Rezumat

În timp ce tot mai multe informaţii în format text devin disponibile, recunoaşterea automată a acestora este dificilă fără o bună indexare şi un rezumat bun al conţinutului documentelor. Catalogarea textului este o soluţie a acestei probleme şi constă în clasificarea documentelor într-un set de categorii predefinite. În ultimi ani au fost aplicate un număr crescând de metode de clasificare (clustering) şi de tehnici de învăţare automată.

Documentele sunt de obicei reprezentate prin vectori rari modelaţi într-un spaţiu vectorial, unde fiecărui cuvânt din vocabular îi corespunde o axă de coordonate şi numărul de apariţii ale cuvântului în document reprezintă valoarea componentei din vector în reprezentarea documentului. Dimensiunea mare a spaţiului trăsăturilor este o problemă majoră în catalogarea textului. Spaţiul trăsăturilor constă din termeni unici care apar în documente. Când modelăm o colecţie de documente spaţiul trăsăturilor poate fi de ordinul zecilor sau sutelor de mii de termeni. Antrenarea clasificatorilor pe o astfel colecţie mare de documente necesită mult timp şi memorie. De aceea se încearcă aplicarea de metode diverse pentru a reduce spaţiul de reprezentare şi timpul de răspuns. Aşa cum vom vedea rezultatele sunt mai bune când lucrăm cu o dimensiune mult redusă a vectorilor. Odată cu creşterea dimensiunii nu va creşte neapărat şi acurateţea clasificării, urmând ca aceasta să scadă substanţial când se utilizează o dimensiune foarte mare sau întreaga dimensiune a spaţiului.

În acest raport voi prezenta un studiu comparativ între câteva metodele de selecţie al trăsăturilor caracteristice (Information Gain, Mutual Information şi Support Vector Machine) precum şi tipul de reprezentare al datelor de intrare în învăţarea folosită pentru clasificarea documentelor. De asemenea voi prezenta o metodă folosită cu mult succes în ultimi ani în cadrul problemelor de clasificare pe date de intrare neliniar separabile. Voi începe prin a prezenta programele dezvoltate pentru partea de procesare a documentelor şi crearea vectorilor caracteristici şi voi continua cu cele dezvoltate pentru partea de clasificare (clusterare) a acestora utilizând tehnicile bazate pe vectori suport şi nuclee.

Am considerat partea de minerit al textului ca o aplicaţie care foloseşte tehnicile de minerit al datelor pentru a extrage semnătura fiecărui document (a crea vectorul de trăsături). Ca şi bază de date am utilizat baza Reuters care este des utilizată în articolele de procesare şi clasificare a documentelor. Baza de date Reuters conţine articolele de presă publicate de agenţia Reuters. Pornind cu o mulţime de d documente şi t termeni (cuvintele din documente) putem modela fiecare document ca un vector v într-un spaţiu cu t dimensiuni $\Re^t$. În faza de clasificare, am utilizat tehnica bazată pe învăţarea cu nuclee şi vectori suport. Aceasta este o tehnică puternică folosită cu mult succes în problemele de clasificare neliniar separabile. Avantajul ei este că poate fi aplicată pe mulţimi mari de date. Astfel putem testa uşor influenţa numărului de trăsături caracteristice asupra acurateţi clasificării. Am implementat clasificatorul pentru 2 tipuri de nuclee, nucleul polinomial şi nucleul Gaussian. Voi prezenta rezultatele obţinute atât pentru clasificarea la două clase cât şi pentru clasificarea în mai multe clase. Pentru clasificarea în două clase am luat în considerare doar documentele care aparţin la o clasă comparativ cu restul documentelor. Pentru clasificarea la mai multe clase am repetat clasificarea la două clase pentru fiecare topic (categorie în care este grupat documentul în faza de învăţare) obţinând mai multe funcţii de decizie (funcţii care fac distincţia între documentele dintr-un topic şi restul documentelor). În pasul de minerit al textului am utilizat această tehnică pentru selecţia trăsăturilor caracteristice. O să prezint şi o vizualizare grafică a rezultatului clasificării (clusterării) utilizând această tehnică.

