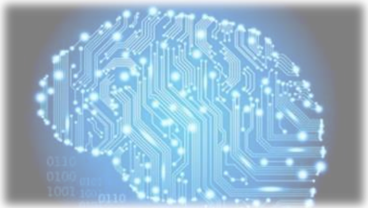
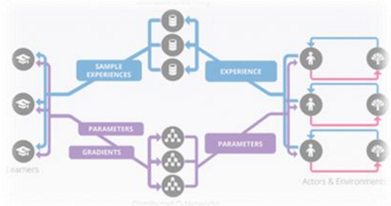
 

Review of Problem Transformation Method of Multi-Label Classification

Hae-Cheon Kim

2018. 03. 28.

Notation & Definition

- A vector $\vec{X} = [x_1, x_2, x_3, \dots, x_d]\ (\in \mathbb{R}^d)$
- A set $Y = \{l_1, \dots, l_q\}, \quad |Y| = q$
- A set $Y' \subseteq Y$
- A vector $\overrightarrow{Y'} = \{0,1\}^{Y'} \in \{0,1\}^q = 2^q$

- When given vector of data, $\vec{X}$, find a most relevant vector of set Y'.

$$\mathrm{MLC}(\vec{X}) = \overrightarrow{Y'}$$

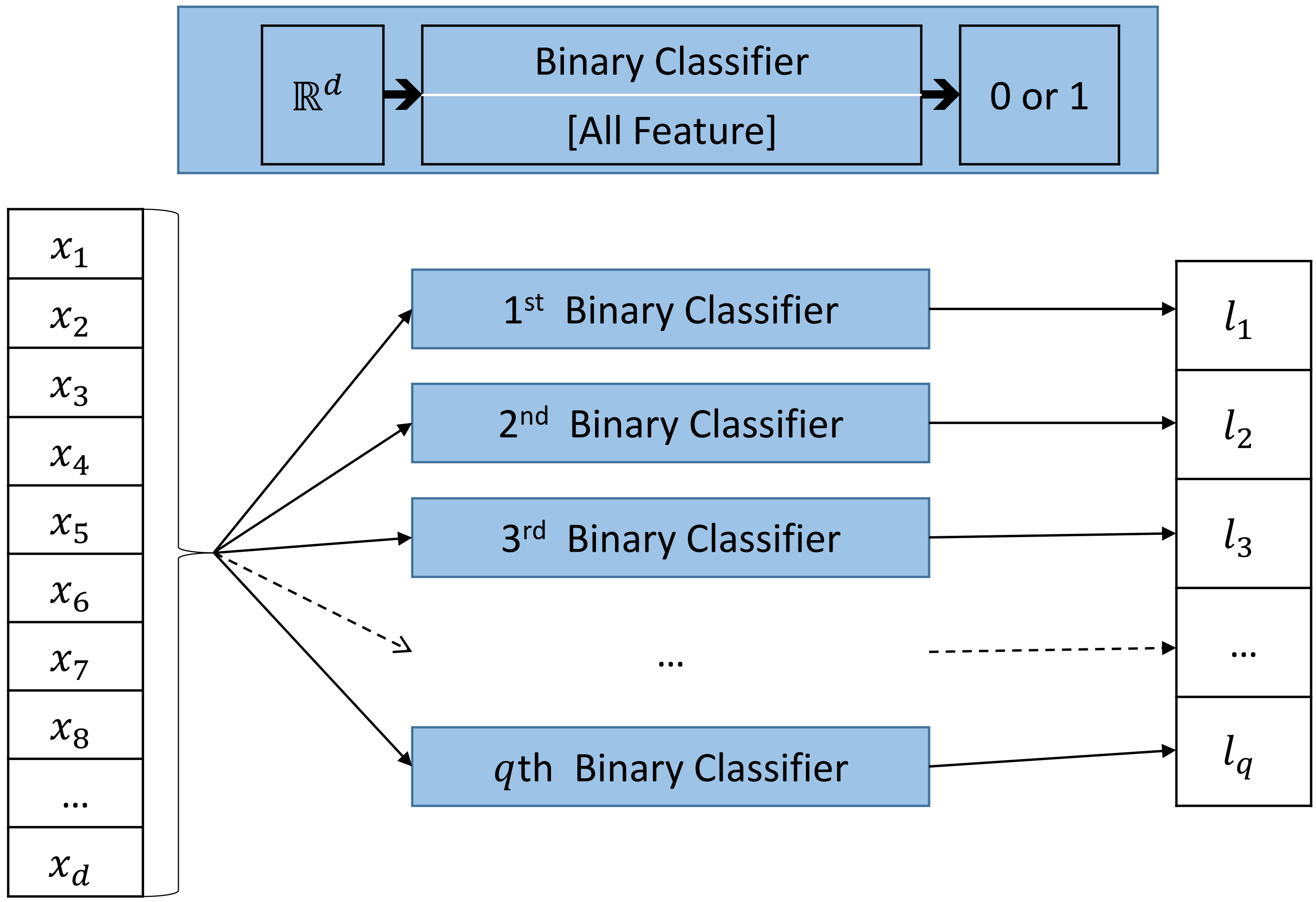
Keyword of Multi-Label Classifier

Algorithm Adaptation	Problem Transformation
Multi-Label Decision Tree	Binary Relevance
ML-kNN	Label Power Set
Ranking Support Vector Machine	Classifier Chain
Backpropagation of Multi Label Learning	Ensemble

