Positive – unlabeled learning

na przykładzie klasyfikacji tekstu

Mateusz Lewandowski
Łukasz Osiński


WANTED TO TRAIN A RAINY WEATHER CLASSIFIER
BUT
IT NEVER RAINS IN THE BAY AREA
quickmeme.com

Klasyfikacja binarna:

- Zbiór treningowy w pełni otagowany,
- Zarówno przykłady pozytywne jak i negatywne
- np. Naive Bayes, Support Vector Machine
- Łatwo zmierzyć skuteczność algorytmu (F-score)

Positive - unlabeled learning:

- Nie wszystkie dane uczące są otagowane
- Zbiór składa się z przykładów pozytywnych i nieotagowanych (mogą być pozytywne lub negatywne)
- np. PEBL, S-EM
- Problemy pomiaru skuteczności algorytmu (F-score)

Positive - unlabeled learning 2 główne etapy:

Positive - unlabeled learning 2 główne etapy:

Etap 1: Wydzielenie negatywnych przykładów ze zbioru nieotagowanych danych U

Positive - unlabeled learning 2 główne etapy:

Etap 1: Wydzielenie negatywnych przykładów ze zbioru nieotagowanych danych U

Etap 2: Klasyfikacja przykładów pozostałych w zbiorze nieotagowanym U

Algorytmy

Etap 1 / Etap 2	Expectation Maximization (NB)	SVM-IS	SVM-I	SVM (alone)
Spy technique				
1-DNF				
Rocchio algorithm				
NB				

Algorytmy

Etap 1 / Etap 2	Expectation Maximization (NB)	SVM-IS	SVM-I	SVM (alone)
Spy technique	S-EM			
1-DNF			PEBL	
Rocchio algorithm		Roc-SVM		
NB	Naive Bayes			

Algorytmy

Etap 1 / Etap 2	Expectation Maximization (NB)	SVM-IS	SVM-I	SVM (alone)
Spy technique	S-EM			
1-DNF			PEBL	
Rocchio algorithm		Roc-SVM		
NB	Naive Bayes			NB-SVM alone

NB - biased SVM

Etap 1: Spy technique

Spy technique

1. $RN = NULL$;
2. $S = Sample(P, s\%)$;
3. $Us = U \cup S$;
4. $Ps = P - S$;
5. Assign each document in Ps the class label 1;
6. Assign each document in Us the class label -1;
7. I-EM(Us, Ps); // This produces a NB classifier.
8. Classify each document in Us using the NB classifier;
9. Determine a probability threshold th using S;
10. **for** each document $d \in Us$
11. **if** its probability $Pr(1|d) < th$ **then**
12. $RN = RN \cup \{d\}$;

U - nieotagowane
S - wylosowane pozytywne
RN - negatywne ze zbioru U
th - próg odcięcia, margines
-----błędu klasyfikacji nr 1
s ~15%

Etap 1: 1-DNF

1-DNF

1. Assume the word feature set be $\{x_1, \ldots, x_n\}$, $x_i \in U \cup P$;
2. Let positive feature set PF = null;
3. **for** $i = 1$ to n
4. **if** $(freq(x_i, P)/|P| > freq(x_i, U)/|U|)$ **then**
5. $PF = PF \cup \{x_i\}$;
6. $RN = U$;
7. **for** each document $d \in U$
8. **if** $\exists x_j\, freq(x_j, d) > 0$ and $x_j \in PF$ **then**
9. $RN = RN - \{d\}$;

U - nieotagowane
P - przykłady pozytywne
S - wylosowane pozytywne
RN - negatywne ze zbioru U
th - próg odcięcia, margines
-----błędu klasyfikacji nr 1
s ~15%
freq(xi, S) - licz. wyst. słowa
----------xi w zbiorze S
PF - zbiór słów (cech
pozytywnych positive feature

Etap 1: Rocchio algorithm

Rocchio algorithm

$$\vec{c}_j = \alpha \frac{1}{|C_j|} \sum_{\vec{d} \in C_j} \frac{\vec{d}}{\|\vec{d}\|} - \beta \frac{1}{|D - C_j|} \sum_{\vec{d} \in D - C_j} \frac{\vec{d}}{\|\vec{d}\|}$$

cj - wektor reprezentujący wszystkie dokumenty zaetykietowane do klasy j (class j)
d - wektor dokumentu (składowymi są słowa kluczowe oraz ich wystąpienia)
|D| - ilość wszystkich przykładów (obiektów)
|Cj| - ilość wszystkich przykładów zaklasyfikowany do klasy j (Class j)

alfa - parametr reguluje wpływ dokumentów pozytywnych na klasyfikację
beta - parametr reguluje wpływ dokumentów nieotagowanych na klasyfikację
ważny jest stosunek alfa i beta!

