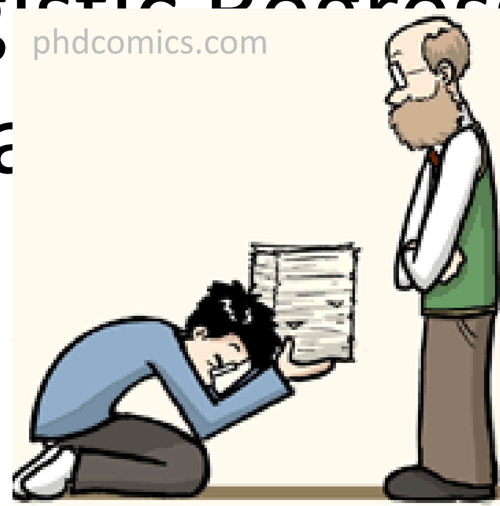
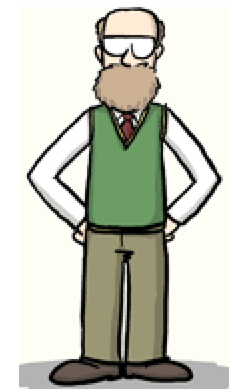
Combining Instance-Based Learning and Logistic Regression for Multilabel Classification

Weiwei Cheng & Eyke Hüllermeier

Knowledge Engineering & Bioinformatics Lab
Department of Mathematics and Computer Science
University of Marburg, Germany

Combining Instance-Based Learning and Logistic Regression for Multilabel Classification