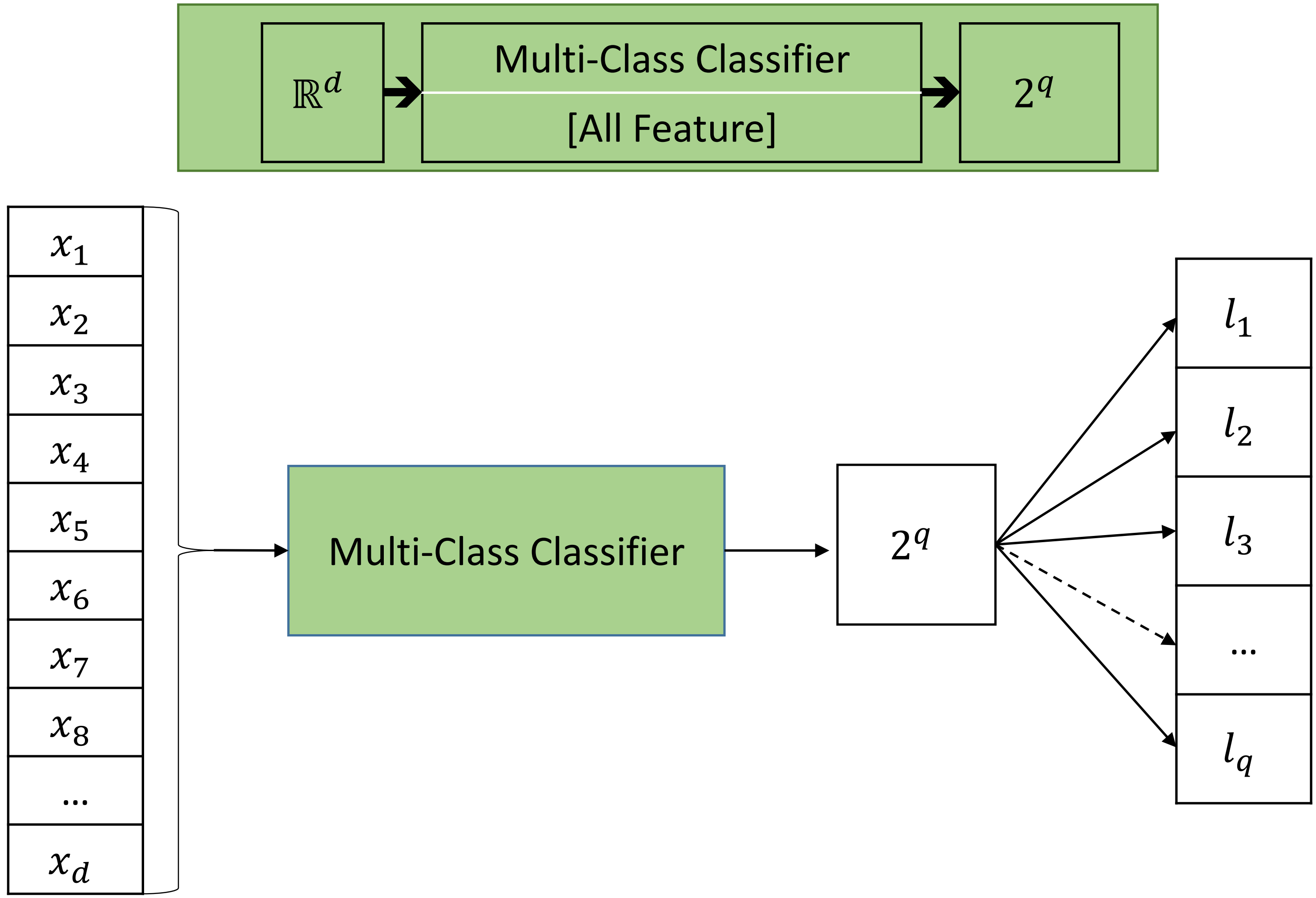
Algorithm Adaptation

- Multi-Label Decision tree (C4.5)
- ML-kNN (Multi-Label Classifier using k-Nearest Neighbor)
- Ranking Support Vector Machine (RSVM)
- Back propagation of multi label learning (BPMLL)
- etc..

