Ελληνική Εταιρία Ανάλυσης Δεδομένων
Πρακτικά 9ου Πανελληνίου Συνεδρίου Ανάλυσης Δεδομένων (2017), σελ. 208–215

Παράμετροι για τον εντοπισμό φύλου αγνώστων χρηστών του Διαδικτύου

Ιωάννης Τσιμπερίδης, Αυγερινός Αραμπατζής, Αλέξανδρος Καράκος
Τμήμα Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών
Δημοκρίτειο Πανεπιστήμιο Θράκης

ΠΕΡΙΛΗΨΗ

Η πλήρης ανωνυμία που μπορεί να διατηρήσει ένα χρήστης στο Διαδίκτυο, εκτός από τα ευεργετικά οφέλη, αποτελεί τη σημαντικότερη αιτία κακόβουλων ενεργειών. Με βάση την παρατήρηση ότι το κείμενο είναι το κύριο μέσο επικοινωνίας μεταξύ χρηστών, η παρούσα έρευνα προτείνει την άρση αυτής της πλήρους ανωνυμίας με δεδομένα που προκύπτουν αποκλειστικά από τον τρόπο που πληκτρολογεί ένας χρήστης, ενώ ταυτόχρονα διασφαλίζει την προστασία των προσωπικών και ευαίσθητων δεδομένων του. Χρησιμοποιώντας μία τεχνική για την επιλογή των καταλληλότερων παραμέτρων και ελέγχοντας την προτεινόμενη διαδικασία με πέντε γνωστά μοντέλα ταξινόμησης, αποδεικνύεται ότι είναι εφικτή η δημιουργία αξιόπιστων και ευέλικτων συστημάτων για την αναγνώριση του φύλου ενός χρήστη του Διαδικτύου.

1. ΕΙΣΑΓΩΓΗ

Οι βιομετρικές τεχνολογίες που βασίζονται στη συμπεριφορά των χρηστών προσφέρουν πολλά πλεονεκτήματα έναντι των φυσικών βιομετρικών τεχνολογιών. Ανάμεσά τους είναι η συλλογή δεδομένων χωρίς παρενόχληση των χρηστών και χωρίς την απαίτηση για πρόσθετο υλικό. Οι προσπάθειες για αναζήτηση των χαρακτηριστικών χρήστη χρησιμοποιώντας συμπεριφορικές βιομετρικές παραμέτρους εστιάζουν κυρίως στο φύλο και την ηλικία, με αντιπροσωπευτικές εργασίες στον τομέα αυτό να παρουσιάζονται ακολούθως.

Οι Li κ.α. (2008) προσδιόρισαν το φύλο ενός ατόμου από τον τρόπο βαδίσματος με ακρίβεια 90%, με τη βοήθεια ενός ταξινομητή SVM. Όσον

αφορά την ανθρώπινη φωνή, οι Barkana και Zhou (2015) χρησιμοποιώντας δύο ταξινομητές πέτυχαν ποσοστό ορθής πρόβλεψης 63%. Οι Peersman κ.α. (2011) με δεδομένα από ένα βελγικό κοινωνικό δίκτυο κατέληξαν σε 66% ποσοστό επιτυχίας στην ταξινόμηση ηλικίας και φύλου. Οι Sboev κ.α. (2016) χρησιμοποίησαν ένα σύνολο κειμένων και προσδιόρισαν το φύλο συγγραφέα με ακρίβεια 86%. Η εύρεση του φύλου ατόμου μπορεί να επιτευχθεί και με άλλες μεθόδους, όπως από φωτογραφίες προσώπου. Έτσι, οι Eidinger κ.α. (2014) και οι Kalansuriya και Dharmaratne (2014), παρουσιάζοντας νέες τεχνικές, πέτυχαν 88% και 86% ακρίβεια, αντίστοιχα.

