Învăţare Automată în Vederea Artificială şi Procesarea Şirurilor de Caractere (Rezumat)

Conducător Ştiinţific:

Prof. Dr. Denis Enăchescu

Radu Tudor Ionescu

Departamentul de Informatică

Facultatea de Matematică şi Informatică

Universitatea din Bucureşti

Bucureşti, Decembrie 2013

Rezumat

Învăţarea automată reprezintă o arie largă de cercetare ce are applicaţii în multe domenii, precum vederea artificială, bioinformatica, regăsirea informaţiei, procesarea limbajului natural, procesarea semnalelor, data mining, şi multe altele. În varietatea metodelor de învăţare automată propuse până în prezent, se numără metodele de învăţare bazate pe similaritate. Învăţarea bazată pe similaritate se referă la procesul de învăţare prin folosirea similarităţii între perechi de exemple de antrenare. Procesul de învăţare bazat pe similaritate poate fi atât supervizat, cât şi nesupervizat, iar relaţia dintre perechi poate fi dată print-o măsură de similaritate sau una de disimilaritate, sau chiar printr-o funcţie de distanţă.

Această teză studiaza o serie de metode de învăţare bazate pe similaritate, cum ar fi metoda celor mai apropiaţi vecini (Nearest Neighbor), metodele de tip nucleu (kernel), şi algoritmii de clustering. În această teză este prezentată o metodă de tip Nearest Neighbor bazată pe o nouă măsură de disimiaritate pentru imagini. Metoda este aplicată pentru recunoaşterea caracterelor scrisă de mână, obţinând rezultate foarte bune din punct de vedere al acurateţii. Metodele de tip kernel sunt utilizate pentru mai multe probleme abordate în această teză. În primul rând, în teză este introdusă o nouă funcţie kernel ce poate fi utilizată pentru histograme de *cuvinte vizuale* (visual words). Folosind această funcţie kernel, metodele de tip *sac de cuvinte vizuale* (bag of visual words) obţin o performanţă foarte bună în cazul recunoaşterii obiectelor în imagini. În al doilea rând, în teză sunt prezentate şi câteva metode de tip nucleu bazate pe reprezentarea imaginilor sub formă de structură piramidală. Aceste metode sunt folosite pentru

recunoaşterea expersiei faciale în imagini cu persoane care exprimă diferite stări de spirit. În al treilea rând, în teză este descrisă o abordare bazată pe *funcţii de tip nucleu pentru şiruri de caractere* (string kernels) cu aplicaţii în identificarea limbii native în texte scrise în limba engleză de către persoane care nu sunt vorbitoare native de limba engleză. Abordarea propusă în teză obţine cea mai bună performanţă până în prezent pentru identificarea limbii native, fiind în acelaşi timp o abordare independentă de limbă. De asemenea, în această teză sunt studiaţi o serie de algoritmi de clustering. Algoritmii de clustering sunt aplicaţi pentru construirea arborilor filogenetici folosind secvenţe de ADN mitocondrial provenit de la mamifere. Se poate observa imediat că problemele de învăţare automată, prezentate în teza de faţă, se împart în două domenii, anume vederea artificială (computer vision) şi procesarea de nivel înalt a şirurilor de caractere (string processing).

În ciuda faptului că vederea artificială şi procesarea de nivel înalt a şirurilor de caractere par a fi domenii de studiu complet diferite, analiza imaginilor şi a textelor este asemănătoare sub multe aspecte. Aşa cum urmează a fi arătat în această teză, conceptul de a trata imaginile şi şirurile de caractere (în special textele) într-un mod asemănător s-a dovedit a fi foarte fertil şi productiv pentru anumite aplicaţii din vederea artificială. De fapt, una din metodele de ultimă oră folosite pentru clasificarea imaginilor este inspirată de reprezentarea sub formă de *sac de cuvinte* (bag of words) a documentelor text, o abordare extreme de răspândită şi folosită în regăsirea informaţiei şi procesarea limbajului natural. Într-adevăr, modelul *sac de cuvinte vizuale*, ce are la bază contruirea unui vocabular de cuvinte vizuale prin folosirea unei metode de clustering pentru a grupa descriptori locali de imagine, a demonstrat un nivel de acurateţe impresionant pentru clasificarea imaginilor, pentru regăsirea imaginilor, sau pentru alte sarcini asemănătoare. Prin adaptarea tehnicilor folosite în procesarea şirurilor de caractere pentru a analiza imagini, sau prin adaptarea tehnicilor folosite în vederea artificială pentru a analiza