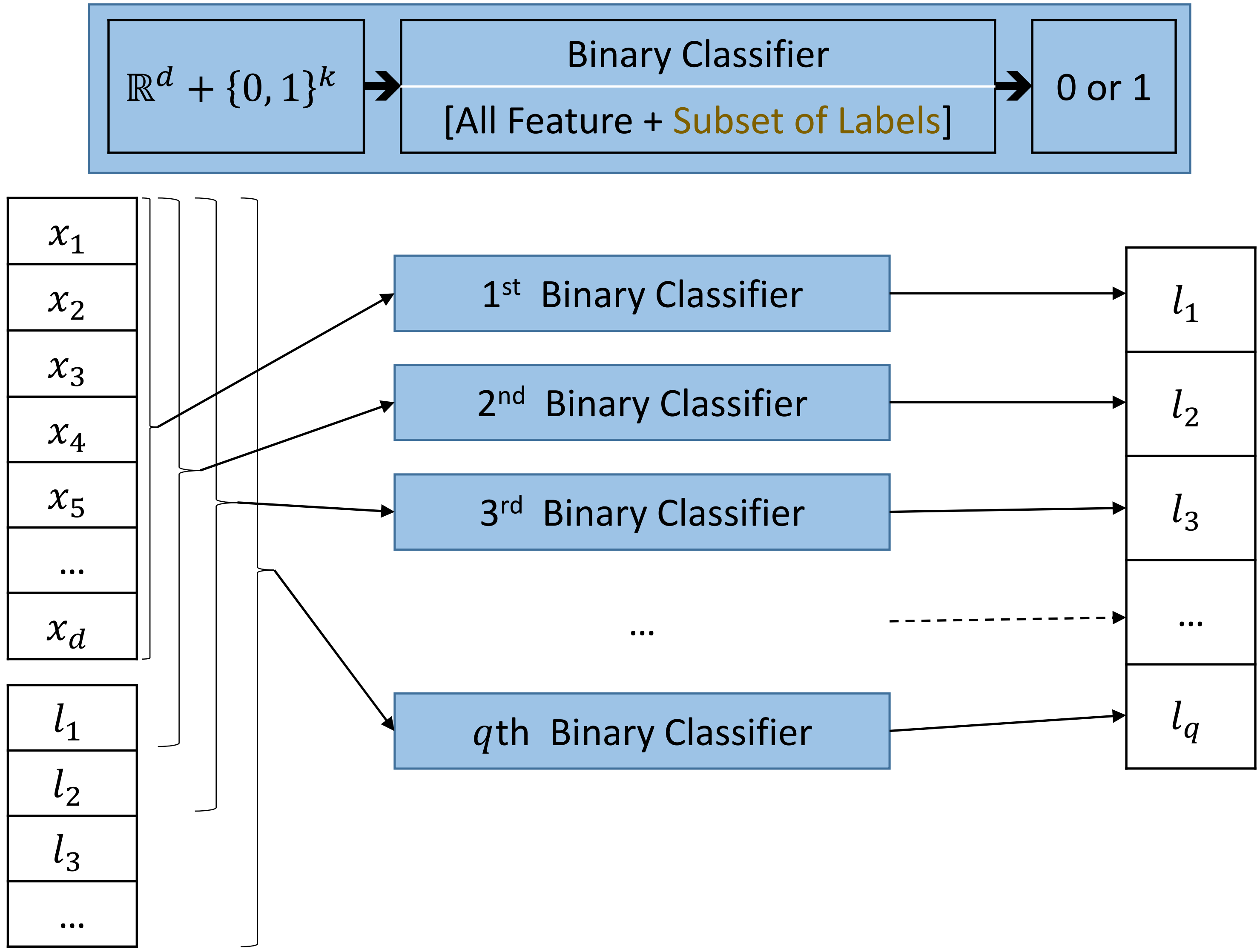
Binary Relevance



Label Power Set



Classifier Chain



Ensemble Leaning [1]

- Ensemble learning combine individual learners
  from heterogeneous or homogeneous modeling
  in order to obtain a combined learner
  that improves
  the generalization ability
  and reduces the overfitting risk of each one of them.

- To make generalization ability and to avoid the overfitting risk,
  - Let D_i be a subset of data set
  - Let L_j be a subset of label set
  - Train the classifier h_k using D_i and L_j.

