Feature and Label Selection for Multi Label Learning

Sebastian Lehnerer
lehnerer@in.tum.de

Technische Universität München

July 14, 2011

Multi-Label Learning

- find a bipartionen of given labels into positive and negative sets (relevant/irrelevant)
- example
  - instance $I = \{X_1 ... X_n \cup Y_1, Y_2, Y_3, Y_4, Y_5\}$ where feature attributes are X and target (label) attributes are noted as Y.
  - aim is to produce a bipartionen $P_I := \{Y_1, Y_3\}, N_I := \{Y_2, Y_4, Y_5\}$ which classifies labels Y_1, Y_3 as positive, the rest negative.

Multi-Label Learning

- labelset $Y := \{forest, desert, city, island, beach, city, hills, boat\}$
- bipartion
  $P := \{city, beach, hills, boat\}, N := \{forest, desert, island\}$

Feature and Label Selection for Multi Label Learning

- multi-label datasets can be huge and contain a lot of information for different labels.
- labels don't need to overlap, eg. attribute X_1 is useless to Y_1 but determining for Y_2
- *multi-label datasets may contain subsets of features and labels which are highly relevant intrinsically with lower relation to the rest of the data*

Feature and Label Selection for Multi Label Learning

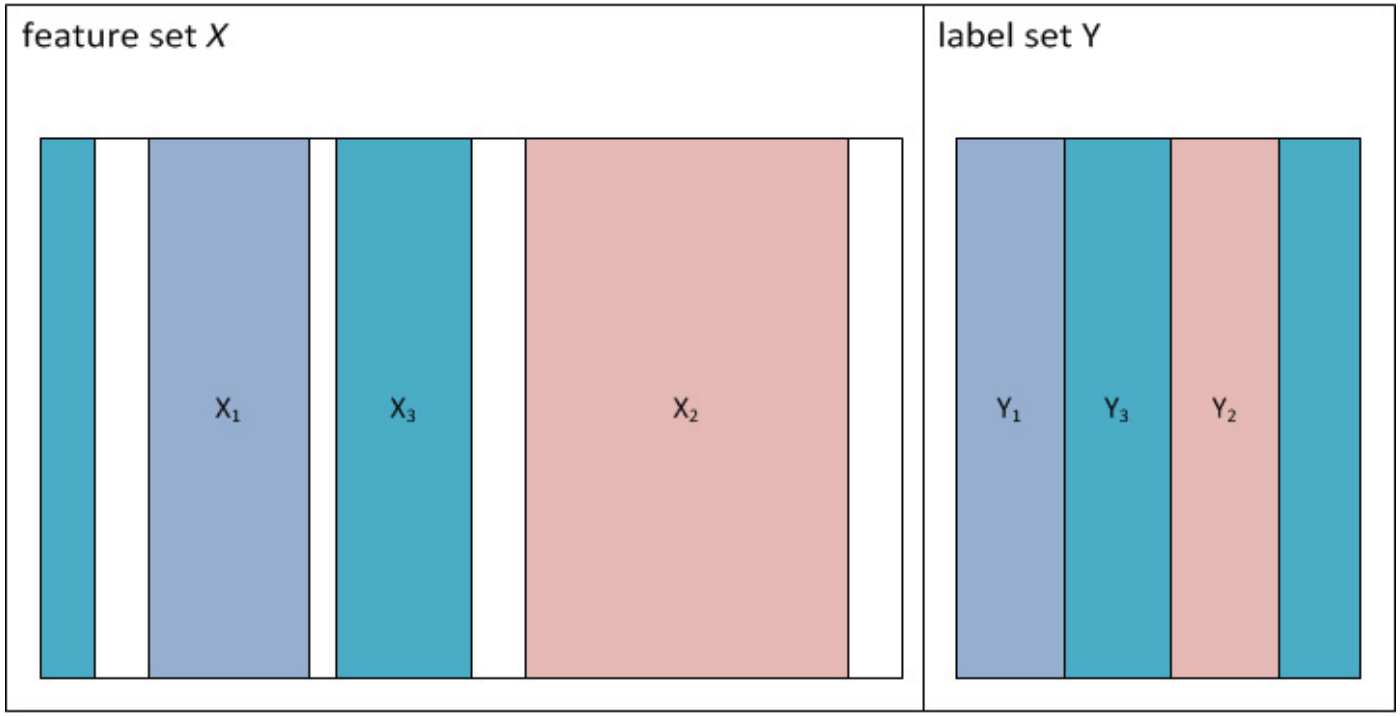


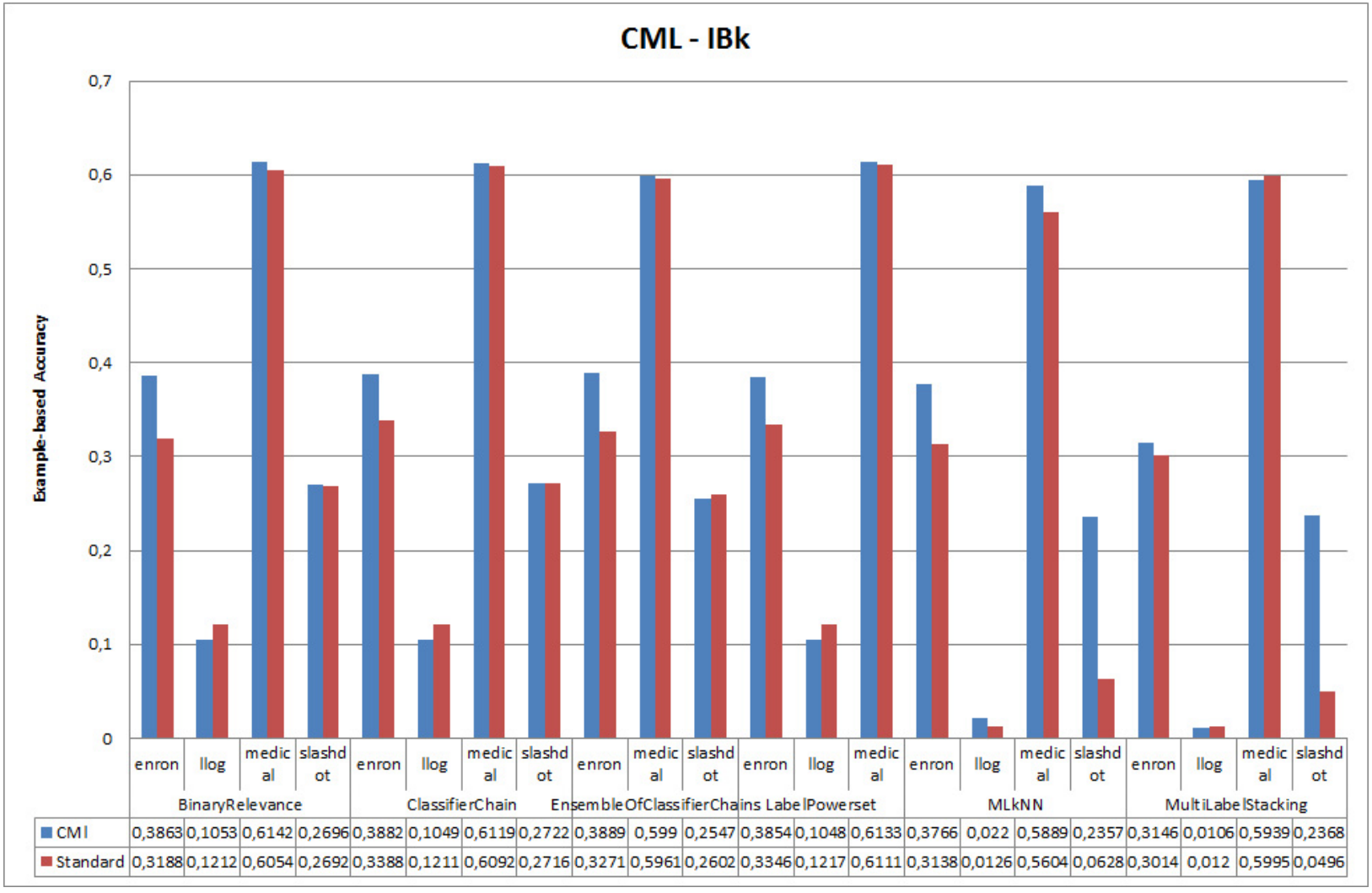
Figure: split of a multi-label dataset in different subsets where the featureset X_i belongs to the labelset Y_i.

Feature and Label Selection for Multi Label Learning

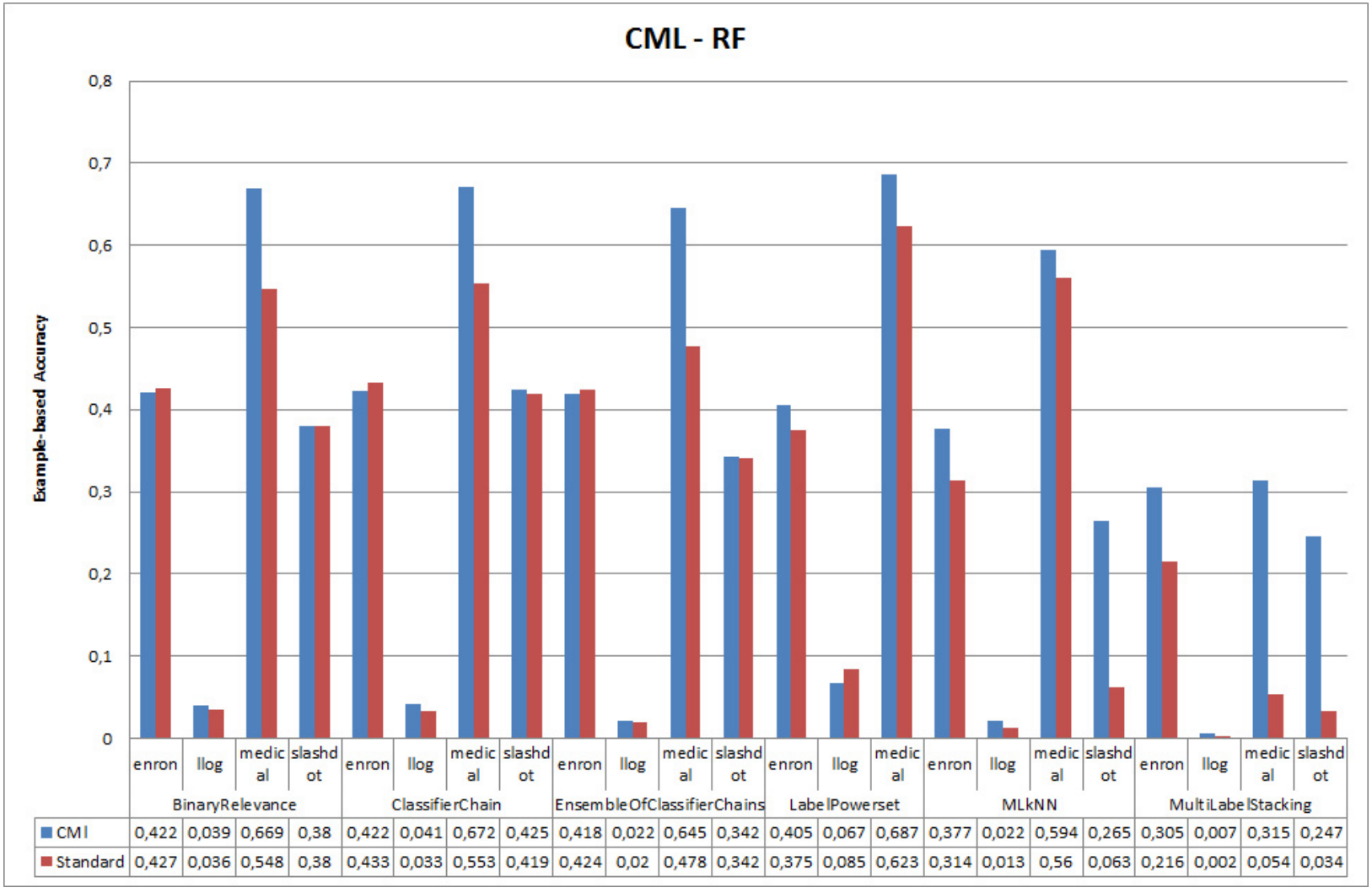
- aim of this work is to find and evaluate those subsets.
- three different methods are developed and tested:
  - FCML (feature score based clustering)
  - TCML (tanimoto distance based clustering)
  - CML (instance based clustering)

