UNIVERSITATEA ”LUCIAN BLAGA” DIN SIBIU

Contract UEFISCSU nr. 27 / 04.08.2010

Programul: Resurse umane
Tipul proiectului: Proiect de cercetare post-doctorală
Cod proiect: PD_670 / 28.07.2010

SINTEZA DE CERCETARE

SISTEM DE CLASIFICARE AUTOMATA A DATELOR NESTRUCTURATE FOLOSIND METACLASIFICATOARE BAZATE PE METODE DE TIP SUPPORT VECTOR MACHINE ŞI NAIVE BAYES

Etapa unica 2010
Anul I

DIRECTOR PROIECT,

Sef lucrari dr. Ing. Daniel MORARIU

decembrie, 2010

1 Rezumatul proiectului

Acest proiect se încadrează în domeniul clasificării automate a documentelor text şi îşi propune îmbunătăţirea rezultatelor clasificării prin realizare unor strategi de combinare a rezultatelor metodelor de clasificare folosite.

Un număr impresionant de documente se găsesc în format electronic, iar utilizatorul are nevoie de componente de clasificare automată a acestora pentru a le putea gestiona. Gestiunea lor a devenit o problema foarte importanta în ultima perioada. Devine esenţială existenta unor programe inteligente de organizare automată a documentelor în categorii pentru a facilita analiza şi prelucrarea acestor documente. Datorită domeniului foarte vas în care ar trebui să lucreze acestea, devine dificil de realizat un singur clasificator cu performanţe foarte bune. Abordarea actuala este de a utiliza mai mulţi clasificatori de diferite tipuri combinaţi într-un meta-clasificator, sau realizarea unei clasificări hibride care se bazează pe predicţia clasificatorului cel mai bun pentru o problema particulară folosind caracteristicile vectorilor de intrare ale documentelor şi istoria clasificărilor. Având mai mulţi clasificatori de baza, de tipuri diferite (SVM - Support Vector Machine, Bayes, reţele neurale, etc), idea este de a învăţa un meta-clasificator care prezice gradul de corectitudine pentru fiecare dintre clasificatorii de bază. Meta-etichetarea unei instanţe indică încrederea în clasificarea făcută de acesta, dacă instanţa este clasificata corect de către acel clasificator dintre toţi ceilalţi clasificatori utilizaţi. Regula de clasificare a meta-clasificatorului este ca fiecare clasificator de baza să atribuie o clasa la instanţa curentă şi apoi meta-clasificatorul să decidă dacă clasificarea este demnă de încredere sau nu. Pe lângă creşterea acurateţei de clasificare, prin exploatarea sinergismului mai multor clasificatoare, un alt avantaj al meta-clasificării constă în posibilitatea de exploatare a paralelismelor funcţionale (multiprocesor).

2 Obiectivele proiectului pentru anul 2010

Etapa	Obiective	Activităţi	Categorii de buget (Valoare lei)
Unică	1. Dezvoltarea unor meta-clasificatoare neadaptive pentru clasificarea documentelor text	1.1. Adaptarea unui clasificator de tip Naive Bayes pentru lucru cu documente foarte mari. Realizare practica.	15500
		1.2 Integrarea clasificatorului Bayes în meta-clasificatorul existent.	13500
			Total etapa: 29000

3 Sinteza activităţilor de cercetare realizate în 2010

3.1 Setul de date utilizat în experimente

Experimentele prezentate sunt efectuate folosind colecţia de date Reuters-2000 [Reuters00], care conţine 984Mbytes de articole de tip ştiri prezentată într-un format comprimat. Această colecţie este de obicei utilizată în cercetare pentru clasificarea automată a documentelor. Colecţia include un total de 806.791 documente, articole de ştiri publicate de agenţia de presă Reuters în perioada 20 august 1996 – 19 august 1997. Analizate, articolele conţin 9.822.391 paragrafe, 11.522.847 propoziţii şi 310.033 rădăcini de cuvinte distincte rămase după eliminarea cuvintelor de legătură (stopword).

