Využitie genetického algoritmu na výber funkcií v oblasti behaviorálnej biometrie

Jozef Varga

Fakulta informatiky a informačných technológií STU v Bratislave, Ilkovičova 2, 842 16 Bratislava 4

Abstrakt Výber funkcií je v strojovom učení náročnou kombinatorickou úlohou. Funkcie, ktoré neskôr vstupujú do rôznych modelov v strojovom učení, majú veľký vplyv na následnú presnosť a rýchlosť samotného modelu. Toto je jeden z problémov, ktorý sa vyskytuje pri autorizovaní užívateľa pomocou behavioralnej biomerie. Práve pri zaznamenávaní dát o užívateľovy, získame veľké množstvo atributov, ktoré znižujú rýchlosť výpočtu a taktiež degradujú kvalitu klasifikácie.
V tomto článku sme vytvorily genetický algoritmus, ktorý používame práve na výber funkcií. Nami vytvorenú metódu porovnávame s bežnými metódami ako sú Variance Inflation Factor (VIF) a vyber funkcií pomocou korelácií. Zvolené metódy boli porovnané pomocou rôznych klasifikátorov ako sú napríklad K-Nearest Neighbors (KNN), SVM, Naive Bayes a iné. V práci sme využili verejný dataset [1,10] ktorý obsahuje biometrické dáta o užívateľoch. Tieto dáta obsahujú informácie o tom v akom stave sa nachádza užívateľ, teda či leží, sedí, kráča, kráča hore schodmy, kráča dole schodmy alebo stojí. V tejto práci sa ukázalo, že použitie genetického algoritmu zlepšilo výsledky jednotlivých klasifikátorov.

Keywords: Behavioralna biometria · Genetický algoritmus · Strojové učenie · Výber funkcií / atributov.

1 Úvod

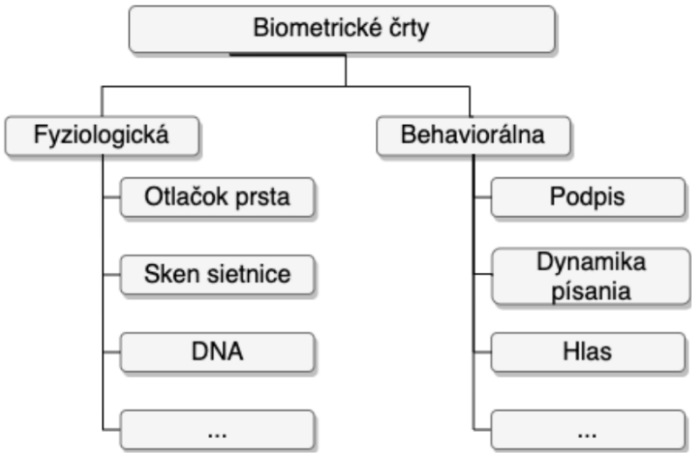
V dnešnej dobe je využitie smartfónov na vzostupe. Veľké množstvo ľudí tieto zariadenia využíva aj na vyhľadávanie informácií, úpravu dokumentov a rôzne ďalšie funkcionality, ktoré tieto zariadenia podporujú [3]. To, že sú smartfóny veľkou súčasťou technologického sveta potvrdzuje aj štatistika z roku 2018, ktorá hovorí, že až 52,2% zobrazení globálnych webových stránok prebiehalo pomocou mobilných zariadení. Zaujímavé je, že v tomto prípade ide o nárast až 51,5% oproti štatistike z roku 2009 [8]. Mobilné zariadenie prináša veľmi veľa nových možností výskumu ktoré sú orientované na správanie človeka. Dôvodom je hlavne veľké množstvo informácií ktoré sú možné vďaka senzorom v mobilnom zariadení zaznamenať.

