Kategorizace textů metodou NBCI

Kučera M.[1], Ježek K.[1], Hynek J.[2]

[1]Katedra informatiky a výpočetní techniky – Západočeská univerzita v Plzni
Univerzitní 22, 306 14 Plzeň
[2]inSITE, s. r. o., Rubešova 29, 326 00 Plzeň (www.insite.cz)
kuceram@kiv.zcu.cz jezek_ka@kiv.zcu.cz jiri.hynek@insite.cz

Abstrakt. Příspěvek uvádí novou metodu pracující na principu induktivního strojového učení, jež je kombinací naivní Bayesovy metody a metody Itemsets. Stručně ji lze popsat buďto jako metodu Itemsets používající ke klasifikaci naivní Bayesův klasifikátor, nebo jako naivní Bayesův klasifikátor využívající aproximace vlastních dokumentů častými množinami položek. Implementace této nové robustní metody je poměrně nenáročná. Metoda byla testována na kolekci Reuters-21578. V článku jsou prezentovány výsledky dokladující kvality nové metody.

Klíčová slova: Kategorizace, klasifikace, taxonomie, text mining, strojové učení, klasifikátor, množina položek, itemsets, Bayesova metoda, Apriori algoritmus, digitální knihovna

1 Úvod

Vytvoření a údržba digitální knihovny představuje časově i finančně náročný úkol. Kategorizace dokumentů musí být prováděna doménovými experty, kteří jsou velmi dobře obeznámeni s jednotlivými tematickými okruhy, do nichž jsou příspěvky zařazovány. Kategorizace prováděná manuálně je navíc zatížena subjektivním hlediskem. Kvalitou zatřídění je determinována i kvalita vyhledání informací relevantních k dotazu uživatele knihovny. Náš předchozí výzkum v oblasti klasifikace dokumentů byl motivován prací na oborové digitální knihovně Západočeské energetiky, a.s., předpokládáme však jeho další využití i v rámci vyhledávacích služeb univerzitního intranetu. Výsledkem byl návrh a testování klasifikačního algoritmu, který jsme nazvali *metodou Itemsets* [3]. Tento algoritmus dává dobré výsledky v přesnosti i úplnosti klasifikace, je však prioritně určen ke klasifikaci krátkých dokumentů (abstraktů) a je relativně citlivý na původní nastavení parametrů. Ve snaze o jeho zdokonalení proto vznikla metoda, kterou jsme nazvali NBCI (*Naive Bayes Combined with Itemsets*). Jde o kombinací naivního Bayesova klasifikátoru a metody *Itemsets*.

Text příspěvku je strukturován následovně: V oddílech 2 a 3 uvádíme stručný popis naivního Bayesova klasifikátoru, resp. metody *Itemsets*. Oddíl 4 je stěžejní, neboť popisuje algoritmus NBCI a úvahy, které vedly k jeho návrhu. Poslední oddíl obsahuje výsledky testů algoritmu NBCI, včetně jeho porovnání s ostatními zmíněnými metodami.

2 Naivní Bayesův klasifikátor

Metoda se zakládá na aplikaci Bayesova teorému, který udává pravděpodobnost platnosti hypotézy podmíněné výskytem určitého atributu:

$$P(h|A) = \frac{P(h) \times P(A|h)}{P(A)} \tag{1}$$

Známe-li apriorní pravděpodobnosti hypotézy h z množiny hypotéz H, tj. $P(h)$, a současně pravděpodobnost výskytu atributu A podmíněného platností h, tedy $P(A|h)$, můžeme určit nejpravděpodobnější hypotézu h_{MAP}[1] v případě výskytu atributu A:

$$h_{MAP} = argmax_{h \in H}\ P(h) \times P(A|h) \tag{2}$$

Naivní Bayesův klasifikátor (NB) se používá v úlohách, kdy každá instance x je popsána kombinací hodnot atributů a cílová klasifikační funkce $f(x)$ nabývá hodnot z konečné množiny V. Po poskytnutí množiny trénovacích příkladů pro $f(x)$ musí klasifikátor určit zařazení nové instance popsané n-ticí $(a_1, a_2, \ldots, a_n)$.