Ωστόσο, οι προαναφερθείσες προσεγγίσεις εμφανίζουν περιορισμούς που εμποδίζουν τη γενίκευση της χρήσης τους. Για παράδειγμα, απαιτούν φωτογραφία προσώπου ή πληκτρολόγηση σε συγκεκριμένη γλώσσα, οπότε έρχονται αντιμέτωπες με την προσπάθεια απόκρυψης χαρακτηριστικών από τον χρήστη, ή με την ετερογένεια του σημερινού Internet. Είναι προφανές ότι η χρήση της δυναμικής της πληκτρολόγησης μπορεί να θεωρηθεί ως καταλληλότερη για ένα τέτοιο πρόβλημα. Αυτό συμβαίνει επειδή η επικοινωνία μέσω κειμένου παραμένει ο κυρίαρχος τρόπος επικοινωνίας, ενώ η επικοινωνία μεταξύ κακόβουλου χρήστη και θύματος, σε περιπτώσεις όπως η αποπλάνηση ανηλίκων, οι απειλές, κλπ, γίνεται μέσω κειμένου.

Δυναμική της πληκτρολόγησης είναι η λεπτομερής καταγραφή των ενεργειών ενός χρήστη επί του πληκτρολογίου. Οι παράμετροί της είναι χρονικές και μη χρονικές. Οι πιο αποδεκτές χρονικές παράμετροι είναι η διάρκεια πατήματος πλήκτρου (ο χρόνος πίεσης ενός πλήκτρου) και ο λανθάνων χρόνος διγράμματος (ο χρόνος για τη χρήση δύο διαδοχικών πλήκτρων). Άλλες χρονικές παράμετροι αναφέρονται στη δουλειά των Giot κ.α. (2011). Στις μη χρονικές παραμέτρους συγκαταλέγονται η ταχύτητα πληκτρολόγησης, η συχνότητα και ο τρόπος διόρθωσης σφαλμάτων, η χρήση ορισμένων πλήκτρων (Monaco κ.α., 2012), κ.α. Γίνεται σαφές ότι η δυναμική της πληκτρολόγησης συνοδεύεται από μεγάλο αριθμό παραμέτρων, και επομένως για να μειωθεί η πολυπλοκότητα και το υπολογιστικό κόστος, είναι απαραίτητο να ακολουθηθεί μια διαδικασία επιλογής παραμέτρων. Ο τύπος του προβλήματος και η επιλογή παραμέτρων συνδέονται άμεσα. Για παράδειγμα, οι

Tsimperidis κ.α. (2015) επέλεξαν τις πιο συχνά εμφανιζόμενες παραμέτρους προσπαθώντας να προσδιορίσουν το φύλο ενός χρήστη από το μικρότερο δυνατό κείμενο.

Η παρούσα εργασία επιχειρεί την αναγνώριση φύλου ενός άγνωστου χρήστη με δεδομένα που προέρχονται από τον τρόπο που πληκτρολογεί. Αναζητείται ο καλύτερος συνδυασμός μεταξύ ποσοστού ορθής πρόβλεψης και χρόνου εκπαίδευσης, κάτι που επιτυγχάνεται τροποποιώντας τον αριθμό των παραμέτρων που εμπλέκονται. Η αποτελεσματικότητα της προτεινόμενης προσέγγισης αποδεικνύεται από ένα σύνολο πειραμάτων και από όσο γνωρίζουμε, παρουσιάζει το υψηλότερο ποσοστό ορθής πρόβλεψης στην αντίστοιχη βιβλιογραφία.

