

WEB TRAFFIC MONITORING BY DESCRIPTIVE STATISTICS AND MULTILAYER NEURAL NETWORKS ²¹

Assoc. Prof. Ivelina Balabanova, PhD

Department of Communications Equipment and Technologies,
Technical University of Gabrovo
Tel.: 0896 640 473
E-mail: ivstoeva@abv.bg

Chief Assist. Prof. Georgi Georgiev, PhD

Department of Communications Equipment and Technologies,
Technical University of Gabrovo
Tel.: 0877 522 029
E-mail: givanow@abv.bg

Eng. Teodora Zhorova, PhD Student

Department of Communications Equipment and Technologies,
Technical University of Gabrovo
Tel.: 08998 690 061
E-mail: teddy.tedun@gmail.com

Abstract: The paper examines one of the main aspects of network traffic management related to traffic load and flexibility of system resources. An approach for the analysis and categorization of WEB traffic accessed by corporate clients in an urban environment, according to registered quantitative indicators of packet transmission is proposed. The approach is based on the application of Descriptive Statistics and Multilayer Feed-Forward Neural Networks. A method of synthesizing multilayer neural structures by experimenting with variations of different ratios between computational neurons in the hidden layers has been introduced. Backpropagation Gradient algorithm training was applied by the Levenberg-Marquardt algorithm. The selection of multilayer networks for the identification of areas of consumption of WEB content was based on accepted criteria, respectively Accuracy, Mean-Squared Error, and Mean Absolute Error. High levels of accuracy in minimizing the target errors for various data samples used have been achieved.

Keywords: Traffic Analysis, Descriptive Statistics, Network Efficiency, Multilayer Networks, MSE, MAE.

ВЪВЕДЕНИЕ

Изкуственият интелект значително повишава своето приложение в системите за управление и QoS на мрежовия трафик (Vashishth, T., Sharma, V., Kumar, B., & Chaudhary, S., 2023). Значителен дял от научните изследвания разглеждат адекватното администриране и планиране на информационни ресурси, свързани с достъп до Интернет услуга за корпоративни клиенти (Guo, A., & Yuan, Ch., 2021). Основни комплексни направления се очертават „повишаване на производителността“, „ефективни технологии за достъп, методи за защита“, „обезпечаване на обслужването на големи трафични потоци и буферното натрупване на информационни масиви“ (Latah, M., & Toker, L., 2016).

Един от тези аспекти засяга процедурния мониторинг на мрежовия Интернет трафик в урбанизирани градски региони. По отношение на това бяха проведени серия от изследвания за категоризация на зони на достъп до WEB съдържание в активни часови пояси за райони на гр. Чикаго чрез анализ на индекси на пакетното предаване на данни на основата на методите на изкуствения интелект и машинното обучение. Едно от изследванията предлага

²¹ Докладът е представен на научната сесия на 27.10.2023 в секция „Комуникационна и компютърна техника“ с оригинално заглавие на български език: МОНИТОРИНГ НА ИНТЕРНЕТ ТРАФИКА ЧРЕЗ ДЕСКРИПТИВЕН АНАЛИЗ И МНОГОСЛОЙНИ НЕВРОННИ МРЕЖИ

идентификация с използване на трислойни невронни мрежи с вариации на активацията в изходните слоеве – линейна “purelin”, тангес-сигмоидална “tansig”, логаритмична-сигмоидална “logsig”, Levenberg-Marquardt обучение с обратно разпространение на грешката и вероятностни невронни мрежи (Balabanova, I., Zhorova, T., & Georgiev, G., 2023). В (Balabanova, I., Zhorova, T., & Georgiev, G., 2022) е разработен комбиниран подход за разпознаване на зони на WEB потребление с включване на softmax функция на активация при Scaled Conjugate Gradient алгоритъм при невронните мрежи, статистическите методи k -най-близки съседи и дърво на решенията. Следващо изследване (Zhorova, T., 2022) разширява обхвата на прилагания невронен апарат с включване на многослойни Feed-Forward структури при фиксирана тангенс-сигмоидална активация в първия скрит и изходния и логаритмича-сигмоидална активационна функция във втория скрит слой.