Example of Ensemble Leaning

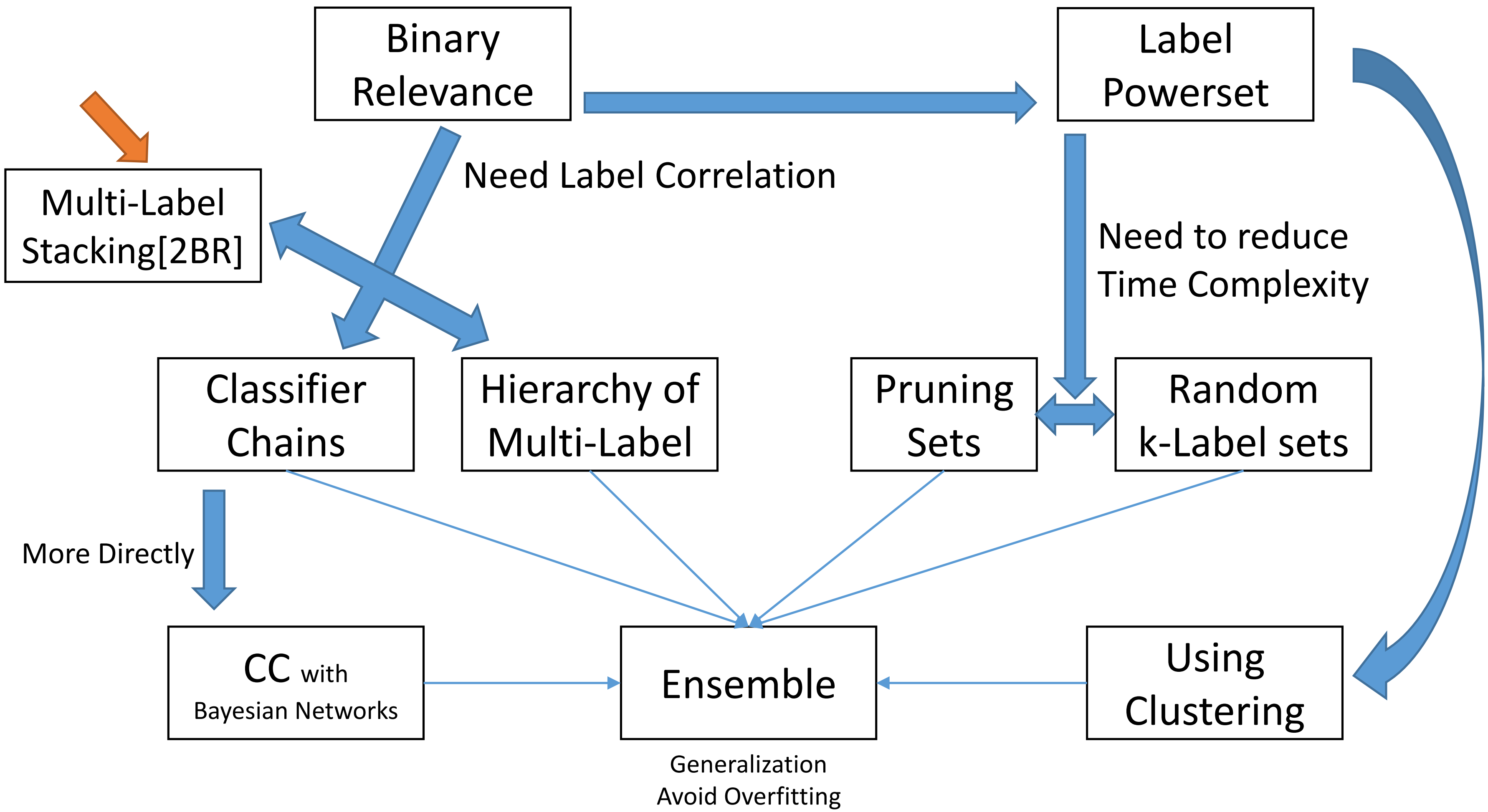
- $X = \mathbb{R}^{m \times d}$, $Y = \{0,1\}^{m \times q}$
  - $D_i = \{x_1, x_2, \dots, x_m\}$, $x_3 \notin D_i$
  - $L_i = \{l_2, \dots, l_q\}$
- Training one classifier $h_i(x)$ using D_i and L_i.

	f_1	f_2	f_3	…	f_d
x_1	0.2	Yes	14	…	0.2
x_2	0.1	Yes	20	…	0.1
x_3	0.5	No	32	…	0.2
…	…	…	…	…	…
x_m	0.1	Yes	31	…	0.2

