

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 behaviorálnej biometrie. Práve pri zaznamenávaní dát o užívateľovi, získame veľké množstvo atributov, ktoré znižujú rýchlosť výpočtu a taktiež degradujú kvalitu klasifikácie.

V tomto článku sme vytvorili 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 KNN, SVM, Naive Bayes a iné. V práci sme využili verejný dataset [1,13] ktorý obsahuje biometrické dáta o užívateľoch. Tieto dáta obsahujú informácie o tom v akom stave sa nachádza užívateľ, teda či sedí: 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

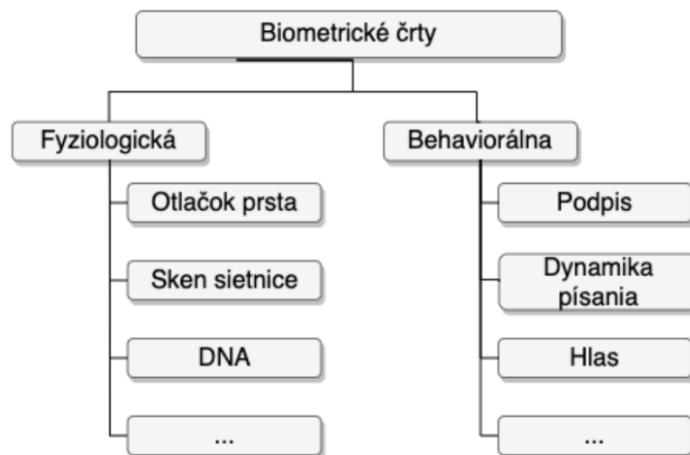
V dnešnej dobe je využitie smartfónov na vzostupe. Veľké množstvo ľudí tieto zariadenia využíva aj na vyhľadávanie informácií, úpravu dokumentov a rôzne ďalšie funkcionality, ktoré tieto zariadenia podporujú [3]. To, že sú smartfóny veľkou súčasťou technologického sveta potvrdzuje aj štatistika z roku 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 [11]. 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 biometrické 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). [12] 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. [12] Výhodou behaviorálnych črt 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 [10].

Jedným z problémov ktorý sa objavuje pri prácach s behaviorálnymi črtami, je množstvo atributov, ktoré sa získavajú zo senzorov. [8] Získané atributy môžu obsahovať irelevantné/zavádzajúce informácie, ktoré môžu mať za následok zníženie kvality klasifikácie modelu. [2,7,8,9,15]

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 črt v datasete ktorý sa venuje skúmaniu stavu užívateľa (leží, chodí, sedí) na základe biometrických črt získaných z mobilného zariadenia. [1,13]



Obr. 1. Typy biometrických črt [12]

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 rôznych datasetov. V práci [8] experimentovali nad dvoma databázami, pričom prvá databáza mala 43 atributov a 231 záznamov. Druhá databáza obsahovala 43 atributov a 7555 záznamov. Dáta si rozdelili na testovaciu a trénovaciu množinu v pomere 66:33. Teda 66% bolo dát určených na trénovanie modelu a 33% dát na jeho trénovanie. V práci využívajú na klasifikáciu modely KNN, SVM a Naive Bayes (viď. Tab. 1). Na konci experimentu zhodnotili že využitie genetického algoritmu na výber 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 (presnosť ACC) modelov v percentách výberu funkcií. [8]

| Databáza | Bez výberu funkcií | | |
|----------------------|---|------------------|------------------|
| | KNN | SVM | Naive Bayes |
| Da Silva et al. 2016 | 72.73 \pm 2.74 | 73.55 \pm 2.63 | 55.54 \pm 3.54 |
| Giot et al. 2009 | 86.72 \pm 0.58 | 78.28 \pm 0.67 | 69.33 \pm 0.67 |
| Databáza | Využitie genetického algoritmu na výber funkcií | | |
| Da Silva et al. 2016 | 87.85 \pm 0.84 | 88.10 \pm 0.90 | 81.64 \pm 0.97 |
| Giot et al. 2009 | 88.86 \pm 0.23 | 90.05 \pm 0.41 | 77.44 \pm 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[15] využila genetický algoritmus na optimalizáciu nie len výberu funkcií, ale aj výberu parametrov pre SVM klasifikátor. Výsledky im ukázali, že využitie tohoto algoritmu, nie len zlepšilo výsledky klasifikácie, ale oproti využitiu grid searchu aj zrýchlilo výpočet o 2,62 sekundy bez využitiu výberu funkcií a o 1,21 sekundy rýchlejšia ak sa výber funkcií využil. Presnosť klasifikácie sa zvýšila až o 3,57%. 83,14 \pm 7,19 boli najnižšie získané hodnoty s využitím výberu funkcií založených na genetickom algoritme.

Posledná skúmaná práca [9] sa zaoberala predovšetkým klasifikátorom C4.5. Skúmali ako vplýva genetický algoritmus, ktorého fitness funkcia pozostáva na strome, na výber funkcií aj u iných algoritmov. Teda neskúšali využívať vo fitness funkcii len klasifikátor s ktorým neskôr prebieha klasifikácia. Zistili že takáto fitness funkcia, taktiež zlepšuje aj iné modely. Ich výsledky sa ukázali ak rovnakú funkciu použili na Naive Bayes kde sa klasifikácia zlepšila o 4,92 %.

3 Využitie algoritmy

3.1 Variance Inflation Factor

3.2 Výber funkcií na pomocou korelácií

3.3 Genetický algoritmus

4 Experimenty

5 Vyhodnotenie

6 Záver

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. Fröhlich, H. et al.: Feature Selection for Support Vector Machines by Means of Genetic Algorithms. Proc. Int. Conf. Tools with Artif. Intell. 142–148 (2003). <https://doi.org/10.1142/s0218213004001818>.
5. Ghareb, A.S. et al.: Hybrid feature selection based on enhanced genetic algorithm for text categorization. Expert Syst. Appl. 49, 31–47 (2016). <https://doi.org/10.1016/j.eswa.2015.12.004>.
6. Khan, M.A. et al.: Erratum: An optimized method for segmentation and classification of apple diseases based on strong correlation and genetic algorithm based feature selection (EEE Access (2019) 7 (46261-46277)). IEEE Access. 8, 36514 (2020). <https://doi.org/10.1109/ACCESS.2020.2974161>.
7. 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>.
8. 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).
9. 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>.
10. 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>.
11. 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

12. Teh, P.S. et al.: A survey of keystroke dynamics biometrics. *Sci. World J.* 2013, (2013). <https://doi.org/10.1155/2013/408280>.
13. UC Irvine Machine Learning Repository, <https://archive.ics.uci.edu/ml/datasets/human+activity+recognition+using+smartphones>. Posledny pristup 10 Marca 2020
14. Wójcik, M.P. et al.: Measuring Cognitive Workload in Arithmetic Tasks Based on Response Time and EEG Features Architecture and Technology : Proceedings of 38th International Con- Measuring Cognitive Workload in Arithmetic Tasks Based on Response Time and EEG Features. *Proc. 38th Int. Conf. Inf. Syst. Archit. Technol.* 2, November 2017, (2018). <https://doi.org/10.1007/978-3-319-67220-5>.
15. 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>.