Надграждането на дейностите от посочените изследвания в настоящия доклад се постига чрез въвеждане на подход, съчетаващ:

- първичен Дескриптивен анализ на базисни метрики на доставяния WEB трафик, използвани в качеството на информативни признаци при инструментите за разпознаване;
- втори етап на процедурен синтез на четирислойни Feed-Forward Neural Networks при различно процентно съотношение между невроните в междинните слоеве и заложено интегриране на основните активации – “tansig”, “logsig” и “purelin”.

ИЗЛОЖЕНИЕ

Дескриптивен анализ на индикатори на пакетното предаване WEB съдържание

Статистическата оценка на основни индекси на доставяния Интернет трафик, свързани с количествена характеристика на пакетното предаване на данни, е важна задача при анализ на поведението, установяване на тенденции и т.н., по отношение на крайните потребители и цялостното администриране на Интернет услугата. В тази връзка в настоящата секция е направен процедурен Дескриптивен анализ в MATLAB среда на 4 основни трафични индекса, респективно:

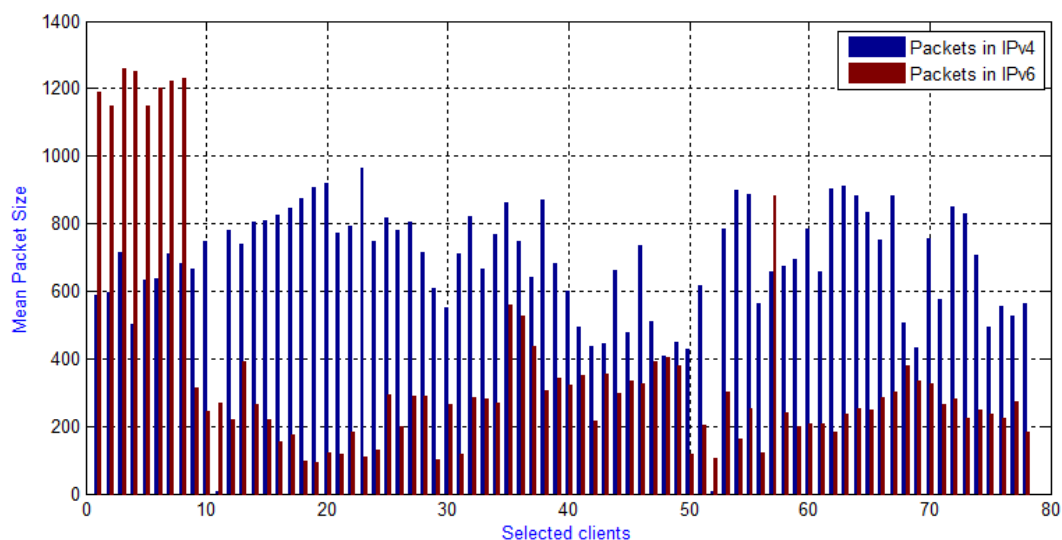
- ✓ Index № 1: Flows, [flows/s] – kilo;
- ✓ Index № 2: Mean Packet Size in IPv4, [pkts];
- ✓ Index № 3: Mean Packet Size in IPv6, [pkts];
- ✓ Index № 4: Mean Transition Rate, [pkts/s] – kilo.

Таблица 1. Статистически индикатори от Descriptive analysis на трафични индикатори на Интернет потребление по региони в градска среда – първи регион