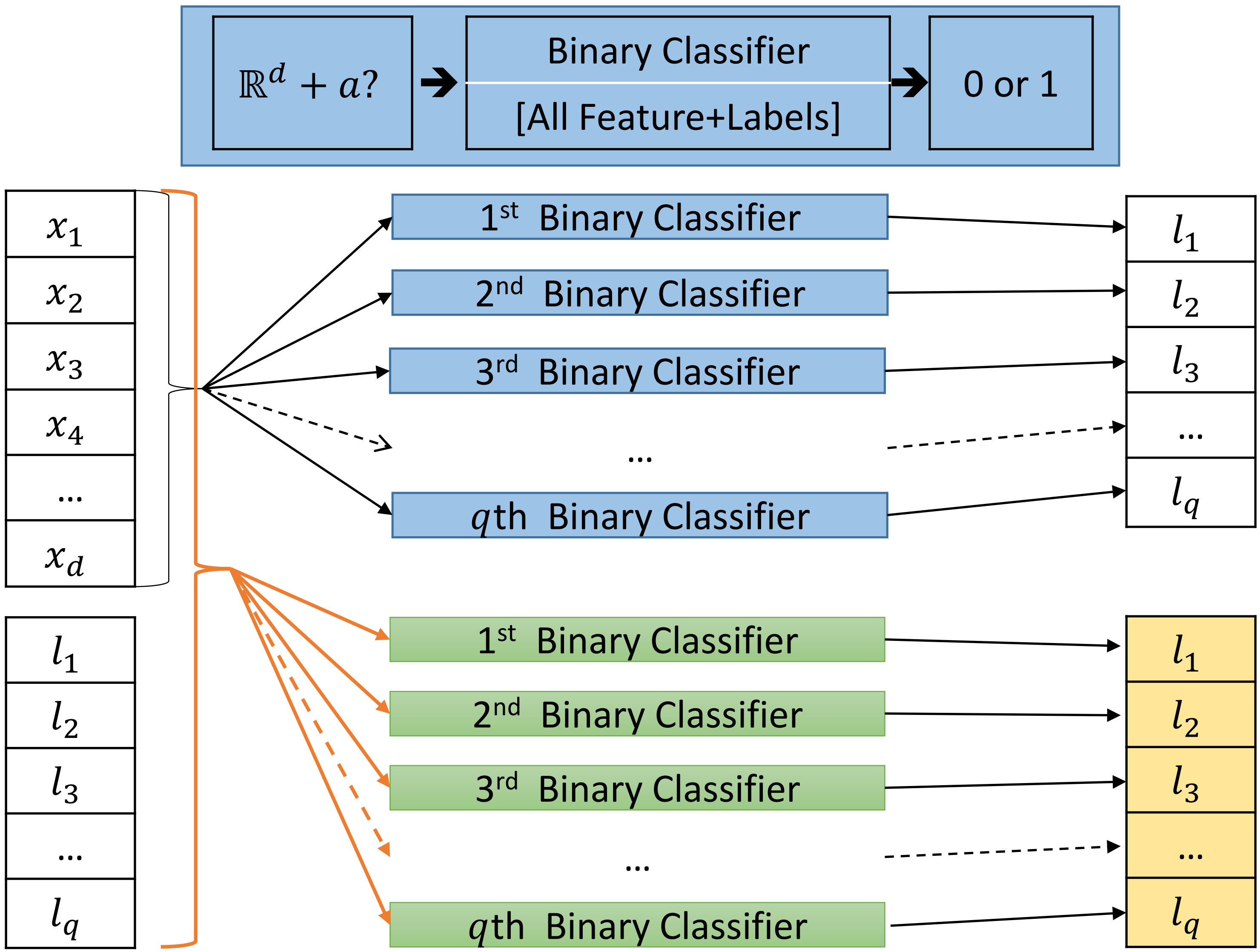
$h_i(x_{new})$

	l_1	l_2	…	l_q
y_1	0	0	…	1
y_2	1	0	…	0
y_3	1	0	…	1
…	…	…	…	…
y_m	1	1	…	1

Map of Problem Transformation



Multi-Label Stacking [2BR]



Hierarchy of Multi-Label Classifiers [HOMER] [6]

- Hierarchy Of Multilabel classifiERs

- Label set ➔ Tree-shaped Hierarchy structure
- One brunch point ≈ One multi-label classifiers
- How to make tree? ➔ Clustering(k-means)

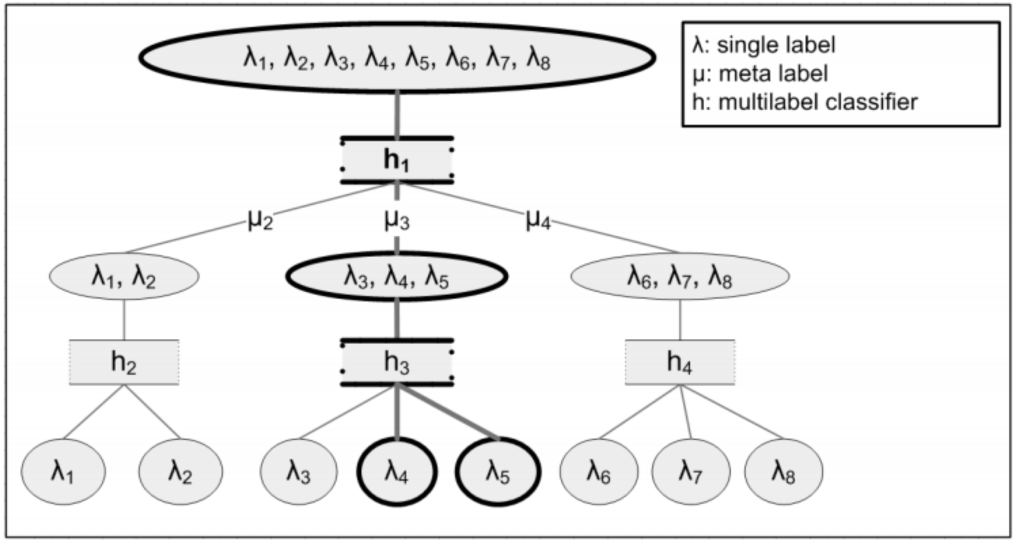
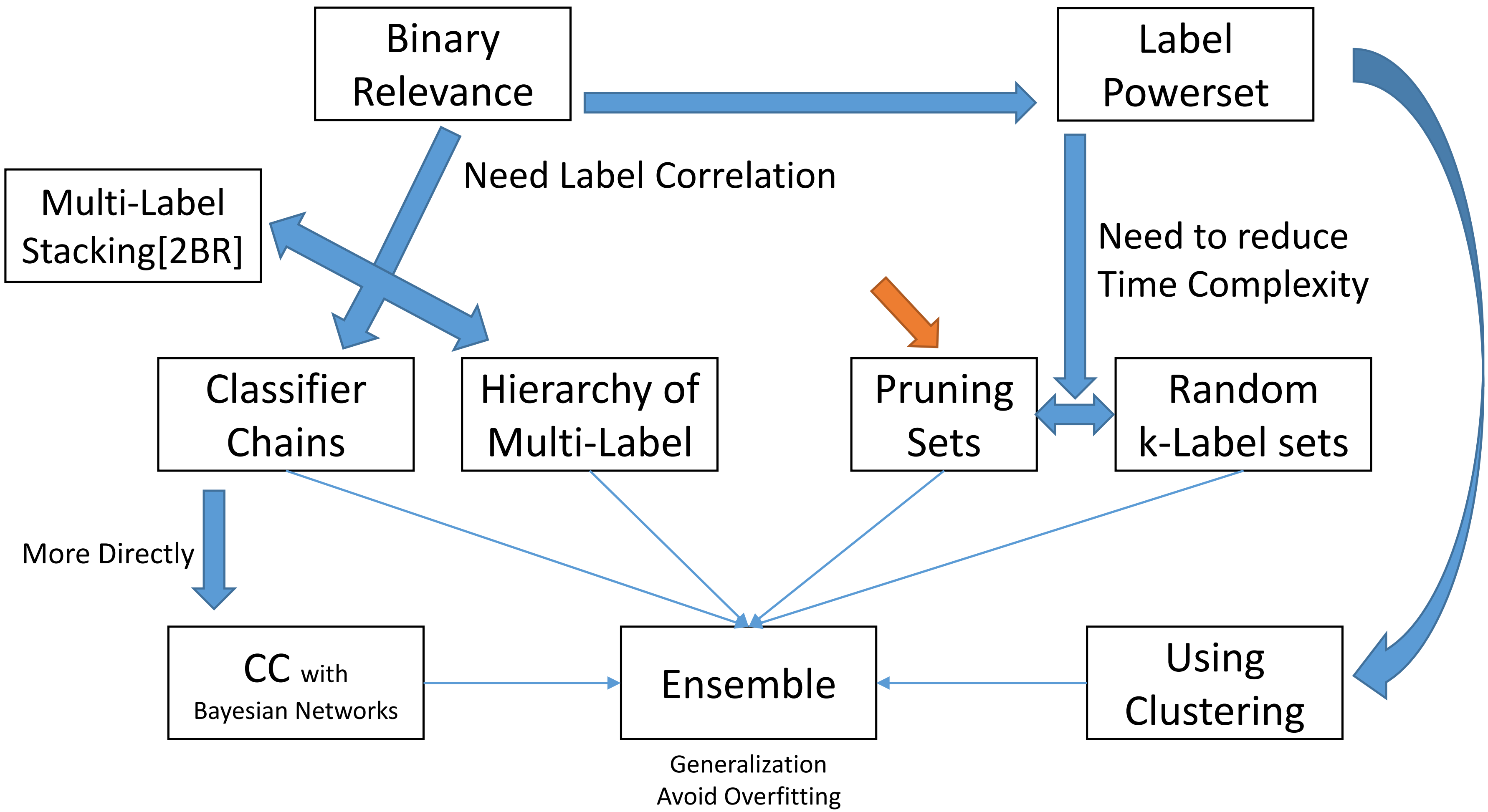