Στο υπόλοιπο της εργασίας, στην Ενότητα 2 περιγράφεται η απόκτηση δεδομένων, η εξαγωγή και επιλογή παραμέτρων. Στην Ενότητα 3 συνοψίζονται τα αποτελέσματα σύγκρισης επιδόσεων πέντε μοντέλων μάθησης μηχανής. Συγκεκριμένα των μηχανή διανυσμάτων υποστήριξης (SVM), τυχαίο δάσος (RF), ταξινομητής Naïve Bayes (NB), νευρωνικό δίκτυο με συνάρτηση ακτινωτής βάσης (RBFN), και πολυστρωματικός perceptron (multi-layer perceptron, MLP). Στην Ενότητα 4 εξετάζονται τα αποτελέσματα και τελικά στην Ενότητα 5 συνοψίζεται το άρθρο.

2. ΜΕΘΟΔΟΛΟΓΙΑ

Η μεθοδολογία αποτελέστηκε από τρεις διαδοχικές φάσεις. Στην πρώτη συλλέχθηκαν δεδομένα, στη δεύτερη επιλέχθηκαν παράμετροι σύμφωνα με την ταξινόμησή τους ως προς το κέρδος πληροφορίας, και στην τρίτη οι επιδόσεις των SVM, RF, NB, MLP και RBFN, συγκρίθηκαν ως προς το ποσοστό ορθής πρόβλεψης και τη χρονική πολυπλοκότητα.

2.1 Σύνολο Δεδομένων Δυναμικής της Πληκτρολόγησης

Η καταγραφή της πληκτρολόγησης ενός εθελοντή εμπεριέχει κινδύνους αποκάλυψης προσωπικών ή και ευαίσθητων δεδομένων. Αυτός είναι ο κύριος λόγος για την έλλειψη τέτοιων συνόλων δεδομένων στη βιβλιογραφία. Για τη δημιουργία ενός νέου σύνολο δεδομένων δυναμικής της πληκτρολόγησης,

σχεδιάσαμε έναν keylogger ελεύθερου κειμένου. Για να μετριαστούν οι επιβλαβείς επιπτώσεις, λήφθηκαν όλα τα απαραίτητα μέτρα και δεσμεύσεις.

Ο keylogger δημιουργεί αρχεία txt, με κάθε ενέργεια του εθελοντή να αποτυπώνεται σε μία εγγραφή. Από αυτά τα αρχεία κειμένου είναι εφικτή η εξαγωγή των περισσότερων παραμέτρων της δυναμικής της πληκτρολόγησης. Για τις ανάγκες της παρούσας έρευνας, υπολογίστηκαν οι διάρκειες πατήματος πλήκτρου και οι λανθάνοντες χρόνοι διγράμματος. Μετά από μια περίοδο επιστράτευσης εθελοντών και συλλογής αρχείων, δημιουργήθηκε ένα σύνολο δεδομένων με 248 αρχεία καταγραφής (125 από άντρες και 123 από γυναίκες).

2.2 Επιλογή Παραμέτρων

Για την εξαγωγή παραμέτρων αναπτύχθηκε λογισμικό το οποίο διαβάζει αρχεία από τον keylogger και εκτελεί τους απαραίτητους υπολογισμούς. Οι παράμετροι που εξήχθησαν τελικώς υπερέβαιναν τις 10.000 και επομένως έπρεπε να ακολουθηθεί μια διαδικασία επιλογής.

Ένας τρόπος για την επιλογή των κατάλληλων παραμέτρων είναι μέσω του υπολογισμού του κέρδους πληροφορίας (information gain, IG) τους. Η διαδικασία περιγράφεται λεπτομερώς στο άρθρο των Sharma και Dey (2012), και εάν εφαρμοστεί για κάθε εκμαιευμένη παράμετρο, θα δημιουργηθεί μία λίστα με το κέρδος πληροφορίας που φέρει κάθε μία από αυτές. Ένα μέρος αυτής της λίστας, με τις 15 πρώτες παραμέτρους με το υψηλότερο IG, παρουσιάζεται στον Πίνακα 1.

Πίνακας 1. Παράμετροι της δυναμικής της πληκτρολόγησης με το υψηλότερο IG.

