Multilabel Attribute Selection

cluster based group selection

Sebastian Lehnerer
lehnerer@in.tum.de

Technische Universität München

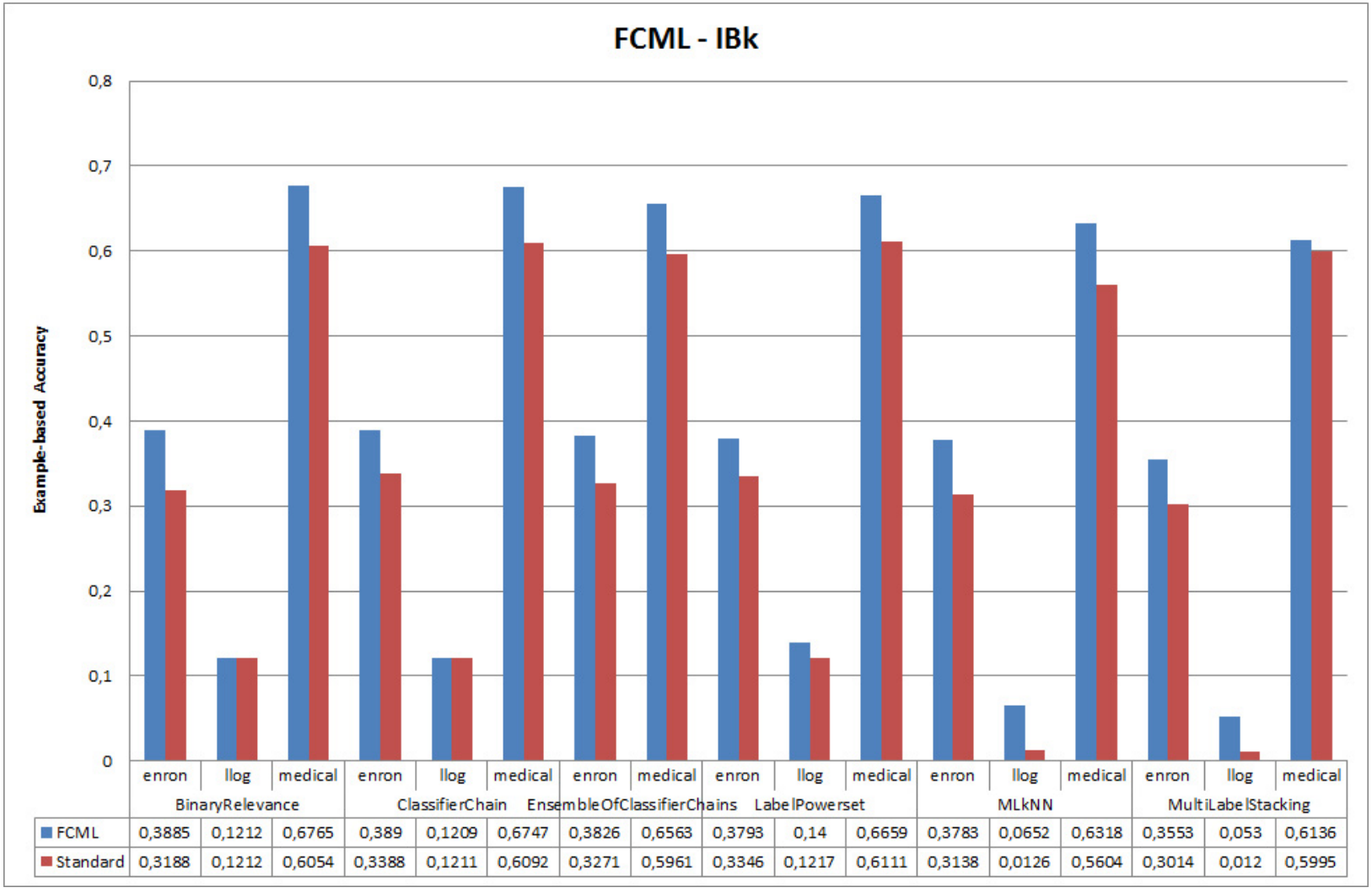
June 14, 2011

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

