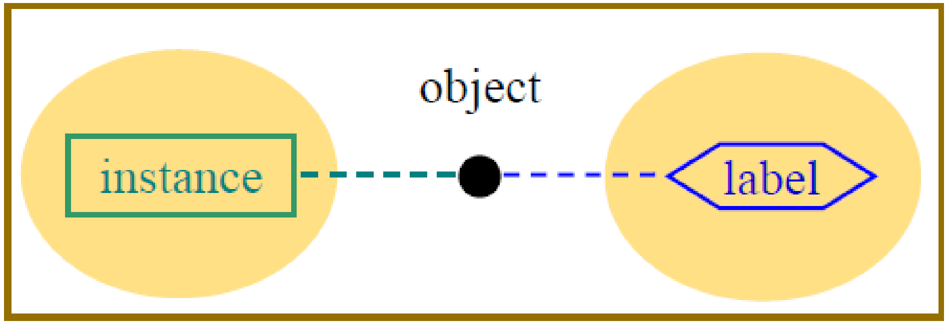
A Brief Introduction to Discriminative Feature Analysis for Multi-label Data Understanding

ZHANG Jia

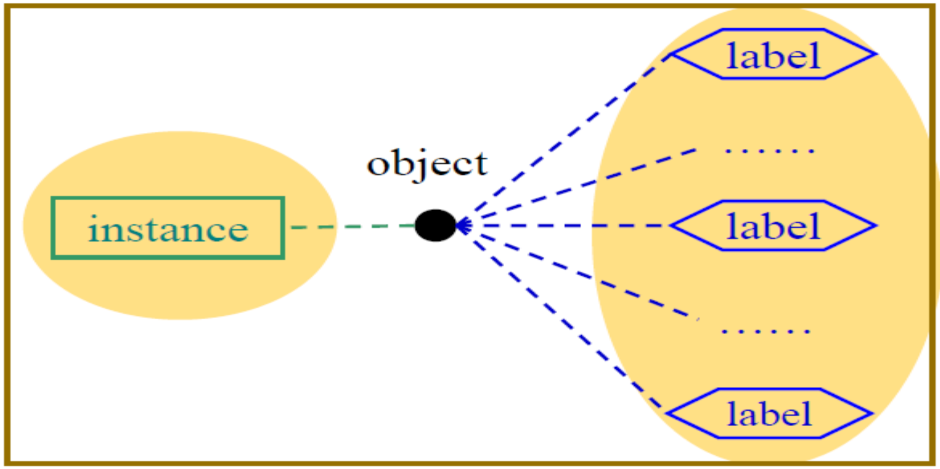
Artificial Intelligence Department
School of Informatics
Xiamen University, China

Multi-label Learning

- For single-label learning, an instance is attributed with a single label characterizing its semantics.



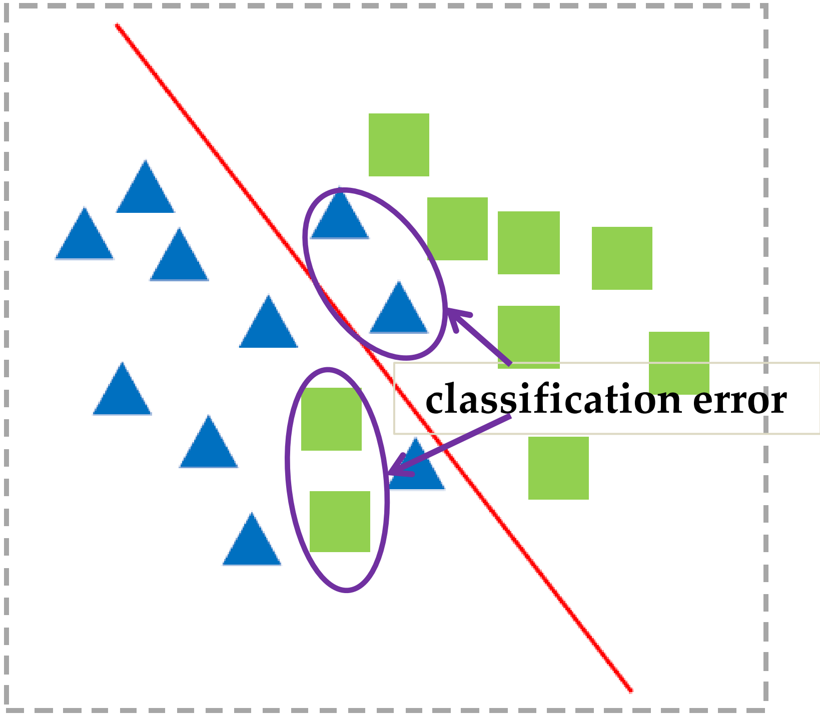
- For multi-label learning, an instance is attributed with multiple labels simultaneously.



