Multilabel Attribute Selection

Tanimoto Distance

Sebastian Lehnerer
lehnerer@in.tum.de

Technische Universität München

June 14, 2011

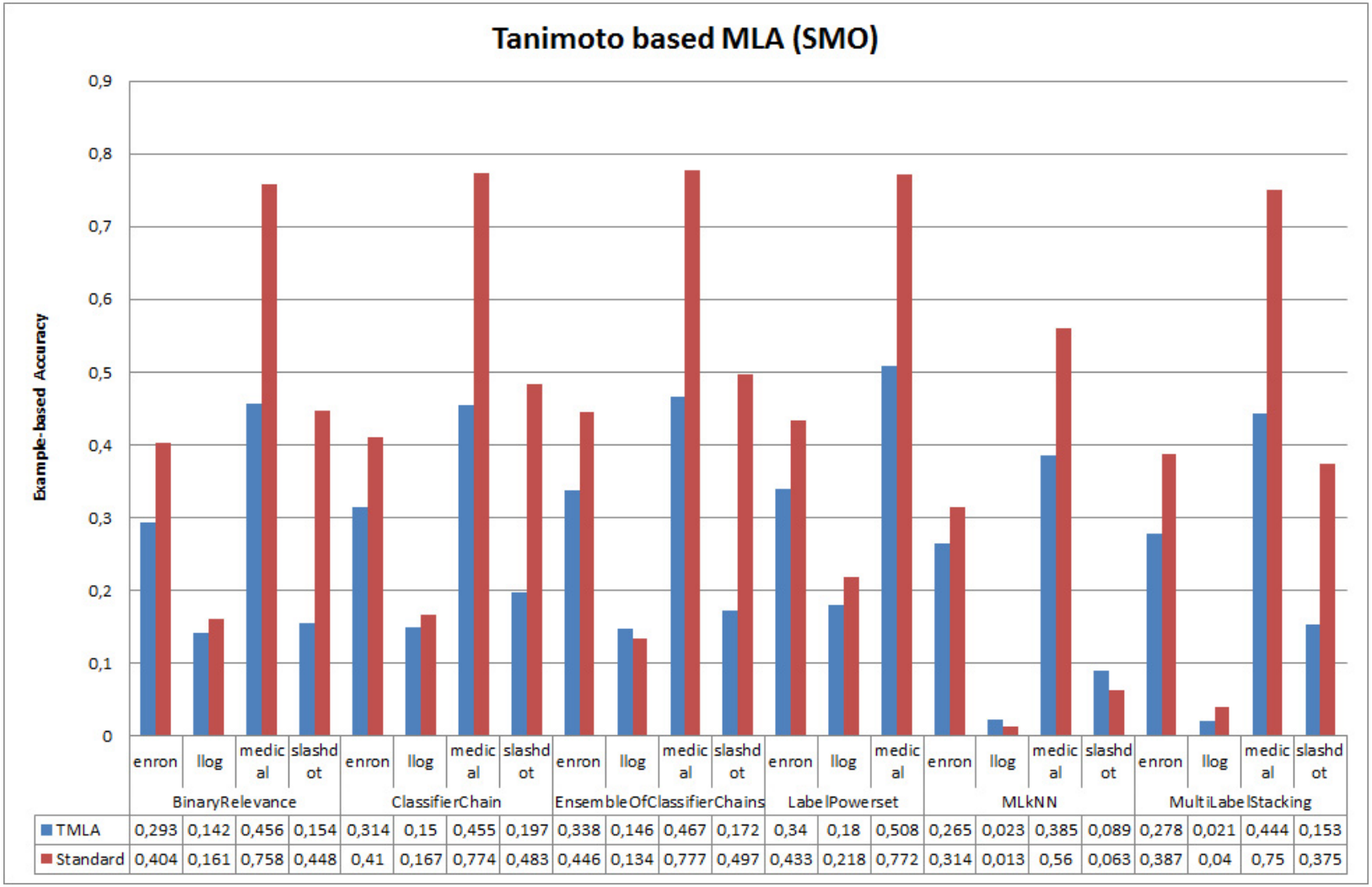
- feature selection for every label, where other labels are treated as normal features

