай бяха поставени опции за вариране на съотношението между изчислителните единици в два структурни скрити слоя, както следва:

- Съотношение №1: „Едно към едно";
- Съотношение №2: „Едно към две";
- Съотношение №3: „Две към едно".

Освен „*точността*" за разлика от предходните случай при трислойните невронни архитектури бяха въведени допълнителни индикатори за качество при невронен синтез, респективно:

- „Средноквадратична грешка" (Mean Squared Error - MSE);
- „Средна абсолютна грешка" (Mean Absolute Error).

Относно задачата за обучение, оценка и синтез на многослойни FFNNs за идентификация на географски райони на Интернет потребление от клиентски корпоративни компании в гр. Чикаго (САЩ) бяха реализирани изследвания при фиксирано разпределение на данните от входния информационен набор, съответно 70 % за обучаващите, 15 % за валидиращите и 15 % за тестовите процедури. Проведените експериментални процедури по „дълбоко обучение" бяха базирани на алгоритъма на Levenberg-Marquardt. Използвани са базисни невронни архитектури при фиксиране на:

- Тангенс-сигмоидална активационна функция в първи скрит слой;
- Логаритмичен-сигмоидален активационен тип във втори междинен слой;
- Тангенс-сигмоидална активационна функция в изходния мрежови слой.

3. СИНТЕЗ НА МНОГОСЛОЙНИ НЕВРОННИ АРХИТЕКТУРИ ЗА РАЗПОЗНАВАНЕ НА ТРАФИЧНИ ЗОНИ НА ПОТРЕБЛЕНИЕ

Резултатите от проучване на вариациите на избраните критерии „точност", „MSE" и „MAE" са поместени от таблица 1 до таблица 3. Във връзка с първото приложено съотношение между невроните в първи и втори скрит слой са регистрирани минимално 56.5 % и максимално показание за точността 100.0 % за модели с 10 и 7 до 9 изчислителни междинни единици.

Таблица 1. *Изследване на FFNNs с дълбоко обучение за идентификация на зони на потребление на Интернет съдържание при съотношение между невроните в скритите слоеве 1:1*

№	Скрит слой №1	Скрит слой №2	Точност, %	MSE	MAE
1.	3	3	95.7	0.0397	0.1007
2.	4	4	78.3	0.1134	0.1821
3.	5	5	65.2	0.1855	0.2070
4.	6	6	95.7	0.0393	0.0627
5.	7	7	100.0	0.0151	0.0436
6.	**8**	**8**	**100.0**	**0.0017**	**0.0200**
7.	9	9	100.0	0.0222	0.1061
8.	10	10	56.5	0.2608	0.2684
9.	11	11	95.7	0.0463	0.0669

Таблица 2. *Изследване на FFNNs с дълбоко обучение за идентификация на зони на потребление на Интернет съдържание при съотношение между невроните в скритите слоеве 1:2*

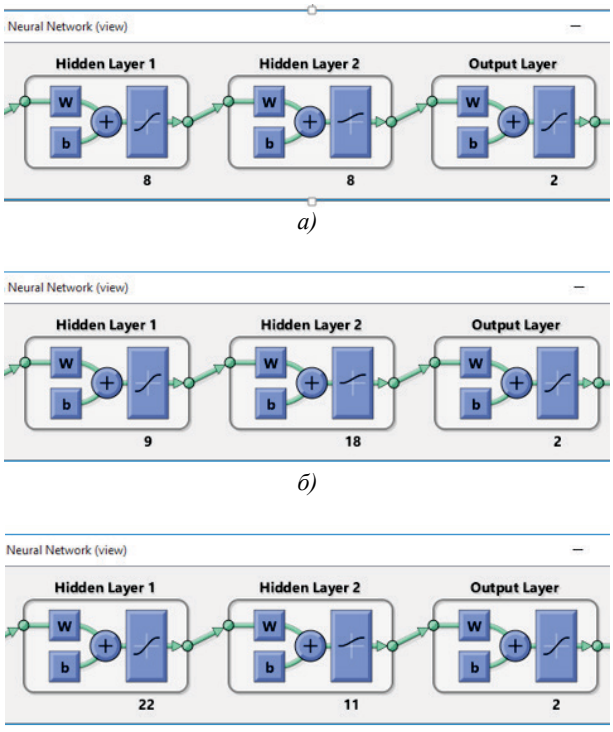
№	Скрит слой №1	Скрит слой №2	Точност, %	MSE	MAE
1.	3	6	95.7	0.0531	0.0609
2.	4	8	95.7	0.0451	0.0641
3.	5	10	47.8	0.2681	0.2848
4.	6	12	95.7	0.0558	0.0900
5.	7	14	87.0	0.1142	0.1811
6.	8	16	43.5	0.4134	0.4874
7.	**9**	**18**	**100.0**	**0.0263**	**0.1145**
8.	10	20	95.7	0.0244	0.0616
9.	11	22	95.7	0.0317	0.1211