Etap 1: Naive Bayes

Naive Bayes

$$\Pr(c_j) = \frac{\sum_{i=1}^{|D|} \Pr(c_j \mid d_i)}{|D|},$$

$$\Pr(x_t \mid c_j) = \frac{\lambda + \sum_{i=1}^{|D|} N(x_t, d_i) \Pr(c_j \mid d_i)}{\lambda |V| + \sum_{s=1}^{|V|} \sum_{i=1}^{|D|} N(x_s, d_i) \Pr(c_j \mid d_i)}$$

$$\Pr(c_j \mid d_i) = \frac{\Pr(c_j) \prod_{k=1}^{|d_i|} \Pr(x_{d_i,k} \mid c_j)}{\sum_{r=1}^{|C|} \Pr(c_r) \prod_{k=1}^{|d_i|} \Pr(x_{d_i,k} \mid c_r)}.$$

N(xt, di) - liczba wystąpień słowa t w dokumencie i
gamma - współczynnik wygładzania
x(di, k) - słowo k w dokumencie i
C - dokumenty (przykłady) zaklasyfikowane do klasy C (tu: pozytywne, negatywne)
V - zbiór wyrazów będących cechami, słownik (vocabulary)

Naive Bayes

1. Assign each document in P the class label 1;
2. Assign each document in U the class label -1;
3. Build a NB classifier using P and U;
4. Use the classifier to classify U. Those documents in U that are classified as negative form the reliable negative set RN.

Etap 2: Expectation Maximization (EM, NB)

Expectation Maximization

1. Each document in P is assigned the class label 1;
2. Each document in RN is assigned the class label -1;
3. Each document $d \in Q$ ($= U - RN$) is not assigned any label initially. At the end of the first iteration of EM, it will be assigned a probabilistic label, $Pr(1|d)$. In subsequent iterations, the set Q will participate in EM with its newly assigned probabilistic classes.
4. Run the EM algorithm using the document sets, P, RN and Q until it converges.

Etap 2: Support Vector Machine (SVM-I)

Iterative SVM (SVM-I)

1. Every document in P is assigned the class label 1;
2. Every document in RN is assigned the class label –1;
3. $i = 1$;
4. **Loop**
5. Use P and RN to train a SVM classifier S_i;
6. Classify Q using S_i;
7. Let the set of documents in Q that are classified as negative be W;
8. **if** $W = \{\}$ **then** *exit-loop*
9. **else** $Q = Q - W$;
10. $RN = RN \cup W$;
11. $i = i + 1$;

Etap 2: Support Vector Machine (SVM-IS)

Iterative SVM with Classifier Selection (SVM-IS)

1. Every document in P is assigned the class label 1;
2. Every document in RN is assigned the class label –1;
3. $i = 1$;
4. **Loop**
5. Use P and RN to train a SVM classifier S_i;
6. Classify Q using S_i;
7. Let the set of documents in Q that are classified as negative be W;
8. **if** $W = \{\}$ **then** *exit-loop*
9. **else** $Q = Q - W$;
10. $RN = RN \cup W$;
11. $i = i + 1$;

Iterative SVM with Classifier Selection (SVM-IS)

1. Every document in P is assigned the class label 1;
2. Every document in RN is assigned the class label –1;
3. $i = 1$;
4. **Loop**
5. Use P and RN to train a SVM classifier S_i;
6. Classify Q using S_i;
7. Let the set of documents in Q that are classified as negative be W;
8. **if** $W = \{\}$ **then** *exit-loop*
9. **else** $Q = Q - W$;
10. $RN = RN \cup W$;
11. $i = i + 1$;

1. Use the last SVM classifier S_{last} to classify P;
2. **if** > 8% positive are classified as negative **then**
3. use S_1 as the final classifier;
4. **else** use S_{last} as the final classifier;

Etap 2: SVM - alone

SVM (alone)

Minimize: $\frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w}$

Subject to: $y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1 \quad , \quad i = 1, 2, ..., n$

- idea podziału przestrzeni wielowymiarowej hiperpłaszczyzną
- maksymalizacja minimalnego marginesu błędu

SVM (alone)

Minimize: $\frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + C\sum_{i=1}^{n}\xi_i$

Subject to: $y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1 - \xi_i, \quad i = 1, 2, ..., n$

- idea podziału przestrzeni wielowymiarowej hiperpłaszczyzną
- maksymalizacja minimalnego marginesu błędu
- 1 - eta - funkcja błędu zawiasowego
- C - wpływa na dopuszczalny poziom błędu na zbiorze uczącym

Etap 2: Biased SVM

Biased SVM

Minimize: $\frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + C\sum_{i=k}^{n}\xi_i$

Subject to: $\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b \geq 1, \quad i = 1, 2, \ldots, k-1$

$-1(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1 - \xi_i, \quad i = k, k+1, \ldots, n$

$\xi_i \geq 0, \; i = k, k+1 \ldots, n$

- założenie: pierwsze k-1 przykładów pozytywnych, pozostałe negatywne
- dla wystarczająco dużego zbioru uczącego minimalizacja liczby przykładów z U do P, przy jednoczesnym ograniczeniu, że wszystkie pozytywne przykłady muszą być zaklasyfikowane jako pozytywne, pozwala stworzyć dobry klasyfikator

Biased SVM

Minimize: $\frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + C_{+}\sum_{i=1}^{k-1}\xi_i + C_{-}\sum_{i=k}^{n}\xi_i$

