ИЗПОЛЗВАНЕ НА ДЪЛБОКИ НЕВРОННИ МРЕЖИ ЗА ОТКРИВАНЕ НА ИЗМАМИ С КРЕДИТНИ КАРТИ[1]

Гл. ас. д-р Ангелин Лалев
Д-р Александрина Александрова

Резюме

Настоящата студия представя изследване върху приложимостта на дълбоките изкуствени невронни мрежи за откриване на неприсъствени измами с кредитни карти. Изследването се базира на отворен набор от данни и цели да определи до каква дълбочина е практично да се изграждат такива мрежи, как архитектурните параметри на такива мрежи (ширина на слоевете, активационна функция и пр.) се отразяват на процеса на обучение и предиктивните способности на мрежата, както и доколко генеративното съперничество може да се използва за преодоляване на ефектите от дисбаланса между легитимни и фалшиви трансакции, който е определяща характеристика на всички набори от данни, свързани с откриването на измами с кредитни карти.

В рамките на изследването бяха проведени множество експерименти, които съчетават обучението на мрежи с различна дълбочина с различни параметри на обучението като активационна функция и дропаут параметър. Също така бяха тествани различни архитектурни подходи към изграждане на мрежата, като особен акцент бе поставен върху възможността за използване на GAN мрежи и принципа на генеративно съперничество, залегнал в тяхната архитектура.

Получените резултати показват, че дълбоките невронни мрежи с умерен брой скрити слоеве, съчетани с техники за овърсамплинг, се справят отчетливо по- добре при откриването на фалшиви трансакции в сравнение с мрежите, притежаващи само един скрит слой. Също така, получените резултати навеждат на заключението, че използването на генеративно съперничество в ролята на техника за овърсамплинг е сравнимо по своята ефективност с други подобни техники.

Ключови думи: измами с кредитни карти, дълбоки невронни мрежи, генеративно съперничество.

JEL: C45.

[1] Участието на авторите е както следва: гл. ас. д-р Ангелин Лалев – т. 2, 3, 4 и заключение; д-р Александрина Александрова – увод, т.1;

DEEP NEURAL NETWORKS FOR DETECTION OF CREDIT CARD FRAUD

Head Assist. Prof. Angelin Lalev, PhD
Alexandrina Alexandrova, PhD

Abstract

This publication presents research on the possibility of using Deep Neural Networks (DNNs) for detection of credit card fraud. The research is based on an open dataset and aims to determine how the varying number of hidden layers and other architectural parameters (width of each layer, activation functions) affects training and predictive potential of the network. It also aims to determine how to use generative-adversarial approach efficiently in order to overcome the problem with the imbalance between legitimate and fraudulent transactions, which is crucial characteristic of the datasets related to detection of credit card fraud.

The publication presents the results of various experiments, conducted to achieve the stated goals. A number of experiments have been carried out combining network training of various depth of the network and with different parameters of the learning process, for example, activation function and dropout values. The experiments focus on using generative-adversarial approach as a method for oversampling, in particular.

The results indicate that deep neural networks with moderate hidden layers, combined with methods for oversampling are performing much better in identifying fraudulent transactions than the classical neural networks, consisting of only one hidden layer. The results also indicate that using generative-adversarial approach as a method for oversampling is comparable in efficiency to other widely used techniques.

Keywords: credit card fraud, deep neural network, GAN.

JEL: C45.

Увод

Измамите с кредитни карти са една от най-често срещаните форми на банкова измама. С разрастването на електронната търговия и навлизането на електронните разплащания тези измами продължават да еволюират и да нанасят все по-големи щети. Най-значимото явление, характеризиращо тази еволюция, е това, че в последните години „присъствените“ форми на измама, при които се атакува самата физическа карта или устройството за четене на такава карта, отстъпиха място по превалентност на измами, при които се атакуват “неприсъствени” трансакции. При тези трансакции, известни като „Card-Not-Present” или „CNP”, картата не може да се верифицира физически от продавача.

Появата на CNP измамите промени коренно мащабите на щетите, които издателите на кредитни карти, големите онлайн ритейлъри, банките и потребителите претърпяват ежегодно от измами. Някои изследвания (European Central Bank, 2018) (Javelin, 2019) оценяват годишните щети от CNP измамите на близо 2 млрд. евро само за страните от SEPA (по данни за 2016 година) и на до 14.7 млрд. долара за САЩ (данни от 2018 г.). Ето защо не е изненадващо, че в последните години засегнатите страни влагат мащабни усилия в ограничаването на този вид измами. Това от своя страна се оказа изключително предизвикателство поради огромните обеми на CNP разплащанията и своеобразната инерция, която предотвратява бързото навлизане на нови технически мерки (Маринова, 2017) за защита откъм страната на потребителя.

Отчитайки мащабите на щетите, които понасят, както и възможността за осъществяване на превантивни мерки, издателите на кредитни карти и банковите институции разчитат все повече на технологии от областта на изкуствения интелект за разпознаване на фалшивите трансакции. Разпознаването на една част от тези трансакции като фалшиви на база информацията за самата трансакция е възможно, тъй като много от тези трансакции имат атрибути или комбинация от атрибути, които се различават достатъчно от тези на легитимните трансакции. За съжаление много подобни комбинации са трудни за откриване от човешко око, а самият брой на трансакциите прави абсолютно непрактично за целта да се използва човешки труд.

Поради изключително чувствителния характер на информацията, свързана с плащанията чрез кредитни карти, усилията за внедряване на изкуствен интелект при откриване на фалшиви трансакции до голяма степен остават извън обхвата на академичния дискурс. Едва в последните години това започна да се променя с появата на силно анонимизирани, но достатъчно обхватни набори от данни, които да позволят по-задълбочени анализи.

Последните години също бележат сериозен прогрес в областта на технологиите за изкуствен интелект и по-специално технологиите за „дълбоки“ невронни мрежи. Появата на достъпни устройства, способни да извършват масивно-паралелни изчисления, стана катализатор за развитието на най-разнообразни архитектури за изграждане на дълбоки невронни мрежи, много от които тепърва предстои да се приложат към проблеми от областта на бизнеса и икономиката.

Настоящата студия представя именно подобно изследване. **Обект** на изследването са CNP измамите с кредитни карти, а негов **предмет** са техниките за предотвратяване и ограничаване на подобни измами с помощта на класификационни модели, базирани на дълбоки невронни мрежи. Главната **цел** на изследването е определянето на ефективността на подобни модели спрямо по-традиционните техники за класификация като

еднослойните невронни мрежи и логистичната регресия. За целта изследването решава следните **изследователски задачи**:

1. Определяне на това до каква дълбочина е практично да се изграждат подобни мрежи.
2. Тестване на методи за обучение на подобни мрежи с цел определяне на най-подходящите от тях, които могат да се прилагат успешно към небалансирани набори от данни като тези, генерирани от CNP трансакциите.
3. Апробиране на нови архитектури на база генеративно съперничество за подобряване на ефективността на мрежата.

За постигане на очертаните цели в рамките на изследването бяха проведени редица експерименти с различни работни параметри на мрежите (включително брой слоеве на мрежата) и различни обучителни функции. Експериментално бе изследвана ефективността от използването на дропаут и овърсамплинг.