Multi-label Learning

The reason to do the research:

- Considering that there are many labels, more training data are needed for distinguishing a label. However, the available training data are limited;
- The positive sample is not sufficient for each label (Even causing class-imbalance problem);
- Traditional supervised methods can't deal with multi-label data well.

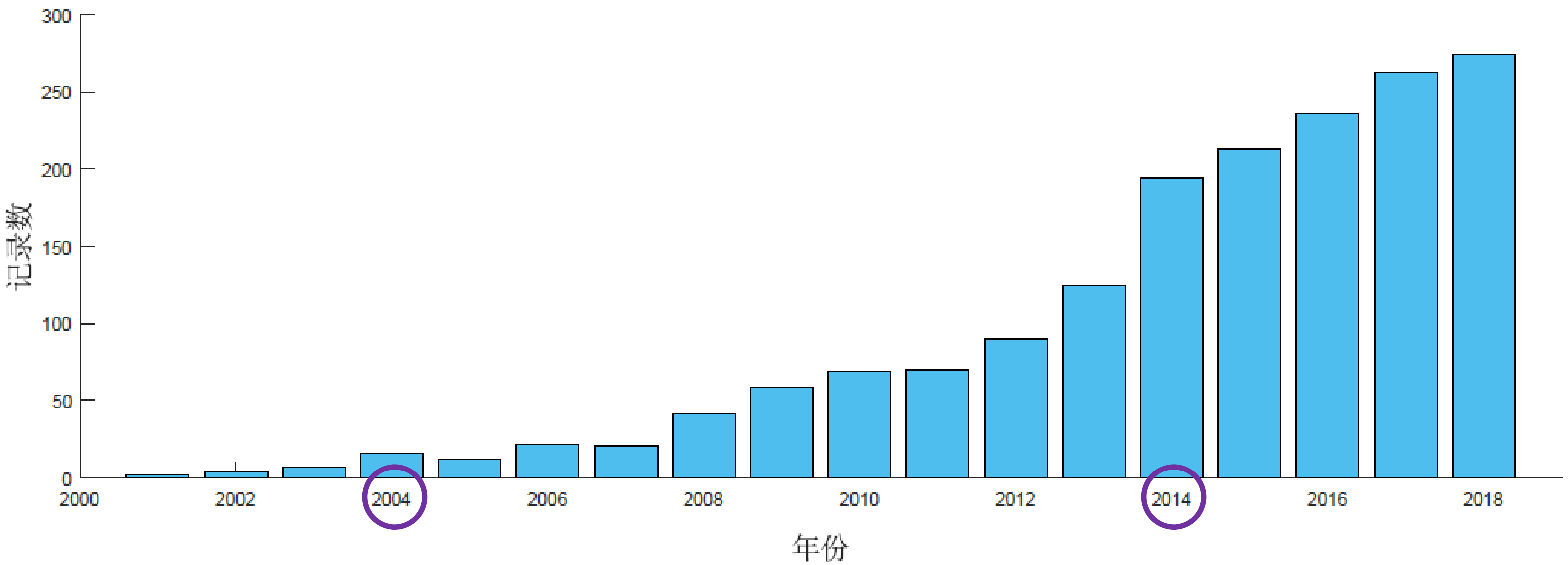


Supervised Learning

- ✓ G. Tsoumakas, I. Katakis, I. P. Vlahavas: Mining Multi-label Data. *Data Mining and Knowledge Discovery Handbook* 2010: 667-685
- ✓ M.-L. Zhang, Z.-H. Zhou: A Review on Multi-Label Learning Algorithms. *IEEE Trans. Knowl. Data Eng.* 26(8): 1819-1837 (2014)
- ✓ E. Gibaja, S. Ventura: A Tutorial on Multilabel Learning. *ACM Comput. Surv.* 47(3): 52:1-52:38 (2015)

Multi-label Learning