CML

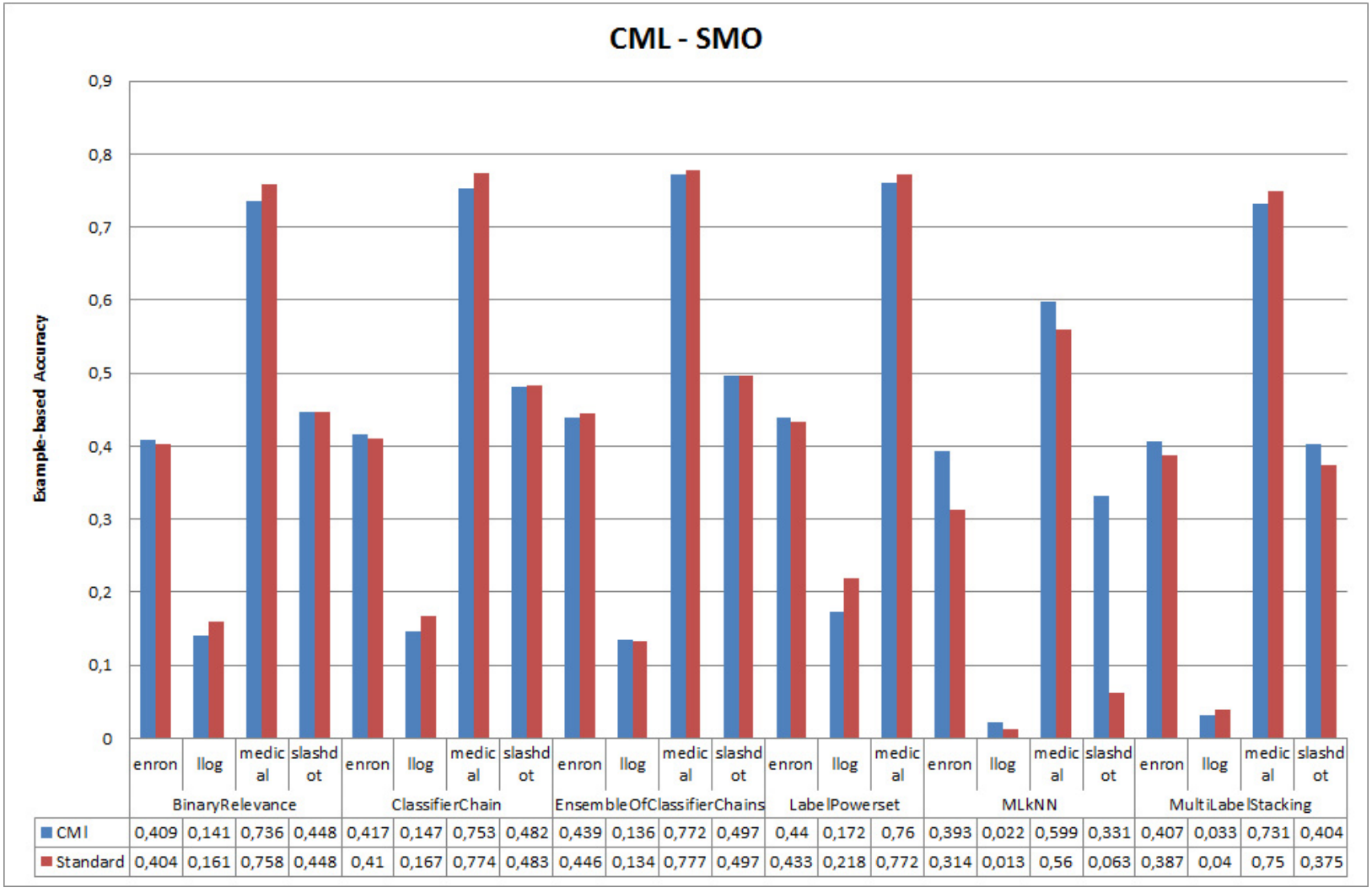
- using transposed dataset, each attribute will become an instance
- clustering over those instances
- clusters resolve directly into groups by splitting into labels and features



	BinaryRelevance				ClassifierChain				EnsembleOfClassifierChains			LabelPowerset			MLkNN				MultiLabelStacking			
	enron	llog	medical	slashdot	enron	llog	medical	slashdot	enron	medical	slashdot	enron	llog	medical	enron	llog	medical	slashdot	enron	llog	medical	slashdot
CMl	0,3863	0,1053	0,6142	0,2696	0,3882	0,1049	0,6119	0,2722	0,3889	0,599	0,2547	0,3854	0,1048	0,6133	0,3766	0,022	0,5889	0,2357	0,3146	0,0106	0,5939	0,2368
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



	enron	llog	medical	slashdot	enron	llog	medical	slashdot	enron	llog	medical	slashdot	enron	llog	medical	enron	llog	medical	slashdot	enron	llog	medical	slashdot
	BinaryRelevance				ClassifierChain				EnsembleOfClassifierChains				LabelPowerset			MLkNN				MultiLabelStacking			
CMI	0,422	0,039	0,669	0,38	0,422	0,041	0,672	0,425	0,418	0,022	0,645	0,342	0,405	0,067	0,687	0,377	0,022	0,594	0,265	0,305	0,007	0,315	0,247
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