Při klasifikaci nové instance se pro danou n-tici atributů vybere nejpravděpodobnější cílová hodnota:

$$v_{MAP} = argmax_{v_j \in V}\ P(v_j)P(a_1, a_2, \ldots, a_n|v_j) \tag{3}$$

Hodnoty $P(a_1, a_2, \ldots, a_n|v_j)$ lze spolehlivě stanovit pouze pro velké množiny trénovacích dat[2]. Proto je naivní Bayesův klasifikátor založen na zjednodušujícím předpokladu, a to že hodnoty jednotlivých atributů jsou **vzájemně nezávislé** vzhledem k cílové hodnotě klasifikační funkce. Pro konkrétní cílovou hodnotu můžeme tudíž určit pravděpodobnost výskytu kombinace atributů a_1, a_2, …, a_n jako součin hodnot pro jednotlivé atributy: $P(a_1, a_2, \ldots, a_n|v_j) = \prod_i P(a_i|v_j)$. Dosazením do vztahu (3) obdržíme tzv. **naivní Bayesův klasifikátor**

$$v_{NB} = argmax_{v_j \in V}\ P(v_j) \prod_i P(a_i|v_j) \tag{4}$$

3 Klasifikátor Itemsets

Principy klasifikační metody *Itemesets* jsme poprvé publikovali před dvěma lety [3]. Možné modifikace klasifikátoru a částečné vyhodnocení dosahovaných výsledků je publikováno v [2]. Klasifikátor pracující na principu induktivního učení má celou řadu aplikací, z nichž některé popisujeme v [1].

V našem případě pracujeme s množinou textových dokumentů, v nichž vyhledáváme množiny významových termů. Termy označujeme jako *položky* a množiny termů jako *množiny položek – itemsets*. Množina položek o mohutnosti k prvků se označuje jako k-množina položek (k-itemset).

[1] Označení MAP pochází z termínu *maximum aposteriori*.

[2] Jejich počet je roven počtu všech možných instancí, násobenému počtem cílových hodnot klasifikační funkce.

Pro množinu položek lze stanovit její *četnost*[3] jako počet dokumentů obsahujících danou množinu položek. Pokud četnost množiny položek přesáhne určitou prahovou hodnotu, hovoříme o *časté množině položek* (frequent itemset).

Časté množiny položek vyhledáváme iterativně pomocí *Apriori algoritmu* [4]. Prahová hodnota, která rozhoduje o zařazení množiny položek mezi časté itemsety, může být volena různě pro množiny položek různé mohutnosti. Výběr vhodných prahů zásadním způsobem ovlivňuje výsledky kategorizace.

Dále je třeba stanovit určitý počet *charakteristických itemsetů* pro každý tematický okruh. Ke každé časté množině položek Π_j můžeme určit množinu dokumentů $D\Pi_j$ obsahujících Π_j. Analogicky pro každé téma T_i doménové taxonomie existuje množina dokumentů DT_i zařazených do této třídy. Váhový koeficient $w_{\Pi_j}^{T_i}$ udává, do jaké míry itemset Π_j charakterizuje třídu T_i.

$$w_{\Pi_{ji}}^{T_i} = \frac{|D\Pi_j \cap DT_i|}{|DT_i|} \tag{5}$$

Ve vlastní implementaci je výpočet podle vztahu (5) normalizován vzhledem k ostatním třídám. Na základě hodnot váhových koeficientů se pro každou třídu vybere soubor charakteristických množin položek. Můžeme buďto vybrat ty množiny položek, jejichž váha přesáhne pro danou třídu jistou prahovou mez, nebo stanovíme pevný počet charakteristických itemsetů a vybíráme vždy prvních n množin položek[4], uspořádaných sestupně podle hodnot jejich váhových koeficientů pro dané téma. Nejlepších výsledků jsme s klasifikátorem *Itemsets* dosahovali použitím druhého způsobu výběru charakteristických itemsetů.