Voi prezenta rezultate pe diferite tipuri de nuclee şi pentru diferite tipuri de reprezentare a datelor de intrare încercând să găsesc o corespondenţă optimă pentru îmbunătăţirea acurateţei clasificării. Am utilizat o formă cât mai simplificată a nucleelor, fără a reduce performanţele, uneori chiar îmbunătăţindu-le, astfel parametrii utilizaţi de acestea să fie cât mai uşor de specificat. Pentru algoritmii de clasificare şi clusterare implementaţi şi prezentaţi în acest raport am folosit o tehnică nouă utilizată cu mult succes în ultimi ani mai ales pentru lucrul cu mulţimi de date neliniar separabile - Support Vector Machine.

Datele de intrare sunt reprezentate în diferite formate şi am analizat influenţa acestora asupra tipului nucleului. Am utilizat trei tipuri de reprezentări. Reprezentarea *binară* în care atributele iau doar valoarea „0" sau „1" („0" dacă cuvântul respectiv nu apare în document şi „1" dacă apare fără să ne intereseze numărul de apariţii al acestuia). Reprezentarea *nominală* unde atributul păstrează numărul de apariţii al cuvântului în vectorul de frecvenţe normalizat folosind norma normală. În reprezentarea *Connel Smart* atributul îşi păstrează numărul de apariţii a cuvântului în vectorul de frecvenţe dar este normalizat utilizând o altă formulă. Rezultatele experimentale sugerează că alegerea unui procent de 4 – 10% din setul iniţial de trăsături conduce la cele mai bune rezultate. Deci după cum se poate observa din experimente nu este nevoie să utilizăm multe trăsături caracteristice în procesul de învăţare. Utilizarea prea multor trăsături poate fi chiar dăunătoare ducând şi la scăderea performanţelor învăţării.

În secţiunea 2 sunt prezentate metodele de selecţie a trăsăturilor şi detaliile de construcţie a mulţimilor de test şi de antrenament care vor fi utilizate în acest raport. Voi prezenta un scurt rezumat al metodelor de selecţie a trăsăturilor în contextul strategiei de antrenare propusă. În secţiunea 3 voi descrie algoritmii de clasificare şi de clusterare bazaţi pe tehnica de învăţare utilizând vectori suport şi nuclee. Aici voi prezenta şi detaliile de implementare a acestor algoritmi. În secţiunea 4 voi prezenta modul de utilizare al aplicaţiilor realizate pentru a creşte acurateţea clasificării. Tot aici voi descrie experimentele realizate şi voi compara rezultatele obţinute cu rezultatele obţinute cu altă aplicaţie (LibSvm). În secţiunea finală voi prezenta concluziile rezultate precum şi stadiul lucrării şi posibile dezvoltări viitoare.

Mulţumiri

Pe lângă părinţii mei o mulţime de oameni merită recunoştinţă, nu-i pot aminti pe toţi, dar le mulţumesc tuturor. Ţin totuşi să le mulţumesc câtorva care mi-au fost dascăli şi care m-au îndrumat în acest proiect. În primul rând ţin să mulţumesc conducătorului meu de doctorat prof. univ. dr. ing. Lucian VINŢAN, căruia îi sunt recunoscător în primul rând pentru responsabilitatea cu care dânsul îşi priveşte meseria, apoi pentru discuţiile extrem de stimulative pe care le-am avut şi pentru sprijinul acordat. De asemenea, rămân îndatorat pentru sprijinul profesional competent acordat mie cu generozitate încă de la începuturile perioadei de pregătire prin doctorat de către profesorii: mat. Ioana MOISIL, ing. Boldur BĂRBAT, ing. Daniel VOLOVICI, ing. Dorin SIMA şi ing. Macarie BREAZU.

Aş vrea să mulţumesc firmei SIEMENS AG, Corporate Technology, Information & Communication Division, MUNCHEN Germania, şi în special vicepreşedintelui Dr.h.c. Hartmut RAFFLER, pentru sponsorizarea acordată acestui program de doctorat. Deosebit de utile mi-au fost sugestiile profesionale precum şi suportul material oferit. De asemenea vreau să-i mulţumesc tutorelui meu de la firma SIEMENS, dr. Volker TRESP, Senior Principal Research Scientist in Neural Computation, pentru sprijinul oferit şi pentru îndrumarea în acest imens şi dificil domeniu al cercetării. Aş vrea de asemenea să-i mulţumesc d-lui Dr. Kai YU de la aceeaşi companie pentru informaţiile furnizate în dezvoltarea ideilor.