Tieto senzory zaznamenávajú takzvané biometricke črty. Biometria sa zakladá na charakteristických črtách konkrétnej osoby. Tieto biometrické črty sa

delia na dva typy, a to na fyziologické a behaviorálne (viď. Obr. 1). [9] Fyziologická biometria, ako napríklad odtlačok prsta, je veľmi často využívanou overovacou technikou, keďže je jedinečná a stála. Táto biometria má však aj svoje nevýhody. DNA, čo je typickým príkladom tejto biometrie, je dosť invazívne na to, aby sa využívalo napríklad na smartfónoch, keďže jej využitie by vyžadovalo určitú podmienenú činnosť užívateľa, napríklad odber krvi. Naopak behaviorálna biometria má výhodu v možnosti práce na pozadí bez akejkoľvek určitej požadovanej činnosti užívateľa. Keďže sa jedná o charakteristiky, ako napríklad rýchlosť písania, nakláňanie smartfónu alebo chôdza, používateľove správanie môže byť overené počas bežných činností. [9] Výhodou behaviorálnych čŕt je teda fakt, že ak už zariadenie disponuje potrebným hardvérom, v prípade mobilných zariadeniach napríklad senzormi, je možné vykonávať rôzne opreácie ako napríklad autentifikáciu užívateľa alebo identifikovanie stavu v akom sa užívateľ nachádza bez toho, aby o tom samotný užívateľ vedel. Naopak nevýhodou tejto biometrie je, že behaviorálne charakteristiky užívateľa sa časom menia, a teda je potrebné častejšie vytvorenie profilu užívateľa [7].

Jedným z problémov ktorý sa objavuje pri prácach s behaviorálnimy črtamy, je množstvo atributov, ktoré sa získavajú zo senzorov. [5] Získané atributy môžu obsahovať irelevantné/zavádzajúce informácie, ktoré môžu mať za následok zníženie kvality klasifikácie modelu. [2,4,5,6,13]

Práve výberom atributov sa budeme v tejto práci bližšie zaoberať. Naša metóda na výber atributov spočíva vo využití genetického algoritmu, ktorý bude aplikovaný na výber čŕt v datasete ktorý sa venuje skúmaniu stavu užívateľa (leží, chodí, sedí) na základe biometrických čŕt získaných z mobilného zariadenia. [1,10]



Obr. 1. Typy biometrických čŕt [9]

2 Podobné práce

Výber atributou je veľmi podstatnou časťou pri strojovom učení. Veľmi veľa prác sa zaoberá rôznymi metódami ktoré by boli nápomocné pri spracovávaní rôznich datasetov. V práci [5] experimentovali nad dvoma databázami, pričom prvá databáza mala 43 atributov a 231 záznamov. Druhá databáza obsahovala 60 atributov a 7555 záznamov. Dáta si rozdelily na testovaciu a trénovaciu množinu v pomere 66:33. Teda 66% bolo dát bolo určených na trénovanie modelu a 33% dát na jeho testovanie. V práci využívajú na klasifikáciu modely K-Nearest Neighbors (KNN), SVM a Naive Bease (viď. Tab. 1). Na konci experimentu zhodnotili že využitie genetického algoritmu na vyber atributov pri rôznych klasifikátoroch mal veľmi priaznivé účinky práve na SVM a to v zlepšení až o 14,55% v prvom datasete a 11,77% v druhom datasete. Keďže zlepšenie bolo viditeľne vidieť pri každom zo spomenutých klasifikátorov, v práci usúdili že využitie genetických algoritmov v tejto problematike zvyšuje presnosť klasifikátora.

Tabuľka 1. Accuracy (pesnosť ACC) modelov v percentách výberu funkcií. [5]

	Bez výberu funkcií		
Databáza	**KNN**	**SVM**	**Naive Bayes**
Da Silva et al. 2016	72.73 ±2.74	73.55 ±2.63	55.54±3.54
Giot et al. 2009	86.72 ±0.58	78.28 ±0.67	69.33 ±0.67
Databáza	**Využitie genetického algoritmu na výber funkcií**		
Da Silva et al. 2016	87.85±0.84	88.10 ±0.90	81.64 ±0.97
Giot et al. 2009	88.86 ±0.23	90.05±0.41	77.44 ±0.27

V ďalšej práci[2] využili na výber atributov tri rôzne algoritmy a to genetický algoritmus, WEKA-CFS a WEKA-ranker. Na vyhodnotenie taktiež použili viacero klasifikátorov a to Multi-Layer Perceptron (MLP), Random Forest (RF), J48, Naive Bayes a regresiu. Najlepšie výsledky boli získané pomocou MLP. Ako najlepšia funkcia na výber atributov sa ukázal práve genetický algoritmus, ktorý má výhodu práve v širokej možnosti nastavovania.