Před vlastní klasifikací je vhodné stanovit váhové faktory, které nám pomohou odlišit množiny položek různé mohutnosti. V zařazovaném dokumentu vyhledáme charakteristické itemsety pro každou třídu a sečtením jejich vah (po vynásobení váhovými faktory) určíme váhu klasifikace do daného tématu. Dokument zařadíme do té třídy, pro kterou dosáhne váha zařazení nejvyšší hodnoty, nebo do těch témat, pro něž překročí jistou prahovou mez.

4 Kombinace klasifikačních metod – NBCI

Základní myšlenka NBCI je relativně přímočará: namísto vlastních termů z textových dokumentů bude klasifikátor pracovat s častými množinami položek, přičemž zůstane zachován algoritmus naivního Bayesova klasifikátoru a změní se pouze charakter dat, pro něž se vypočítávají podmíněné pravděpodobnosti. Zbývá rozhodnout, jakým způsobem určíme množiny položek a zejména to, jak pro ně budeme zjišťovat podmíněné pravděpodobnosti vzhledem ke třídám taxonomie.

[3] Četnost bývá označována též jako *podpora* podle původního anglického *support*.
[4] Určení tohoto parametru má na výsledky klasifikace zcela zásadní vliv.

4.1 Výběr množin položek a přiřazení k dokumentům

Pro natrénování klasifikátoru použijeme pouze *časté množiny položek*. Stanovení charakteristických itemsetů totiž tvoří nejslabší místo metody *Itemsets* a nevhodný výběr vede k výraznému zhoršení výsledků klasifikace. V našem případě bychom se nevhodným výběrem připravili o informace vedoucí k odůvodněnému zařazení dokumentu a klasifikátor by tak častěji „tipoval“ jen na základě apriorních pravděpodobností tříd. Navíc dochází ke konfliktu váhy itemsetu s podmíněnou pravděpodobností slova v závislosti na třídě. Určení $w_{\Pi_j}^{T_i}$ (5) se totiž značně podobá způsobu výpočtu $P(\Pi_k|v_j)$ (6), přesto nelze zaměňovat jedno za druhé. Použitím obojího (tzn. nejprve vybrat charakteristické množiny položek a poté pro ně počítat podmíněné pravděpodobnosti) bychom naivnímu Bayesovu klasifikátoru „podsouvali“ náš názor a nenechali bychom jej učinit si vlastní.

Hledání častých množin položek provedeme *Apriori algoritmem*, stejně jako v případě původní metody *Itemsets*. Dáme přednost množinám položek vybraným na základě relativní četnosti v rámci jednotlivých tříd.

Množinu položek přiřadíme k dokumentu, pokud se v daném dokumentu vyskytuje každý její prvek. Na rozdíl od původní metody *Itemsets*, která pouze rozlišovala, zda se itemset v dokumentu vůbec objevuje, budeme v NBCI pracovat i s počtem výskytů množiny položek v daném dokumentu[5]. Skutečný počet výskytů totiž hraje při výpočtu pravděpodobnostních hodnot významnou roli, zejména v případě tříd s velkým počtem dokumentů.

Použití počtu výskytů množiny položek vedlo již při prvních pokusech během vývoje NBCI k lepším výsledkům, než jaké byly dosahovány při pouhém rozlišování, zda se množina položek v dokumentu vyskytuje či nikoli. Proto byla při dalším vývoji metody uvažována pouze alternativa sledující četnost itemsetu v rámci dokumentu.