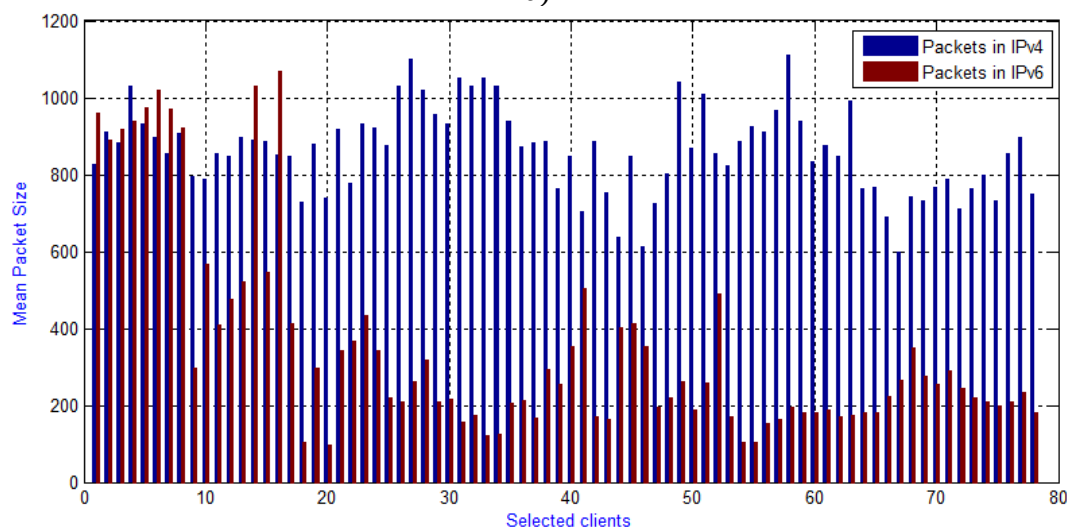
Flows, [flows/s] – kilo							
max	mean	median	Min	mode	std	var	range
32.7700	9.2324	8.3500	2.4000		4.5040	20.2860	30.3700
Mean Packet Size in IPv4, [pkts]							
max	mean	median	min	mode	std	var	range
963	680.7176	711	7.8400	494	179.9733	3.2390e+04	955.1600
Mean Packet Size in IPv6, [pkts]							
max	mean	median	min	mode	std	var	range
1260	360.7692	268	92	118	310.5407	9.6436e+04	1168
Mean Transition Rate, [pkts/s] – kilo							
max	mean	median	min	mode	std	var	range
698.8900	241.5795	233.3400	31.0900		145.6683	2.1219e+04	667.8000

Таблица 2. Статистически индикатори от Descriptive analysis на трафични индикатори на Интернет потребление по региони в градска среда – втори регион

Flows, [flows/s] – kilo							
max	mean	median	min	mode	std	var	range
75.3300	27.1341	25.3700	8.8000		14.2871	204.1219	66.5300
Mean Packet Size in IPv4, [pkts]							
max	mean	median	min	mode	std	var	range
1110	861.9615	870.5	598	1030	110.2515	1.2155e+04	512
Mean Packet Size in IPv6, [pkts]							
max	mean	median	min	mode	std	var	range
1070	349.9103	248.5	97	182	262.5382	6.8926e+04	973
Mean Transition Rate, [pkts/s] - kilo							
max	mean	median	min	mode	std	var	range
1140	594.8574	570.605	290.7300		198.3779	3.9354e+04	849.2700



a)



б)

Фиг. 1. Диаграми на разпределението на информационни пакети при IPv4 и IPv6 за избрани корпоративни клиенти при първи (а) и втори регион на Интернет потребление (б)

Указаните индекси са регистрирани относно диагностика на Интернет потреблението за два градски района. Спецификата на извършения Дескриптивен информационен анализ предвижда разделяне в две фази, съответно:

- Фаза № 1: Извличане на базисни статистически индикатори по региони на трафично потребление – Таблица 1 и Таблица 2;
- Фаза № 2: Представяне на разпределението на доставените пакети при IPv4 и IPv6 за селектирани корпоративни клиенти по региони на трафично потребление (Фиг. 1).

Многослойни Feed-Forward Neural Networks за идентификация на зони на разпределение на Интернет трафик

Първият етап в процеса на невронен синтез е насочен към използване на многослойни FFNN структури за идентификация на трафични зони на разпределение на потребителско Интернет съдържание в градска среда. Изследванията са базирани на различни въведени процентни съотношения между изчислителните невронни единици в първи и втори междинен структурен слой, респективно 1 – 2, 2 – 1 и 1 – 1. Въведена е класова принадлежност с използване на дискретни кодови комбинации, респективно „1 0“ за зона № 1 и „0 1“ за втори географски регион. Спрямо всички модели бяха фиксирани идентични типове невронна активация:

- Тангенс-сигмоидална за първи скрит слой;
- Логаритмична-сигмоидална за втори междинен слой;
- Линейна функция в изходните слоеве.