#	Παράμ.	IG	#	Παράμ.	IG	#	Παράμ.	IG
1	N-A	0.0897	6	T-O	0.0593	11	E-I	0.0543
2	M-O	0.0815	7	A	0.0584	12	O-N	0.0536
3	K-A	0.0706	8	A-S	0.0553	13	E-(spacebar)	0.0526
4	R-I	0.0647	9	D	0.0550	14	P-A	0.0503
5	M-A	0.0612	10	I-A	0.0545	15	T-E	0.0458

3. ΠΕΙΡΑΜΑΤΑ ΚΑΙ ΑΠΟΤΕΛΕΣΜΑΤΑ

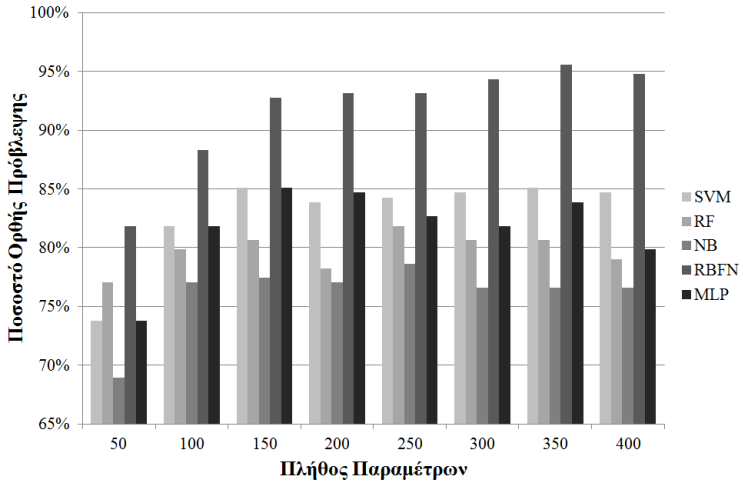
Οι αποδόσεις των μοντέλων μάθησης μηχανής, αξιολογήθηκαν με τη χρήση συνόλων δεδομένων με διαφορετικό αριθμό παραμέτρων, που προέκυψαν από το αρχικό σύνολο δεδομένων. Τα κριτήρια για να βρεθεί το σύνολο παραμέτρων που οδηγεί σε σύστημα με τις καλύτερες επιδόσεις ήταν η ακρίβεια (accuracy) και η χρονική πολυπλοκότητα (TBM).

Εκτελέστηκαν πολυάριθμα πειράματα για κάθε σύνολο δεδομένων από 50 έως και 400 παραμέτρους, με βήμα των 50 παραμέτρων, και εντοπίστηκαν οι βέλτιστες διαμορφώσεις για όλα τα μοντέλα σε κάθε μία από τις οκτώ περιπτώσεις. Στον Πίνακα 2 παρουσιάζεται η απόδοση των 5 μοντέλων στα 8 διαφορετικά σύνολα δεδομένων, με το υψηλότερο ποσοστό ορθής πρόβλεψης να εμφανίζεται έντονο και υπογραμμισμένο, για κάθε ένα από τα μοντέλα.

Πίνακας 2. Η απόδοση των 5 μοντέλων στα 8 διαφορετικά σύνολα δεδομένων.

	SVM		**RF**		**NB**		**RBFN**		**MLP**	
Feat.	**Acc.**	**TBM**	**Acc.**	**TBM**	**Acc.**	**TBM**	**Acc.**	**TBM**	**Acc.**	**TBM**
50	73,8%	0,16	77,0%	1,00	69,0%	0,03	81,9%	0,53	73,8%	8,55
100	81,9%	0,13	79,8%	2,65	77,0%	0,02	88,3%	0,73	81,9%	31,33
150	**85,1%**	0,16	80,7%	2,03	77,4%	0,18	92,7%	2,43	**85,1%**	73,47
200	83,9%	0,19	78,2%	4,60	77,0%	0,02	93,2%	2,95	84,7%	120,43
250	84,3%	0,22	**81,9%**	6,65	**78,6%**	0,09	93,2%	3,68	82,7%	181,93
300	84,7%	0,33	80,7%	6,15	76,6%	0,08	94,4%	4,31	81,9%	274,20
350	**85,1%**	0,31	80,7%	8,22	76,6%	0,02	**95,6%**	4,89	83,9%	373,90
400	84,7%	0,42	79,0%	8,14	76,6%	0,02	94,8%	5,46	79,8%	509,65