4.2 Pravděpodobnosti a váhové faktory

Apriorní pravděpodobnosti jednotlivých témat taxonomie stanovujeme stejným způsobem, jako v případě původní metody *Naïve Bayes*. Vlastní termy dokumentů ale nyní nahrazují časté množiny položek. Hodnotu $P(\Pi_k|v_j)$, která vystihuje pravděpodobnost výskytu itemsetu Π_k v dokumentech třídy v_j, je možné určit jako

$$P(\Pi_k|v_j) = \frac{n_k}{n} \tag{6}$$

přičemž n udává počet výskytů častých itemsetů stejné mohutnosti jako Π_k v dokumentech třídy v_j, zatímco n_k je počet výskytů Π_k mezi těmito n případy.

Tento jednoduchý způsob však nese značné riziko: pokud se objeví nulové n_k, dominuje Bayesově klasifikátoru (4), neboť všechny ostatní hodnoty násobíme

[5] Pro k-itemsety, kde $k \geq 2$, se počet výskytů v dokumentu stanoví jako minimum z počtů výskytů jednotlivých termů, tvořících itemset.

nulou. Tomuto problému se vyhneme použitím tzv. ***m*-odhadu pravděpodobnosti**[6]:

$$P(\Pi_k|v_j) = \frac{n_k + mp}{n + m} \tag{7}$$

kde n_k a n se definují stejně jako v předchozím případě, p je apriorní odhad pravděpodobnosti, kterou chceme určit, a konstanta m, nazývaná *ekvivalentní velikost vzorku*[7], určuje váhu p vzhledem k trénovacím datům. Při nedostatku dalších informací volíme typicky $p = \frac{1}{r}$, kde r značí počet různých hodnot odhadovaného atributu.

Původní vztah (6) tedy rozšíříme použitím *m-odhadu*:

$$P(\Pi_k|v_j) = \frac{n_k + 1}{n + |itemsets|} \tag{8}$$

Hodnota $|itemsets|$ značí celkový počet častých množin položek nalezených v trénovacích dokumentech[8].

Stejně jako se při klasifikaci metodou *Itemsets* používají váhové faktory pro odlišení významu množin položek různé mohutnosti, tak i NBCI dovoluje rozlišovat pravděpodobnost výskytu itemsetu podle počtu prvků, které jej tvoří. Zavádí se váhový faktor wf_L pro každou mohutnost L jako číslo od nuly do jedné. Tímto koeficientem se umocní příslušná hodnota $P(\Pi_k|v_j)$ předtím, než se započte do celkové pravděpodobnosti zařazení do příslušného tematického okruhu.

4.3 Postup klasifikace metodou NBCI

Činnost NBCI se dělí do dvou kroků – *fáze trénování* a *fáze vlastní klasifikace*. Během trénování se nejprve vyhledají časté množiny položek, jimiž se nahradí obsah dokumentů. Pro ně se poté počítají podmíněné pravděpodobnosti vzhledem ke všem třídám taxonomie. Při vlastní klasifikaci se v novém dokumentu naleznou množiny položek stanovené během trénování, a poté se dokument vyhodnotí mechanizmem naivního Bayesova klasifikátoru s uplatněním váhových koeficientů jednotlivých množin položek.

Algoritmus 1. Trénování NBCI

> **vstupní data:** *Examples* – množina textových dokumentů spolu s jejich zařazením do témat; V – množina všech klasifikačních témat.
> **výstup:** množina častých množin položek *itemsets*; podmíněné pravděpodobnosti $P(v_j)$, $P(\Pi_k|v_j)$ pro všechny $v_j \in V$ a všechny $\Pi_k \in itemsets$

1. *itemsets* ← všechny časté množiny položek ze všech dokumentů v *Examples*

[6] Tato metoda bývá rovněž označována termínem *Laplace smoothing*.
[7] Vztah (7) můžeme chápat jako rozšíření existujících n vzorků o dalších m případů rozdělených podle p.
[8] Při dosazování do (7) bylo stanoveno $m = |itemsets|$ a $p = \frac{1}{|itemsets|}$.