Fig. 1. Sample hierarchy for a multilabel classification task with 8 labels

Map of Problem Transformation



Pruning set

- If the label set L_i is infrequently occurring,

  change L_i to $L_i'(\subseteq L_i)$ [Pruning]

- Reduce cardinality of set of Label-set (4 → 3)

Y_1	Y_2	Y_3
0	0	1
0	0	1
0	0	0
0	1	0
0	0	0
0	1	0
0	1	0
0	0	0
0	0	1
0	1	1

$D_{train\text{-}LP}$	Y_{LP}
$n=1$	1
$n=2$	1
$n=3$	0
$n=4$	2
$n=5$	0
$n=6$	2
$n=7$	2
$n=8$	0
$n=9$	1
$n=10$	3

Y_1	Y_2	Y_3
0	0	1
0	0	1
0	0	0
0	1	0
0	0	0
0	1	0
0	1	0
0	0	0
0	0	1
0	0	1
0	1	0

$D_{train\text{-}PPT}$	Y_{PPT}
$n=1$	1
$n=2$	1
$n=3$	0
$n=4$	2
$n=5$	0
$n=6$	2
$n=7$	2
$n=8$	0
$n=9$	1
$n=10$	1
$n=11$	2

Random k-Labelsets for Multi-Label Classification [4]

- The main idea in this work is to randomly break
  a large set of labels into a number of small-sized labelsets,
- And
  for each of them train a multilabel classifier using the LP method.

- RAkEL (RAndom k labELsets) k is cardinality of subset.

- In using LP, we must think about computational efficiency.
- This proposed is very efficiency a point of time complexity.