	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
CMl	0,409	0,141	0,736	0,448	0,417	0,147	0,753	0,482	0,439	0,136	0,772	0,497	0,44	0,172	0,76	0,393	0,022	0,599	0,331	0,407	0,033	0,731	0,404
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

example cluster characteristics (∅ over folds)

- enron
  ∅ number of clusters : 8
  ∅ number of cluster ($>$ 2 labels) : 2
  ∅ number of labels per cluster: 13.3
- llog
  ∅ number of clusters : 4
  ∅ number of cluster ($>$ 2 labels) : 4
  ∅ number of labels per cluster: 37.5
- medical
  ∅ number of clusters : 8.4
  ∅ number of cluster ($>$ 2 labels) : 2
  ∅ number of labels per cluster: 10.7
- slashdot
  ∅ number of clusters : 10
  ∅ number of cluster ($>$ 2 labels) : 2
  ∅ number of labels per cluster: 4.4

TCML

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

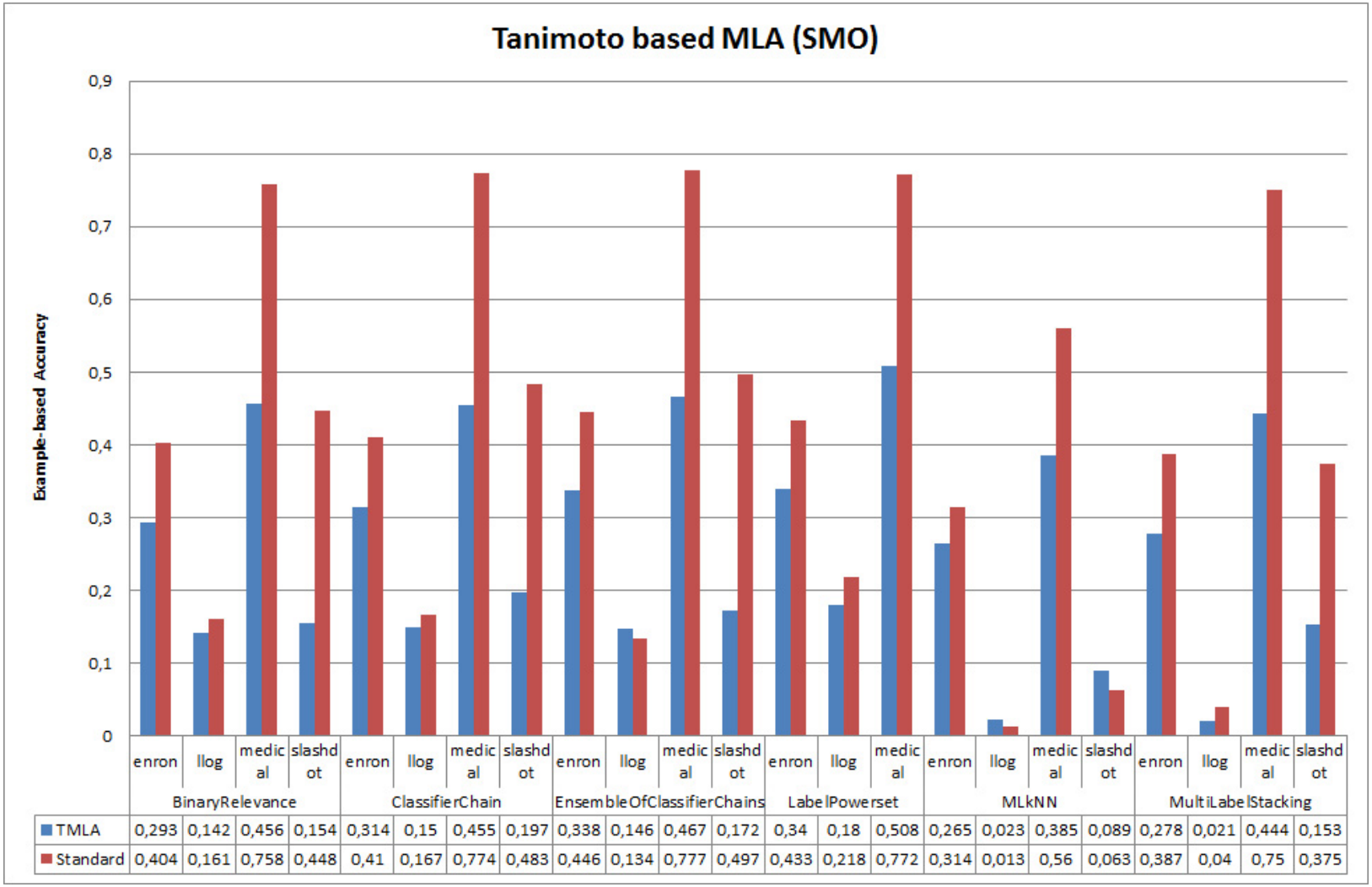
$$
\begin{aligned}
Y_1 &\leftarrow \{X_1 \ldots X_n \cup Y_2 \ldots Y_n | X_i, Y_i \in \{0,1\}\} \\
Y_2 &\leftarrow \{X_1 \ldots X_n \cup Y_1, Y_3 \ldots Y_n | X_i, Y_i \in \{0,1\}\} \\
&\vdots \\
Y_n &\leftarrow \{X_1 \ldots X_n \cup Y_1 \ldots Y_{n-1} | X_i, Y_i \in \{0,1\}\}
\end{aligned}
$$