2. Pro každou třídu $v_j \in V$
   (a) $docs_j \leftarrow$ podmnožina dokumentů z *Examples* zařazených do v_j
   (b) $P(v_j) \leftarrow \frac{|docs_j|}{|Examples|}$
   (c) Pro každý itemset $\Pi_k \in itemsets$
      (1) $n \leftarrow$ počet výskytů různých $\Pi_n \in itemsets$ v $docs_j$, kdy $|\Pi_n| = |\Pi_k|$
      (2) $n_k \leftarrow$ počet výskytů Π_k v $docs_j$
      (3) $P(\Pi_k|v_j) \leftarrow \frac{n_k+1}{n+|itemsets|}$

Algoritmus 2. Klasifikace NBCI

vstupní data: *Doc* – nový dokument určený ke klasifikaci; V, $P(v_j)$, $P(\Pi_k|v_j)$, *itemsets* – viz alg. 1
výstup: zařazení do třídy v_{NBCI}

1. $itemsets_{Doc} \leftarrow$ všechny $\Pi_k \in itemsets$, jež jsou obsaženy v *Doc*
2. $v_{NBCI} \leftarrow argmax_{v_j \in V}\ P(v_j) \prod_{\Pi_k \in itemsets_{Doc}} P(\Pi_k|v_j)^{wf_L}$, $L = |\Pi_k|$

5 Testování klasifikační metody NBCI

Pro testování byla použita kolekce **Reuters-21578** [5]. Z kolekce jsou odstraněny dokumenty bez asociovaných témat a dokumenty bez významových termů. V tabulce 1 jsou uvedeny základní parametry kolekce spolu s počtem *vyhovujících* témat. Empiricky jsme stanovili, že vyhovující jsou kategorie obsahující alespoň 50 dokumentů. Třídy s menším počtem dokumentů nejsou dostatečně charakterizovány a zpravidla neumožňují spolehlivé natrénování klasifikátoru.

Tabulka 1. Parametry kolekce Reuters-21578

počet dokumentů kolekce	10202
počet různých významových slov (před / po lemmatizaci)	31490 / 25648
průměrná délka dokumentu (počet významových slov)	70,16
počet tematických okruhů	116
maximální počet zařazení dokumentu	16
průměrný počet zařazení dokumentu	1,26
počet vyhovujících témat (s alespoň 50 dokumenty)	28
počet dokumentů zařazených do vyhovujícíh témat	9662 (94,7 %)
maximální počet zařazení do vyhovujícího tématu	8
průměrný počet zařazení do vyhovujícího tématu	1,13

Při indexaci se používá **anglický slovník nevýznamových slov** čítající celkem **137 slov**. Dále se provádí **algoritmická lemmatizace**, která sníží počet různých významových slov téměř o 20 %. Během indexace se rovněž dokumenty rozdělují na trénovací a testovací ve zvoleném poměru **3:1**.

Tabulka 2. Rozdělení dokumentů Reuters na trénovací a testovací v poměru 3:1

dokumenty	**trénovací**	**testovací**	**celkem**	
všechny	7652	2550	10202	100,0 %
vyhovující třídy	7251	2411	9662	94,7 %
10 nejčastějších tříd	6319	2138	8457	82,9 %
2 nejčastější třídy	4372	1510	5882	57,7 %
třída *earn*	2779	941	3720	36,5 %
třída *acq*	1605	575	2180	21,4 %

5.1 Testovací případy

Klasifikátor testujeme na celé kolekci **Reuters-21578**. Dokumenty dělíme na trénovací a testovací podle tabulky 2. Během trénování se zaměřujeme pouze na **1-itemsety** označené za časté na základě relativní četnosti v jednotlivých třídách. Uvádíme vždy nejlepší výsledky z hlediska přesnosti, úplnosti a míry F_1, doplněné údaji o přibližné době potřebné k natrénování klasifikátoru a vlastní klasifikaci[9].