$$
\begin{aligned}
Y_1 &\leftarrow \{X_1 \ldots X_n \cup Y_2 \ldots Y_n | X_i, Y_i \in \{0,1\}\} \\
Y_2 &\leftarrow \{X_1 \ldots X_n \cup Y_1, Y_3 \ldots Y_n | X_i, Y_i \in \{0,1\}\} \\
&\vdots \\
Y_n &\leftarrow \{X_1 \ldots X_n \cup Y_1 \ldots Y_{n-1} | X_i, Y_i \in \{0,1\}\}
\end{aligned}
$$

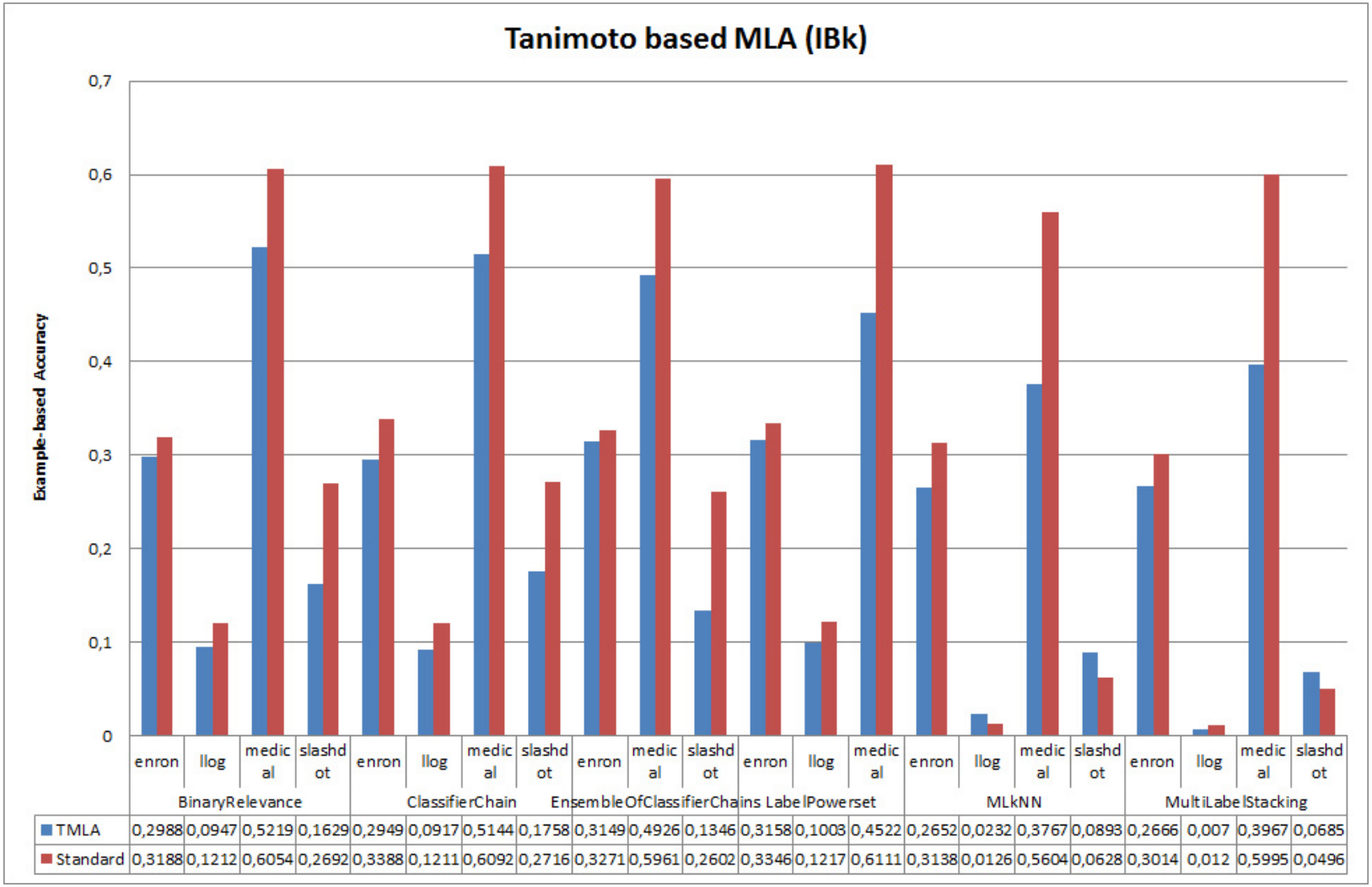
- using label feature sets as vectors $< 0, 1, 0, 0, 1, 0, \cdots, 1, 0 >$
- Hierachical Clustering using the Tanimoto Distance.