В рамките на изследването специално бяха проведени експерименти относно възможността за използване на генеративно съперничество, вдъхновено от дизайна на GAN мрежите (Goodfellow, et al., 2014) като техника за овърсамплинг.

Дълбочината на невронната мрежа е свързана с нейната предиктивна способност, както и с обема на данните, нужни за нейното обучение. В областите, където технологиите на дълбоките невронни мрежи бяха приложени успешно, като например разпознаването на изображения и лексикалния анализ, дълбочината на тези мрежи варира от 5 до 180 скрити слоя (Szegedy, et al., 2014). Обемът на данните, които се обработват от тези мрежи обаче, е вероятно няколко степени по-голям от данните, които се обработват при анализирането на CNP разплащанията. Много малко изследвания до момента се фокусират върху въпроса, доколко мрежи с подобни дълбочини биха допринесли за подобряване откриването на CNP измами. Въпросът не е тривиален, тъй като дълбоките невронни мрежи имат вроден недостатък, който се изразява в тенденцията към прекомерно нагаждане (overfitting). При използването им за откриване на фалшиви трансакции тази тенденция ще бъде засилена от диспропорцията между легитимни и фалшиви трансакции (небалансираност на набора от данни), която се изразява в това, че под една от хиляда трансакции всъщност е фалшива[2]. Ефектите от този дисбаланс при обучението на невронните

[2] За по-нататъшна илюстрация на проблемите с прекомерното нагаждане и небалансираните данни авторът предлага абстрактен пример. Ако приемем, че точно една от 1000 трансакции е фалшива, евентуален класификатор, който трябва да определи коя трансакция е фалшива, но всъщност винаги определя тестваната трансакция като легитимна, ще е прав в 99.9% от случаите, но ще бъде напълно безполезен за откриване на фалшивите трансакции. За съжаление повечето

мрежи могат да бъдат преодолени до някаква степен, ако се използват различни методи за овърсамплинг, но е реалистично да се очаква, че с увеличаването на броя на слоевете на обучаваната мрежа на някакъв етап ефектът от тези техники ще бъде недостатъчен, за да противодейства на засилващата се тенденция към прекомерно нагаждане.

Отвореният характер на използваните набори от данни, както и фактът, че проблемът отскоро е обект на академична дискусия, вероятно допринасят за това, че голям брой от най-популярните и видими публикации по темата имат практически характер. Подобни публикации се концентрират повече върху разработването на програмен код отколкото върху сравнение между различни подходи. Прави силно впечатление, че много от тези публикации измерват получените резултати в понятията на потенциално несравними (Kaggle, 2019a) и дори евентуално неподходящи (Kaggle, 2019a) за целта показатели, което придава дискусионен характер на проблема за избор на показатели за ефективност на разработените модели за класификация. Ето защо немалка част от настоящата студия е посветена на проблема за измерване ефективността на двоичните класификатори и аргументация на избрания от авторите подход.

За да се постигне възможно по-голяма прозрачност и възпроизводимост на получените резултати, разработените от авторите инструментални средства (програмен код) също са направени публично достъпни на адрес http://www.github.com/lalev-angelin/creditcardfraudexperiments/.

1. Особености на набора от данни

Изследването използва публично достъпен набор от данни с 280 хил. трансакции, осъществени в рамките на 48 часа от европейски картодържатели[3]. Фалшивите трансакции в набора са 492 или 0.17% от целия набор.

Преди публикуването на набора той е преминал през процес на анонимизация чрез използване анализ на главните компоненти (PCA) за намаляване на размерността на данните. Резултатът от тази операция е таблица с 31 колони. 28 от тези колони съдържат нормализирани стойности, попадащи основно между -1 и +1. Докато тези стойности не могат лесно да се съпоставят с оригиналните и да бъдат деанонимизирани, те запаз-

от метриките, които се използват в хода на обучение на невронната мрежа, биха приели 99.9% точност като изключително добър резултат, което би довело до това, че обучената мрежа ще има силна тенденция да обявява фалшивите трансакции за истински.

[3] Наборът от данни може да бъде свободно изтеглен от https://www.kaggle.com/mlg-ulb/creditcardfraud.

ват до голяма степен основните характеристики и релации между данните, съществуващи в оригиналната таблица.

Останалите три колони съдържат съответно време на осъществяване на трансакцията, сума на трансакцията и число, което обозначава легитимна или фалшива трансакция.

Оригиналният набор от данни посочва стойностите в колоната за време като секунди, изминали от началото на 48-часовия период. Числото, което обозначава легитимна трансакция, е 0, докато числото, което обозначава фалшива трансакция, е 1.

Като част от подготовката на данните за обработка върху данните бяха извършени две допълнителни трансформации. На база хипотезата, че честотата на измамите може да има слаба връзка с времето от денонощието, колоната за време бе превърната в стойности от 0 до 1, представляващи времето от началото на всеки от двата 24 часови периода.

Втората трансформация касае избягването на грешки при закръгляване при обучаването на невронната мрежа. Тъй като всички колони на невронната мрежа трябва да бъдат приведени към стойности между -1 и 1 или 0 и 1, за да може да се осъществи смислено обучение на мрежата, колоната със сумата на трансакцията също трябва да бъде обработена. Докато повечето стойности в тази колона варират от 0 до 100, тя съдържа и екстремни стойности от порядъка на 25000. Евентуална нормализация, при която 25000 се приравнява на единица, би довела до това, че останалите стойности в колоната биха били малки дробни числа с няколко разряда след запетаята. При изчисляването на градиентите това би довело до още по-малки междинни резултати, които евентуално ще бъдат закръглени поради ограниченото място, отделено за представяне на дробната част на числата в компютъра, а това на свой ред ще предотврати правилното обучение на мрежата.

За да се избегне този проблем, настоящото изследване избира подхода на „отрязване“ на стойностите. Стойности, по-големи от 5000, са заменени с 5000 в колоната за стойността, преди тя бъде нормализирана. Аргументацията на този подход е свързана с това, че вероятността трансакция на сума 25000 да е фалшива, трябва да бъде много близка до вероятността, трансакция от 5000 да бъде фалшива. Казано по друг начин, за целите на разпознаването на фалшиви трансакции, невронната мрежа, обучена върху така трансформирания набор, ще третира всички трансакции над 5000 като еднакво големи. 55 трансакции в целия набор имат стойности над 5000 и съответно са засегнати от тази трансформация.

За да се измери производителността на всяка тествана архитектура, така нормализираният набор от данни е разделен на обучително и тестово множество. ¾ от трансакциите в оригиналния набор са разпределени в обучителното множество, а останалите формират тестовото множество, като разпределението е извършено на случаен принцип.

2. Показатели за ефективност на моделите за откриване на фалшиви трансакции

По същество откриването на фалшиви трансакции представлява задача за двоична класификация, при която евентуален класификатор трябва да постави всяка от трансакциите в една от двете категории – легитимни или нелегитимни трансакции. Подобни задачи се появяват в най-различни предметни области и като резултат съществуват редица различни методи за сравняване на класификаторите, които се препокриват частично в понятията на терминология и подход.

Всички тези методи идват със специфични предимства и недостатъци и трябва да бъдат съобразени с естеството на задачата, за да се избегне неправилна интерпретация на получените резултати. Ето защо един от най-важните въпроси, свързани с използването на невронни мрежи за откриване на фалшиви трансакции с кредитни карти, опира до това, какви показатели ще бъдат използвани за сравняването на ефективността на различните тествани мрежови архитектури.