V případě zařazování do **vyhovujících tříd** trénujeme klasifikátor pouze na třídy, které obsahují alespoň 50 dokumentů (celkem 28 témat). Výsledky vyhodnocujeme pouze pro dokumenty, které jsou asociovány s některou vyhovující třídou a hodnoty přesnosti a úplnosti počítáme vzhledem k vyhovujícím třídám. Sledujeme vliv meze četnosti častých jednic a prahu pro zařazení do třídy.

Tabulka 3. Klasifikace NBCI do vyhovujících tříd kolekce Reuters-21578

práh jednic[10]	práh klas.[11]	P [%]	R [%]	F_1	počet jednic	čas výpočtu
1 %	5 %	91,85	*88,74*	***90,27***	7258	1 m 15 s
	10 %	92,12	88,10	**90,07**		
	15 %	92,16	87,70	**89,87**		
	20 %	92,27	87,51	**89,83**		
	argmax	*92,62*	86,14	**89,26**		
3 %	5 %	91,37	88,64	**89,98**	2837	45 s
	10 %	91,59	88,00	**89,76**		
5 %	5 %	91,12	*88,74*	**89,91**	1782	34 s
	10 %	91,35	88,04	**89,66**		

Z tabulky 3 vidíme, že vyšší přesnosti dosahujeme vždy na úkor úplnosti. Nejlepších výsledků jsme v tomto testu dosáhli při použití meze 1 % pro výběr častého itemsetu. Zároveň je patrné, že srovnatelných výsledků dosahujeme při

[9] Testováno v konfiguraci Intel Celeron @ 400 MHz, 128 MB RAM, MS Windows 2000.

[10] Jednice je považována za častou, vyskytuje-li se alespoň v daném procentu dokumentů některé třídy taxonomie.

[11] Dokument zařadíme do všech tříd, pro něž pravděpodobnost zařazení přesáhne stanovené procento nejlepšího zařazení; *argmax* označuje klasifikaci do jediné (nejpravděpodobnější) třídy.

prahu 3 % nebo 5 %, přičemž dramaticky poklesne počet nalezených jednic. Tím se zmenší paměťové nároky i doba potřebná pro výpočet.

V dalším testu se klasifikátor natrénuje na **deset nejčastějších tříd** v kolekci, které obsahují více než 80 % všech dokumentů, přičemž do každého tématu spadá více než 150 trénovacích dokumentů. Toto nastavení se často uvádí v literatuře při srovnávání různých klasifikačních algoritmů.

Tabulka 4. Klasifikace NBCI do 10 největších tříd kolekce Reuters-21578

práh jednic	práh klas.	P [%]	R [%]	F_1	počet jednic	čas výpočtu
3 %	5 %	94,31	*93,31*	**93,81**	1461	29 s
5 %	10 %	94,54	93,14	***93,83***	953	24 s
	20 %	94,60	92,27	**93,42**		
	argmax	*95,04*	90,86	**92,90**		

Výborných výsledků dosahuje metoda NBCI při klasifikaci dokumentů do **dvou tříd**. V našem případě použijeme třídy *earn* a *acq*, jež představují dvě největší témata z kolekce Reuters (viz tabulku 2). V tomto případě je vhodné zařazovat dokument vždy do jediné třídy, neboť obě témata obsahují minimální počet společných dokumentů.

Tabulka 5. Klasifikace NBCI dvou největších tříd kolekce Reuters-21578

práh jednic	práh klas.	P [%]	R [%]	F_1	počet jednic	čas výpočtu
1 %	5 %	97,35	*98,44*	**97,89**	939	22 s
	15 %	97,65	98,25	***97,95***		
	argmax	*97,95*	97,75	**97,85**		

5.2 Srovnání s původními metodami

V tabulce 6 uvádíme výsledky metody *Itemsets* ve srovnání s nejlepšími výsledky klasifikátorů NB[12] a NBCI pro jednotlivé testovací případy.