Таблица 3. *Изследване на FFNNs с дълбоко обучение за идентификация на зони на потребление на Интернет съдържание при съотношение между невроните в скритите слоеве 2:1*

№	Скрит слой №1	Скрит слой №2	Точност, %	MSE	MAE
1.	6	3	52.2	0.2394	0.2468
2.	8	4	87.0	0.0774	0.1402
3.	10	5	39.1	0.3102	0.3378
4.	12	6	95.7	0.0327	0.0792
5.	14	7	91.3	0.0859	0.1137
6.	16	8	95.7	0.0506	0.1303
7.	18	9	91.3	0.0671	0.1124
8.	20	10	82.6	0.1600	0.2364
9.	**22**	**11**	**100.0**	**0.0073**	**0.0345**

При средноквадратичната грешка бяха констатирани най-ниска 0.0017 и най-висока стойност 0.2628, респективно при зададени 8 и 10 скрити неврона. Намерените минимална и максимална MAE се равняват на 0.0200 и 0.2684, установени при същите FFNNs относно предходния индикатор за качество. Съобразно изложените резултати е избрана невронна архитектура с наличие на 8 невронна в структурните междинни слоеве.

По отношение на второто разгледано съотношение са намерени две архитектури, при които е получена точност под ниво от 50.0 %, както следва 43.5 % при 8 в първи и 16 неврона във втори скрит слой и 47.8 % за FFNN с 5 в първи и 10 скрити неврона във втори слой. Констатирано е еднократно достигане на пълно коректно разпознаване на обработваните информационни трафични еталони – точност 100.0 %, при структура с 9 и 18 неврона в първия и втория междинен слой. В рамките на описания случай бяха регистрирани и минимални MSE = 0.0263 и MAE = 0.1145. Максимални нива на средноквадратичната грешка 0.4134 и средната абсолютна грешка 0.4874 са наблюдавани за невронна архитектура при „8 в първия и 16 неврона във втория междинен слой“. Анализът на резултатите показва предимство на FFNN със съдържание на девет в първи и осемнадесет междинни неврона във втори скрит слой. Но имайки предвид в пъти по-ниските степени на MSE и MAE при първия разгледан случай в процедурния синтез, по-подходяща се оказва неговата употреба.



а)

б)

в)

Фиг. 1. *Селектирани FFNNs за идентификация на географски зони на корпоративно Интернет потребление при съотношения между невроните в скритите слоеве а) „1:1“, б) „1:2“ и в) „2:1“*

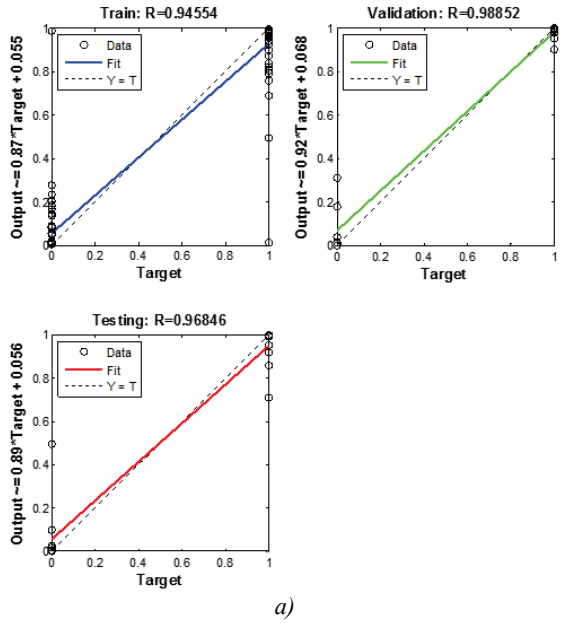
Съгласно последното приложено количествено съотношение между невроните в скритите слоеве отново е установена една комбинация, при която е постигната коректна идентификация на целевите зони на корпоративно Интернет потребление. Това е FFNN с наличие на 22 и 11 неврона, респективно в първи и втори скрит слой. По