Το σχήμα 1 απεικονίζει το ποσοστό ορθής πρόβλεψης των πέντε μοντέλων στα διάφορα σύνολα δεδομένων με διαφορετικό αριθμό παραμέτρων, από όπου φαίνεται ότι το RBFN έχει πάντα την υψηλότερη ακρίβεια και ότι το NB έχει πάντα τη χαμηλότερη.



Σχήμα 1. Ποσοστό ορθής πρόβλεψης των 5 μοντέλων για διαφορετικό σύνολο παραμέτρων.

4. ΑΠΟΤΙΜΗΣΗ

Από τα αποτελέσματα εξήχθησαν τρία βασικά συμπεράσματα. Πρώτον, φαίνεται ότι όλα τα μοντέλα επιτυγχάνουν την υψηλότερη ακρίβεια πριν χρησιμοποιήσουν το μέγιστο αριθμό παραμέτρων. Αυτή είναι μια σημαντική ένδειξη αφού δεν είναι απαραίτητο για ένα σύστημα να χρησιμοποιεί πολύ μεγάλο αριθμό παραμέτρων δυναμικής της πληκτρολόγησης για να φτάσει στη μέγιστη ακρίβεια.

Δεύτερον, τα δοκιμασμένα μοντέλα φαίνεται να έχουν σχεδόν σταθερή ακρίβεια για τα σύνολα δεδομένων από 150 έως 350 παραμέτρους. Το SVM έχει ακρίβεια 84,5 ±0,6%, το RF έχει 80,0 ±1,8%, το NB έχει 77,6 ±1,0%, το RBFN έχει 94,2 ±1,4%, και το MLP έχει 83,5 ±1,6%. Αυτό υποδεικνύει ότι είναι δυνατή η υλοποίηση συστημάτων που θα λειτουργούν αξιόπιστα, ακόμη και αν κάποιες παράμετροι απουσιάζουν από τα διαθέσιμα δεδομένα.

Τρίτον, στην περίπτωση των 350 παραμέτρων, το μοντέλο RBFN προβλέπει σωστά το φύλο ενός άγνωστου χρήστη σε 19 από τις 20 φορές. Σύμφωνα με όσα γνωρίζουμε, το ποσοστό του 95,6% είναι το υψηλότερο στην πρόβλεψη φύλου χρηστών χρησιμοποιώντας τη δυναμική της πληκτρολόγησης, στη βιβλιογραφία. Αυτό σημαίνει ότι είναι δυνατόν να αναπτυχθούν αρκετά ακριβή συστήματα τα οποία εκτελούν αναγνώριση φύλου με δεδομένα που

προέρχονται μόνο από την απλούστερη μορφή επικοινωνίας μεταξύ χρηστών, το κείμενο.