Tabulka 6. Srovnání klasifikačních metod Itemsets, NB a NBCI

Použité	Itemsets			NB			NBCI		
třídy	P[%]	R[%]	F_1	P[%]	R[%]	F_1	P[%]	R[%]	F_1
všech 116	52,36	50,40	**51,36**	85,93	81,02	**83,40**	89,91	86,14	**87,98**
28 vyhovujících	82,59	79,60	**81,07**	89,92	85,88	**87,85**	91,85	88,74	**90,27**
10 nejčastějších	89,57	87,59	**88,57**	94,77	91,67	**93,29**	94,54	93,14	**93,83**
earn, *acq*	93,97	94,74	**94,35**	97,62	97,42	**97,52**	97,95	97,75	**97,85**

Z údajů v tabulce 6 vidíme, že NBCI vykazuje ve všech případech lepší výsledky, než původní metody. Výrazný je rozdíl zejména při klasifikaci do všech

[12] Jedná se o re-implementaci NB určenou pro základ klasifikátoru NBCI.

tříd. Změnou parametrů u metody *Itemsets* můžeme sice touto metodou dosáhnout lepších výsledků (zejména pokud jde o přesnost zařazení), ovšem pouze za cenu, že velká část dokumentů zůstane nezařazena.

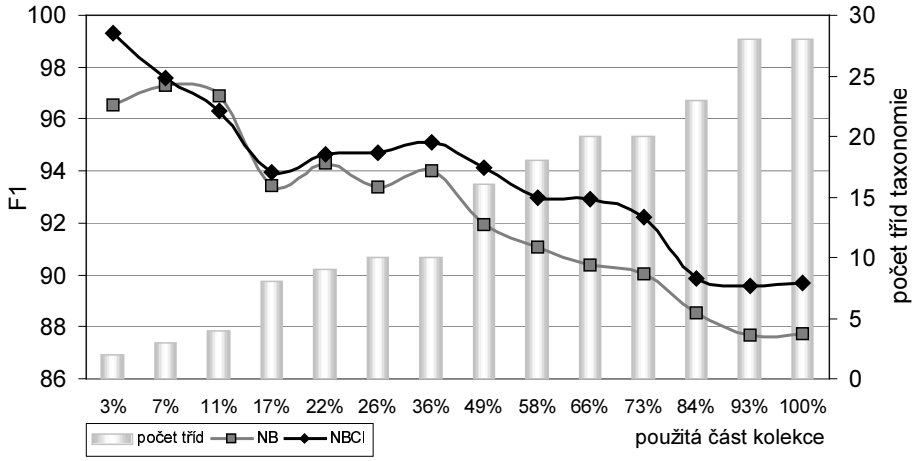
Klasifikátor *Itemsets* je původně určen pro kategorizaci krátkých textových dokumentů. Velmi dobrých výsledků dosahuje při klasifikaci extrémně krátkých dokumentů (do 30 významových termů). Tabulka 7 ukazuje, že se tato vlastnost přenáší i na metodu NBCI.

Tabulka 7. Srovnání pro krátké dokumenty (do 30 významových termů)

Použité	Itemsets			NB			NBCI		
třídy	P[%]	R[%]	F_1	P[%]	R[%]	F_1	P[%]	R[%]	F_1
všech 56	85,90	85,53	**85,71**	95,77	95,11	**95,44**	96,89	96,96	**96,93**
10 nejčastějších	94,93	94,57	**94,75**	97,96	97,56	**97,76**	98,16	98,23	**98,19**
4 vyhovující	96,81	96,50	**96,65**	98,16	98,31	**98,23**	98,44	98,31	**98,37**
earn, acq	98,17	98,30	**98,24**	98,83	98,83	**98,83**	99,09	99,09	**99,09**

Ve všech testovacích případech jsme při výpočtu podmíněných pravděpodobností v rámci klasifikátoru NB i NBCI používali *Laplace smoothing* podle vztahu (7), neboť měl příznivý vliv na výsledky klasifikace. Zlepšení bylo patrné zejména u původního klasifikátoru NB (až o 15 % vyšší hodnoty F_1).