отношение на указаната невронна структура бяха регистрирани най-ниски нива на MSE = 0.0073 и MAE = 0.0345. Най-възходящи вариации на грешките MSE = 0.3102 и MAE = 0.3378 са намерени в архитектура, при която са използвани 10 изчислителни единици за първия и 5 междинни неврона за втория скрит слой. От разгледните случай на невронен синтез тук е констатирана най-ниска класификационна точност, попадаща в порядъка 39.1 %.

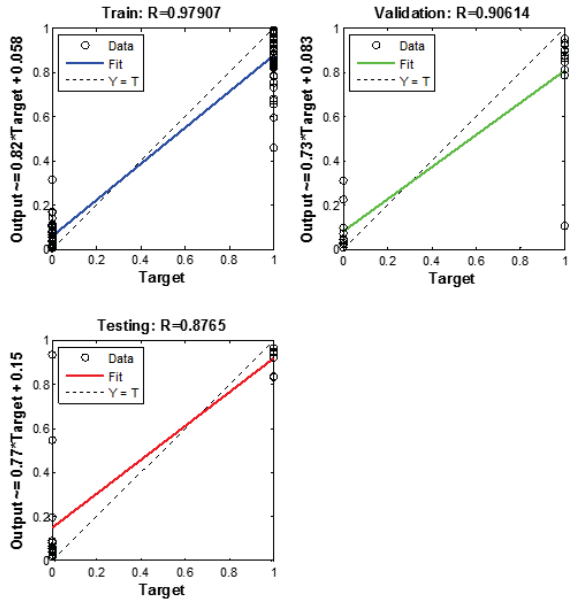
Сравнявайки нивата на грешките, може да се каже, че FFNNs при съотношение „2:1“ притежават по-добри качества спрямо тези, изградени на основата на съотношение „1:2“. Давайки оценка на синтезираните крайни модели на идентификация на трафични зони, дадени на фиг. 1 - архитектурата с идентично количество невронни изчислителни единици в междинните слоеве се определя като модел с най-висока мрежова производителност.

4. АНАЛИЗ НА ФУНЦИОНАЛНОСТТА НА МНОГОСЛОЙНИТЕ НЕВРОННИ МРЕЖИ ЗА ИДЕНТИФИКАЦИЯ ТРАФИЧНИ ЗОНИ НА ПОТРЕБЛЕНИЕ

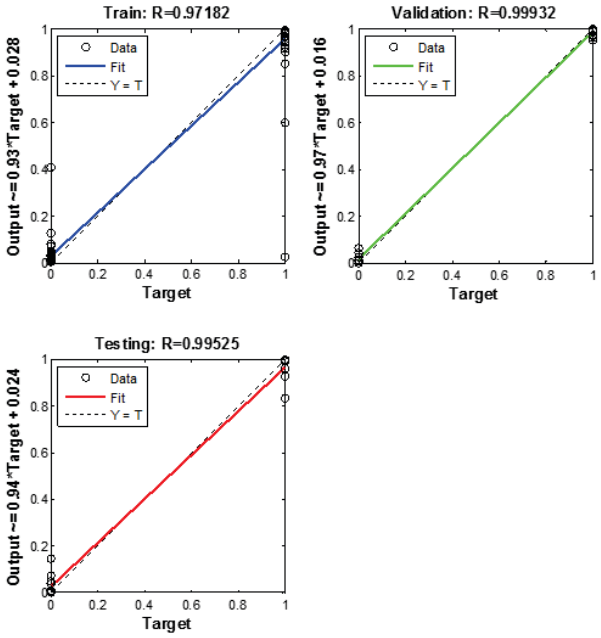
Фигура 2 представя линейните регресионни зависимости за базисните мрежови процеси на селектираните многослойни FFNNs. При идентичен брой на невроните в междинните слоеве се наблюдава известно отклонение между теоретичните и емпиричните линии на регресия за посочените случаи. Констатирани са високи стойности на корелационните коефициенти над нива „0.94“, „0.98“ и „0.96“ при обучаващите, валидиращите и тестовите процедури.