Research trend (Web of Science, searching for “multi-label”, ” multilabel”):



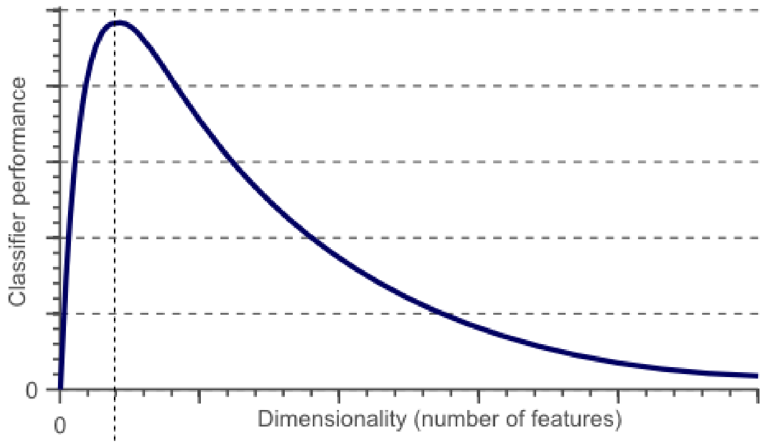
Related issues for multi-label learning:

- Multi-label Feature Selection
- Label-specific Feature Learning
- Extreme Multi-label Learning
- Multi-label Learning with Missing Labels
- Semi-supervised Multi-label Learning
- Hierarchical Multi-label Learning
- Label Distribution Learning
- Multi-label Learning with Streaming Labels
- Multi-source Multi-label Learning
- Large-scale Multi-label Learning

Contents

Multi-label Feature Selection

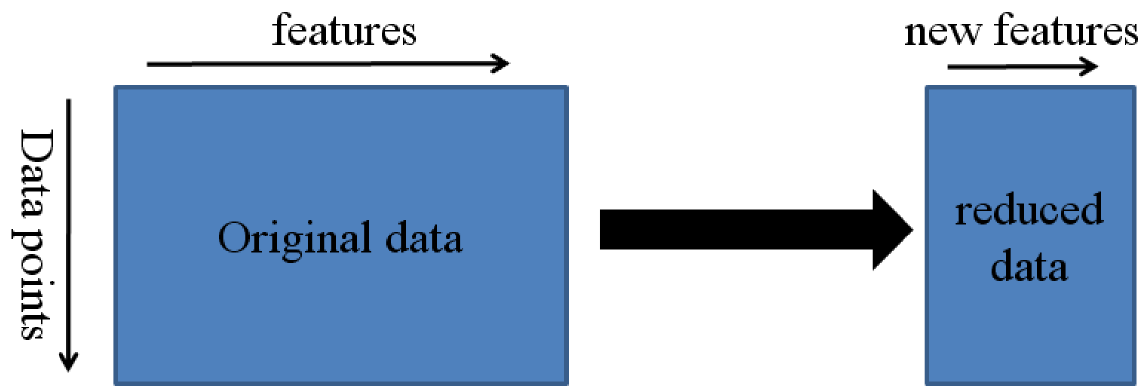
In practice, the curve of learning performance w.r.t. the feature dimension looks like this



Optimal number of features

http://www.visiondummy.com/2014/04/curse-dimensionality-affect-classification/

For a fixed sample size, there is an optimal number of features to use.