Nasledujúca práca[13] využila genetický algoritmus na optimalizáciu nie len výberu funkcií, ale aj výberu parametrov pre SVM klasifikátor. Výsledky im ukázali , že využitie tohto algoritmu, nie len zlepšilo výsledky klasifikácie, ale oproti grid searchu aj zrýchlilo výpočet o 2,62 sekundy. Presnosť klasifikácie sa zvýšila až o 3,57%. 83,14 ± 7,19 boli najnižšie získané hodnoty s využitím výberu funkcií založených na genetickom algoritme.

Posledná skúmaná práca [6] sa zaoberala predovsetkým klasifikátorom C4.5. Skúmali ako vplýva genetický algoritmus, ktorého fitnes funkcia pozostáva zo stromu, na výber funkcií aj u iných algoritmov. Teda neskúšali využívať vo fitnes funkcii len klasifikátor s ktorým neskôr prebieha klasifikácia. Zistili že takáto fitnes funkcia, taktiež zlepšuje aj iné modely. Ich výsledky ukázali že ak rovnakú funkciu použili na Naive Bayes tak jeho klasifikácia sa zlepšila o 4,92 %.

Naš vytvorený genetický algoritmus chceme porovnať aj s existujúcimi algoritmamy na výber funkcií a to variance inflation factor (VIF) a výber na základe korelácií. Využiť chceme na fitnes funkciu len jeden algoritmus ako buďe spomenuté v nasledujúcej kapitole.

3 Využívané algoritmy na výber funkcií

V tejto práci sme sa rozhodli porovnať genetický algoritmus na výber funkcií s algoritmamy Variance Inflation Factor (VIF) a Výber funkcií pomocou korelácií. Tieto algoritmy sa radia medzi základné algoritmy na výber funkcií. [12]

3.1 Variance Inflation Factor

Tento algoritmus odstranuje funkcie ktoré majú veľmmi silnú kolinearitu. VIF odhaduje do akej miery je rozptyl koeficientu zväčšený kôli linearnej závislosti s inim prediktorom. Výpočet VIF pre každý stĺpec sa dá získať vykonaním lineárnej regresie tohoto stĺpca so všetkými ostatnými stĺpcami.[12] Vzorec pre výpočet VIF vyzerá nasledujúco:

$$VIF_i = \frac{1}{1 - R_i^2} \tag{1}$$

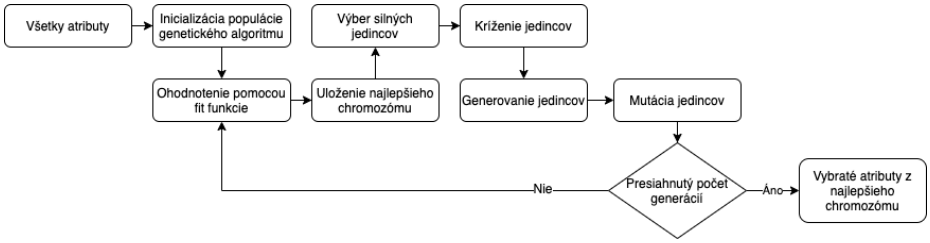
V uvedenom vzorci R_i^2 predstavuje koeficint regresie medzi i-tou premennou a všetkými ostatnými premennými. VIF sa pri výbere funkcí porovnáva a odstranujú sa hraničné hodnoty.[12]