Оценката на качеството на класификация на заложените изходни групи е направено спрямо получените нива на три основни показатели, респективно Accuracy, Mean-Squared Error и Mean Absolute Error. Резултатите относно индикаторите за определяне на степента на невронна ефективност съобразно зададените съотношения между скритите неврони на анализирания невронни структури са обобщени от Таблица 3 до Таблица 5.

Таблица 3. Оценка на качеството на FFNNs за идентификация на зони на Интернет потребление при съотношение 1:2

Скрит слой № 1	Скрит слой № 2	Точност, %	MSE	MAE
2	4	91.3	0.0726	0.1105
3	6	95.7	0.0397	0.1265
4	8	91.3	0.0544	0.1427
5	10	100.0	0.0214	0.1004
6	12	91.3	0.0747	0.1842
7	14	91.3	0.0776	0.1572
8	16	100.0	0.0247	0.0995
9	18	100.0	0.0098	0.0574
10	20	95.7	0.0489	0.0858
11	22	100.0	0.0228	0.0996
12	24	100.0	0.0056	0.0573
13	26	100.0	0.0257	0.1231
14	28	95.7	0.0343	0.1159
15	30	100.0	0.0230	0.0918

При първото заложено процентно съотношение беше наблюдавано вариране на точността на разпознаване от 91.3 %, регистрирана при невронни комбинации „2:4“, „4:8“, „6:12“ и „7:14“, до 100.0 %, установени при съотношения „5:10“, „11:22“ и „12:24“. Бяха констатираи максимални MSE = 0.0776 и MAE = 0.1842 при комбинации „7:14“ и „6:12“. Докато минималните показания на грешките MSE = 0.0056 и MAE = 0.0573 са получени за модела при тестово съотношение „12:24“, оценен с най-добра адекватност.

Таблица 4. Оценка на качеството на FFNNs за идентификация на зони на Интернет потребление при съотношение 2:1

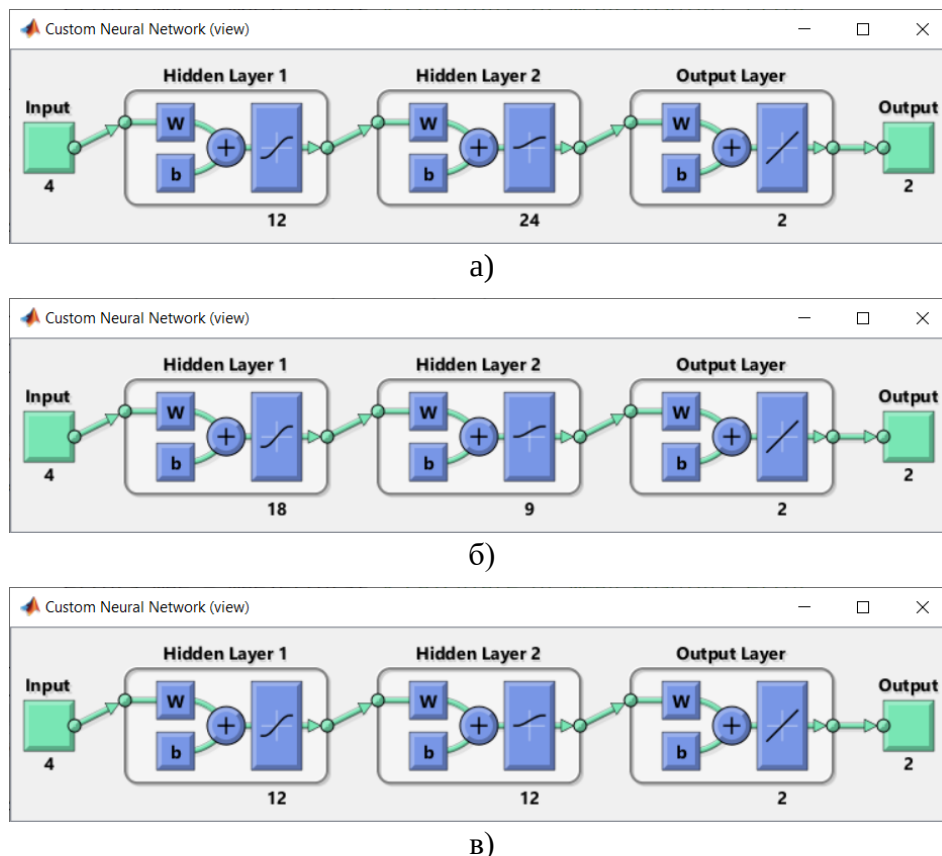
Скрит слой № 1	Скрит слой № 2	Точност, %	MSE	MAE
4	2	95.3	0.0303	0.1122
6	3	87.0	0.0887	0.1663
8	4	100.0	0.0174	0.0799
10	5	91.3	0.0344	0.1107
12	6	95.7	0.0387	0.1070
14	7	95.7	0.0381	0.0929
16	8	91.3	0.0874	0.1372
18	9	100.0	0.0165	0.0696
20	10	95.7	0.0540	0.1027
22	11	95.7	0.0395	0.1482
24	12	95.7	0.0524	0.1024
26	13	87.0	0.0892	0.2098
28	14	87.0	0.1142	0.2464
30	15	91.3	0.0684	0.1281