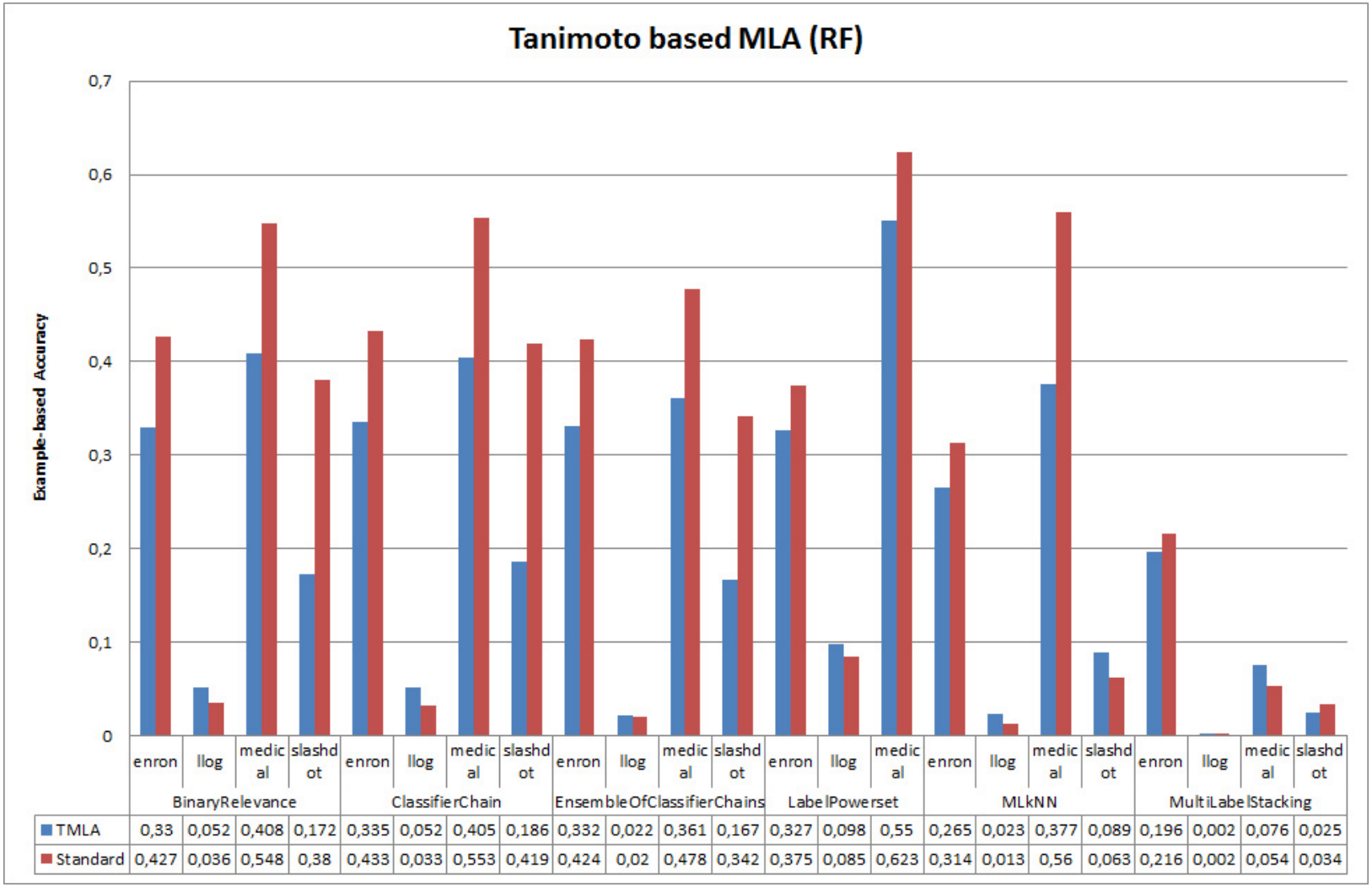
$$T_s(X, Y) = \frac{\sum_i (X_i \wedge Y_i)}{\sum_i (X_i \vee Y_i)}$$