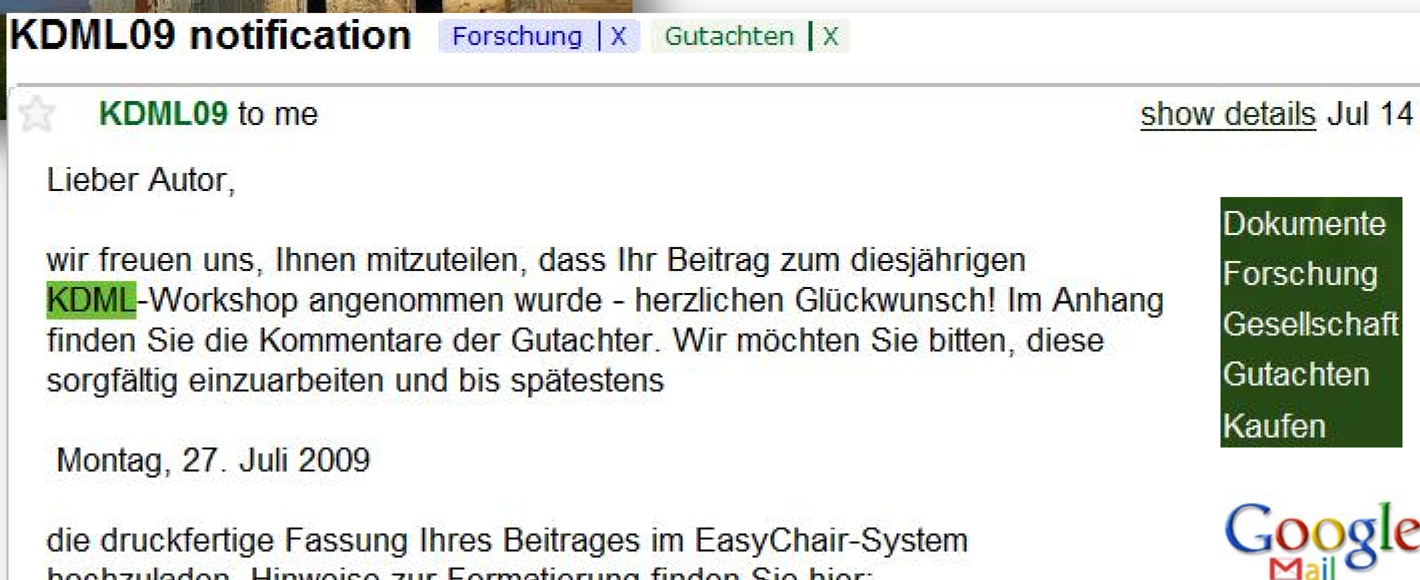
Weiwei Cheng & Eyke Hüllermeier

Knowledge Engineering & Bioinformatics Lab

Department of Mathematics and Computer Science

University of Marburg, Germany

Multilabel Classification


KDML09 notification Forschung | X Gutachten | X

KDML09 to me show details Jul 14

Lieber Autor,

wir freuen uns, Ihnen mitzuteilen, dass Ihr Beitrag zum diesjährigen KDML-Workshop angenommen wurde - herzlichen Glückwunsch! Im Anhang finden Sie die Kommentare der Gutachter. Wir möchten Sie bitten, diese sorgfältig einzuarbeiten und bis spätestens

Montag, 27. Juli 2009

die druckfertige Fassung Ihres Beitrages im EasyChair-System hochzuladen. Hinweise zur Formatierung finden Sie hier:

Dokumente
Forschung
Gesellschaft
Gutachten
Kaufen

Google Mail

What is Multilabel Classification?

- Conventional classification
  - Instances are associated with a single label λ from as set $\mathcal{L}$ of finite labels
    - if $|\mathcal{L}| = 2,$ binary classification;
    - if $|\mathcal{L}| > 2,$ multi-class classification.
- Multilabel classification
  - Instances are associated with a set of labels $L \subseteq \mathcal{L}$.

Existing Methods

- Quite a number of methods for multilabel classification have been proposed, most of them being model-based approaches (training a global model for prediction).

- Our work is especially motivated by MLKNN:

  Zhang & Zhou. ML-kNN: A lazy learning approach to multi-label learning. *Pattern Recognition*, 2007, 40(7): 2038-2048.

  In a number of practical problems, MLKNN shows very strong performance and even outperforms RankSVM and AdaBoost.MH.

- Still, many methods ignore the correlation between labels.
  A paper with label *CS* is more likely having label *Math*, than *Law*.

Our Contributions

- A new multilabel learning method,
- which is based on a formalization of instance-based classification as logistic regression (combination of model-based and instance-based learning),
- takes the correlation between labels into account and represents it in an easily interpretable way.

IBL as Logistic Regression

Key idea:

Consider the labels of neighbors as "*extra features*" of an instance

	age	*weight*	*height*	*sex*	*w.child*	
nearest neighbors	26	62	1.83	male	no	1
	16	45	1.65	female	no	0
	28	85	1.90	male	yes	1
				...	...	
test instance	27	50	1.63	male	yes	?

Does he like basketball?

IBL as Logistic Regression

Extended representation:

age	*weight*	*height*	*sex*	*w.child*	# 	
26	62	1.83	male	no	1/3	1
16	45	1.65	female	no	0	0
28	85	1.90	male	yes	2/3	1
			...	...		
27	50	1.63	male	yes	2/3	?

Extra feature: 2 among 3 neighbors like basketball

IBL as Logistic Regression (binary case)

Consider query instance $\mathbf{x}_0$, distance $\delta_i \stackrel{\text{df}}{=} \Delta(\mathbf{x}_0, \mathbf{x}_i)$,
posterior probability $\pi_0 \stackrel{\text{df}}{=} \mathbf{P}(y_0 = +1 \,|\, y_i)$:

$$\frac{\pi_0}{1-\pi_0} = \frac{\mathbf{P}(y_i \,|\, y_0 = +1)}{\mathbf{P}(y_i \,|\, y_0 = -1)} \cdot \frac{p_0}{1-p_0} = \rho \cdot \frac{p_0}{1-p_0}$$

$$\log\left(\frac{\pi_0}{1-\pi_0}\right) = \log(\rho) + \underbrace{\log(p_0) - \log(1-p_0)}_{\omega_0}$$

For example, we can define $\rho = \rho(\delta) \stackrel{\text{df}}{=} \exp\left(y_i \cdot \frac{\alpha}{\delta}\right)$.