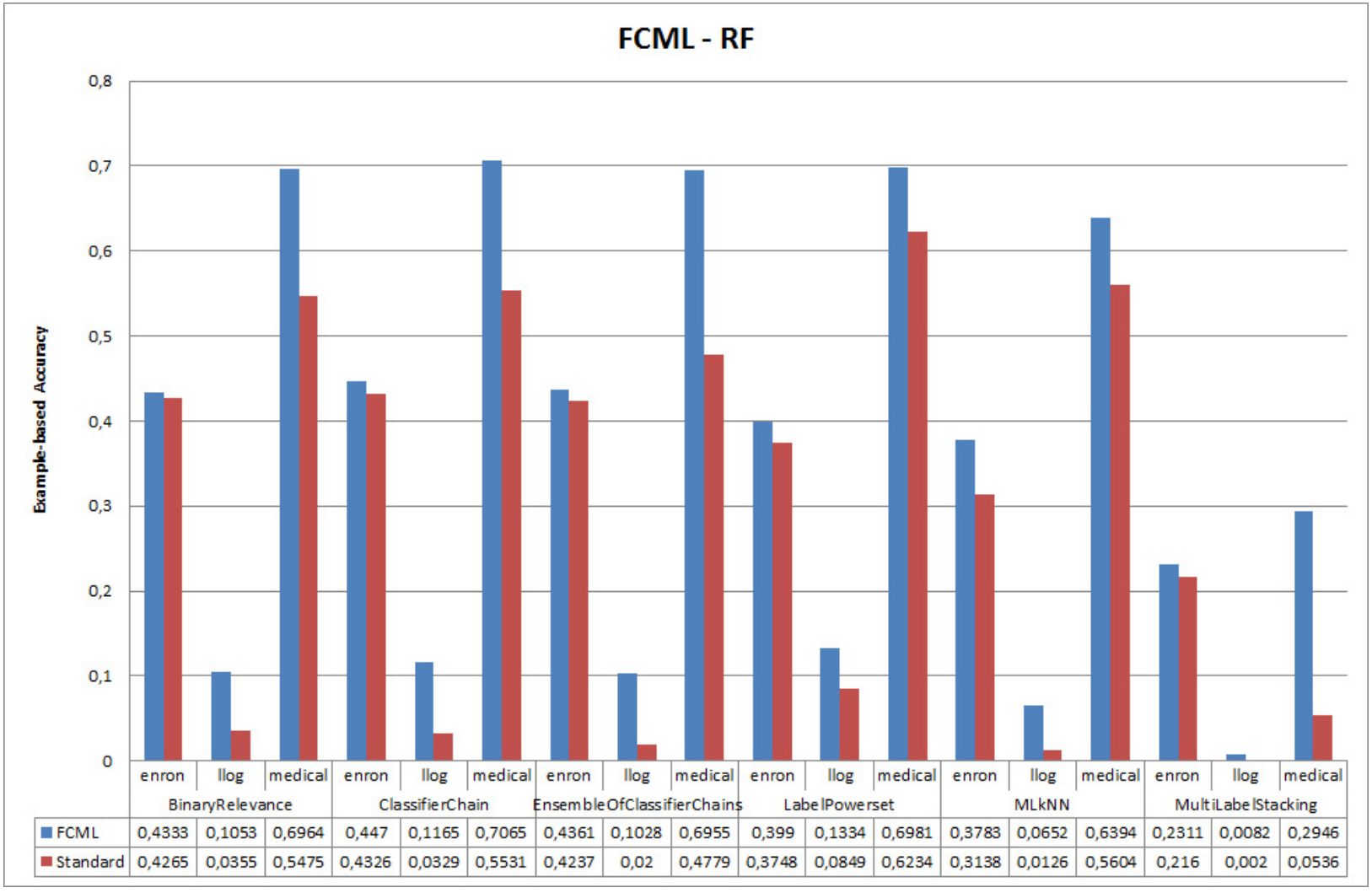
$$\begin{aligned} Y_1 &\leftarrow \{X_1 \ldots X_n \cup Y_1 \ldots Y_n | X_i, Y_i \in \mathbb{R}\} \\ &\vdots \\ Y_n &\leftarrow \{X_1 \ldots X_n \cup Y_1 \ldots Y_n | X_i, Y_i \in \mathbb{R}\} \end{aligned}$$