Таблица 5. Оценка на качеството на FFNNs за идентификация на зони на Интернет потребление при идентично съотношение

Скрит слой № 1	Скрит слой № 2	Точност, %	MSE	MAE
2	2	91.3	0.0824	0.1447
3	3	87.0	0.1287	0.1508
4	4	82.6	0.1174	0.2146
5	5	82.6	0.1092	0.2291
6	6	95.7	0.0349	0.1157
7	7	95.7	0.0230	0.0861
8	8	91.3	0.0694	0.1016
9	9	87.0	0.0627	0.1828
10	10	95.7	0.0647	0.1413
11	11	87.0	0.0590	0.1439
12	12	100.0	0.0035	0.0389
13	13	91.3	0.0948	0.1455
14	14	95.7	0.0165	0.0716
15	15	100.0	0.0202	0.1109
16	16	100.0	0.0193	0.1021
17	17	95.7	0.0339	0.1429
18	18	95.7	0.0539	0.1639
19	19	91.3	0.0732	0.1692
20	20	91.3	0.0642	0.1599

Спрямо второто изследвано съотношение между скритите неврони беше установена точност под прага от 90.0 %, респективно 87.0 % при комбинации „6:3“, „26:13“ и „28:14“. Най-висока степен на показателя 100.0 % е отчетена за невронни структури със съдържание на „8:4“ и „18:9“ невронни единици. Предвид по-ниските постигнати MSE и MAE, както следва 0.0165 спрямо 0.0174 и 0.0696 спрямо 0.0799, в качеството на по-адекватен модел е избрана невронна мрежа. Тук е констатиран по-висок праг на максималните MSE и MAE в сравнение с предходното базисно невронно съотношение, равняващи се на 0.1142 и 0.2464 за FFNN със съдържание на 28 в първи и 14 скрити неврони във втори структурен слой.