3.2 Výber funkcií pomocou korelácií

Funkcie sú vyberaná pomocou korelačného koeficientu. Koeficient korelácie je miera linearnej intenzity korelácie. Najčastejšie je využívaný Pearsonov korelačný koeficient. Tento koeficient nadobúda hodnoty medzi 1 až -1. 1 vo výsledku predstavuje perfektnú pozitívnu koreláciu (podobnosť). Naopak -1 predstavuje perfektńu negatívnu koreláciu (podobnosť). 0 označuje, že medzi prvkami nie je žiadna korelácia / podobnosť. Ak prebieha výbere funkcií touto metódou, vďaka výsledkom korelácie sa odstráňuju veľmi podobné funkcie.[12]

3.3 Genetický algoritmus

Genetické algoritmy sú inšpirované evolúciou. Teda ich základ sa skladá z generácií, ktoré sa vyvíjajú a obsahujú určité množstvo chromozómv, ktoré sa spolu volajú populácia. Tieto chromozómy obsahujú kritické informácie potrebné na vyriešenie problému. Tieto algoritmy sa využívaju najme na optimalizáciu rôznich problémov. Ich uplatnenie je v praxi veľmi široké.[2,11] Nami vytvorený algoritmu začína inicializáciou populácie genetického algoritmu (viď. Obr. 2).



Obr. 2. Schéma genetického algoritmu na výber funkcií

Inicializácia populácie genetického algoritmu. Pri inicializácií sa nastavia prvotné atributy genetického algoritmu (viď. Tab. 2). V algoritme je chromozóm reprezentovaný n génmy. Jeden gén prezentuje 1 alebo 0 ktorá hovorý či vo výslednom vektore funkcií sa táto funkcia nachádza alebo nie. Počet tých to génov n je počtom funkcií ktoré sú vo vstupnom vektore. Teda koľko funkcií má daný dataset.

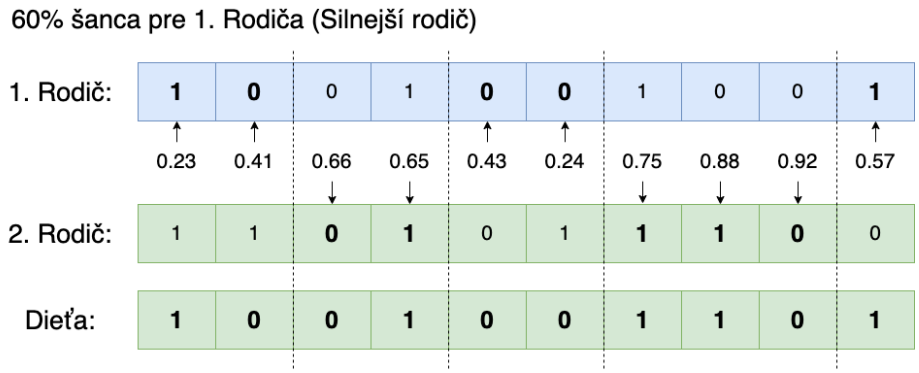
Tabuľka 2. Nastavenie genetického algoritmu

Parametre GA	**hodnota**
Fit funkcia založená na modely	K-Nearest Neighbors
Reprezentácia génu	Boolean
Maximálny počet generácií	20
Veľkosť populácie	100
Chromozómy vytvorené krížením	30
Chromozómy vytvorené generovaním	30
Najlepšie chromozómy z minulej generácie	20
Percentuálna šanca mutácie a kríženia	60%

Ohodnotenie populácie pomocou fit funkcie. Populácia ktorá sa dostane do tohoto kroku, musí byť ohodnoťená pomocou fit (fitnes) funkcie. Táto funkcia musí odlišovať jednotlivé chromozómy tak, aby bolo možné ich porovnať a zistiť, ktorý z týchto chromozómov je najúspešnejší. Táto fit funkcia pozostáva z priemeru ďesať násobnej krížovej validácie ktorej jadro je vytvorené z KNN algoritmu. Najlepšie hodnotený chromozóm sa uloží a algoritmus pokračuje na výber jedincov (chromozómov) do ďalšej generácie.