$\delta \to +\infty$ $\quad$ $\rho \to 1$

$\delta \to 0$ $\quad$ $y_i = +1 \to \rho \uparrow$ $\quad$ $y_i = -1 \to \rho \downarrow$

Now consider the whole neighborhood of $\mathbf{x}_0$:

$$\log\left(\frac{\pi_0}{1-\pi_0}\right) = \omega_0 + \alpha \sum_{\mathbf{x}_i \in \mathcal{N}(\mathbf{x}_0)} \frac{y_i}{\delta_i} = \omega_0 + \alpha \cdot \omega_+(\mathbf{x}_0)$$

bias term (prior probability) $\quad$ evidence for positive class

IBL as Logistic Regression (binary case)

$$\log\left(\frac{\pi_0}{1-\pi_0}\right) = \omega_0 + \alpha \cdot \omega_+(\mathbf{x}_0) = \omega_0 + \alpha \sum_{\mathbf{x}_i \in \mathcal{N}(\mathbf{x}_0)} \frac{y_i}{\delta_i}$$

From *distance* to *similarity*

$$= \omega_0 + \alpha \sum_{\mathbf{x}_i \in \mathcal{N}(\mathbf{x}_0)} \kappa(\mathbf{x}_0, \mathbf{x}_i) \cdot y_i$$

The standard *KNN* classifier is recovered as a special case:

- Set $\omega_0 = 0$, and
- $\kappa(\mathbf{x}_0, \mathbf{x}_i) = \begin{cases} 1 & \text{if } \mathbf{x}_i \in \mathcal{N}_k(\mathbf{x}_0) \\ 0 & \text{otherwise} \end{cases}$.

IBL as Logistic Regression

Same idea for multilabel case:

Consider the labels of neighbors as "*extra features*" of an instance

	age	weight	height	sex	w.child			
NN	26	62	1.83	male	no	1	0	1
	16	45	1.65	female	no	0	1	0
	28	85	1.90	male	yes	1	0	1
				...	...			
test inst.	27	50	1.63	male	yes	?	?	?

Does he like basketball?

IBL as Logistic Regression

Extended representation:

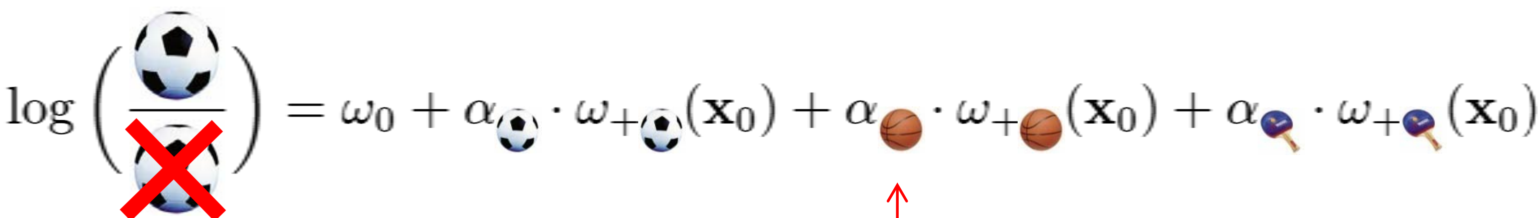
age	*weight*	*height*	*sex*	*w.child*	# 	# 	# 			
26	62	1.83	male	no	1/3	0	1	1	0	1
16	45	1.65	female	no	0	1	1/3	0	1	0
28	85	1.90	male	yes	2/3	0	1	1	0	0
			...	...						
27	50	1.63	male	yes	2/3	*1/3*	*1/3*	?	?	?

Extra feature: 1 among 3 neighbors like table tennis

IBL as Logistic Regression (multilabel case)

We solve one logistic regression problem for each label!