TCML

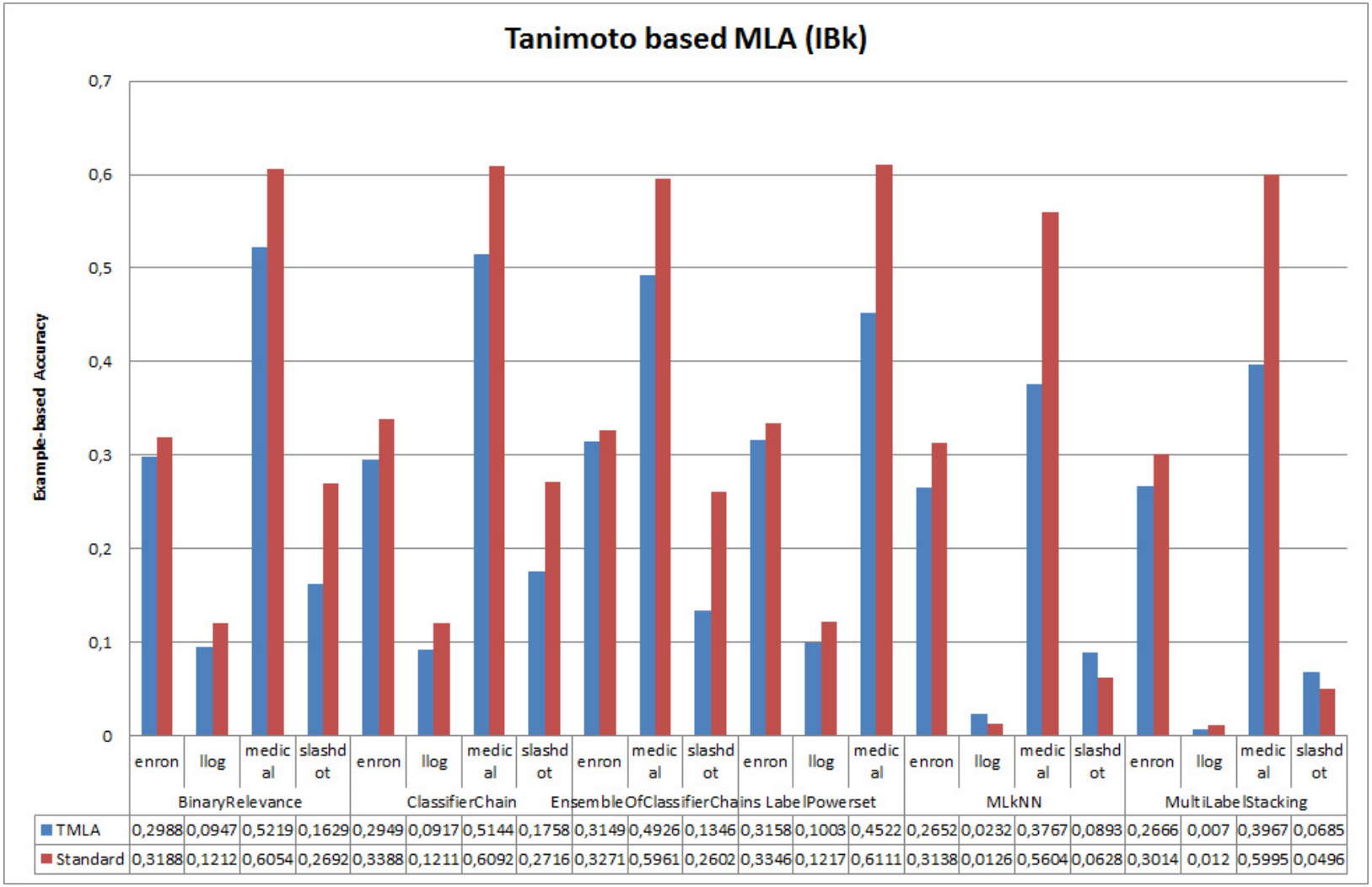
- using label feature sets as vectors $< 0, 1, 0, 0, 1, 0, \cdots, 1, 0 >$
- Hierachical Clustering using the Tanimoto Distance.

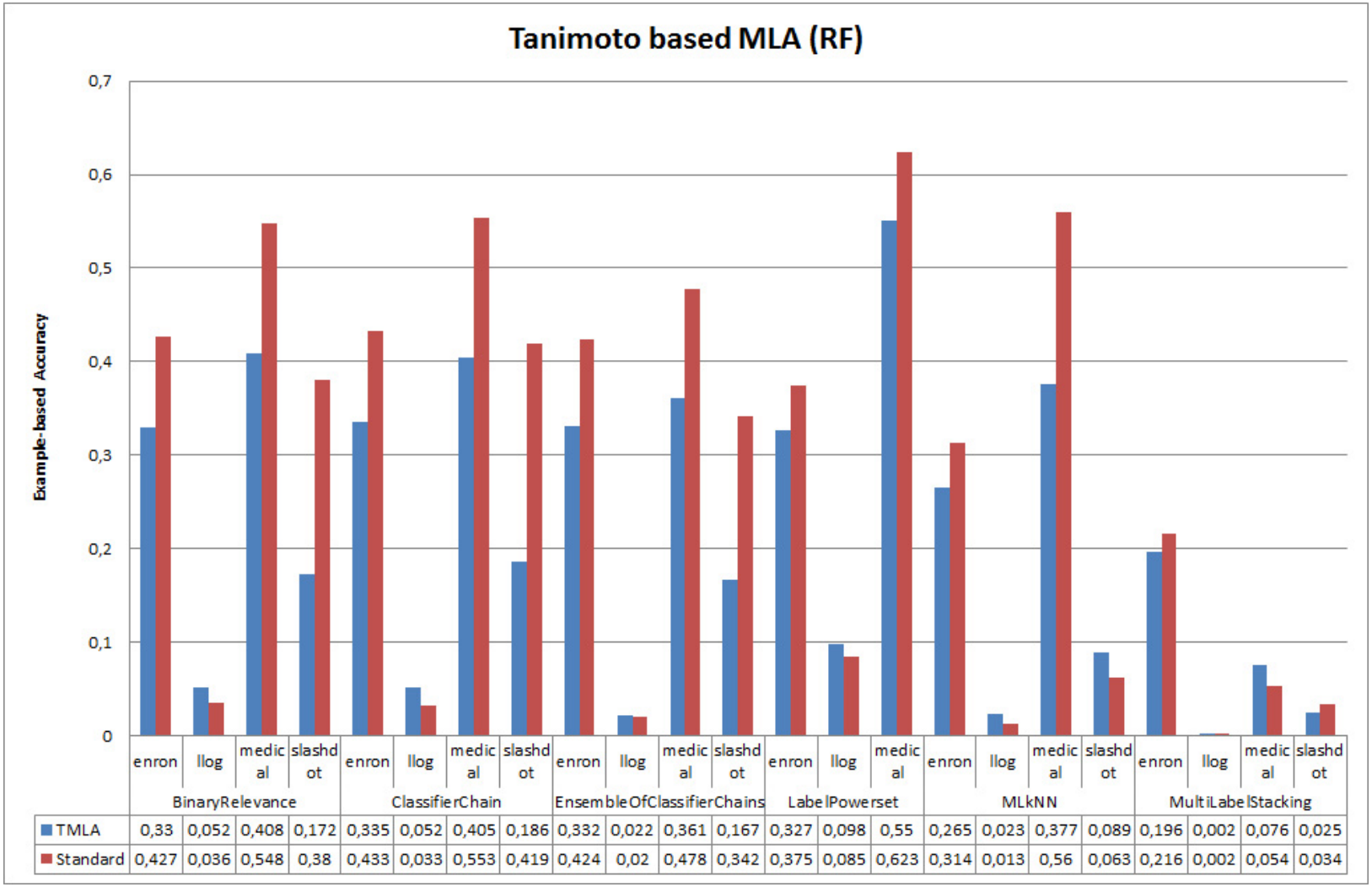
$$T_s(X, Y) = \frac{\sum_i (X_i \wedge Y_i)}{\sum_i (X_i \vee Y_i)}$$