- Single, Complete, Average and Mean Clustering
- no. of clusters: $2, 4, 6$



	BinaryRelevance				ClassifierChain				EnsembleOfClassifierChains				LabelPowerset			MLkNN				MultiLabelStacking			
	enron	llog	medical	slashdot	enron	llog	medical	slashdot	enron	llog	medical	slashdot	enron	llog	medical	enron	llog	medical	slashdot	enron	llog	medical	slashdot
TMLA	0,293	0,142	0,456	0,154	0,314	0,15	0,455	0,197	0,338	0,146	0,467	0,172	0,34	0,18	0,508	0,265	0,023	0,385	0,089	0,278	0,021	0,444	0,153
Standard	0,404	0,161	0,758	0,448	0,41	0,167	0,774	0,483	0,446	0,134	0,777	0,497	0,433	0,218	0,772	0,314	0,013	0,56	0,063	0,387	0,04	0,75	0,375


Tanimoto based MLA (IBk)
Example-based Accuracy
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
enron llog medic al slashd ot enron llog medic al slashd ot enron medic al slashd ot enron llog medic al enron llog medic al slashd ot enron llog medic al slashd ot
BinaryRelevance ClassifierChain EnsembleOfClassifierChains LabelPowerset MLkNN MultiLabelStacking
TMLA 0,2988 0,0947 0,5219 0,1629 0,2949 0,0917 0,5144 0,1758 0,3149 0,4926 0,1346 0,3158 0,1003 0,4522 0,2652 0,0232 0,3767 0,0893 0,2666 0,007 0,3967 0,0685
Standard 0,3188 0,1212 0,6054 0,2692 0,3388 0,1211 0,6092 0,2716 0,3271 0,5961 0,2602 0,3346 0,1217 0,6111 0,3138 0,0126 0,5604 0,0628 0,3014 0,012 0,5995 0,0496



	BinaryRelevance				ClassifierChain				EnsembleOfClassifierChains				LabelPowerset			MLkNN				MultiLabelStacking			
	enron	llog	medical	slashdot	enron	llog	medical	slashdot	enron	llog	medical	slashdot	enron	llog	medical	enron	llog	medical	slashdot	enron	llog	medical	slashdot
TMLA	0,33	0,052	0,408	0,172	0,335	0,052	0,405	0,186	0,332	0,022	0,361	0,167	0,327	0,098	0,55	0,265	0,023	0,377	0,089	0,196	0,002	0,076	0,025
Standard	0,427	0,036	0,548	0,38	0,433	0,033	0,553	0,419	0,424	0,02	0,478	0,342	0,375	0,085	0,623	0,314	0,013	0,56	0,063	0,216	0,002	0,054	0,034

example cluster characteristics (∅ over folds) s

- enron
  ∅ number of clusters : 2
  ∅ number of clusters ($>$ 2 labels) : 2
  ∅ number of labels per cluster: 31
- llog
  ∅ number of clusters : 2
  ∅ number of cluster ($>$ 2 labels) : 1.8
  ∅ number of labels per cluster: 40,60
- medical
  ∅ number of clusters : 6
  ∅ number of cluster ($>$ 2 labels) : 2
  ∅ number of labels per cluster: 8,9
- slashdot
  ∅ number of clusters : 2
  ∅ number of cluster ($>$ 2 labels) : 1.8
  ∅ number of labels per cluster: 14.90