Когато става въпрос за сравняване на ефективността на класификатори, които дават дискретен резултат, отговарящ на една от двете категории, стандартно се използва т.нар. „матрица на объркването“ (фиг. 1) и базирани на нея показатели.

Матрицата на объркването е формулирана в понятията на търсене на обекти от един клас (позитивен клас) измежду набор, който се състои от два класа (позитивен и негативен клас). Тя има четири категории, които отговарят на релацията между изхода от двоичния класификатор и реалния клас на обекта, подлежащ на класификация.

В категория „истински позитивни“ (TP) попадат обекти от търсената категория, за които класификаторът правилно е предсказал, че попадат в нея. В категория „истински негативни“ (TN) попадат всички обекти, за които класификаторът правилно е предсказал, че не попадат в търсената категория. Останалите две категории представляват грешки при класификацията. В категория „фалшиви негативни“ (FN) попадат обекти от търсената категория, за които класификаторът неправилно е предсказал, че не попадат в търсената категория (известно като тип 2 грешка). По същия начин в категория „фалшиво позитивни“ попадат всички обекти, за които класификаторът неправилно е предсказал, че принадлежат към търсената категория (тип I грешка).

Класически мерки за сравняване на двоични класификатори, произлизащи от областта на медицината, са т.нар. чувствителност (TPR[4]) и специфичност (TNR[5]):

$$TPR = \frac{TP}{TP+FN},\ TNR = \frac{TN}{FP+TN}$$

		Реално	
		Позитивен	Негативен
Предсказано	Позитивен	Истински позитивни (TP)	Фалшиво позитивни (FP)
	Негативен	Фалшиво негативни (FN)	Истински негативни (TN)

Фигура 1. Матрица на объркването

В контекста на откриването на фалшиви трансакции чувствителността измерва какъв дял от всички фалшиви трансакции са коректно разпознати от класификатора като такива, докато специфичността съответно измерва какъв дял от легитимните трансакции са разпознати като такива.

Трудността при сравняване на класификатори на базата на чувствителност и специфичност е свързана с това, че двата показателя трябва да бъдат тълкувани винаги заедно, което затруднява извършването на сравнения на тяхна база. Така например не е достатъчно да се посочи само чувствителността на даден класификатор, тъй като класификатор, който обявява всички трансакции за фалшиви, ще има 100% чувствителност за сметка на 0% специфичност, която отразява факта, че всички легитимни трансакции също ще бъдат неправилно класифицирани като фалшиви.

Друг възможен подход идва от областта на извличане на документи и се базира на измерване на т.нар. “Recall” (TPR) и “Precision” (PPV[6]).

[4] Чувствителността е известна още като „True Positive Rate” – “Дял на истински позитивните“, откъдето идва и съкращението.

[5] Специфичността е известна и като „True Negative Rate” – “Дял на истински негативните”, поради което тук е съкратена като “TNR”.

[6] Този показател е известен още като “Positive Predictive Value”, откъдето идва и съкращението.

$$TPR = \frac{TP}{TP+FN},\ PPV\ = \frac{TP}{TP+FP}$$

При търсене на документи на определена тема показателят „Recall" отразява каква част от всички документи на темата ще бъдат показани в резултатите от търсенето. Тъй като търсенето на документи разделя всички документи в една информационна система на две групи – отговарящи и неотговарящи на критериите на търсенето, на практика този показател е еквивалентен на чувствителността. Показателят "Precision" от своя страна показва каква част от показаните от търсенето документи са релевантни към темата. В понятията на търсене на фалшиви трансакции, Precision показва каква част от класифицираните като фалшиви трансакции наистина са такива.

За разлика от чувствителността и специфичността Precision и Recall показателите могат да бъдат усреднени по смислен начин при небалансирани съвкупности от данни. Среднохармоничната стойност на двата показателя е известна като F1 критерий. F1 критерият е стандартизиран начин за сравняване на двоични класификатори, като по-големи стойности на F1 отразяват по-добри стойности за Precision и Recall компонентите. Забележително за този показател е това, че броят на елементите на истински негативния клас (TN) не взема участие в изчисленията, което е недостатък в някои ситуации и предимство – в други.

Всички дискутирани до момента показатели касаят дискретни класификатори. Това прави директното им приложение към невронните мрежи проблематично, тъй като невронните мрежи обикновено извеждат число между 0 и 1, което се интерпретира като вероятност, обектът да принадлежи към търсения клас. За да бъдат превърнати в дискретни класификатори, невронните мрежи трябва да бъдат комбинирани с някаква прагова стойност (threshold). Изход от мрежата под праговата стойност се интерпретира като „негативен", докато съответно изход над праговата стойност следва да се интерпретира като „позитивен".

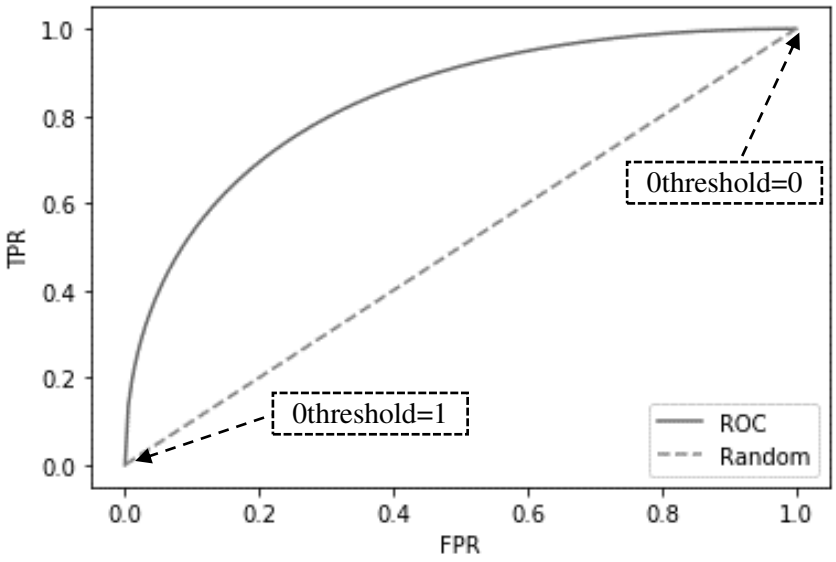
Всяка възможна избрана прагова стойност генерира собствена матрица на объркването, което прави директното сравнение на класификационните способности проблематично, защото или трябва да се избере конкретна прагова стойност, или трябва да се адресира проблемът за сравняване на множество (дори безброй) матрици на объркването, произтичащи от различни избрани прагови стойности.

Възможен подход, който обобщава показателите чувствителност и специфичност за всички възможни прагови стойности е т.нар. „Receiver Operating Characteristic"[7] (ROC). ROC е замислен като визуален инстру-

[7] Показателят идва от още една различна предметна област – радиотехниката, което обяснява наименованието му.