- 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

TCML

example cluster characteristics (∅ over folds) s

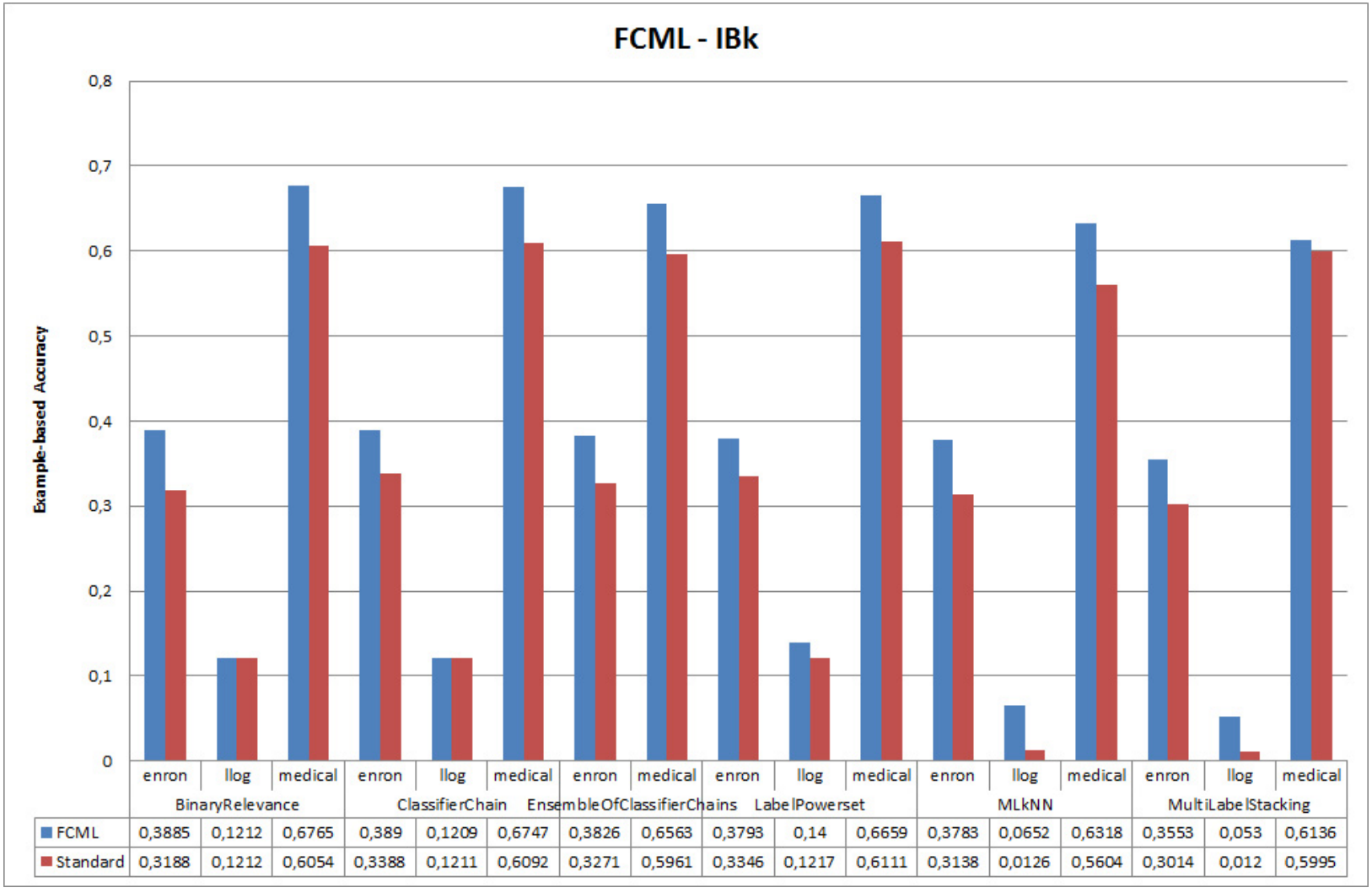
- enron
  ∅ number of clusters : 2
  ∅ number of clusters (> 2 labels) : 2
  ∅ number of labels per cluster: 31
- llog
  ∅ number of clusters : 2
  ∅ number of cluster (> 2 labels) : 1.8
  ∅ number of labels per cluster: 40,60
- medical
  ∅ number of clusters : 6
  ∅ number of cluster (> 2 labels) :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