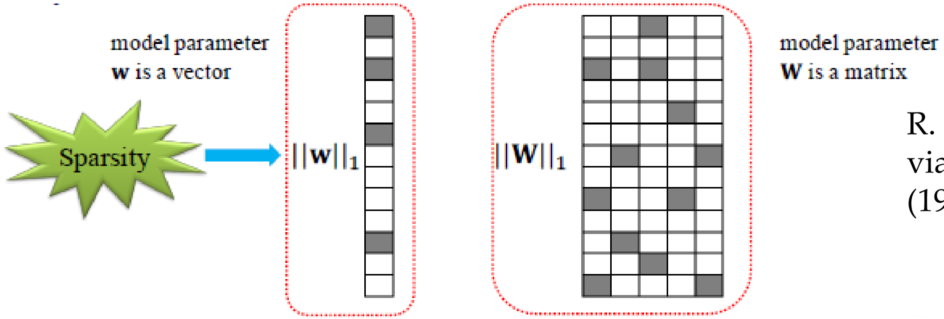
Y. Li, T. Li, H. Liu: Recent advances in feature selection and its applications. *Knowl. Inf. Syst.* 53(3): 551-577 (2017)

Sparse Learning based Methods

Suppose **W** is defined as a feature coefficient matrix.

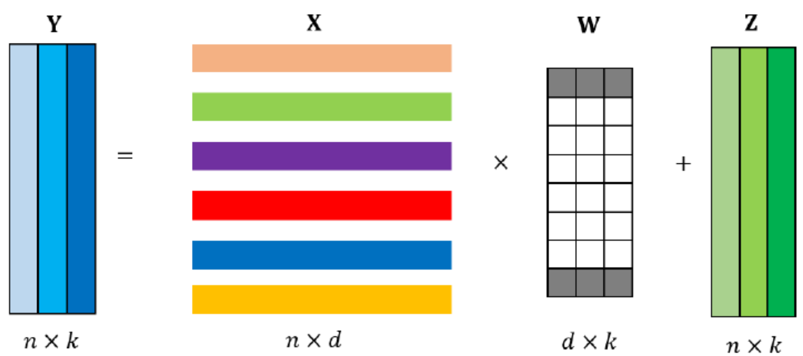
$||\mathbf{W}||_2$: It is capable for feature discriminability, commonly used to control the complexity.

$||\mathbf{W}||_1$: It is beneficial to obtain a strictly sparse solution.



R. Tibshirani: Regression shrinkage and selection via the lasso, *J. Royal Statistical Soc.* 58: 267–288 (1994)

$||\mathbf{W}||_{2,1}$: It is beneficial to obtain a strictly sparse solution shared by multiple labels.



F. Nie, H. Huang, X. Cai, C. H. Q. Ding: Efficient and Robust Feature Selection via Joint ℓ2, 1-Norms Minimization. *NIPS* 2010: 1813-1821

$||\mathbf{W}||_{2,0}$: T. Pang, F. Nie, J. Han, X. Li: Efficient Feature Selection via ℓ2, 0-norm Constrained Sparse Regression. *IEEE Trans. Knowl. Data Eng.* 31(5): 880-893 (2019)

Optimization Solution

Optimization Scheme for $||\mathbf{W}||_1$:

$$\min_{\mathbf{w}} loss(\mathbf{w}; \mathbf{X}, \mathbf{y}) + \alpha \|\mathbf{w}\|_1$$