мент за сравнение на различни класификатори, които извеждат вероятността, даден обект да принадлежи към търсения клас, и представлява крива, изчертана на координатна система. Кривата се получава при нанасяне на двойките стойности за TPR (чувствителност, "recall") и FPR (дял на фалшиво позитивните) за всяка стойност на прага, докато той варира от 1 до 0.

$$FPR = 1 - TNR = \frac{FP}{FP + TN}$$



Фигура 2. ROC крива на примерен класификатор (непрекъсната линия) срещу ROC крива на класификатор, който определя резултата на случаен принцип (пунктир)

Тъй като FPR директно зависи от специфичността, ROC кривата може да бъде интерпретирана като вариант за обобщаване на показателите чувствителност и специфичност, пригоден за класификатори, които извеждат вероятност, обектът да принадлежи към търсения клас, какъвто е случаят с невронните мрежи.

В контекста на откриване на фалшиви трансакции, приближение на кривата за вече обучена невронна мрежа ще се изчисли, като мрежата бъде инструктирана да класифицира тестовото множество. За всеки ред в тестовото множество е предварително известно дали описва легитимна или фалшива трансакция. При това условие изходът от невронната мрежа за всеки ред (стойности в интервала от 0 до 1) ще се интерпретира спрямо избрания праг и на тази база ще могат да бъдат изчислени четирите ком-

понента на матрицата на объркването, откъдето ще бъдат получени и стойностите за TPR и FPR.

За изчертаване на кривата евентуален алгоритъм ще постави първоначално прага на 1, което ще има този ефект, че независимо какъв е изходът от невронната мрежа, всички трансакции ще бъдат интерпретирани като легитимни. Това ще произведе $TPR = 0$ и $FPR = 0$, т.е. точката, отговаряща на този праг, ще бъде началото на координатната система. С бавното сваляне на прага, малка част от трансакциите, за които невронната мрежа е определила най-висока вероятност да бъдат фалшиви, ще бъдат класифицирани като такива. За правилно обучена мрежа тези трансакции наистина ще бъдат фалшиви, което ще доведе до бързо покачване на *TPR* и значително по-слабо покачване на *FPR*, което ще направи *ROC* кривата „изпъкнала нагоре“.

Свалянето на прага ще доведе и до това, че все повече трансакции ще бъдат класифицирани като фалшиви при все по-малка вероятност, определена от мрежата за това. Неминуемо като резултат някои от легитимните трансакции ще бъдат маркирани от мрежата като фалшиви и *FPR* ще започне да расте. Кривата ще започне да се „движи“ надясно, достигайки евентуално точката (1,1) при праг 0, която отговаря на максимална чувствителност (100% от фалшивите трансакции са открити) и 100% дял на фалшивите позитивни (респ. специфичност 0%), което означава, че всички трансакции без изключение са маркирани като фалшиви.

ROC кривата е предпочитан инструмент за анализ и сравняване на класификатори поради нейната висока интуитивност. Тя дава ясна визуална представа, как се променят параметрите на матрицата на объркването при промяна на праговата стойност. Тя позволява визуалното търсене на компромис между чувствителност и специфичност (точка от кривата, която интуитивно се намира максимално наляво и нагоре на координатната система), но по-интересно свойство от гледна точка сравняването на класификатори е това, че по-добрите класификатори имат по-изпъкнала крива, чиито точки се намират „над“ и „вляво“ от кривите на по-лошите класификатори, което позволява визуално сравнение на качествата на двата класификатора като цяло, без да трябва да бъде определян конкретен праг.

Фигура 2 демонстрира сравнение между ROC крива на хипотетичен класификатор и възможно най-лошия класификатор[8], който определя напълно случайно стойността между 0 и 1, която бива интерпретирана като вероятност, трансакцията да е фалшива.

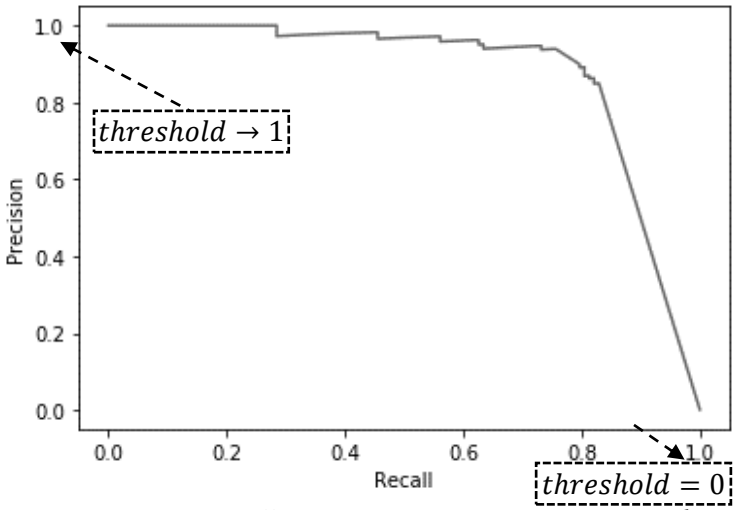
Фактът, че по-добрите класификатори имат криви, разположени „над“ кривите на по-лошите класификатори, позволява към въпроса за

[8] Всеки „по-лош“ от този класификатор може да бъде превърнат в по-добър, като неговият изход се извади от единица и резултатът се интерпретира като вероятност, дадената трансакция да бъде фалшива.

намиране на единна количествена мярка за качество на класификатора да се подходи чрез числова интеграция. Показателят AUROC или „Area Under the ROC Curve" отразява площта, разположена под ROC кривата на класификатора. Поради споменатите отношения между техните криви, по-добрите класификатори имат по-висока стойност на AUROC показателя, като перфектният класификатор има AUROC = 1.

AUROC не е идеален показател, тъй като винаги трябва да се помни, че реалните класификатори ще работят в определен диапазон на прага, който отговаря на висок *TRP* и нисък *FPR*. В този смисъл, ако по-голяма част от разликата в два *AUROC* показателя е генерирана от разлики между площите в лявата част на диаграмата, *AUROC* наистина ще отразява съществени разлики в качеството на двата класификатора. В противен случай разликата ще касае прагове, които са толкова ниски, че произвеждат неприемливо голям дял на фалшиво позитивни трансакции.

Алтернативен подход за сравняване на класификатори с агрегиран показател се базира на т.нар. „Precision-Recall" (PRC) крива, която е аналогичен опит за визуално сравняване на класификаторите, базиран на Precision и Recall параметрите на класификатора (вж. фиг. 3).



Фигура 3. Precision-Recall (PRC) крива на примерен класификатор

PRC кривата се изчертава, като прагът варира между 0 и 1. При праг 0 стойностите за Recall (TPR) и Precision (PPV) са съответно 1 и 0 (100% от фалшивите трансакции са открити; всички трансакции са маркирани като фалшиви). С покачване на прага PPV расте и TPR намалява. При праг, наближаващ 1, TPR се доближава до 0, а PPV се доближава до 1.

По същия начин като при ROC кривата по-добрите класификатори имат криви, разположени над тези на по-лошите класификатори, което

позволява изчисляването на подобен показател – *AUPRC*[9], касаещ площта под кривата.

И *AUROC*, и *AUPRC* могат да се използват за оценка на класификатори и вчастност – на дълбоки невронни мрежи. *AUROC* е предпочитаната мярка за сравняване на резултатите от обучението на невронни мрежи, тъй като представлява може би по-интуитивен показател. Също така в негова полза е и аргументът, че той включва в изчисленията категорията на истински негативните елементи. Немалка част от практикуващите специалисти по анализ на данни дори предпочитат *ROC* по отношение откриването на фалшиви трансакции с кредитни карти[10].