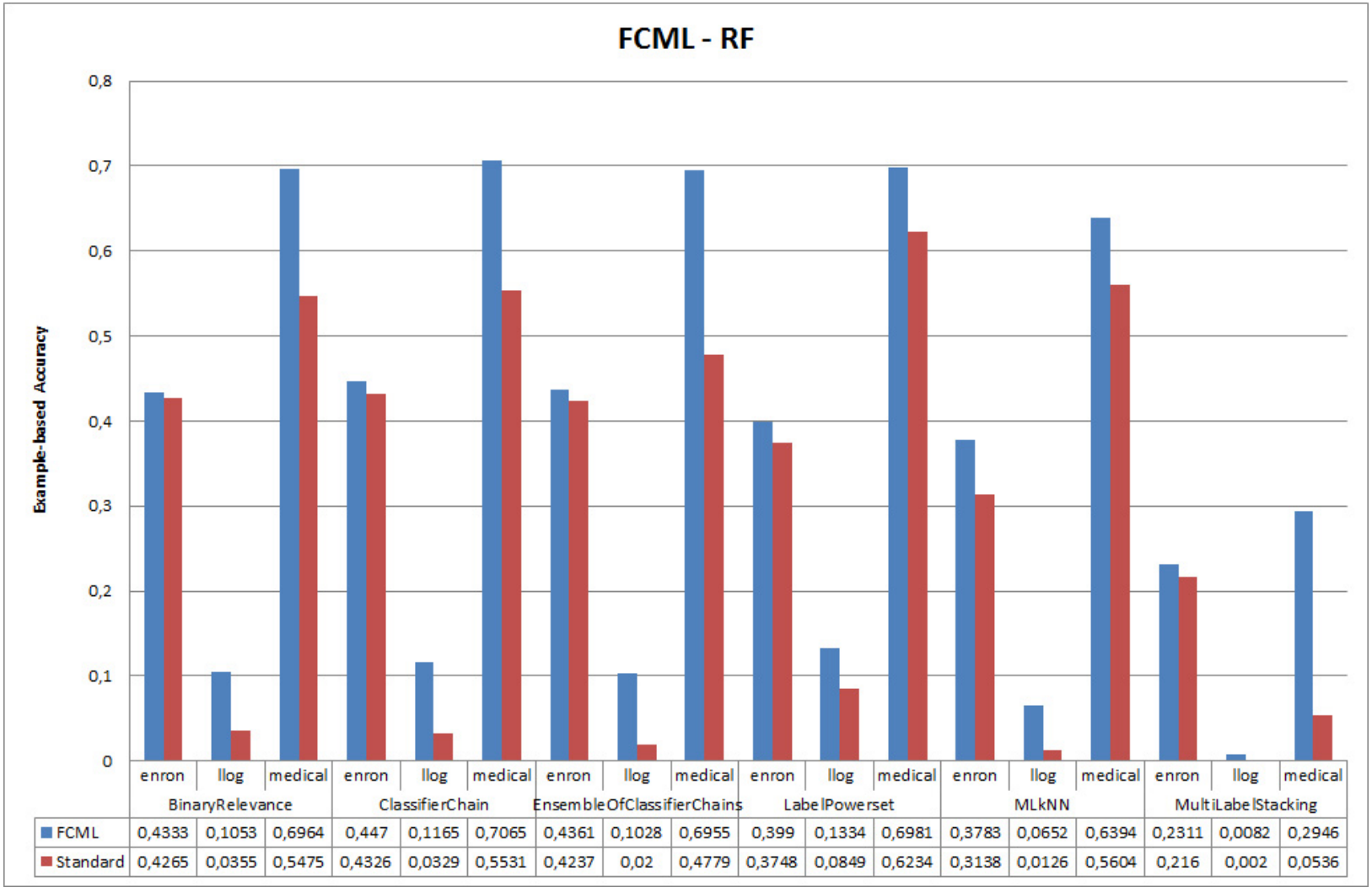
FCML

- feature selection for each label
- using log-scores for further processing

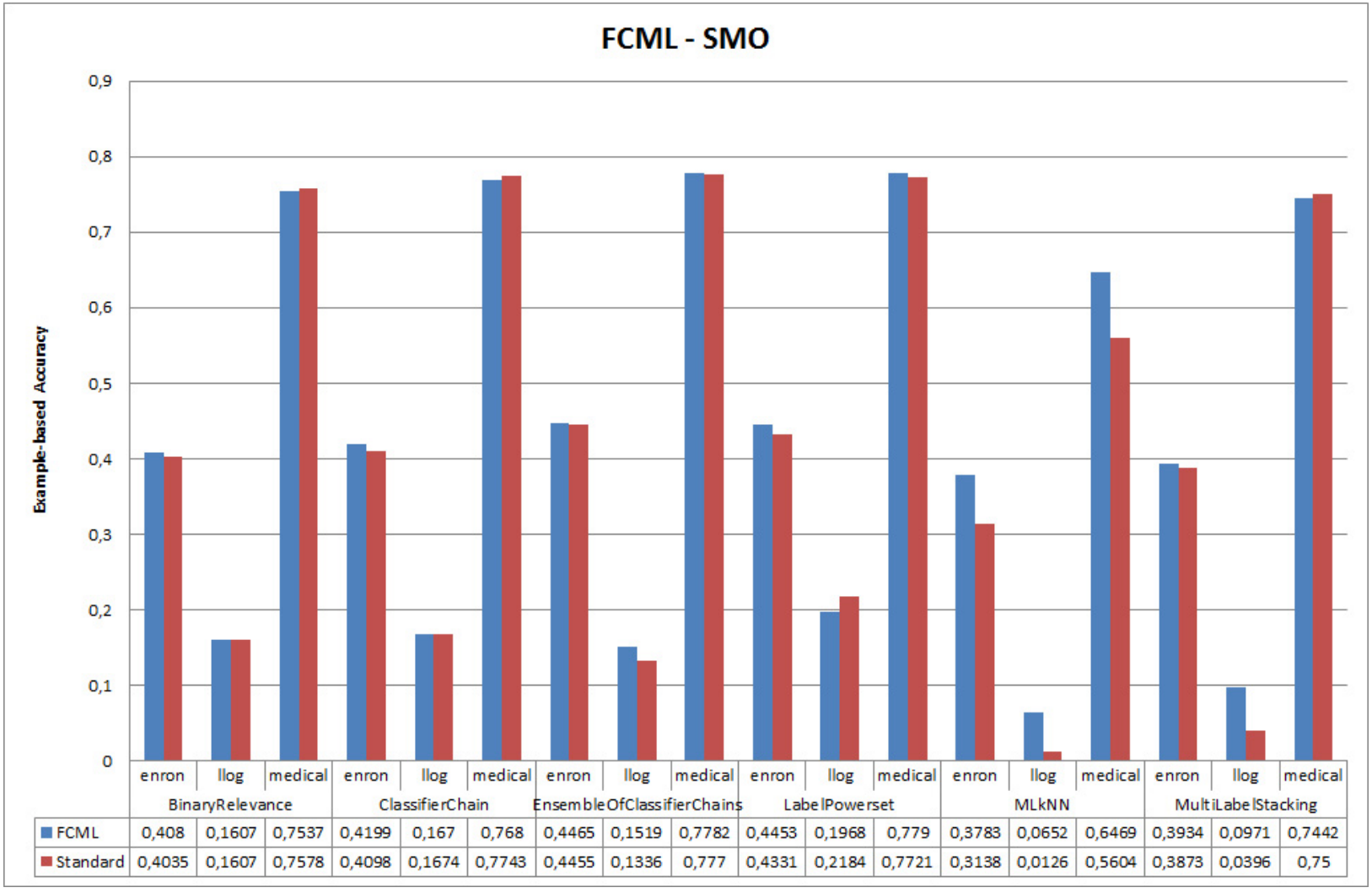
$$\begin{aligned} Y_1 &\leftarrow \{X_1 \ldots X_n \cup Y_1 \ldots Y_n | X_i, Y_i \in \mathbb{R}\} \\ &\vdots \\ Y_n &\leftarrow \{X_1 \ldots X_n \cup Y_1 \ldots Y_n | X_i, Y_i \in \mathbb{R}\} \end{aligned}$$

- Hierachical Clustering using Chebyshev-, Euclidean-, Manhatten-, Mikowski-Distance
- Single, Complete, Average and Mean Clustering
- no. of clusters: $2, 4, 6$


FCML - IBk
Example-based Accuracy
0,8
0,7
0,6
0,5
0,4
0,3
0,2
0,1
0
enron llog medical enron llog medical enron medical enron llog medical enron llog medical enron llog medical
BinaryRelevance ClassifierChain EnsembleOfClassifierChains LabelPowerset MLkNN MultiLabelStacking
FCML 0,3885 0,1212 0,6765 0,389 0,1209 0,6747 0,3826 0,6563 0,3793 0,14 0,6659 0,3783 0,0652 0,6318 0,3553 0,053 0,6136
Standard 0,3188 0,1212 0,6054 0,3388 0,1211 0,6092 0,3271 0,5961 0,3346 0,1217 0,6111 0,3138 0,0126 0,5604 0,3014 0,012 0,5995



	enron	llog	medical	enron	llog	medical	enron	llog	medical	enron	llog	medical	enron	llog	medical	enron	llog	medical
	BinaryRelevance			ClassifierChain			EnsembleOfClassifierChains			LabelPowerset			MLkNN			MultiLabelStacking		