şiruri de caractere, cunoştinţele dintr-un domeniu pot fi transferate către celălalt domeniu. De fapt, multe dintre cele mai importante descoperiri ştiinţifice au fost realizate în urma transferului de cunoştinţe între domenii de studiu diferite. Teza de faţă se încadrează în această direcţie de cercetare, urmărind fie să prezinte noi abordări, fie să dezvolte abordările deja existente, prin transferarea şi adaptarea metodelor între cele două domenii studiate (vederea artificială şi procesarea şirurilor de caractere).

Mai întâi, este prezentată o nouă măsură de disimilaritate pentru compararea imaginilor. Această măsură de disimilaritate este inspirată de distanţa rang (rank distance), folosită pentru compararea şirurilor de caractere. Principala sarcină care trebuie îndeplinită presupune extinderea şi adaptarea distanţei rang pentru a lucra cu intrări bidimensionale (imagini digitale), în loc de intrări unidimensionale (şiruri de caractere). În timp ce distanţa rang este o masură foarte precisă pentru compararea şirurilor de caractere, rezultatele empirice prezentate în această teză sugerează faptul că măsura de disimilaritate ce extinde distanţa rang la imagini are o acurateţe foarte bună atât pentru recunoaşterea cifrelor scrise de mână, cât şi pentru clasificarea şi analiza imaginilor cu texturi.

În al doilea rând, sunt prezentate câteva metode de îmbunătăţire a modelului sac de cuvinte vizuale. Aşa cum s-a menţionat mai sus, acest model este inspirat de modelul sac de cuvinte folosit în procesarea limbajului natural şi în regăsirea informaţiei. Printre îmbunătăţirile aduse acestui model se numără o nouă metodă de tip kernel (denumită PQ kernel), reprezentarea piramidală a imaginilor folosind vectori de prezenţă, dar şi aplicarea *învăţării locale* (local learning).

În al treilea rând, în teza de faţă este introdusă o nouă distanţă pentru şiruri de caractere. Această distanţă este inspirată de noua măsura de disimilaritate pentru imagini descrisă mai sus. Construită astfel încât să se conformeze unor principii mai generale, dar adaptată în acelaşi timp pentru secvenţe de ADN, noua distanţă obţine rezultate

mai bune în comparaţie cu alte metode de ultimă oră folosite pentru analiza secevenţelor ADN. Mai mult, această distanţă este aplicată cu succes şi în altă direcţie, cea a comparării documentelor text. Mai exact, o metodă de tip nucleu, ce are la bază acestă nouă distanţă, este folosită pentru identificarea limbii native în text. În concluzie, toate contribuţiile prezentate în această teză vin să susţină ideea că imaginile şi şirurile de caractere pot fi tratate într-un mod asemănător, cu scopul de a îmbunătăţi rezultatele metodelor de învăţare.

Înainte de a încheia, trebuie atrasă atenţia asupra faptului că metodele studiate şi descrise în această teză obţin o performanţă comparabilă sau uneori mai bună decât metodele de ultimă oră din cele două domenii abordate. În continuare, sunt prezentate câteva argumente care să confirme acest fapt. În primul rând, un model de tip sac de cuvinte vizuale, îmbunătăţit din mai multe puncte de vedere, a obţinut locul al patrulea la competiţia Facial Expression Recognition (FER) Challenge organizată în cadrul workshop-ului ICML 2013 Workshop in Challenges in Representation Learning (WREPL). În al doilea rând, sistemul ce are la bază metode de tip nucleu pentru şiruri de caractere, prezentat în această teză, s-a clasat pe locul al treilea la competiţia Native Language Identification Shared Task organizată în cadrul workshop-ului BEA-8 Workshop of NAACL 2013. În al treilea rând, articolul care introduce funcţia de tip nucleu, denumită PQ kernel, folosită pentru a calcula similaritatea între histograme de cuvinte vizuale, a obţinut premiul Caianiello Best Young Paper Award la conferinţa ICIAP 2013.