Авторите на настоящото изследване се присъединяват към мненията на критиците на използването на *ROC* при небалансирани множества[11], които подчертават, че включването на истинските негативи в изчисленията, когато негативният клас е много по-превалентен, води до твърде оптимистични оценки на разликата между два класификатора. В този смисъл използването на *ROC* и *AUROC* за оценка на класификационните способности на невронните мрежи за откриване на фалшиви трансакции може да се счете за *грешка*.

В полза на това могат да бъдат приведени следните аргументи:

1. Докато съществуват резултати, които подсказват, че ако ROC кривата на един класификатор е стриктно над ROC кривата на друг класификатор, PRC на първия класификатор ще бъде над PRC кривата на втория класификатор[12], тези резултати не се отнасят за ситуации, когато двете ROC криви се пресичат, което често се случва на практика.

2. Когато позитивният клас е много по-рядко срещан от негативния, малки промени в ефективността на класификатора (например една транспозиция в тоталния ред на трансакциите, дефиниран от генерираните от мрежата прогнозни вероятности, те да са фалшиви) ще доведат до по-голяма абсолютна промяна на TPR и много малка промяна на FPR в сравнение със ситуацията, когато двата класа са балансирани. Това ще доведе до това, че скромни по възможности класификатори ще имат по-изпъкнали ROC криви, а разликите между две ROC криви на два класификатора ще бъдат по-малки в абсолютно отношение.

3. Истинските негативни не участват в изчислението на PRC кривата, поради което подобно явление при дисбаланс от описания тип не се наблюдава при тях.

---

[9] Съкращението идва от „Area Under Precision-Recall Curve".

[10] Вж. например https://www.kaggle.com/joparga3/in-depth-skewed-data-classif-93-recall-acc-now или https://www.kaggle.com/varunsharma3/credit-card-fraud-detection-using-smote-0-99-auc

[11] Вж. например http://pages.cs.wisc.edu/~jdavis/davisgoadrichcamera2.pdf

[12] Вж. https://www.biostat.wisc.edu/~page/rocpr.pdf

Кумулативният ефект от това може да бъде илюстриран със следния пример, който фиксира само една точка на ROC кривите на двата класификатора, отговаряща на конкретно ниво на чувствителност:

Ако в класификаторите трябва да различат 100 фалшиви трансакции измежду група от 1000000 такива трансакции и първият от тях трябва да върне 100 маркирани като фалшиви, за да постигне брой на истинските позитивни 90, а вторият трябва да маркира 1000 трансакции като фалшиви, за да постигне същия брой на истинските позитивни, то и двата ще имат $TPR = 0.9$. Първият класификатор ще има $FPR \approx 0.0001$, докато вторият ще има $FPR \approx 0.0009$, което означава, че по линията $TPR = 0.9$ точките от двете ROC криви ще са разположени много близко една до друга. По отношение на показателя *Precision* обаче нещата стоят по различен начин. Първият класификатор ще има $PPV = 0.9$, докато вторият ще има $PPV = 0.09$, което означава, че по линията *TPR=0.9* двете точки от Precision-Recall кривата ще са значително отдалечени.

Докато тази разлика в мащаба не е сама по себе си критична, не е ясно доколко евентуални грешки при закръгляване на дробната част в компютърните системи биха имали допълнителен негативен ефект при изчисляването на *FPR* и *AUROC*.

Тъй като по въпросите за използване на ROC и PRC е възможно да няма пълен консенсус, представяното изследване изчислява и двата показателя (*AUROC* и *AUPRC*), но интерпретира резултатите в понятията на *AUPRC*.

3. Тествани архитектури

Една от главните цели на извършеното проучване е да определи дали допълнителни скрити слоеве с различни размери могат да помогнат за подобряване откриването на фалшиви трансакции. Това е валиден въпрос, тъй като за разлика от областта на разпознаването на изображения, където дълбоките невронни мрежи се използват с голям успех, размерността на финансовите данни е значително по-малка. Освен това за разлика от разпознаването на изображения, при които първите слоеве следва да научат дребни детайли като контури, цветове и т.н., а от следващите се очаква да направят генерализации, при финансовите данни няма ясни „рецепти“ за избор на архитектура.

Въпреки това, на база общите свойства на набора от данни могат да се направят някои обосновани предположения. Най-важното от тях вече бе споменато и касае факта, че небалансираният характер на данните ще тласка мрежата към прекомерно нагаждане. Добавянето на допълнителни слоеве и увеличаването на размерите на всеки от тях допълнително ще засили тази тенденция.

Това на свой ред най-вероятно ще означава, че тестваните невронни мрежи трябва да бъдат съчетани с допълнителни мерки за предотвратяване на прекомерното нагаждане.

В допълнение на експериментите с различни брой слоеве и брой неврони във всеки слой, в рамките на изследването бяха проведени и експерименти с две активационни функции, подходящи за целта – ReLU (Rectified Linear Unit) и tanh (хиперболичен тангенс).

$$relu(x) = \begin{cases} x, & x > 0 \\ 0, & x \leq 0 \end{cases}, tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$$

ReLU се използва сравнително отскоро и има репутацията на подобрата от двете активационни функции. Тя е особено пригодна за работа с дълбоки невронни мрежи, тъй като не страда от проблема с изчезващите градиенти. ReLU има обаче и недостатъци. Функцията е диференцируема при 0 и в редки ситуации може да доведе до това, че част от мрежата ще „забие“ и няма да участва в следващите итерации на обучението.

Хиперболичният тангенс, от друга страна, е традиционна активационна функция, която обаче страда от проблема с изчезващите градиенти.

Всички тестове са проведени чрез използване на ADAM (Kingma & Ba, 2014) оптимизатор за обучение на мрежата.

Таблица 1 обобщава повечето от тестваните в рамките на изследването архитектури:

Таблица 1

Тествани архитектури на дълбоки невронни мрежи

Съкращение	Слоеве									
	0	1	2	3	4	5	6	7	8	9
	Ширина и активация	Дропаут	Ширина и активация	Дропаут	Ширина и активация	Дропаут	Ширина и активация	Дропаут	Ширина и активация	Дропаут
32-28-20-12-1	32 ReLU	0.3	28 ReLU	0.3	20 ReLU	0.3	12 ReLU	0.3	1 Sigmoid	-
32-24-16-1	32 ReLU	0.3	24 ReLU	0.3	16 ReLU	0.3	1 Sigmoid	-	-	-