Two conditions need to be met: (1) Empirical loss function *loss,* defined as *f(W),* is convex; (2) *Lipschitz* constant L_f of ∇f satisfies : $\|\nabla f(W_1) - \nabla f(W_2)\| \le L_f \|W_1 - W_2\|$

$$G^{(t)} = W^{(t)} - \frac{1}{L_f} \nabla f(W^{(t)}).$$

$W^{t+1} = O_\epsilon[G^{(t)}]$, where $O_\epsilon[w] = sign(w)(|w| - \epsilon)_+$

Optimization Scheme for $||\mathbf{W}||_{2,1}$:

$$\min_{\mathbf{W}} loss(\mathbf{W}; \mathbf{X}, \mathbf{Y}) + \alpha \|\mathbf{W}\|_{2,1}$$

$$h(W) = \|W\|_{2,1} = \sum_{i=1}^{d} \sqrt{\sum_{j=1}^{c} W_{ij}^2}$$

Note: the regularization is convex, $\nabla h = QW$, where Q is a diagonal matrix whose elements involve W.

$$\nabla f(W^{t+1}) + Q^t W^{t+1} = 0$$

Related References

Some references which use Sparse Learning based Method for multi-label feature selection:

- Y. Zhu, J. T. Kwok, Z.-H. Zhou: Multi-Label Learning with Global and Local Label Correlation. *IEEE Trans. Knowl. Data Eng.* 30(6): 1081-1094 (2018)
- J. Huang, G. Li, Q. Huang, X. Wu: Joint Feature Selection and Classification for Multilabel Learning. *IEEE Trans. Cybern.* 48(3): 876-889 (2018)
- T. Ren, X. Jia, W. Li, L. Chen, Z. Li: Label distribution learning with label-specific features. *IJCAI* 2019: 3318-3324
- A. Braytee, W. Liu, D. R. Catchpoole, P. J. Kennedy: Multi-Label Feature Selection using Correlation Information. *CIKM* 2017: 1649-1656
- P. Zhu, Q. Xu, Q. Hu, C. Zhang, H. Zhao: Multi-label feature selection with missing labels. *Pattern Recognit.* 74: 488-502 (2018)
- J. Wang, J. Wei, Z. Yang: Supervised Feature Selection by Preserving Class Correlation. *CIKM* 2016: 1613-1622

Information Theoretical based Methods

- Intuitively, with more selected features, the effect of feature redundancy should gradually decrease;
- Meanwhile, pairwise feature independence becomes stronger.

mRMR:

maximum relevance between features and labels

$$score(f_k) = \boxed{I(f_k;Y)} - \boxed{\frac{1}{|\mathcal{S}|}\sum_{f_j \in \mathcal{S}} I(f_k;f_j)}$$

reduced effect of feature redundancy

H. Peng, F. Long, C. H. Q. Ding: Feature Selection Based on Mutual Information: Criteria of Max-Dependency, Max-Relevance, and Min-Redundancy. *IEEE Trans. Pattern Anal. Mach. Intell.* 27(8): 1226-1238 (2005) **Cited by 6743**

Optimization Formulation for Minimum Redundancy Maximum Relevance:

mRMR-opt:

$$\max_x J_x = c^T x - x^T D x \quad \text{s.t.} \quad x_1, \ldots, x_N \geq 0, \sum_{i=1}^{N} x_i = 1$$

H. Lim, J.-S. Lee, D.-W. Kim: Optimization approach for feature selection in multi-label classification. *Pattern Recognit. Lett.* 89: 25-30 (2017)

Note: x is a feature weight vector, which can access the importance of all the features.

Discussion on mRMR-opt

mRMR-opt vs. mRMR: $\max_x J_x = c^T x - x^T D x$ s.t. $x_1, \ldots, x_N \geq 0, \sum_{i=1}^{N} x_i = 1$

- mRMR-opt is a constrained quadratic programming problem, which can be solved efficiently for a global optimal solution. mRMR is a filter method, and the feature subset is obtained as a local search.
- mRMR-opt: all features are involved for the global optimization; mRMR needs to specify the number of required features in the selection process.