Datorită dimensiunii mari a bazei de date, voi prezenta rezultatele obţinute utilizând o submulţime a acesteia. Din toate cele 806.791 documente, s-au selectat acelea care sunt grupate de Reuters în categoria „System Software" din punct de vedere al codul industrial. După această selecţie, s-a obţinut un număr de 7.083 documente, care sunt reprezentate utilizând un număr de 19.038 trăsături şi 68 clase (categorii) diferite din punct de vedere al grupării după conţinut făcută de Reuters. Pentru a reduce numărul de trăsături de la 19.038 s-a folosit o metodă de selecţie a trăsăturilor caracteristice numită „Information Gain" - câştigul informaţional [Morariu07]. Un document este reprezentat ca un vector de cuvinte, după eliminarea cuvintelor de legătură (folosind un fişier standard de 510-stop de cuvinte) şi extragerea rădăcinii cuvântului [Chakrabarti03]. Setul de 7083 documente este reprezentat ca o matrice de frecvenţe de cuvinte în care fiecare rând reprezintă un document unic şi fiecare coloană reprezintă un singur cuvânt [Engler10]. Din cele 68 topicuri extrase am eliminat acele topicuri care sunt slab sau excesiv reprezentate. Astfel, am eliminat acele topicuri care conţin mai puţin de 1% (mai mult de 99%) de documente din totalul de 7083 documente. După această eliminare am obţinut 24 topicuri diferite şi 7053 documente, care au fost împărţite în mod aleatoriu în setul de antrenament (4702 eşantioane) şi setul de test (2351 eşantioane). În partea de extragere a trăsăturilor caracteristice am luat în considerare atât conţinutul articolului şi titlul acestuia.

3.2 Evaluarea clasificatorilor de tip SVM

În [Morariu07] este prezentat un meta-clasificator bazat pe 8 clasificatoare de tip SVM care era folosit pentru îmbunătăţirea acurateţei de clasificare a documentelor de tip text. Maximul acurateţei de clasificare obţinut de către un singur clasificatorul de tip SVM este 87.11% şi a fost obţinut de clasificatorul SVM cu nucleu de tip polinomial cu grad 2 şi reprezentare Cornell Smart. În [Morariu07] sunt prezentaţi şi testaţi mai mulţi clasificatori de tip SVM bazaţi atât pe nucleul polinomial cât şi pe cel Gaussian cu diferite forme de reprezentare. Dintre toţi clasificatorii testaţi şi prezentaţi, s-au inclus în meta-clasificator 8 clasificatori SVM distincţi. Alegerea celor 8 clasificatori s-a făcut pe baza acurateţei de clasificare obţinută separat de fiecare dintre aceştia. Clasificatorii astfel selectaţi şi prezentaţi în [Morariu07] sunt:

- 4 de tip Polinomial cu gradele 1 (reprezentare nominală), gradul 2(reprezentare binară şi reprezentare Cornell Smart) şi gradul 3 cu reprezentare Cornell Smart;
- 4 de tip Gaussian cu gradele 1.8, 2.1 2.8 şi 3.0 şi cu reprezentarea Cornell Smart a datelor.

Utilizând aceşti clasificatori s-a ajuns la o acurateţe maximă de clasificare de 92,04% în cazul meta-clasificatorului bazat pe distanţa euclidiană şi după un număr de 14 paşi de învăţare. Tot în [Morariu07] s-a prezentat şi o analiză în care se calcula şi limita teoretică maximă la care ar putea să ajungă meta-clasificatorul astfel creat. Limita calculată era de 94.21%. Această limită a clasificării s-a obţinut, deoarece din 2351 de documente de test, 136 de documente nu au putut fi clasificate corect de nici un clasificator selectat în cadrul meta-clasificatorului.
În prima fază am încercat găsirea unui nou clasificator care să reuşească să clasifice corect documentele, ce s-au dovedit imposibil de clasificat de către toţi clasificatorii selectaţi în meta-clasificator din [Morariu07].