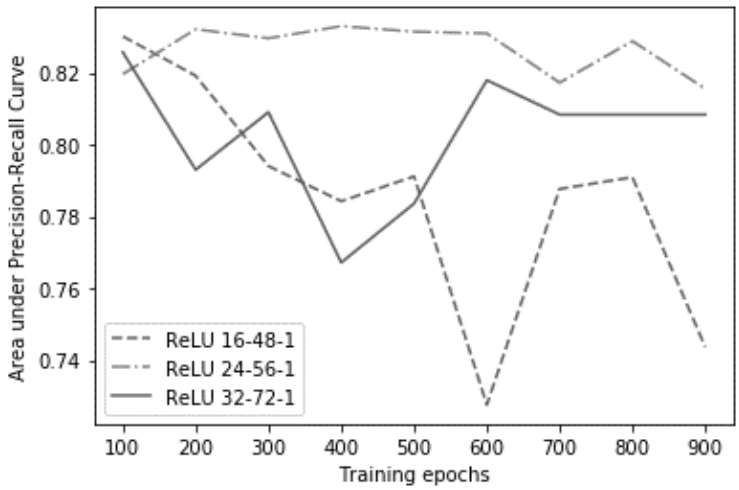
32-72-1	32-72-1	32-72-1	32-72-1	32-72-1	32-72-1	32-72-1	32-72-1	24-56-1	16-48-1
32 tanh	32 ReLU	32 tanh	32 ReLU	32 tanh	32 ReLU	32 tanh	32 ReLU	24 ReLU	16 ReLU
0.4	0.4	0.3	0.3	0.2	0.2	-	-	-	-
72 tanh	72 ReLU	72 tanh	72 ReLU	72 tanh	72 ReLU	72 tanh	72 ReLU	56 ReLU	48 ReLU
0.4	0.4	0.3	0.3	0.2	0.2	-	-	-	-
1 Sigmoid	1 Sigmoid	1 Sigmoid	1 Sigmoid	1 Sigmoid	1 Sigmoid	1 Sigmoid	1 Sig-moid	1 Sig-moid	1 Sig-moid
-	-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-	-
-	-	-	-	-	-	-	-	-	-

В допълнение на по-конвенционалните архитектури, в рамките на изследването бе тествана и архитектура, която наподобява специален вид невронни мрежи – т.нар. „Generative Adversarial Networks" (GANs). Идеята зад тези архитектури е да се обучават двойка мрежи – генератор и дискриминатор. Генераторът има за цел да генерира фалшиви данни, които да заблудят дискриминатора, докато дискриминаторът има за цел обратното – да идентифицира фалшивите данни, когато те са добавени към истински набор от данни. Експериментите в рамките на настоящото изследване използваха обучената генеративна мрежа като източник на допълнително генерирани трансакции със статистическите свойства на позитивния клас, т.е. като техника за овърсамплинг.

4. Резултати

От проведените тестове на дълбоки невронни мрежи без дропаут слоеве, най добрите резултати върху тестовото множество бяха постигнати от мрежата 24-56-1, т.е. мрежа с два скрити слоя, съответно с 24 и 56 неврона. Тази мрежа постигна *AUPRC* от около 0.832. Нейните резултати бяха почти достигнати от мрежата с параметри 32-72-1, която в рамките на първите 100 епохи на обучение достигна *AUPRC* от 0.826.

Последната тествана мрежа – 16-48-1 – не постигна конвергенция в рамките на същия брой епохи на обучение и постигна най-добрите си резултати рано в периода на обучение, което е възможна индикация, че мрежата страда от прекомерно нагаждане.



Фигура 4. AUPRC на невронни мрежи с различна архитектура без дропаут слоеве. Показаните резултати са за 900 епохи (т.е. 900 пълно преминаване на обучителния набор през мрежата)

Таблица 2

AUROC и AUPRC на невронни мрежи с различна архитектура без дропаут слоеве

Епоха	32-72-1		24-56-1		16-48-1	
	AUROC	**AUPRC**	**AUROC**	**AUPRC**	**AUROC**	**AUPRC**
100	0.914374	0.82566	0.914351	0.819608	0.934536	0.830124
200	0.898185	0.792962	0.918538	0.832111	0.906322	0.819154
300	0.90227	0.809008	0.902314	0.829592	0.902226	0.794012
400	0.881918	0.767143	0.902324	0.832983	0.898154	0.784205
500	0.894121	0.783554	0.890156	0.831443	0.902245	0.791125
600	0.882019	0.817907	0.902327	0.830965	0.869674	0.727522
700	0.890112	0.808326	0.886074	0.817228	0.861683	0.787625
800	0.890112	0.808326	0.902324	0.828813	0.886027	0.790888
900	0.890112	0.808326	0.906352	0.815527	0.857543	0.743642

Добавянето на дропаут слоеве, т.е. слоеве, които при всяка итерация симулира изключването на определен брой случайно избрани неврони от предходния слой на мрежата, е основен метод за противодействие на прекомерното нагаждане при дълбоките невронни мрежи. Добавянето на такива слоеве в рамките на проведените експерименти бе осъществено, като след всеки скрит слой на тестваните мрежи бе добавен дропаут слой. Всички тествани архитектури използваха дропаут слоеве с еднакъв относителен дял на блокираните неврони във всеки слой на мрежата, като бяха тествани различни стойности за този относителен дял.

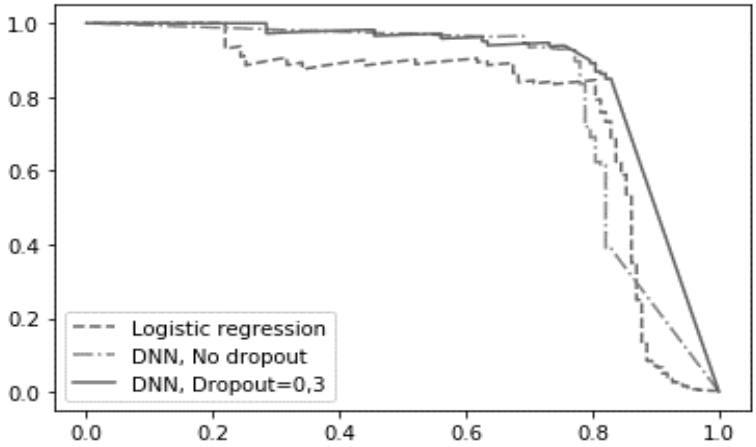
Добавянето на дропаут слоеве към мрежата 32-72-1 произведе най-добри резултати от всички проведени експерименти. От тестваните сценарии, с добавяне на 20%, 30% и 40% дропаут, най-добри резултати произведе експериментът с 30% дропаут. Тази стойност често се използва с успех при други задачи, поради което получените резултати бяха до голяма степен очаквани.

Мрежата 32-72-1 с 30% дропаут постигна *AUPRC* 0.877 върху тестовото множество в рамките на първите 200 епохи от обучението. През оставащите 700 епохи класификационната способност на мрежата постепенно деградира, вероятно поради засилване на тенденцията към прекомерно нагаждане.

Увеличаването на процента на дропаут до 40% не доведе до особен спад на класификационната способност на мрежата, която постигна AUPRC 0.874 отново в рамките на първите 200 епохи на обучението. Намаляването на дропаут слоевете до 20% доведе до *AUPRC* от 0.853.

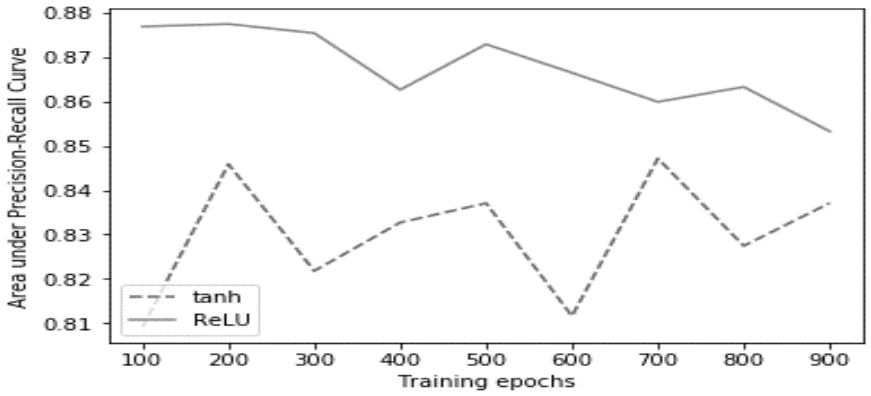
За да се сравнят получените резултати с получените от по-традиционни методи, на този етап бяха извършени експерименти с използването на традиционна „плитка“ невронна мрежа с един скрит слой от 32 неврона, както и с използването на логистична регресия.