Limitation of mRMR-opt: It's designed for multi-label feature selection, but unfriendly for multi-label data understanding.

- Label relationship;
- Extension like binary relevance: class-imbalance, relative labeling-importance…

Related References

Some references which use Information Theoretical based Method for multi-label feature selection:

- J. Lee, I. Yu, J. Park, D.-W. Kim: Memetic feature selection for multilabel text categorization using label frequency difference. *Inf. Sci.* 485: 263-280 (2019)
- P. Zhang, G. Liu, W. Gao: Distinguishing two types of labels for multi-label feature selection. *Pattern Recognit.* 95: 72-82 (2019)
- J. Gonzalez-Lopez, S. Ventura, A. Cano: Distributed multi-label feature selection using individual mutual information measures. *Knowl.-Based Syst. (*in press).
- J. Wang, J.-M. Wei, Z. Yang, S.-Q. Wang: Feature Selection by Maximizing Independent Classification Information. *IEEE Trans. Knowl. Data Eng.* 29(4): 828-841 (2017)
- J.-S. Lee, D.-W. Kim: SCLS: Multi-label feature selection based on scalable criterion for large label set. *Pattern Recognit.* 66: 342-352 (2017)
- Y. Lin, Q. Hu, J. Liu, J. Duan: Multi-label feature selection based on max-dependency and min-redundancy. *Neurocomputing* 168: 92-103 (2015)

Contents

1 Multi-label Feature Selection

2 Label-specific Feature Learning

3 Learning Deep Features for MLC

Label-specific Feature Learning

Label-specific features are exploited to benefit the discrimination of different class labels.

- Color-based features would be preferred in discriminating sky and non-sky images.
- Texture-based features would be preferred in discriminating desert and non-desert images.

Method 1: L1-norm Regularization

Example 1:

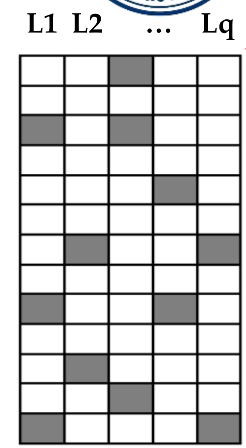
Learning label manifold for missing label complement

Search for discriminative features shared for each label

$$\min_{\mathbf{W},\mathbf{C}} \frac{1}{2}||\mathbf{XW}-\mathbf{YC}||_F^2 + \boxed{\frac{\lambda_1}{2}||\mathbf{YC}-\mathbf{Y}||_F^2 + \lambda_2||\mathbf{C}||_1} + \lambda_3||\mathbf{W}||_1 + \lambda_4\mathrm{tr}(\mathbf{WLW}^T)$$

$$s.t. \quad \mathbf{C} \succeq 0$$

Generate the classifier *W*



The mapping from feature space to the generated label space

J. Huang, F. Qin, X. Zheng, Z. Cheng, Z. Yuan, W. Zhang, Q. Huang: Improving multi-label classification with missing labels by learning label-specific features. *Inf. Sci.* 492: 124-146 (2019)

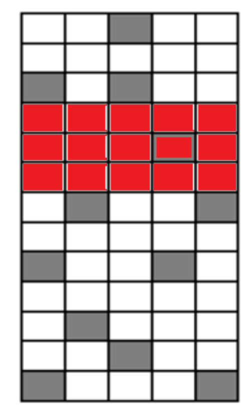
Example 2:

Term 2: Search for discriminative features for each label

Term 3: Search for discriminative features shared by all labels

$$\min_{W,M} \frac{1}{2}\|X(W+M)-D\|_F^2 + \lambda_1\|W\|_1 + \lambda_2\|M\|_{2,1}$$
$$+ \lambda_3 tr(X(W+M)(P-R)(X(W+M))^T)$$
$$s.t. \; X(W+M)\times 1_{l\times 1} = 1_{n\times 1}$$
$$X(W+M) \geq 0_{n\times l},$$