а)



б)

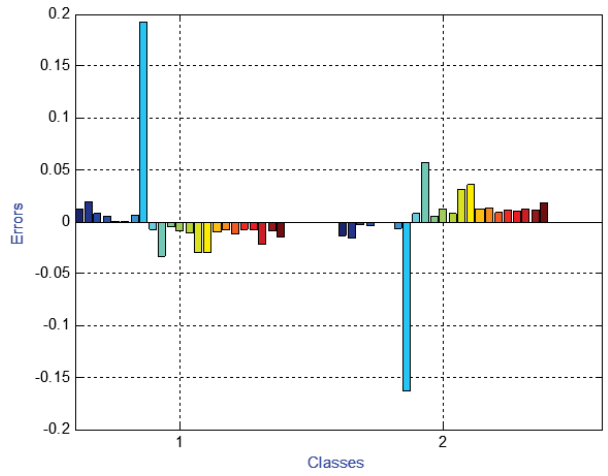


в)

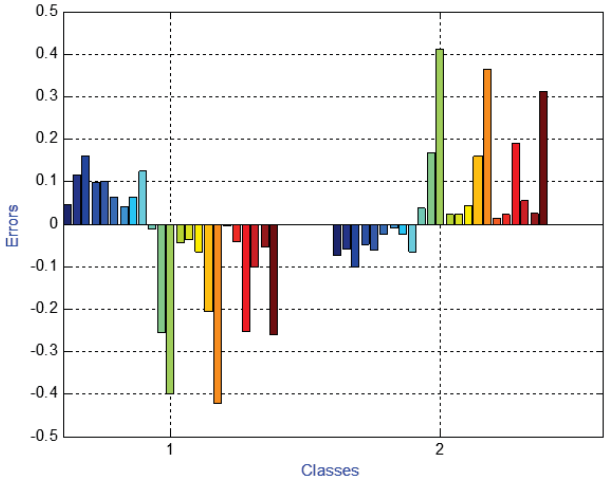
Фиг. 2. *Линии на регресия за селектираните FFNNs за идентификация на географски зони на Интернет потребление при съотношения между невроните в скритите слоеве а) „1:1“, б) „1:2“ и в) „2:1“*

Разглеждайки зависимостите при второто приложено съотношение се виждат по-ниските степени на корелация при валидиране, тестване и изходите за синтезираната многослойна изкуствена невронна мрежа. Като при тестовите процедури R попада дори под изискваното минимално ниво от „0.9“ – тук R = 0.8765. Изключение е регистрирано при обучаващия процес, където R се равнява на „0.97907“. Забелязва се по-из-

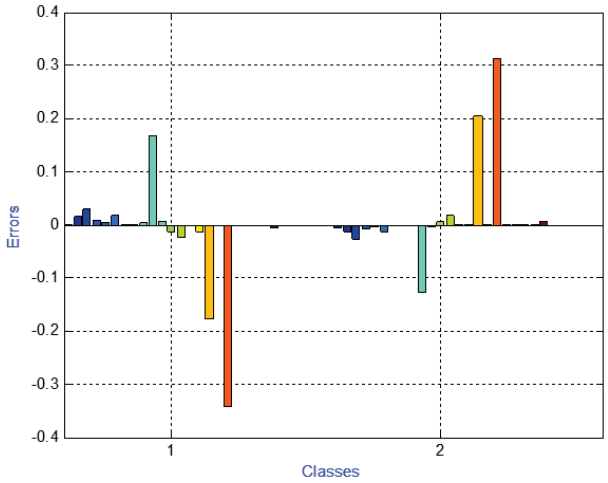
разеното отклонение между теоретичните и опитните линии на регресия.