Резултатите от използването на плитка невронна мрежа достигнаха AUPRC от 0.81 за 900 епохи на обучение. Логистичната регресия се справи по-лошо с класификационната задача, постигайки *AUPRC* 0.76.



Фигура 5. Сравнение на Precision-Recall кривите на логистичната регресия (AUPRC 0.76), на дълбока невронна мрежа 32-72-1 с дропаут 0.3 (AUPRC 0.877 и на същата мрежа без дропаут (AUPRC 0.826)

В рамките на изследването на различни активационни функции, описаните по-горе експерименти бяха проведени и с използването на хиперболичен тангенс. Получените резултати във всички случаи бяха *по-лоши* от тези, получени при използването на ReLU. На Фигура 6 са показани стойностите на *AUPRC* на най-добрата (според резултатите с ReLU активация) архитектура 32-72-1, когато за активационни функции са използвани ReLU и tanh.



Фигура 6. Сравнение на ReLU (AUPRC 0.877) с tanh (AUPRC 0.847) в ролята на активационни функции на мрежа 32-72-1 с дропаут 0.3 и активационни функции съответно ReLU и tanh

Таблица 3

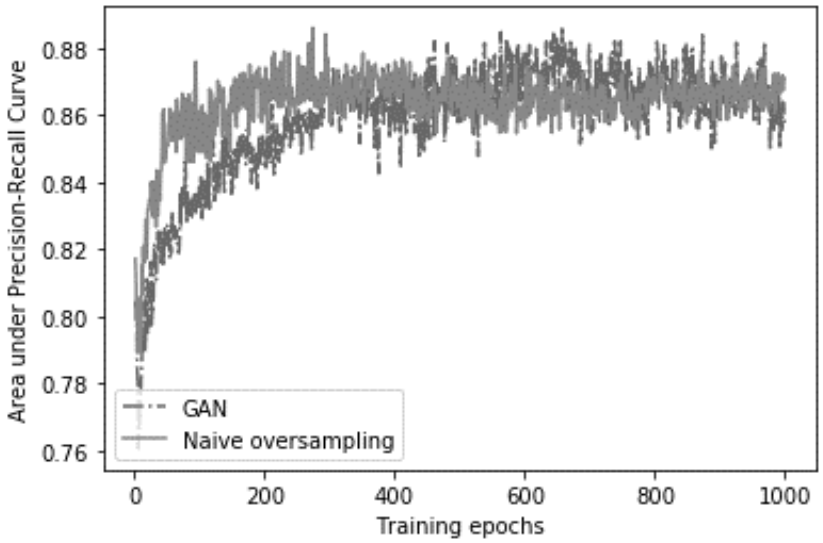
AUROC и AUPRC на невронна мрежа 32-72-2 с 0.3 дропаут и различни активационни функции

Епохи	ReLU		tanh	
	AUROC	**AUPRC**	**AUROC**	**AUPRC**
100	0.914579	0.876852	0.972447	0.809117
200	0.914583	0.87742	0.964365	0.845838
300	0.910522	0.87536	0.948091	0.821799
400	0.902386	0.862593	0.941748	0.832697
500	0.906457	0.872853	0.942297	0.837039
600	0.906452	0.866421	0.930932	0.811621
700	0.902379	0.859869	0.934543	0.847108
800	0.898325	0.863231	0.922654	0.827431
900	0.902368	0.853224	0.914584	0.83707

За да бъде направен опит за по-нататъшно подобрение на получените резултати, най-добрата архитектура бе комбинирана с два подхода за овърсамплинг. За да се получи база за сравнение, първо бяха извършени експерименти с „наивен“ овърсамплинг, при който членовете на позитивния клас (т.е. фалшивите трансакции) бяха повторени многократно, като по този начин на практика бе извършено балансиране на набора от данни. За да се постигне балансирането, наборът от данни бе разделен на пакети от 200 легитимни и 200 фалшиви трансакции, избрани на случаен принцип от обучителното множество. За да бъде постигнат приблизително същият брой итерации като при другите експерименти, обучението на мрежата по този начин бе извършено в 1000 епохи от по 1000 пакета.

Като алтернативен подход бе тестван, повлиян от GAN, модел, който прилага принципа на генеративното съперничество по същия начин. За целта бяха генерирани две допълнителни мрежи. Т.нар. „мрежа–генератор“ бе отговорна за генерирането на трансакции, които наподобяват позитивния клас (фалшивите трансакции). Бе обучена и „мрежа–дискриминатор“, която бе отговорна за разпознаване на „истинските“ фалшиви трансакции от тези, генерирани с помощта на мрежата–генератор, като за целта периодично двете мрежи бяха захранвани с истинска информация от набора от данни.

Двете мрежи бяха обучавани заедно в хода на множество епохи, което би следвало да помогне на генератора да създава все по-реалистично изглеждащи членове на позитивния клас. В края на обучението генераторът бе използван за генериране на допълнителни трансакции, които да бъдат вмъкнати наред с фалшивите в оригиналния набор от данни, постигайки по този начин търсения баланс.

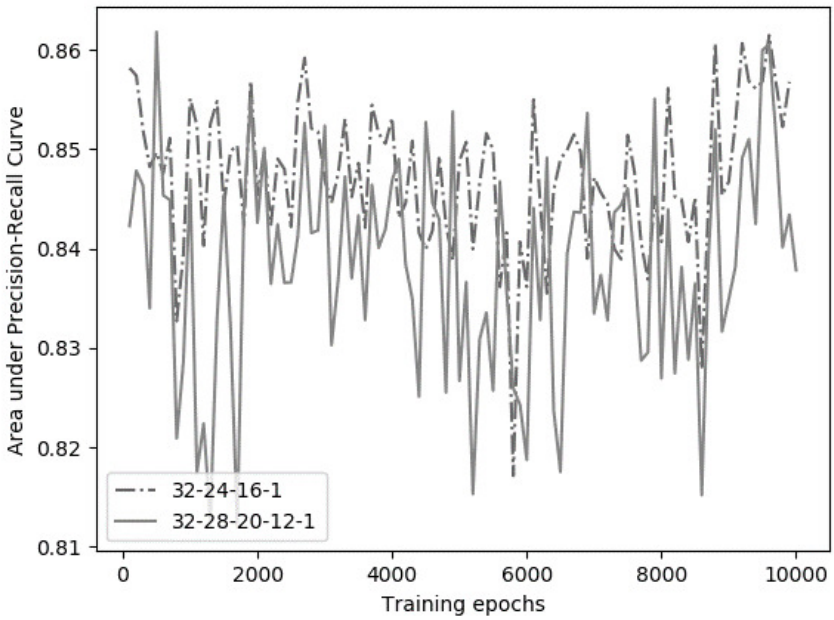


Фигура 6. Сравнение на наивен овърсамплинг (AUPRC 0.886) с повлиян от GAN подход (AUPRC 0.8856) при мрежа 32-72-1 с активационна функция ReLU и дропаут 0.3

Комбинирана с наивен овърсамплинг, мрежата 32-72-1 постигна максимален *AUPRC* от 0.886. (фиг. 7). Резултатите, получени чрез генеративния подход, постигнаха максимален *AUPRC* от 0.8856. И двата резултата представляват подобрение на максималните резултати, получени без използването на овърсамплинг (*AUPRC* 0.877), но разликата е твърде малка, което най-малко поставя под въпрос по-добрата ефективност на тези техники, когато те бъдат използвани с конкретни фиксирани стойности на Recall/Sensitivity параметъра.