3.3 Soluţii pentru îmbunătăţirea meta-clasificatorului cu SVM

O primă abordare în acest proiect de cercetare a fost aceea de a introduce noi clasificatori în meta-clasificator, care să încerce să clasifice corecte acele documente pe care clasificatoarele curente nu reuşesc. Astfel s-a introdus un clasificator te tip Bayes care foloseşte o metodă probabilistică pentru clasificarea documentelor şi acesta v-a rula în paralel cu clasificatoarele de tip SVM existente.
Cercetările actuale folosesc idea ca în cazul unui meta-clasificator să nu se mai ia în considerare doar prima clasă întoarsă de fiecare clasificator în parte ci să se considere şi clasa de pe a doua poziţie în cazul în care aceasta este suficient de aproape de cea de pe prima poziţie. Astfel o a doua abordare pe care am implementat-o în acest proiect de cercetare constă în alegerea unei alte categorii pentru un document dificil de clasificat, pentru care, de asemenea, clasificatorul întoarce un răspuns pozitiv mare. Dacă nici un clasificator nu va fi selectat pentru clasificarea unui document conform regulilor prezentate în [Morariu07] atunci se va alege clasificatorul cu cea mai

mare probabilitate de reuşită (distanţa între documentul curent şi toate documentele din coada clasificatorului este maximă, chiar dacă este mai mică decât pragul stabilit). De la clasificatorul astfel selectat nu se va mai alege prima clasă propusă ci următoarea dacă ea este suficient de aproape faţă de prima clasă.

3.4 Clasificatorul Naive Bayes

Clasificatorul Bayes Naive [Lewis98] realizat primeşte ca date de intrare toate fişierele care urmează a fi clasificate urmând ca alegerea datelor de antrenament şi a datelor de test să se facă după metoda "*n-fold crossvalidation*". Idea acestei metode este de a împărţi un set de date în *n* sub-seturi, urmând ca *n-1* de sub-seturi să se folosească la antrenare iar testarea să se facă pe sub-setul nefolosit la antrenament, adică antrenarea şi testarea să se execute pe seturi de date disjuncte. Algoritmul se va executa de n ori astfel încât fiecare sub-set de date va fi o singură dată un sub-set de test pentru verificarea antrenării.

Fie Y o variabilă pentru o clasă (categorie) care poate lua valorile {y1, y2, ..., y_m}.

Fie X o instanţă a unui vector cu *n* atribute <X1, X2, ..., Xn> şi x_k o valoare posibilă pentru X şi x_{ij} o valoare posibilă pentru X_i. Pentru clasificarea de tip Bayes calculăm probabilităţile $P(Y = y_i | X = x_k)$, *pentru* $i = \overline{1,m}$. Asta ar însemna calcularea tuturor probabilităţilor pentru fiecare categorie pentru fiecare instanţă posibilă din spaţiul de instanţe – ceea ce este foarte greu de calculat pentru un set rezonabil de date.

Practic pentru a determina categoria lui x_k, trebuie să determinăm pentru fiecare y_i probabilitatea:

$$P(Y = y_i \mid X = x_k) = \frac{P(Y = y_i)P(X = x_k \mid Y = y_i)}{P(X = x_k)} \qquad (3.1)$$

Probabilitatea $P(Y = y_i)$ poate fi uşor aproximată având în vedere faptul că dacă n_i exemple din *D* se regăsesc în y_i atunci $P(Y = y_i) = \frac{n_i}{|D|}$, unde D reprezintă mulţimea documentelor din setul de antrenament.

Probabilitatea $P(X = x_k \mid Y = y_i)$ trebuie estimată (deoarece există 2^n posibile instanţe pentru a calcula probabilitatea). De aceea, dacă presupunem că atributele unei instanţe sunt independente (condiţional independente), atunci:

$$P(X \mid Y) = P(X_1, X_2, \cdots X_n \mid Y) = \prod_{i=1}^{n} P(X_i \mid Y) \qquad (3.2)$$

Astfel trebuie să calculăm doar $P(X_i \mid Y)$ pentru fiecare posibilă pereche "valoare atribut" - "categorie"

Dacă Y şi toate X_i sunt binare, atunci trebuie să calculăm doar 2n valori

$P(X_i = true | Y = true)$ şi $P(X_i = true | Y = false)$ pentru fiecare X_i

$P(X_i = false) | Y) = 1 - P(X_i = true | Y)$

faţă de 2^n valori, dacă nu am presupune independenţa atributelor.

Practic, dacă setul de date D conţine n_k exemple din categoria y_k şi n_{ij_k} din aceste n_k exemple au a *j*-a valoare pentru atributul X_i pe x_{ij} atunci estimăm că:

$$P(X_i = x_{ij} \mid Y = y_k) = \frac{n_{ijk}}{n_k} \qquad (3.3)$$

Această estimare poate genera erori la seturi foarte mici de date deoarece un atribut rar într-un set de antrenament face ca X_i să fie fals în setul de antrenament $\forall y_k \; P(X_i = true | Y = y_k) = 0$.