Example:



To what extent does the presence of label basektball in the neighborhood increase the probability that football is relevant for the query?

IBL as Logistic Regression (multilabel case)

Multilabel prediction rule

$$L \;=\; \left\{ \lambda \in \mathcal{L} \;\middle|\; \log\left(\frac{\pi_0(\lambda)}{1-\pi_0(\lambda)}\right) > 0 \right\}$$

Ranking rule

$$\lambda_i \succ \lambda_j \iff \log\left(\frac{\pi_0(\lambda_i)}{1-\pi_0(\lambda_i)}\right) > \log\left(\frac{\pi_0(\lambda_j)}{1-\pi_0(\lambda_j)}\right)$$

Experiments

dataset	domain	#inst.	#attr.	#labels	card.
emotions	music	593	72	6	1,87
image	vision	2000	135	5	1,24
genbase	biology	662	1186(*n*)	27	1,25
mediamill	multimedia	5000	120	101	4,27
reuters	text	7119	243	7	1,24
scene	vision	2407	294	6	1,07
yeast	biology	2417	103	14	4,24

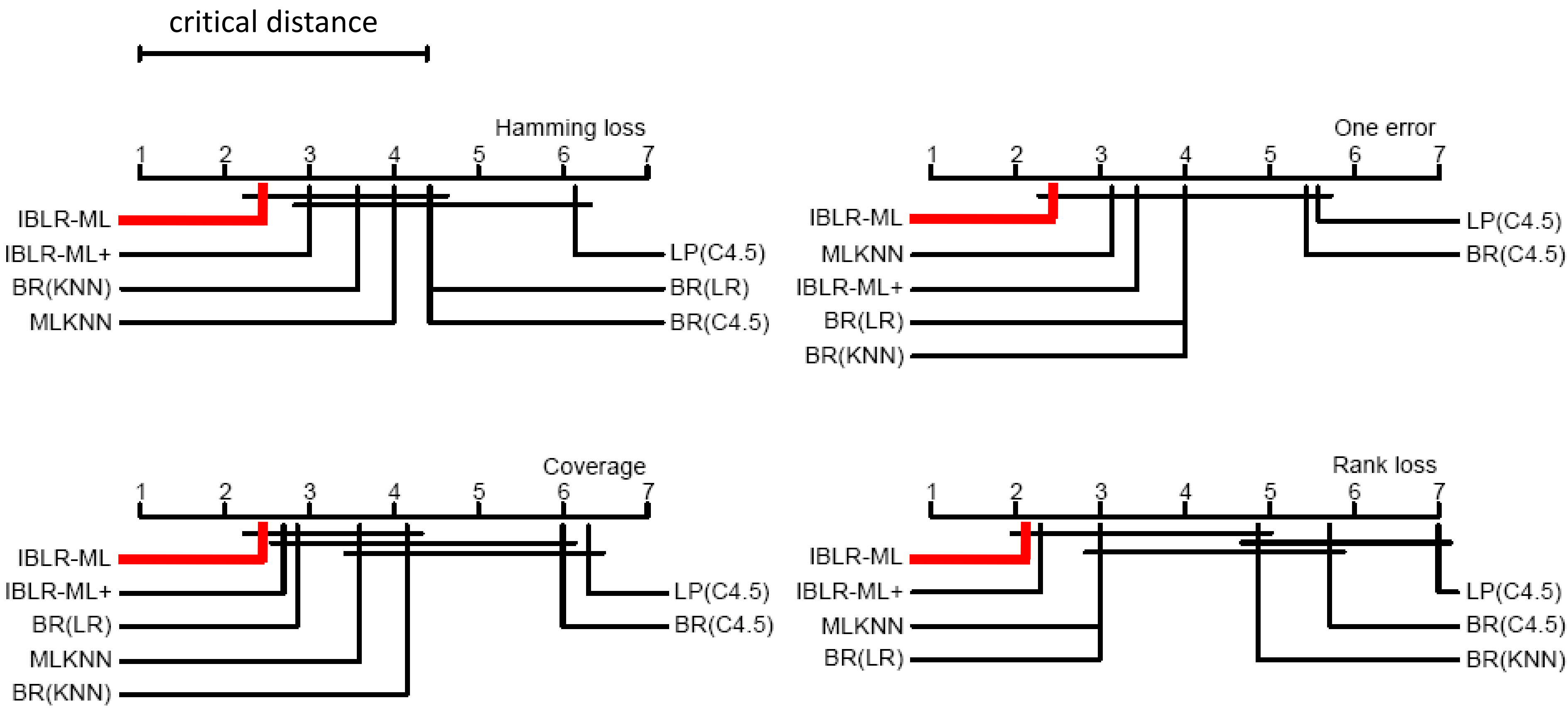
Tested methods:

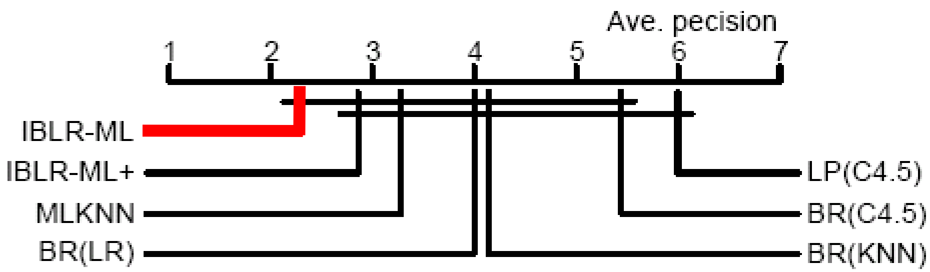
- MLKNN
- Binary relevance learning (BR) with logistic regression, C4.5 and KNN
- Label powerset (LP) with C4.5
- Our method: IBLR-ML

Evaluation metrics

- Hamming loss $= \frac{1}{|\mathcal{L}|}\left|h(\mathbf{x}) \,\Delta\, L_{\mathbf{x}}\right|$
- one error $= \begin{cases} 1 & \text{if } \arg\max_{\lambda\in\mathcal{L}} f(\mathbf{x},\lambda) \notin L_{\mathbf{x}} \\ 0 & \text{otherwise} \end{cases}$
- coverage $= \max_{\lambda\in L_{\mathbf{x}}} \operatorname{rank}_f(\mathbf{x},\lambda) - 1$
- rank loss $= \dfrac{\#\{(\lambda,\lambda') \mid f(\mathbf{x},\lambda) \le f(\mathbf{x},\lambda'), (\lambda,\lambda') \in L_{\mathbf{x}} \times \overline{L_{\mathbf{x}}}\}}{|L_{\mathbf{x}}||\overline{L_{\mathbf{x}}}|}$
- average precision

$$= \frac{1}{|L_{\mathbf{x}}|} \sum_{\lambda\in L_{\mathbf{x}}} \frac{|\{\lambda' \mid \operatorname{rank}_f(\mathbf{x},\lambda') \le \operatorname{rank}_f(\mathbf{x},\lambda), \lambda' \in L_{\mathbf{x}}\}|}{\operatorname{rank}_f(\mathbf{x},\lambda)}$$





Nemenyi test with *p*=0.05

Contributions of Our Work

- Novel approach to IBL, applicable to classification in general and multilabel classification in particular.
- Key idea: Consider label information in the neighborhood of a query as “extra features” of that query.
- Balance between global and local inference automatically optimized via fitting a logistic regression function.
- Interdependencies between labels estimated by regression coefficients.
- Extension: Logistic regression combining “normal features” with “extra features”.

IBLR-ML is available in the MULAN Java library, maintained by the
Machine Learning & Knowledge Discovery Group,
Aristotle University of Thessaloniki.

Check www.chengweiwei.com for more info.

Thanks!