FCML	0,4333	0,1053	0,6964	0,447	0,1165	0,7065	0,4361	0,1028	0,6955	0,399	0,1334	0,6981	0,3783	0,0652	0,6394	0,2311	0,0082	0,2946
Standard	0,4265	0,0355	0,5475	0,4326	0,0329	0,5531	0,4237	0,02	0,4779	0,3748	0,0849	0,6234	0,3138	0,0126	0,5604	0,216	0,002	0,0536



	enron	llog	medical	enron	llog	medical	enron	llog	medical	enron	llog	medical	enron	llog	medical	enron	llog	medical
	BinaryRelevance			ClassifierChain			EnsembleOfClassifierChains			LabelPowerset			MLkNN			MultiLabelStacking		
FCML	0,408	0,1607	0,7537	0,4199	0,167	0,768	0,4465	0,1519	0,7782	0,4453	0,1968	0,779	0,3783	0,0652	0,6469	0,3934	0,0971	0,7442
Standard	0,4035	0,1607	0,7578	0,4098	0,1674	0,7743	0,4455	0,1336	0,777	0,4331	0,2184	0,7721	0,3138	0,0126	0,5604	0,3873	0,0396	0,75

FCML

example cluster characteristics (∅ over folds)

- enron
  ∅ number of clusters : 2
  ∅ number of cluster ($>$ 2 labels) : 2
  ∅ number of labels per cluster: 26.5
- llog
  ∅ number of clusters : 2
  ∅ number of cluster ($>$ 2 labels) : 2
  ∅ number of labels per cluster: 37.5
- medical
  ∅ number of clusters : 4
  ∅ number of cluster ($>$ 2 labels) : 1
  ∅ number of labels per cluster: 11.25

Future Work

- use ranking instead of scores (Spearman Correlation)
- remove outliers to reduce noice and find better groups (covering at least 10% of all labels)