Nikola Mrksic (nm480@cam.ac.uk)

SSL methods for data augmentation: current results

KDD2001, Binding to Thrombin

	Precision	Recall	F-Score
SVM Performance on full data set, c = 0.1:	0.2941	0.0333	0.0599
c is the kkt_threshold, i.e. the number of entries in the training data allowed to stay on the wrong side of the hyperplane while training the SVM.			
Reduced set 1(10% uniformly sampled from the original training data):			
SVM Performance on the reduced data set:	0.1429	0.0067	0.0127
SVM Performance on the augmented data sets:			
SVMAugmented Bayes:	0.1818	0.12	0.1446
SVM Spy-EM:	0	0	0
Quality of Entity Set Expansion:			
Bayesian Sets:	0.23	0.6053	0.3333
Spy-EM:	0.27	0.7105	0.3913
Reduced set 2 (33%):			
SVM Performance on the reduced data set:	0.1579	0.02	0.0355
SVM Performance on the augmented data sets:			
SVMAugmented Bayes:	0.1667	0.1133	0.1349
SVM Spy-EM:	0.1183	0.0733	0.0905
Quality of Entity Set Expansion:			
Bayesian Sets:	0.1684	0.5517	0.2581
Spy-EM:	0.2	0.6552	0.3065
Reduced set 3 (80%):			
SVM Performance on the reduced data set:	0.1905	0.0267	0.0468
SVM Performance on the augmented data sets:			
SVMAugmented Bayes:	0.1649	0.1067	0.1296
SVM Spy-EM:	0.1348	0.08	0.1004
Quality of Entity Set Expansion:			
Bayesian Sets:	0.0854	0.7778	0.1538
Spy-EM:	0.0854	0.7778	0.1538

KDD2001, Binding to Thrombin

	Precision	Recall	F-Score
SVM Performance on full data set, c = 0.9:			
	0.2941	0.0333	0.0599
Reduced set 1(10%):			
SVM Performance on the reduced data set:	0.1111	0.0067	0.0126
SVM Performance on the augmented data sets:			
SVMAugmented Bayes:	0.2143	0.18	0.1957
SVM Spy-EM:	0.2143	0.18	0.1957
Quality of Entity Set Expansion:			
Bayesian Sets:	0.2637	0.6316	0.3721
Spy-EM:	0.2857	0.6842	0.4031
Reduced set 2 (33%):			
SVM Performance on the reduced data set:	0.25	0.0267	0.0482
SVM Performance on the augmented data sets:			
SVMAugmented Bayes:	0.2366	1	0.3827
SVM Spy-EM:	0.2366	1	0.3827
Quality of Entity Set Expansion:			
Bayesian Sets:	0.1939	0.6552	0.2992
Spy-EM:	0.2245	0.7586	0.3465
Reduced set 3 (80%):			
SVM Performance on the reduced data set:	0.3571	0.0333	0.061
SVM Performance on the augmented data sets:			
SVMAugmented Bayes:	0.1707	0.0933	0.1207
SVM Spy-EM:	0	0	0
Quality of Entity Set Expansion:			
Bayesian Sets:	0.0602	0.5556	0.1087
Spy-EM:	0.0723	0.6667	0.1304

Reuters 21578, text classification:

	Precision	Recall	F-Score
SVM Performance on full data set, c = 0.1:			
	0.9601	0.8935	0.9256

Same trends for c = 0.5, for 0.9 both fail to achieve much augmentation.