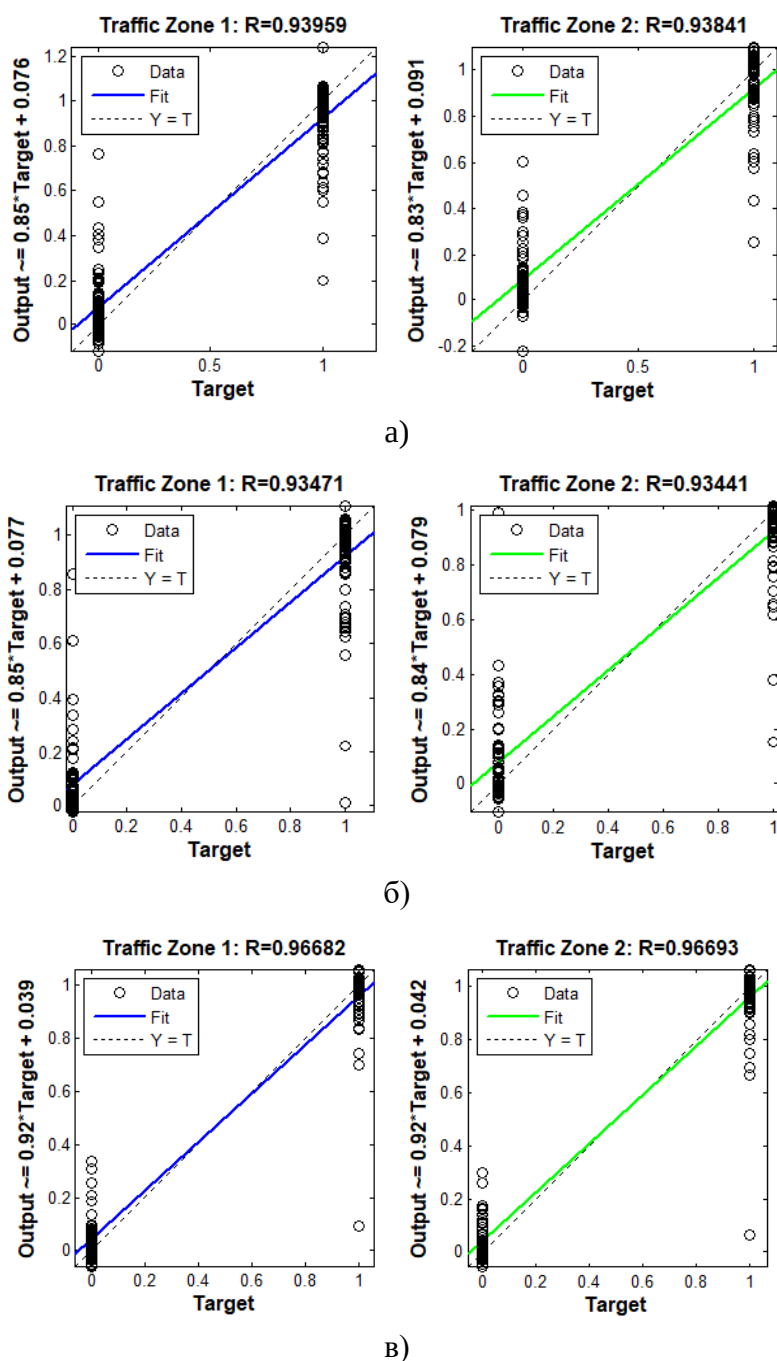
За третото използвано количествено съотношение между междинните неврони са постигнати минимална 82.6 % при „4:4“ и „5:5“ и максимална точност 100.0 % за комбинации „12:12“, „15:15“ и „16:16“. Най-високи грешки MSE = 0.1287 и MAE = 0.2291 бяха получени при невронни съотношения „3:3“ и „5:5“. Респективно най-ниски степени на индикаторите за качество MSE = 0.0035 и MAE = 0.0389, не само за разглежданото, а спрямо трите базисни съотношения 1 – 2, 2 – 1 и 1 – 1, са регистрирани за архитектура с 12 неврони в двата междинни слоя. Посочената FFNN архитектура може да бъде характеризирана с най-висока степен на пригодност.