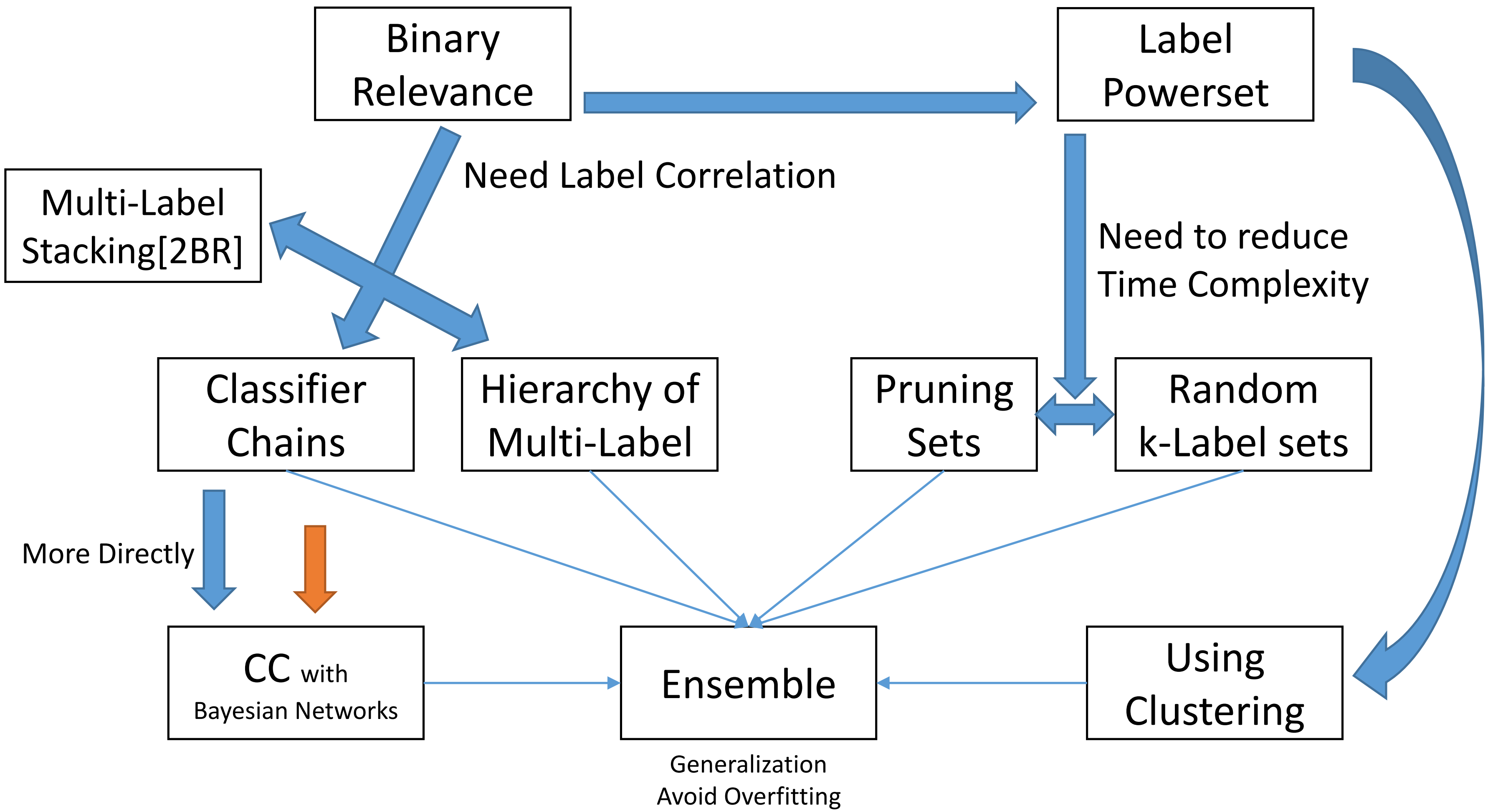
Random k–Labelsets for Multi–Label Classification [4]

- Overlapping (⇔Disjoint) [randomly break]
- Time Complexity: 2^q ➔ $2^k \times \left\lceil \frac{2q}{k} \right\rceil$, ($2^6$ ➔ $2^3 \times 6$)

TABLE 1
An Example of the Classification Process of $\mathrm{RA}k\mathrm{EL}_o$ Run with $k = 3$ and $m = 7$ on a Multilabel Training Set with Six Labels

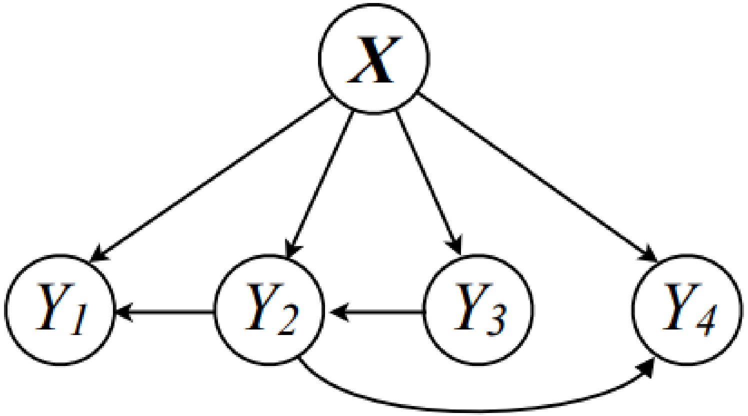
model	labelset	predictions λ_1	λ_2	λ_3	λ_4	λ_5	λ_6
h_1	$\{\lambda_1, \lambda_2, \lambda_6\}$	1	0	-	-	-	1
h_2	$\{\lambda_2, \lambda_3, \lambda_4\}$	-	1	1	0	-	-
h_3	$\{\lambda_3, \lambda_5, \lambda_6\}$	-	-	0	-	0	1
h_4	$\{\lambda_2, \lambda_4, \lambda_5\}$	-	0	-	0	0	-
h_5	$\{\lambda_1, \lambda_4, \lambda_5\}$	1	-	-	0	1	-
h_6	$\{\lambda_1, \lambda_2, \lambda_3\}$	1	0	1	-	-	-
h_7	$\{\lambda_1, \lambda_4, \lambda_6\}$	0	-	-	1	-	0
	average votes	3/4	1/4	2/3	1/4	1/3	2/3
	final prediction	1	0	1	0	0	1