Dacă X_i=*true* într-un exemplu de test atunci $\forall y_k \; P(X | Y = y_k) = 0$ şi $\forall y_k \; P(Y = y_k | X) = 0$

Pentru a evita acest lucru se utilizează uniformizarea (normalizarea) lui Laplace. Această normalizare pleacă de la premisa că fiecare atribut are o probabilitate *p* observată într-un exemplu virtual de dimensiune *m*.

Astfel

$$P(X_i = x_{ij} \mid Y = y_k) = \frac{n_{ijk} + mp}{n_k + m} \qquad (3.4)$$

unde *p* este o constantă. De exemplu pentru atribute binare p=0,5.

Pentru clasificarea de text, clasificatorul Bayes generează pentru un document dintr-o anumită categorie un "bagaj de cuvinte" dintr-un vocabular V = {w_1, w_2,…w_m} calculând probabilitatea $P(w_j|c_i)$. Pentru normalizarea Laplace se presupune existenţa unei distribuţii uniforme a tuturor cuvintelor (adică ar fi echivalentul unui exemplu virtual în care fiecare cuvânt apare doar o singură dată).

$p = \frac{1}{|V|}$ şi m = |V|

3.5 Rezultate obţinute cu clasificatorul de tip Naive Bayes

3.5.1 Modelele meta-clasificatorului

În [Morariu06] am prezentat un meta-clasificator bazat pe 8 clasificatorilor SVM care a fost folosit pentru a îmbunătăţi precizia de clasificare pentru documente de tip text. Acolo am prezentat 3 modele pentru a testa acurateţea de clasificare a meta-clasificatorului. Acestea au fost: un model neadaptiv bazat pe votul majoritar numit Majority Vote (MV), un model adaptiv bazat pe o selecţie folosind distanta euclidiana notat SBED şi un model adaptiv bazat pe distanta cosinus notat (SBCOS). Am utilizat toate aceste modele pentru a testa îmbunătăţirea adusă meta-clasificatorului în cazul introducerii unui nou clasificator Naive Bayes.

Pentru reprezentarea datelor am folosit toate cele 3 reprezentări care au fost utilizate în [Morariu07] anume reprezentarea binară, Nominală şi Cornell Smart. In ceea ce priveşte clasificatori de tip SVM pentru aceştia s-a folosit atât nucleul Polinomial cât şi cel Gaussian.

3.5.2 Rezultate experimentale

Ca urmare a introducerii unui nou clasificator, noul meta-clasificator are acum 9 clasificatoare şi am recalculat limita de clasificare maximă teoretică care ar putea fi atins de acesta. Astfel, introducerea clasificatorului Bayes creşte limita teoretică maximă a meta-clasificatorului la 98.63% (faţă de 94.21% aşa cum a fost fără clasificator Bayes). Acest fapt oferă o oportunitate de a obţine o mai bună precizie de clasificare. Toate aceste rezultate prezentate aici au fost publicate în articolele [Morariu10] şi [Cretulescu10].

Utilizând metoda Votului Majoritar noua acurateţe de clasificare obţinută este de 86.09%. Precizia de clasificare a scăzut faţă de valoarea obţinută cu 8 clasificatorilor (86.38%), conducând la o scădere de 0,29%. Aceasta se datorează clasificatorului Bayes care - pe întregul set de test – are o precizie de doar 81.32%, clasificând incorect destul de multe documente (439). Rezultatele sunt prezentate în Fig. 1.

In cazul metodei SBED sunt prezentate primi 14 paşi, deoarece după acest număr de paşi acurateţea de clasificare nu se modificată în mod substanţial. Ca şi în [Morariu06] pragul pentru primele 7 etape a fost ales egal cu 2.5 şi pragul pentru ultimele 7 etape a fost ales egal cu 1.5. Un pas reprezintă un proces de antrenare, urmat de un proces de testare. Rezultatele pentru votul majoritar cât şi pentru SBED cu 8 şi cu 9 clasificatoare sunt prezentate în Fig.1.