5. ΣΥΝΟΨΗ

Πολλές φορές είναι απαραίτητο να γνωρίζουμε κάποια χαρακτηριστικά ενός χρήστη του Διαδικτύου, για θέματα ασφάλειας ή για καλύτερη εκμετάλλευση των προσφερόμενων υπηρεσιών. Οι υπάρχουσες μέθοδοι αναγνώρισης φύλου είτε απαιτούν συγκεκριμένα δεδομένα, είτε είναι παρεμβατικές. Αντίθετα, η δυναμική της πληκτρολόγησης παρέχει μια μη παρεμβατική μέθοδο χαμηλού κόστους από δεδομένα που προέρχονται μόνο από τον απλούστερο και συνηθέστερο τρόπο επικοινωνίας μεταξύ χρηστών. Με βάση αυτή την ιδέα, η παρούσα μελέτη παρουσιάζει μια διαδικασία στην οποία επιλέγονται οι καταλληλότερες παράμετροι για τον προσδιορισμό φύλου. Για να επιτευχθεί ο στόχος δημιουργήθηκε ένα νέο σύνολο δεδομένων από την καταγραφή χρηστών κατά την καθημερινή χρήση των υπολογιστών τους. Στη συνέχεια υπολογίστηκε το κέρδος πληροφορίας για κάθε παράμετρο και βάσει της κατάταξής τους δημιουργήθηκαν νέα σύνολα δεδομένων με διαφορετικό αριθμό παραμέτρων. Με αυτά τα δεδομένα και με πέντε γνωστά μοντέλα ταξινόμησης, τα αποτελέσματα έδειξαν ότι είναι δυνατή η δημιουργία αρκετά αξιόπιστων συστημάτων που αναγνωρίζουν το φύλο ενός άγνωστου χρήστη με ακρίβεια 95,6%, με μόνο μερικές εκατοντάδες παραμέτρους.

ABSTRACT

The complete anonymity of a user on the Internet sometimes is beneficial, but in the same time is the main cause of malicious actions. Based on the observation that text is the primary means of communication between users, this research proposes to remove this complete anonymity with data derived exclusively from the way a user types, while ensuring the protection of his/her personal and sensitive data. Using a technique to select the most appropriate features and testing the proposed process with five well-known classification models, it turns out that it is possible to create quite reliable systems for identifying the gender of an Internet user.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Barkana, B.D., & Zhou, J. (2015). A new pitch-range based feature set for a speaker's age and gender classification. *Appl Acoust*, 98, 52–61.

Eidinger E., Enbar, R., & Hassner, T. (2014). Age and gender estimation of unfiltered faces. *IEEE Trans Inf Forensics Secur*; *9*(12), 2170–9.

Giot, R., El-Abed, M., & Rosenberger, C. (2011). Keystroke dynamics overview. In: Yang J, editor. In *Biometrics*. InTech, 157–182.

Kalansuriya, T.R., Dharmaratne, A.T. (2014). Neural network based age and gender classification for facial images. *Int J Adv ICT Emerg Reg*, *7*(2), 1–10.

Li, X., Maybank, S.J., Yan, S, Tao, D., & Xu, D. (2008). Gait components and their application to gender recognition. *IEEE Trans Syst, Man, Cybern — Part C: Appl Rev*, *38*(2), 145–55.

Monaco, J.V., Bakelman, N., Cha, S., & Tappert, C.C. (2012). Developing a keystroke biometric system for continual authentication of computer users. In: *Proceedings of the 2012 European Intelligence and Security Informatics Conference*. Washington, DC, USA, pp. 210–6.

Peersman, C., Daelemans, W., & Van Vaerenbergh, L. (2011). Predicting age and gender in online social networks. In: *Proceedings of the 3rd International Workshop on Search and Mining User-Generated Contents*. Glasgow, Scotland, UK, 37–44.

Sboev, A., Voronina, I., Litvinova, T., Dmitry Gudovskikh, D., & Rybka, R. (2016). Deep learning network models to categorize texts according to author's gender and to identify text sentiment. In: *Proceedings of the 2016 International Conference on Computational Science and Computational Intelligence*. Las Vegas, NV, USA, 1101–6.

Sharma A, & Dey, S. (2012). Performance investigation of feature selection methods and sentiment lexicons for sentiment analysis. *Int J Comput Appl Special Issue Adv Comput Commun Technol HPC Appl*, 3, 15–20.

Tsimperidis I, Katos V, & Clarke N. (2015). Language-independent gender identification through keystroke analysis. *Inf Comput Secur, 23*(3), 286–301.