Map of Problem Transformation



Classifier Chains with Bayesian Network

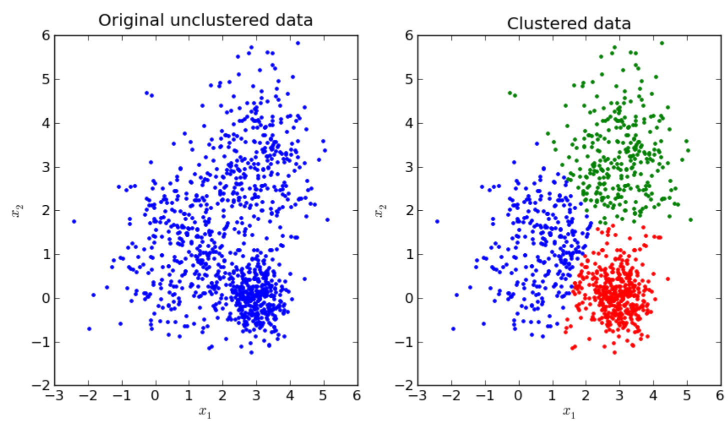
- Bayesian Network : Directed acyclic graph (DAG)
- It marks the associated label in the DAG.
- They are researching how to find DAGs that extract optimal efficiency.



$$P(L|X) \Rightarrow \prod_{i=1}^{q} P(L_i|X \cap (\text{parents}(L_i) = \{0\ or\ 1\}))$$

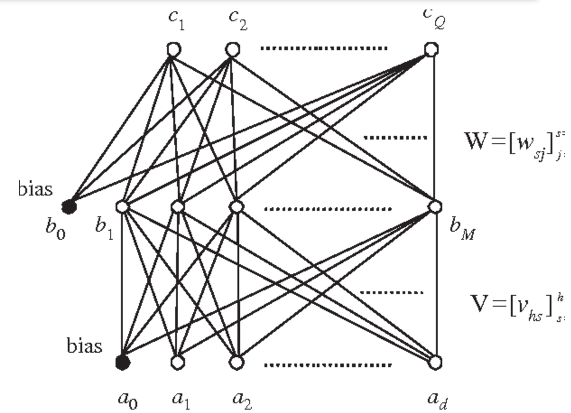
Using Clustering

- Clustering: grouping a set of objects that in same group.



- Clustering-Based for Multi-Label Classification (CBMLC)
  - Divide feature set using c clusters and training similar group of dataset.
  - Distance: k-means
- Predictive Clustering Trees (PCTs)
  - Make decision trees that cluster variation is minimized.
  - Distance: Gini Indices

Direction of Recent Research [2, 5, 7, 8]



- Neural Network
- But almost one hidden layer.

- Back–Propagation Multi–Label Learning (BPMLL)
  - Gradient descent + Error backpropagation strategy
  - Recommend the number of hidden neuron is 20% of input dimensions.

- Semi–supervised Multilabel Learning With Joint Dimensionality Reduction
  - Semi–supervised Learning
    - Some data that don't have label–set exist in training dataset

- Based on kernel extreme learning machine(ML–KELM)
  - The weight of hidden layer is the variable the kernel of input
  - Similar to support vector machine

Direction of Recent Research [8]

- Dimensionality Reduction

$$\min_{W} l(Y, XW) + \lambda R(W)$$

- Want to find W that $Y \approx XW$

- Linear parametrization ➔ $W = UV^T$
  - Using Maximum-Margin Matrix Factorization(MMMF).
- $U \in \mathbb{R}^{d \times k}$, $V \in \mathbb{R}^{L \times k}$, $k \ll d$
- U → Embedding of the dataset X into a k-dimensional latent space.
- V → Linear classifier on this space.

- MLC-HMF(Multi-Label Classification, Hierarchical Matrix Factorization)

  ➔ Divide dataset to hierarchical trees

  (using clustering)

Reference

1. Moyano, Jose M., et al. "Review of ensembles of multi-label classifiers: Models, experimental study and prospects." *Information Fusion* 44 (2018): 33-45.
2. Zhang, Min-Ling, and Zhi-Hua Zhou. "Multilabel neural networks with applications to functional genomics and text categorization." *IEEE transactions on Knowledge and Data Engineering* 18.10 (2006): 1338-1351.
3. Zhang, Min-Ling, and Zhi-Hua Zhou. "ML-KNN: A lazy learning approach to multi-label learning." *Pattern recognition* 40.7 (2007): 2038-2048.
4. Tsoumakas, Grigorios, Ioannis Katakis, and Ioannis Vlahavas. "Random k-labelsets for multilabel classification." *IEEE Transactions on Knowledge and Data Engineering* 23.7 (2011): 1079-1089.
5. Elisseeff, André, and Jason Weston. "A kernel method for multi-labelled classification." *Advances in neural information processing systems*. 2002.

Reference

6. Tsoumakas, Grigorios, Ioannis Katakis, and Ioannis Vlahavas. "Effective and efficient multilabel classification in domains with large number of labels." *Proc. ECML/PKDD 2008 Workshop on Mining Multidimensional Data (MMD' 08)*. Vol. 21. sn, 2008.
7. Luo, Fangfang, et al. "A multi-label classification algorithm based on kernel extreme learning machine." *Neurocomputing* 260 (2017): 313-320.
8. Kumar, Vikas, et al. "Multi-label classification using hierarchical embedding." *Expert Systems with Applications* 91 (2018): 263-269.
9. Yu, Tingzhao, and Wensheng Zhang. "Semisupervised Multilabel Learning With Joint Dimensionality Reduction." *IEEE Signal Processing Letters* 23.6 (2016): 795-799.