Pentru meta-clasificatorul cu 9 clasificatori, rezultatele sunt mai slabe decât cele obţinute utilizând meta-clasificator cu 8 clasificatorilor. Această acurateţe slabă se explică datorită acurateţei scăzute a clasificatorului Bayes (81.32%), comparativ cu SVM, şi datorită faptului că clasificatori sunt selectaţi la întâmplare după cum am explicat în [Morariu07].

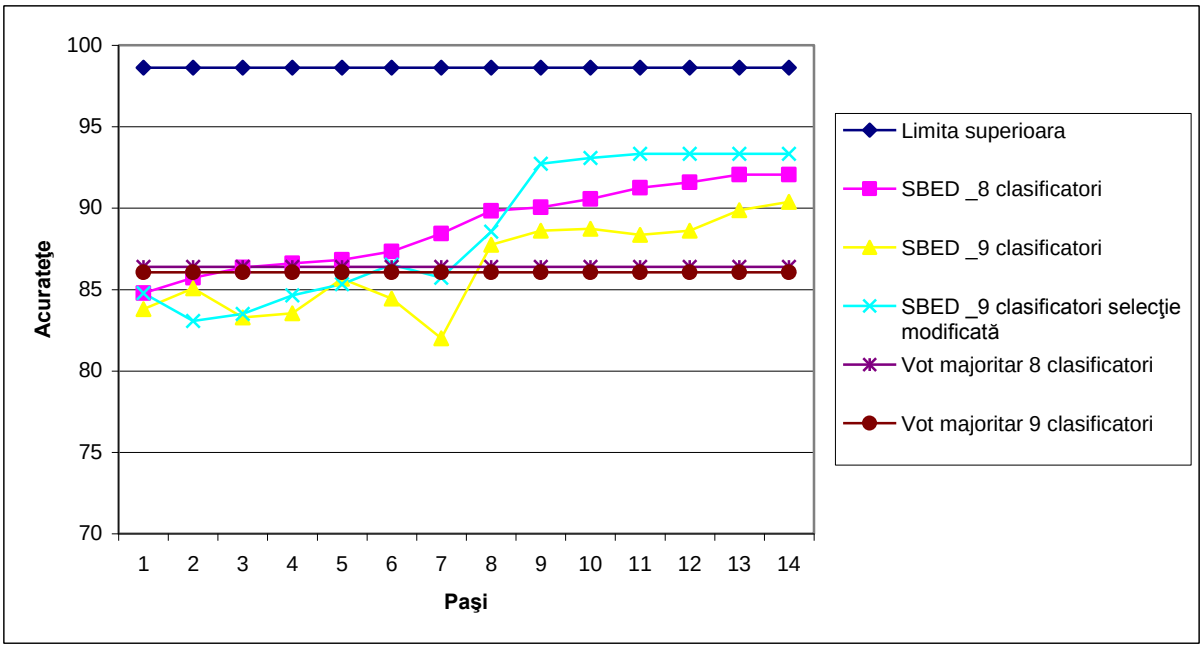

Fig. 1 Acurateţea clasificării – Selecţie folosind votul majoritar sau distanţă euclidiană (SBED)

În cazul utilizării metodei SBCOS de asemenea prezentăm primii 14 paşi. La fel ca în [Morariu07], pragul pentru primi 7 paşi s-au ales egali cu 0,8 şi pragul pentru ultimii 7 paşi s-a ales egal cu 0,9.
În cazul acestui meta-clasificator, rezultatele obţinute arată că acurateţea de clasificare s-a îmbunătăţit de la 89,74% la **93,10%** prin adăugarea clasificatorului Bayes (Fig. 2).