а)



б)



в)

Фиг. 3. *Диаграми на грешките за селектираните FFNNs за идентификация при съотношения между невроните в скритите слоеве а) „1:1“, б) „1:2“ и в) „2:1“*

Във връзка с последната разгледана многослойна невронна архитектура са регистрирани най-добри предимства по отношение на корелационните коефициенти спрямо предходните мрежи, респективно R = 0.97182 от обучение, R = 0.99932 при валидация и R = 0.99525 за тестовия процес. Забелязва се по-доброто близко разположение на емпиричните спрямо идеалните линии на регресия. Имайки предвид изложените показатели от проведения линеен регресионен анализ може да се каже, че при FFNN с приложено съотношение „2:1“ се наблюдават изразени предимства в сравнение с предходните две категории мрежови модели.

Диаграмите на мрежовите грешки, показани на фиг. 3, представляват вариациите на разликите между теоретично заложените и калкулираните посредством синтезираните многослойни модели резултати при манипулации с данните, включени в състава на тестовия поднабор. По дефиниция при пълно коректно разпознаване на тестовите еталони получените грешки следва да попадат в границите от „-0.5“ до „0.5“, каквито са регистрираните за конкретните случаи.

При анализираните многослойни невронни модели за идентификация на зони на корпоративно Интернет потребление са установени следните вариационни диапазони:

- -0.1627 до 0.1928 при FFNN с приложено съотношение между невроните в междинните слоеве “1:1”;
- -0.4241 до 0.4119 относно модела при задаване на 9 в първия и 18 изчислителни единици във втория скрит слой;
- -0.3428 до 0.3130 за многослойната невронна архитектура с използвано съотношение между невроните в скритите слоеве „2:1“.

Според оценка на констатираните вариации на мрежовите грешки се откроява предимството на модела при идентично количество неврони в двата междинни слоя, следвано от невронни архитектури с приложено съотношение „1:2“ и „2:1“. Представените грешки потвърждават предимствата на първата алтернатива за разпределение на невронните изчислителни единици относно разглежданата задача за идентификация.

5. ЗАКЛЮЧЕНИЕ

Представените резултати показват много добра приложимост на невронните мрежи с дълбоко обучение за разпознаване на географски зони на потребление на Интернет съдържание. Следваща фаза от изследванията е насочена към анализ на функционалността на вероятностни и рекурентни невронни мрежи, както и известни методи и алгоритми на машинното обучение. Някои от тях са k – най-близки съседи, дърво на решенията и т.н. Предвижда се обобщените резултати да бъдат използвани при систематизация на методика за идентификация, прогнозен анализ, оценка на тенденциите, установяване на аномалии и други по отношение на зони на потребление на мрежовия трафик на основата на изкуствен интелект, машинно обучение и приложна статистика.

REFERENCE

[1] Shahraki A, Abbasi M, Taherkordi A, Jercut A. Active Learning for Network Traffic Classification: A Technical Study. IEEE Transactions on Cognitive Communications and Networking 2021;8(1):422-439.

[2] Lopez-Martin M, Carro B, Sanchez-Esguevillas A, Lloret J. Network traffic classifier with convolutional and recurrent neural networks for Internet of Things. IEEE Access 2017;5:18042-18050.

[3] Salman O, Elhajj I, Kayssi A, Chehab A. Data representation for CNN based internet traffic classification: a comparative study. Multimedia Tools and Applications 2020;80: 16951–16977.

[4] Revendran R, Menon R. An efficient method for internet traffic classification and identification using statistical features. International Journal of Engineering Research & Technology 2015;4(7):297-303.

[5] Khater N, Overill R. Network traffic classification techniques and challenges. In: Proceedings of the tenth international conference on digital information management, vol. I, 2015, p. 43-48.

[6] Toupas P, D. Chamou D, K. M. Giannoutakis K, A. Drosou A, Tzovaras D. An intrusion detection system for multi-class classification based on deep neural networks. In: Proceedings of the eighteenth IEEE international conference on machine mearning and applications, vol. I, 2019, p. 1253-1258.

[7] Al-Turaiki I, Altwaijry N, Agil A, Aljodhi H, Alharbi S, Alqassem L. Anomaly-based network intrusion detection using bidirectional long short term memory and convolutional neural network. ISeSure 2020;12(3):37-44.

[8] Aouedi O, Piamrat K, Parrein B. Intelligent traffic management in next-generation networks. HAL Open Science Feature Internet 2022;14(44):1-35.