Фиг. 2. FFNNs за разпознаване на градски зони на разпределение на трафикни пакети от глобалната Интернет мрежа при съотношения: а) 1 – 2, б) 2 – 1 и в) 1 – 1

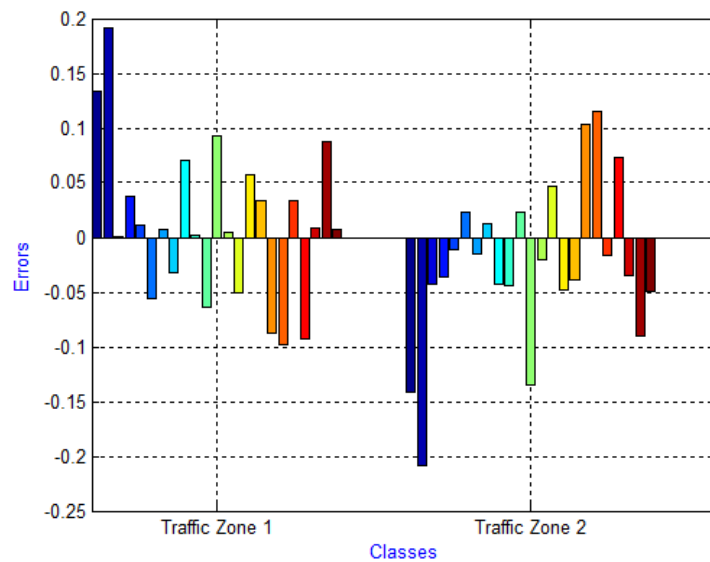
Фиг. 2 представя архитектурите на селектираните концептуални невронни модели за идентификация на географски региони на Интернет потребление в урбанизирана градска среда, спрямо които беше направена оценка на мрежовата производителност. Анализът беше базиран на изследване на поведението на средноквадратичната грешка за основните мрежови процеси – обучение, валидиране и тестване. Обща тенденция за всички невронни съотношения 1 – 2, 2 – 1 и 1 – 1 е достигането на най-добра валидираща производителност 0.011378 и 0.045399 при $7^{-\text{ма}}$, и 0.0063876 при $10^{-\text{та}}$ итерация преди спиране на мрежовото обучение. В хода на указаните процедури не са наблюдавани резки нараствания на грешките от обучение и валидация – индикация за коректност на невронните процеси.

Допълнително бяха изследвани линейните регресионни зависимости за мрежовите изходи на синтезираните FFNN архитектури за разпознаване на трафикно пакетно разпределение между обособените географски зони (Фиг. 3). Налице са удовлетворяващи нива на корелационните коефициенти над прага от 0.90000. Изразено предимство се наблюдава за невронния модел с приложено идентично съотношение между невронните изчислителни единици, където са отчетени $R = 0.96682$ и $R = 0.96693$ за първа и втора изходна група.

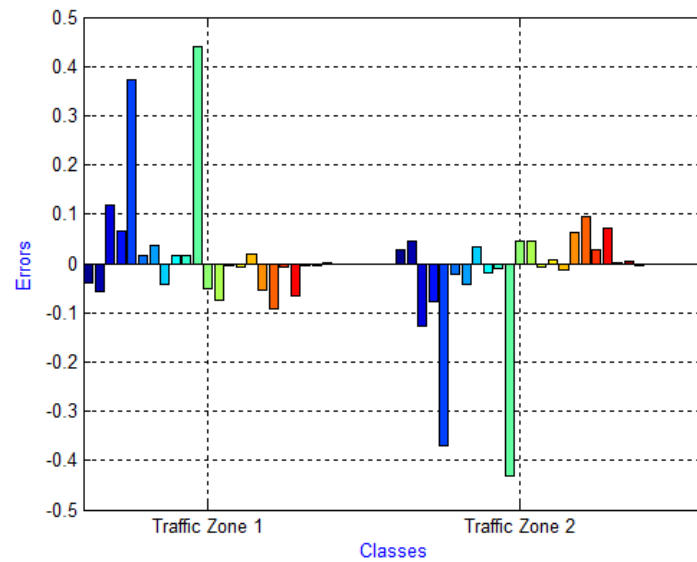


Фиг. 3. Линии на регресия за избраните многослойни FFNN модели за разпознаване на зони на Интернет потребление на пакетни данни: а) 1 – 2, б) 2 – 1 и в) 1 – 1