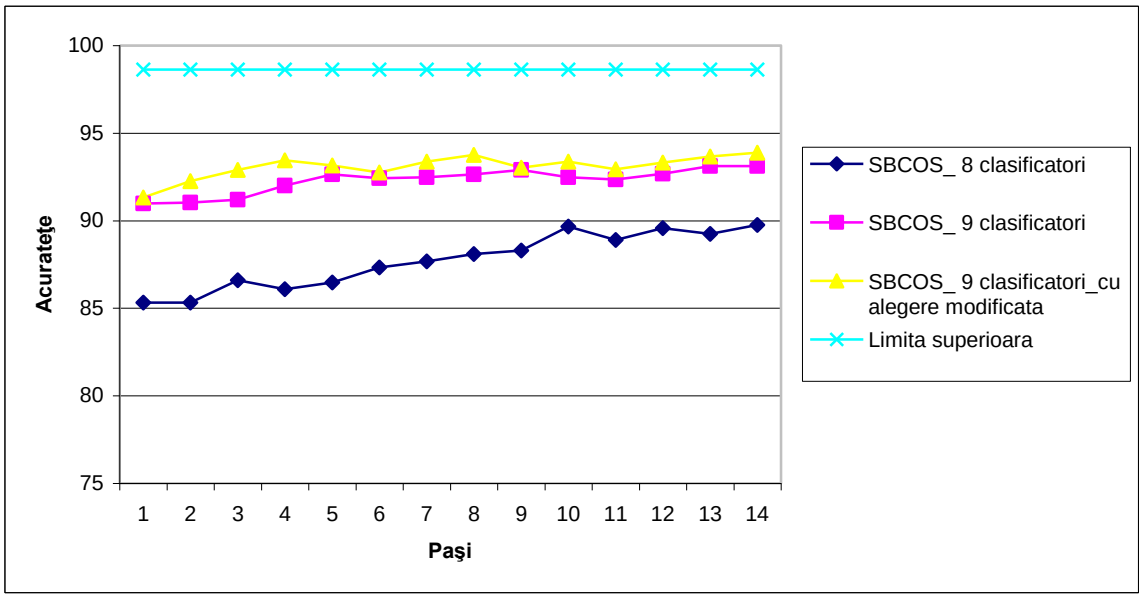

Fig. 2 Acurateţea clasificării – selecţia folosind distanţa cosinus (SBCOS)

3.5.3 Rezultate obţinute modificând alegerea clasei

Continuând tendinţele actuale din domeniu a altă îmbunătăţire adusă meta-clasificatorului se referă la modul de selecţie a clasificatorului. După cum am explicat şi în [Morariu10] în cazul unui nou document d_n care trebuie clasificat, se ia pe rând fiecare clasificator şi se calculează distanta între d_n şi fiecare document din coada de erori a clasificatorului respectiv. Dacă cel puţin o distanţă calculată este mai mică decât pragul stabilit, meta-clasificatorul nu va folosi pentru a clasifica

documentul d_n. În cazul în care se rejectează astfel toţi clasificatorii, meta-clasificatorul totuşi va alege pe cel care are distanţa cea mai mare obţinută (chiar dacă este mai mică decât pragul stabilit). Actualmente, meta-clasificatorul va prezice clasa specificată clasificatorul ales chiar dacă există şanse mari ca acesta să clasifice prost. Modificarea adusă meta-clasificatorului ar fi că, în acest caz, clasificatorul ales să nu mai selecteze clasa cu valoarea cea mai mare (pentru că oricum va da greş deoarece este predispus să clasifice eronat tipul respectiv de documente), ci să aleagă clasa imediat următoare din lista de clase pe care le prezice. Se va alege următoarea clasă prezisă doar dacă valoarea pentru aceasta este suficient de apropiată de valoarea maxim obţinută de clasificator (cu un ε=0,5 ales în experimentele efectuate). În acest caz, clasificatorul ar specifica o altă clasă pentru documentul curent d_n. Efectuând aceste modificări, rezultatele meta-clasificatorului cu 9 clasificatori s-au îmbunătăţit substanţial.

Rezultatele obţinute de către meta-clasificatorul cu 9 clasificatori modificat care se bazează pe distanţa euclidiană s-au îmbunătăţit, obţinând o acurateţe a clasificării de **93,32%** faţă de cel cu 9 clasificatori nemodificat care a obţinut doar 90,38%. Reamintim faptul că în aceleaşi condiţii meta-clasificatorul cu 8 clasificatori de tip SVM din [Morariu07] a obţinut o acurateţe de clasificare de doar 92,04%, maximul obţinut pe 8 clasificatoare.

În cazul distanţei bazată pe cosinus acurateţea de clasificare a meta-clasificatorului s-a îmbunătăţit de la 93,10% la **93,87%** (Fig. 2)**.** Reamintim că meta-clasificatorul cu 8 clasificatori de tip SVM în aceleaşi condiţii a obţinut o acurateţe de clasificare de 89,74%.