Reduced set 1(10%):

SVM Performance on the reduced data set:	0.9525	0.679	0.7928
SVM Performance on the augmented data sets:			
SVMAugmented Bayes:	0.4762	0.8887	0.6201
SVM Spy-EM:	0.7707	0.9919	0.8674
Quality of Entity Set Expansion:			
Bayesian Sets:	0.6636	1	0.7978
Spy-EM:	0.6626	0.9985	0.7965

Reduced set 2 (33%):

SVM Performance on the reduced data set:	0.9488	0.8065	0.8718
SVM Performance on the augmented data sets:			
SVMAugmented Bayes:	0.5673	0.8839	0.691
SVM Spy-EM:	0.8904	0.9565	0.9222
Quality of Entity Set Expansion:			
Bayesian Sets:	0.5866	1	0.7394
Spy-EM:	0.5854	0.9979	0.7379

Reduced set 3 (80%):

SVM Performance on the reduced data set:	0.9582	0.8871	0.9213
SVM Performance on the augmented data sets:			
SVMAugmented Bayes:	0.7766	0.8581	0.8153
SVM Spy-EM:	0.9265	0.8742	0.8996
Quality of Entity Set Expansion:			
Bayesian Sets:	0.2816	1	0.4395
Spy-EM:	0.2816	1	0.4395

20 Newsgroups, text classification, c = 0.5 (c = 0.9 produces excessively bad classifiers)

SVM Performance on full data set, c = 0.5:	0.9601	0.8935	0.9256

Same trends for c = 0.5, for 0.9 both fail to achieve much augmentation.

Reduced set 1(10%):

SVM Performance on the reduced data set:	0.9525	0.679	0.7928
SVM Performance on the augmented data sets:			
SVMAugmented Bayes:	0.4762	0.8887	0.6201
SVM Spy-EM:	0.7707	0.9919	0.8674
Quality of Entity Set Expansion:			
Bayesian Sets:	0.6636	1	0.7978
Spy-EM:	0.6626	0.9985	0.7965

Reduced set 2 (33%):

SVM Performance on the reduced data set:	0.9488	0.8065	0.8718
SVM Performance on the augmented data sets:			
SVMAugmented Bayes:	0.5673	0.8839	0.691
SVM Spy-EM:	0.8904	0.9565	0.9222
Quality of Entity Set Expansion:			
Bayesian Sets:	0.5866	1	0.7394
Spy-EM:	0.5854	0.9979	0.7379

Reduced set 3 (80%):

SVM Performance on the reduced data set:	0.9582	0.8871	0.9213
SVM Performance on the augmented data sets:			
SVMAugmented Bayes:	0.7766	0.8581	0.8153
SVM Spy-EM:	0.9265	0.8742	0.8996
Quality of Entity Set Expansion:			
Bayesian Sets:	0.2816	1	0.4395
Spy-EM:	0.2816	1	0.4395

20 Newsgroups, c = 0.1

SVM Performance on full data set, c = 0.1:	0.7542	0.9118	0.8256
Reduced set 1(10%):			
SVM Performance on the reduced data set:	0.9547	0.2641	0.4137
SVM Performance on the augmented data sets:			
SVMAugmented Bayes:	0.986	0.0803	0.1484
SVM Spy-EM:	0.9809	0.0876	0.1609
Quality of Entity Set Expansion:			
Bayesian Sets:	0.7558	0.8631	0.8059
Spy-EM:	0.7681	0.8771	0.819
Reduced set 2 (33%):			
SVM Performance on the reduced data set:	1	0.0028	0.0057
SVM Performance on the augmented data sets:			
SVMAugmented Bayes:	0.7047	0.9289	0.8014
SVM Spy-EM:	0.7101	0.9311	0.8057
Quality of Entity Set Expansion:			
Bayesian Sets:	0.6543	0.9031	0.7588
Spy-EM:	0.6619	0.9136	0.7676
Reduced set 3 (80%):			
SVM Performance on the reduced data set:	0.7161	0.9317	0.8098
SVM Performance on the augmented data sets:			
SVMAugmented Bayes:	0.6208	0.9596	0.7539
SVM Spy-EM:	0.6071	0.9664	0.7457
Quality of Entity Set Expansion:			
Bayesian Sets:	0.2737	0.9632	0.4262
Spy-EM:	0.2672	0.9404	0.4161

N.B: The number of new positives extracted by Bayesian Sets is at this moment taken from the number of positives that SEM identifies. At this moment (and as the results given show), it is quite a good threshold: using that value, Bayesian Sets frequently outperform Spy-EM!