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

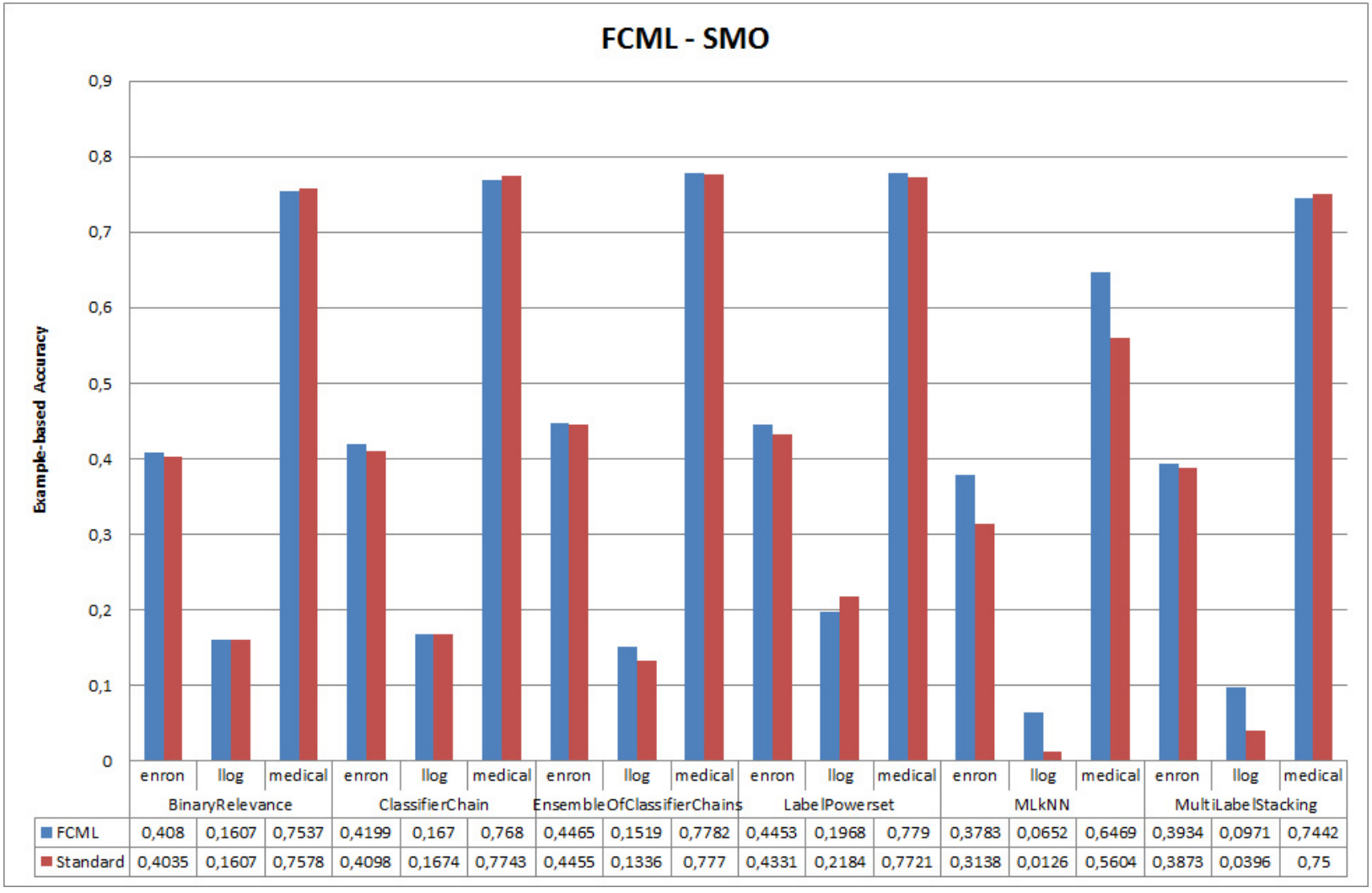
- **function** FINDLABELFEATURESETS(clusters)
  groups ← *Map* < *labelSet*, *featureSet* >
  for $C \in clusters$ **do** ▷ every cluster
    currentLabelSet, *currentFeatureSet*
    for $I \in C$ **do** ▷ every instance
      for $A \in I$ **do** ▷ every attribute
        if $score(A) > threshold$ **then**
          if $A \in labelSet$ **then**
            $currentLabelSet \cap \{A\}$
          else
            $currentFeatureSet \cap \{A\}$
    $put(currentLabelSet, currentFeatureSet) \rightarrow groups$
  return *groups*



	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

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