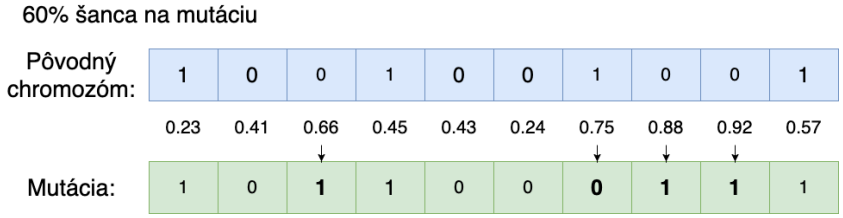
Výber silných jedincov. Tento spôsob sa nazýva elitárstvo. Teda koľko najlepších chromozómov z predchádzajúcej generácie prejde do novej generácie. Výhodou je najme to, že silné jedince sa nadalej reprodukujú a predávajú svoje vlastnosti novo vytvoreným chromozómom.

Kríženie jedincov. Pri kŕžení sa spájajú vlastnosti dvoch jedincov do jedného (viď. Obr. 3). Do kríženia vstupujú dva chromozómy, ktoré majú zrkadlové postavenie v zoradenom liste podľa fit funkcie. Jeden chromozóm je teda z časti kde je fit funkcia najsilnejšia a jeden z časti kde je fit funkcia naopak slabšia. Tento spôsob by mal pomôcť algoritmu aby nezastal na lokálnom maxime. Následne sa prechádza po génoch jednotlivých chromozómov a podľa percentuálnej náhody sa vyberie jeden z génov do budúceho jedinca. Šanca s akou sa jednotlivé gény volia, ako aj počet takto vytvorených jedincov sa v genetickom algoritme nastavuje pomocou parametrov (viď. Tab. 2).



Obr. 3. Kríženie chromozómov

Zvyšné chromozómy sa náhodne vygenerujú ako pri inicializácií aby počet chromozómov v každej generácií bol rovnaky. Ako posledné prejde algoritmus k mutácií. Tá prejde všetky chromozómy v zadanej populácií a prechádza jednotlivé gény. Nad každým génom vygenerujé náhodné číslo na základe ktorého sa s 60% šancov zmení gén na opačný (viď. Obr. 4).



Obr. 4. Mutovanie chromozómov

4 Experimenty

Zvolená konfigurácia genetického algoritmu zobrazená v tabuľke 2 bola zvolená na základe pilotných testov nad rôznimy datami. V experimente využijeme dataset, ktorý obsahuje behaviorálne záznamy z mobilného zariadenia [1,10]. Tento dataset slúži na klasifikovanie stavu používateľa. Teda či používateľ leží, sedí, kráča, kráča hore schodmy, kráča dole schodmy alebo stojí. Dataset obsahuje informácie o 30 dobrovolníkoch vo veku 19 až 48 rokov. Každý používateľ vykonal spomenutých 6 aktivít pri ktorých mal na páse pripevnený smartfón Samsung Galaxy S II. Dataset obsahuje informácie z gyroskopu a akcelerometra. Z nich sa zaznamenávalo 3-osové lineárne zrýchlenie a 3-osové uhlové rýchlosti pri konštantnej rýchlosti 50 Hz. V datasete sa nachádza 3609 záznamov. Pre každý záznam v datasete sa poskytuje vektorom s 561 funkciami, ktoré obsahujú:

- Trojosové zrýchlenie z akcelerometra a odhadované zrýchlenie tela
- Trojosová uhlová rýchlosť z gyroskopu
- Označenie činnosti
- Označenie konkrétneho subjektu

Nad týmto datasetom sa snažíme overiť a porovnať silu genetického algoritmu oproti štandardným metódam. Porovnávame jeho získané accurancy skóre so skóre, ktoré získa algoritmus VIF alebo analýza na základe korelácií. Skóre sa vypočítavalo ako fit funkcia. Teda 10 násobnou krížovou validaciou. Modely ktoré sme využili v experimente boli AdaBoost, Decision Tree, Extra Trees, Naive Bayes, Nearest Neighbors, Linear Discriminant Analysis, Logistic Regression, Neural Net, Random Forest, SVM Sigmoid, SVM RBF, QDA.