Detailní porovnání klasifikátorů NB a NBCI je uvedeno na obrázku 1. Z průběhu úspěšnosti klasifikace je znát silný vliv původního algoritmu NB na metodu NBCI. Rovněž je patrné, že NBCI dosahuje lepších výsledků než NB, a to zejména při větším množství vstupních dat.



Obr. 1. Srovnání klasifikátorů NB a NBCI v závislosti na množství použitých dokumentů kolekce Reuters-21578 při klasifikaci do vyhovujících tříd

Přestože vlastní fáze klasifikace probíhá u NBCI pomaleji než u původní metody NB[13], celkové paměťové a časové nároky hovoří ve prospěch klasifiká-

[13] V klasifikovaných dokumentech třeba nejprve nalézt itemsety získané během trénování.

toru NBCI. Během trénování dochází u NBCI ke značné redukci dimenze prostoru atributů (analogicky metodě *Itemsets*) podmíněné nahrazením signifikatních termů v dokumentech častými itemsety. Při klasifikaci do vyhovujících tříd taxonomie (tj. tříd s 50 a více dokumenty) vystačíme zhruba s 20 MB operační paměti a celý proces trénování a klasifikace trvá asi jednu minutu (při stejné konfiguraci, viz dříve).

Z dosažených výsledků je zřejmé, že metoda NBCI je životaschopná. Náš další výzkum bude zaměřen zejména na další praktické aplikace metod induktivního strojového učení. Za zajímavou aplikační oblast považujeme např. vyhledávání odborníků v rámci podnikového intranetu, tvořící součást podnikového znalostního portálu. Zjednodušená verze tohoto systému je již implementována ve znalostním portálu fy inSITE. Pozornost bude věnována i další práci na algoritmu kombinujícím shlukování (učení bez supervize) se supervizovaným trénováním (klasifikátor Itemsets v kombinaci s některou vektorovou metodou). Náš výzkum je rovněž zaměřen na automatickou sumarizaci plnotextových dokumentů, zejména za účelem klasifikace a shlukování anotací na výstupu tohoto procesu.

Reference

1. Hynek J., Ježek K. *Use of Text Mining Methods in a Digital Library.* Mezinárodní konference o elektronickém publikování elpub 2002 Karlovy Vary, 6.–8. 11. 2002, přijatý příspěvek
2. Hynek J., Ježek K. *Automatic document classification using Itemsets Method, its modifications and evaluation.* Sborník mezinárodní databázové konference Datakon 2001 Brno, Mária Bieliková (Ed.), ISBN: 80-227-1597-2
3. Hynek J., Ježek K. *Document Classification Using Itemsets.* Sborník mezinárodní konference MOSIS 2000 Rožnov pod Radhoštěm, ISBN: 80-85988-45-3
4. Agrawal R. et al. *Advances in Knowledge Discovery and Data Mining*, MIT Press, 1996, pp. 307–328
5. The Reuters-21578 Text Categorization Test Collection (retrieved 1/2002) URL: `http://www.research.att.com/~lewis/reuters21578.html`

Annotation. Research in methods of text categorization and retrieval in digital libraries represents a challenging task. This conference paper introduces a new method based on the principles of inductive machine learning, combining the Naive Bayes classifier and our original Itemsets classification method. We can describe it briefly either as the Itemsets method utilizing Naive Bayes classifier, or as the Naive Bayes Method approximating documents by frequent itemsets. This new and robust method is relatively easy to implement. We have tested the method on Reuters-21578 document collection. Results presented in the paper demonstrate quality of the method being proposed.

Partly supported by grant No. MSM235200005