4 Concluzii

Bazându-mă pe meta-clasificator prezentate în [Morariu07] bazate pe 8 clasificatoare de tip SVM, am adăugat la acesta un nou clasificator de tip Bayes, care conduce la o îmbunătăţire semnificativă a limitei superioare la care poate ajunge meta-clasificatorul. Astfel, limita superioara a meta-clasificatorului a crescut de la 94.21%, atunci când se utilizează 8 clasificatorilor SVM [Morariu07] la 98.63%, atunci când se utilizează 8 clasificatoarele SVM plus clasificatorul Naive Bayes.

Mai mult, am prezentat rezultatele pentru toate cele trei modele ale meta-clasificatorului: cu votul cu majoritate (MV), selecţie pe baza distanţei euclidiene (SBED) şi de selecţie pe baza Cosinusul (SBCOS). În caz de MV, am obţinut o precizie de clasificare doar 86.09%, care este cu 0.29% mai mic decât atunci când se utilizează numai 8 clasificatori. Mai mult decât atât, în cazul celor 9 clasificatoare meta-clasificatorul SBED a obţinut rezultate chiar mai mici, în medie, scăzând de la 92.04% la 90.38%. În cazul SBCOS cu 9 clasificatoare, exactitatea de clasificarea a meta-clasificatorului a crescut de la 89.74% la 93.10%.

În cele din urmă, am considerat că dacă există orice suspiciune că clasa nu va fi cea corectă, atunci meta-clasificatorul va prezice o clasă diferită. Aceasta va fi clasa următoare din listă numai dacă este suficient de aproape de prima clasa prezisă. Această schimbare a condus la o îmbunătăţire substanţială a meta-clasificatorului cu 9 clasificatoare. Am efectuat experimente numai cu meta-clasificator cu 9 clasificatorilor deoarece numai în această situaţie, o acurateţe maximă de 98.63% ar putea fi atinsă. În cazul meta-clasificatorului SBED, am obţinut o precizie de clasificare medie de 93.32%. Această acurateţe este cu 2.94% mai mare decât în cel mai bun caz obţinut când nu schimbăm clasa. În cazul meta-clasificatorului SBCOS în mod similar s-a îmbunătăţit acurateţea de clasificare de la 93.10% la 93.87%.

5 Referinţe bibliografice

[Chakrabarti03] S. Chakrabarti, *Mining the Web- Discovering Knowledge from hypertext data*, Morgan Kaufmann Press, 2003.

[Cretulescu10] R. Cretulescu, D. Morariu, L. Vintan, I. Coman – *An Adaptive Meta-classifier for Text document*, 16th international Conference on Information Systems Analysis, pp. 372-377, ISBN-13: 978-1-934272-86-2(Collection), ISBN-13: 978-1-934272-88-6(Volume II) ,Florida, USA, 2010

(http://www.iiis.org/CDs2010/CD2010IMC/ISAS_2010/Abstract.asp?myurl=UA277CD.pdf)

[Engler10] J. Engler, A. Kusiak, *Mining Authoritativeness of Collaborative Innovation Partners*, International Journal of Computers, Communications and Control, Vol. V, No. 1, pp. 42-51, 2010.

[Lewis98] D. Lewis, *Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval*, ATT Lab Research, NJ, Vol. 1398, pp. 4-15, USA, 1998.

[Morariu06] D. Morariu, L. Vintan, V. Tresp, *Meta-classification using SVM classifiers for Text Documents*, Proceedings of the 3rd International Conference on Machine Learning and Pattern Recognition (MLPR 2006), Barcelona, ISSN 1305-5313, pag. 222-227, Octombrie 2006;
(http://www.waset.org/journals/ijamcs/v1/v1-1-10.pdf)

[Morariu07] D. Morariu - Contributions to Automatic Knowledge Extraction from Unstructured Data, PhD Thesis, Sibiu, 2007

[Morariu10] D. Morariu, R. Cretulescu, L. Vintan – *Improving a SVM Meta-classifier for Text Documents by using Naive Bayes*, International Journal of Computers, Communications & Control, Vol. V, No. 3, pp. 351-361, ISSN 1841-9836, E-ISSN 1841-9844, 2010 (cotata ISI, factor de impact 0.373 pe 2010 - http://journal.univagora.ro/?page=article_details&id=418)

[Reuters00] Misha Wolf and Charles Wicksteed - Reuters Corpus: http://www.reuters.com/researchandstandards/corpus/ lansat în noiembrie 2000, accesat în septembrie 2009