5 Vyhodnotenie

Výsledky experimentu sú zobrazené v tabuľke 4 a tabuľke 3 . Využitie genetického algoritmu nám v tomto experimente ukázalo, že časovo najnáročnejšie na predspracovanie bol algoritmus VIF. Oproti GA bol tento algoritmus až 13 krát pomalší. Ak sa poredpokladá zlepšenie GA zvýšením počtom generácií, vieme povedať, že výsledky GA je ešte možné zlepšiť zvýšením jeho časového zaťaženia. Čo sa týka počtu výberu atributov, Každý z použitých algoritmuv znížil počet atributov o viac ako polovicu, čím sa znížil čas následného trénovania. Najviac atributov vyfiltroval VIF a to až o 70 atributov viac ako GA. Zaujímavosťou je práve výsledné skóre GA, ktoré potvrdzuje, že výberom správneho algoritmu do fit funkcie sa výrazne zvýši výsledná klasifikácia modelu, ktorý je použitý ako základ tejto funkcie. Získané výsledky taktiež ukazujú, že nie je potrebné aby sa zhodovalo jadro fit funkcie s modelom, ktorý je klasifikátorom. Toto je vidieť hlavne pri modeloch Neural Net a QDA, kde aj pri použití fit funkcie s jadrom Nearest Neighbors sa zvýšilo skóre klasifikácie týchto modelov o viac ako 15% oproti klasifikácii nad neupravenými atributmi. Najväčšie zlepšenie je vidieť na algoritme Nearest Neighbors, kde sa klasifikácia zlepšila o 78%. Dôvodom je práve využitie tohoto algoritmu vo fit funkcii. Výsledky taktiež ukazujú, že ak

bol VIF lepší oproti GA, tak rozdiel bol najviac 1% okrem QDA kde bol rozdiel 4% a Naive Bayes s rozdielom 8%. Tieto rozdieli môžu byť zanedbateľné v prípade ak je potrebné znížiť časové žaťaženie na predspraccovanie keďže VIF je 13 krat pomalší.

Tabuľka 3. Počet vybratých atributov

Algoritmus	GA	VIF	CORR	Bez funkcie výberu
Počet atributov	260	190	255	562

Tabuľka 4. Výsledky genetického algoritmu

Model	**ACC**			**Najlepší**
	GA	**VIF**	**CORR**	**výsledok**
AdaBoost	0.54	0.55	0.54	VIF
Decision Tree	0.84	0.85	0.84	VIF
Extra Trees	0.77	0.70	0.68	GA
Naive Bayes	0.79	0.87	0.81	VIF
Nearest Neighbors	0.92	0.14	0.14	GA
Linear Discriminant Analysis	0.95	0.95	0.94	GA, VIF
Logistic Regression	0.95	0.96	0.95	VIF
Neural Net	0.94	0.71	0.69	GA
Random Forest	0.73	0.70	0.70	GA
SVM Sigmoid	0.19	0.19	0.19	Všetky
SVM RBF	0.21	0.19	0.19	GA
QDA	0.87	0.91	0.87	VIF

6 Záver

Ako už bolo uvedené vo vyhodnotení, experiment sa uskutočnil nad existujúcim datasetom, ktorý obsahoval behaviorálne dáta o stavoch používateľov [1,10]. Dataset bol vyhodnocovaný pomocou 11 modelov, a 10 násobnej krížovej validácie. Podľa výsledkov uvedených v tabuľke 4 a tabuľke 3 je možné pozorovať, že využitie rovnakého klasifikátora ako je použitý vo fit funkci je veľmi účinné a zvýšuje klasifikáciu. Taktiež znižuje počet atributov, čo má za následok zrýchlenie následnej klasifikácie. Taktiež je vidieť že je možné zvýšiť klasifikáciu využitím genetického algoritmu oproti klasifikácii nat neupravenými dátamy.