Както бе споменато, един от най-важните въпроси при изследването бе свързан с това, как увеличаването на дълбочината на мрежата повлиява нейните предиктивни способности. За установяване на това в рамките на изследването бяха проведени експерименти с увеличаването на дълбочината на слоевете. Тъй като експериментите с активационните функции и добавянето на дропаут слоеве при мрежите с дълбочина два скрити слоя до голяма степен потвърдиха теоретичните очаквания, експериментите върху мрежи с дълбочина три и четири слоя се концентрираха върху мрежи, имащи ReLU активационна функция и добавени слоеве с 0.3 дропаут.

За да се компенсира възможно забавяне в процеса на обучение, броят на епохите при тези експерименти бе увеличен до 10000, като производителността на обучените мрежи бе тествана всеки 100 епохи. Получените резултати (вж. фиг. 7) бяха много близки и не демонстрираха съществено подобрение спрямо мрежите с 2 скрити слоя.



Фигура 7. Мрежа с дълбочина от 3 (AUPRC 0.8614) и мрежа с дълбочина от 4 (AUPRC 0.8612) скрити слоя

Заключение

Направените експерименти показват, че използването на дропаут слоеве, които блокират между 30 и 40 процента от невроните във всеки слой в хода на всяка итерация, дава най-добри резултати при борба с прекомерното нагаждане, характерно за наборите от данни, свързани с измамите с кредитни карти. В допълнение, използването на ReLU за активационна функция води до отчетливо по-добри резултати при обучението на мрежата, което вероятно се дължи на това, че функцията е неподатлива на проблема с изчезващите градиенти.

Получените резултати за техниките за овърсамплинг са доста противоречиви. Използването на наивен подход, при който позитивният клас се повтаря механично, дава известно малко подобрение, което е сравнимо с по-сложни техники, като описаните в студията. Получените резултати обаче валидират използването на генеративното съперничество като техника за овърсамплинг. Разбира се, за да се твърди това с по-голяма степен на сигурност, трябва да бъдат проведени още множество изследвания върху различни набори от данни, които по необходимост ще имат известна разнородност в понятията на обхванати характеристики на трансакциите (колони на обучителния набор).

По най-важния въпрос, който касае основната архитектурна характеристика на невронните мрежи – тяхната дълбочина, получените резултати водят до заключението, че дълбоките невронни мрежи с малко до среден брой слоеве изглеждат най-подходящи при решаване на задачи, свързани с разпознаването на фалшиви трансакции с кредитни карти. Генералният характер на този извод произтича от факта, че ефективната дълбочина на невронните мрежи е ограничена от характера и количеството на данните, като особено голямо значение има тяхната размерност, т.е. броят на колоните на набора от данни.

Направените експерименти засягат таблица от данни с 32 колони и приблизително 280000 реда, която се доближава по мащаби до най-големите набори от данни, използвани при разпознаване на фалшиви трансакции. Поради експоненциалната зависимост между увеличаването на броя на скритите слоеве и нужните данни, необходими за тяхното обучение, може да се направи обоснованото предположение, че невронните мрежи, използвани за откриването на измами с кредитни карти, никога няма да достигнат дълбочините, използвани например в областта на разпознаването на изображения, където малък набор от изображения със скромна разделителна способност с лекота би генерирал стотици пъти по-голям входен набор от данни.

Извършените експерименти и направеният анализ повдигат въпроси от по-общ характер, като например това дали в областта на бизнеса и икономиката въобще има област (с възможното изключение на финансовите пазари), която би генерирала достатъчно големи набори от данни, върху които могат да бъдат използвани ефективно невронни мрежи от голяма дълбочина. Тези и подобни въпроси, свързани със съдържанието на понятието „големи данни“, тепърва трябва да бъдат изследвани.

Използвани източници

(Kaggle 2019a). Извлечено от Kaggle: https://www.kaggle.com/joparga3/in-depth-skewed-data-classif-93-recall-acc-now

(Kaggle 2019b). Извлечено от Kaggle: https://www.kaggle.com/arathee2/achieving-100-accuracy

European Central Bank. (2018). *Fifth report on card fraud, September 2018.* EuroSystem. Извлечено от https://www.ecb.europa.eu/pub/cardfraud/html/ecb.cardfraudreport201809.en.html

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Bengio, Y. (2014). Generative Adversarial Networks. *https://arxiv.org/abs/1406.2661*.

Javelin. (2019). *2019 Identity Fraud Study: Fraudsters Seek New Targets and Victims Bear the Brunt.* Извлечено от

https://www.javelinstrategy.com/coverage-area/2019-identity-fraud-report-fraudsters-seek-new-targets-and-victims-bear-brunt

Kingma, D., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. *https://arxiv.org/abs/1412.6980*.

KPMG. (2019). *Global Banking Fraud Survey.*

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Rabinovich, A. (2014). Going Deeper with Convolutions. *arXiv:1409.4842*.

Маринова, К. (2017). Състояние и перспективи на мобилните и електронните разплащания в България. *Бизнес управление*(1), 42-59. Извлечено от http://bm.uni-svishtov.bg/title.asp?title=621

Парушева, С. (2015). Картово-безналичните измами – предизвикателства и противодействие. *Народностопански архив*(2).

ИНСТИТУТ ЗА НАУЧНИ ИЗСЛЕДВАНИЯ
ПРИ СТОПАНСКА АКАДЕМИЯ
„Д. А. ЦЕНОВ" - СВИЩОВ

АЛМАНАХ
НАУЧНИ ИЗСЛЕДВАНИЯ

ИНСТИТУЦИИ, ПОЛИТИКИ И ПРЕДИЗВИКАТЕЛСТВА ПРЕД ДИГИТАЛНАТА ТРАНСФОРМАЦИЯ

том 28, 2020 г.

Академично издателство „ЦЕНОВ"
Свищов - 2020 г.

СТОПАНСКА АКАДЕМИЯ „Д. А. ЦЕНОВ”

АЛМАНАХ НАУЧНИ ИЗСЛЕДВАНИЯ

ТОМ 28

ИНСТИТУЦИИ, ПОЛИТИКИ И ПРЕДИЗВИКАТЕЛСТВА ПРЕД ДИГИТАЛНАТА ТРАНСФОРМАЦИЯ

Даден за печат на 27.02.2020 г., излязъл от печат на 30.03.2020 г.
Поръчка № 18460, тираж: 100 бр.

Издателство и печат: Академично издателство „Ценов”
Свищов, ул. Градево № 24

ISSN 1312-3815

СЪДЪРЖАНИЕ

Раздел III
Финансова стабилност, икономически политики, регулации и устойчиво развитие