Subject to: $y_i(\mathbf{w}^{\mathrm{T}}\mathbf{x}_i + b) \geq 1 - \xi_i, \quad i = 1, 2, ..., n$

$\xi_i \geq 0, i = 1, 2, \ldots, n$

- poprawka związana z tolerancją szumu
- parametry C+ oraz C- wpływają na ważenie tolerancji błędów: false negative (C+) oraz false positive (C-)
- wartości C+ oraz C- dobierane są doświadczalnie
- intuicja: duże wartości C+, małe C-
- często strojenie odbywa się na podstawie miary F-score

F-score

$$F = 2pr/(p+r)$$

- p - precyzja (ang. precision)
- r - czułość (ang. recall)
- preferowane wysokie, wyrównane wartości
- miara od 0.0 do 1.0

Wyniki

$$F = 2pr/(p+r)$$

Table 1: Average F scores on Reuters collection

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Step1	1-DNF	1-DNF		1-DNF		Spy	Spy	Spy	Rocchio	Rocchio	Rocchio		NB	NB	NB	NB	
Step2	EM	SVM	PEBL	SVM-IS	S-EM	SVM	SVM-I	SVM-IS	EM	SVM	SVM-I	Roc-SVM	EM	SVM	SVM-I	SVM-IS	NB
0.1	0.187	0.423	0.001	0.423	0.547	0.329	0.006	0.328	0.644	0.589	0.001	0.589	0.547	0.115	0.006	0.115	0.514
0.2	0.177	0.242	0.071	0.242	0.674	0.507	0.047	0.507	0.631	0.737	0.124	0.737	0.693	0.428	0.077	0.428	0.681
0.3	0.182	0.269	0.250	0.268	0.659	0.733	0.235	0.733	0.623	0.780	0.242	0.780	0.695	0.664	0.235	0.664	0.699
0.4	0.178	0.190	0.582	0.228	0.661	0.782	0.549	0.780	0.617	0.805	0.561	0.784	0.693	0.784	0.557	0.782	0.708
0.5	0.179	0.196	0.742	0.358	0.673	0.807	0.715	0.799	0.614	0.790	0.737	0.799	0.685	0.797	0.721	0.789	0.707
0.6	0.180	0.211	0.810	0.573	0.669	0.833	0.804	0.820	0.597	0.793	0.813	0.811	0.670	0.832	0.808	0.824	0.694
0.7	0.175	0.179	0.824	0.425	0.667	0.843	0.821	0.842	0.585	0.793	0.823	0.834	0.664	0.845	0.822	0.843	0.687
0.8	0.175	0.178	0.868	0.650	0.649	0.861	0.865	0.858	0.575	0.787	0.867	0.864	0.651	0.859	0.865	0.858	0.677
0.9	0.172	0.190	0.860	0.716	0.658	0.859	0.859	0.853	0.580	0.776	0.861	0.861	0.651	0.846	0.858	0.845	0.674

Table 2: Average F scores on 20Newsgroup collection

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Step1	1-DNF	1-DNF		1-DNF		Spy	Spy	Spy	Rocchio	Rocchio	Rocchio		NB	NB	NB	NB	
Step2	EM	SVM	PEBL	SVM-IS	S-EM	SVM	SVM-I	SVM-IS	EM	SVM	SVM-I	Roc-SVM	EM	SVM	SVM-I	SVM-IS	NB
0.1	0.145	0.545	0.039	0.545	0.460	0.097	0.003	0.097	0.557	0.295	0.003	0.295	0.368	0.020	0.003	0.020	0.333
0.2	0.125	0.371	0.074	0.371	0.640	0.408	0.014	0.408	0.670	0.546	0.014	0.546	0.649	0.232	0.013	0.232	0.611
0.3	0.123	0.288	0.201	0.288	0.665	0.625	0.154	0.625	0.673	0.644	0.121	0.644	0.689	0.469	0.120	0.469	0.674
0.4	0.122	0.260	0.342	0.258	0.683	0.684	0.354	0.684	0.671	0.690	0.385	0.682	0.705	0.610	0.354	0.603	0.704
0.5	0.121	0.248	0.563	0.306	0.685	0.715	0.560	0.707	0.663	0.716	0.565	0.708	0.702	0.680	0.554	0.672	0.707
0.6	0.123	0.209	0.646	0.419	0.689	0.758	0.674	0.746	0.663	0.747	0.683	0.738	0.701	0.737	0.670	0.724	0.715
0.7	0.119	0.196	0.715	0.563	0.681	0.774	0.731	0.757	0.660	0.754	0.731	0.746	0.699	0.763	0.728	0.749	0.717
0.8	0.124	0.189	0.689	0.508	0.680	0.789	0.760	0.783	0.654	0.761	0.763	0.766	0.688	0.780	0.758	0.774	0.707
0.9	0.123	0.177	0.716	0.577	0.684	0.807	0.797	0.798	0.654	0.775	0.798	0.790	0.691	0.806	0.797	0.798	0.714

Dziękujemy