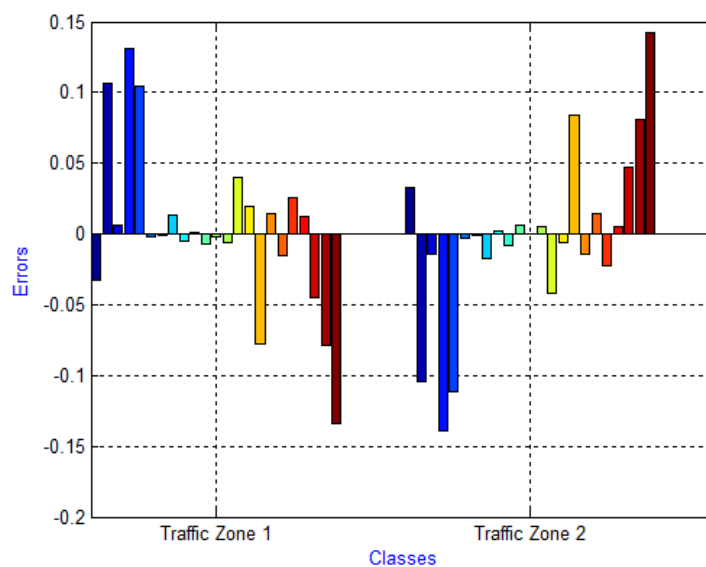
Фиг. 4 показва вариационните мрежови грешки, получени за част от еталоните в тестовата извадка, или отклоненията спрямо заложените теоретични класификационни показатели при обучение. При трите базисни невронни структури положителните и отрицателните вариации на грешките попадат в удовлетворяващия теоретичен диапазон ± 0.5 . Констатирани са отклонения в рамките на гранични диапазони „-0.2088 до 0.1914“, „-0.4324 до 0.4397“ и „-0.1387 до 0.1421“, респективно при базисни съотношения 1 – 2, 2 – 1 и 1 – 1. Тук отново е установено ясно изразено предимство на многослойния модел с идентично невронно съотношение между междинните неврони.



a)



б)



в)

Фиг. 4. Диаграми на мрежовите грешки за избраните FFNNs за идентификация на зони на потребление на Интернет съдържание: а) 1 – 2, б) 2 – 1 и в) 1 – 1

ЗАКЛЮЧЕНИЕ

Ефективността на информационните системи за мониторинг и администриране на мрежовия трафик до голяма степен е зависеща от наличието на софтуерно заложиени модули или приложения за категоризация и идентификация. Посредством интеграцията на технологията на изкуствения интелект чрез предложени подход значително може да бъде подобрена дейността по оценка и оптимизация на трафичното натоварване. Подобна аналогия следва относно гъвкавостта при разпределение на ресурсите, обезпечаващи обслужването на системния трафик.

REFERENCES

- Vashishth, T., Sharma, V., Kumar, B., & Chaudhary, S. (2023). *Artificial Intelligence-Enabled Traffic Optimization: A Comprehensive Survey*. Journal of Industrial Engineering, 52(5), pp. 26-34.
- Guo, A., & Yuan, Ch. (2021). *Network Intelligent Control and Traffic Optimization Based on SDN and Artificial Intelligence*. Electronics, 10(6), pp. 1-20.
- Latah, M., & Toker, L. (2016). *Application of Artificial Intelligence to Software Defined Networking: A Survey*. Indian Journal of Science and Technology, 9(46), pp. 1-7.
- Balabanova, I., Zhorova, T., & Georgiev, G. (2023). *Internet Traffic Zone Identification by Backpropagation and Probabilistic Neural Networks*. Paper presented at the XIVth International Scientific and Practical Conference, 15th-16th June 2023, Rezekne.
- Balabanova, I., Zhorova, T., & Georgiev, G. (2022). *Internet Traffic Analysis by FFNN, k-Nearest Neighbors and Decision Tree*. Paper presented at the LXIst Annual Scientific Conference, 27th-29th October 2022, Ruse.
- Zhorova, T. (2022). *Traffic Information Analysis Using Deep Learning Artificial Neural Networks*. Paper presented at the XXIst International Scientific Conference, 18th-19th November 2022, Gabrovo.