Proces výberu atribútov je jedným z najdôležitejších krokov v predspracovaní údajov. Datasety obsahujúce behaviorálne dáta obsahujú veľmi veľa atribútov, čo môže vyžadovať vysokú výpočtovú silu. V našom experimente je taktiež vidieť že GA znížilo počet atributov o viac ako polovicu s tým že zvýšil presnosť klasifikátoru. Preto je práve výber atributov, ktoré sú podstatné pre klasifikáciu,

môže zvýšiť presnosť klasifikácie ako je ukázane vyžšie. Použitie genetických algoritmov sa zdá byť lepšie, aj oproti VIF vďaka nižšej záťaži. Najlepší klasifikátor pre náš dataset bol Logistic Redression s 96% úspešnosťou. Medzi najlepšie tiež patria modely Linear Discriminant Analysis s 95%, Neural Net s 94% a Nearest Neighbors s úspešnosťou 92%.

Ďalšie pokračovanie tejto práci by bolo zamerané na zmeny fit funkcie, a sledovať možné ovplivnenie ostatných modelov.

Literatúra

1. Anguita, D. et al.: A public domain dataset for human activity recognition using smartphones. ESANN 2013 proceedings, 21st Eur. Symp. Artif. Neural Networks, Comput. Intell. Mach. Learn. April, 437–442 (2013).
2. Babatunde, O. et al.: A Genetic Algorithm-Based Feature Selection. Int. J. Electron. Commun. Comput. Eng. 5, 4, 899–905 (2014).
3. Bomhold, C.R.: Educational use of smart phone technology: A survey of mobile phone application use by undergraduate university students. Program. 47, 4, 424–436 (2013). https://doi.org/10.1108/PROG-01-2013-0003.
4. Lu, H. et al.: A hybrid feature selection algorithm for gene expression data classification. Neurocomputing. 256, 2017, 56–62 (2017). https://doi.org/10.1016/j.neucom.2016.07.080.
5. Nascimento, L.D.O. et al.: An investigation of genetic algorithm-based feature selection techniques applied to keystroke dynamics biometrics An investigation of genetic algorithm-based feature selection techniques applied to keystroke dynamics biometrics. SBSeg, 2–5 (2019).
6. Smith, M.G., Bull, L.: Genetic programming with a genetic algorithm for feature construction and selection. Genet. Program. Evolvable Mach. 6, 3, 265–281 (2005). https://doi.org/10.1007/s10710-005-2988-7.
7. Syed, Z. et al.: Touch gesture-based authentication on mobile devices: The effects of user posture, device size, configuration, and inter-session variability. J. Syst. Softw. 149, 158–173 (2019). https://doi.org/10.1016/j.jss.2018.11.017.
8. statista.com, 2018. Percentage of all global web pages served to mobile phones from 2009 to 2018. `https://www.statista.com/statistics/241462/global-mobile-phone-website-traffic-share/`. Posledny prístup 10 Marca 2020
9. Teh, P.S. et al.: A survey of keystroke dynamics biometrics. Sci. World J. 2013, (2013). https://doi.org/10.1155/2013/408280.
10. UC Irvine Machine Learning Repository, `https://archive.ics.uci.edu/ml/datasets/human+activity+recognition+using+smartphones`. Posledny prístup 10 Marca 2020
11. Whitley, D.: A genetic algorithm tutorial. Stat. Comput. 4, 2, 65–85 (1994). https://doi.org/10.1007/BF00175354.
12. Xu, J. et al.: Methods for performing dimensionality reduction in hyperspectral image classification. (2018). https://doi.org/10.1177/0967033518756175.
13. Zhao, M. et al.: Feature selection and parameter optimization for support vector machines: A new approach based on genetic algorithm with feature chromosomes. Expert Syst. Appl. 38, 5, 5197–5204 (2011). https://doi.org/10.1016/j.eswa.2010.10.041.