Generate the classifier *W+M*



Term4: Label correlation exploitation

T. Ren, X. Jia, W. Li, L. Chen, Z. Li: Label distribution learning with label-specific features. *IJCAI 2019*: 3318-3324

Method 2: Cluster

For one class label $l_k \in \mathcal{Y}$, the set of positive training instances $\mathcal{P}_k$ as well as the set of negative training instances $\mathcal{N}_k$ correspond to:

$$\mathcal{P}_k = \{\boldsymbol{x}_i \mid (\boldsymbol{x}_i, Y_i) \in \mathcal{D}, l_k \in Y_i\}$$
$$\mathcal{N}_k = \{\boldsymbol{x}_i \mid (\boldsymbol{x}_i, Y_i) \in \mathcal{D}, l_k \notin Y_i\}$$

k-means algorithm

$\mathcal{P}_k$ is partitioned into m_k^+ disjoint clusters whose centers are denoted $\{\boldsymbol{p}_1^k, \boldsymbol{p}_2^k, \ldots, \boldsymbol{p}_{m_k^+}^k\}$

$\mathcal{N}_k$ is partitioned into m_k^- disjoint clusters whose centers are denoted $\{\boldsymbol{n}_1^k, \boldsymbol{n}_2^k, \ldots, \boldsymbol{n}_{m_k^-}^k\}$

Label-specific feature space construction:

$$\phi_k(x) = [d(\boldsymbol{x}, \boldsymbol{p}_1^k), \ldots, d(\boldsymbol{x}, \boldsymbol{p}_{m_k}^k), d(\boldsymbol{x}, \boldsymbol{n}_1^k), \ldots, d(\boldsymbol{x}, \boldsymbol{n}_{m_k}^k)]$$

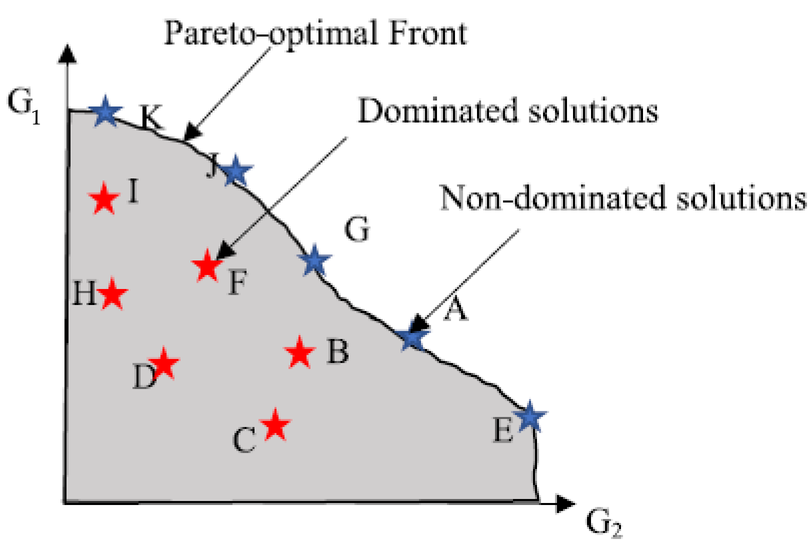
- ✓ M.-L. Zhang, L. Wu: LIFT: Multi-Label Learning with Label-Specific Features. *IEEE Trans. Pattern Anal. Mach. Intell.* 37(1): 107-120 (2015)
- ✓ Y. Guo, F. Chung, G. Li, J. Wang, J. C. Gee: Leveraging Label-Specific Discriminant Mapping Features for Multi-Label Learning. *ACM Trans. Knowl. Discov. Data* 13(2): 24:1-24:23 (2019)

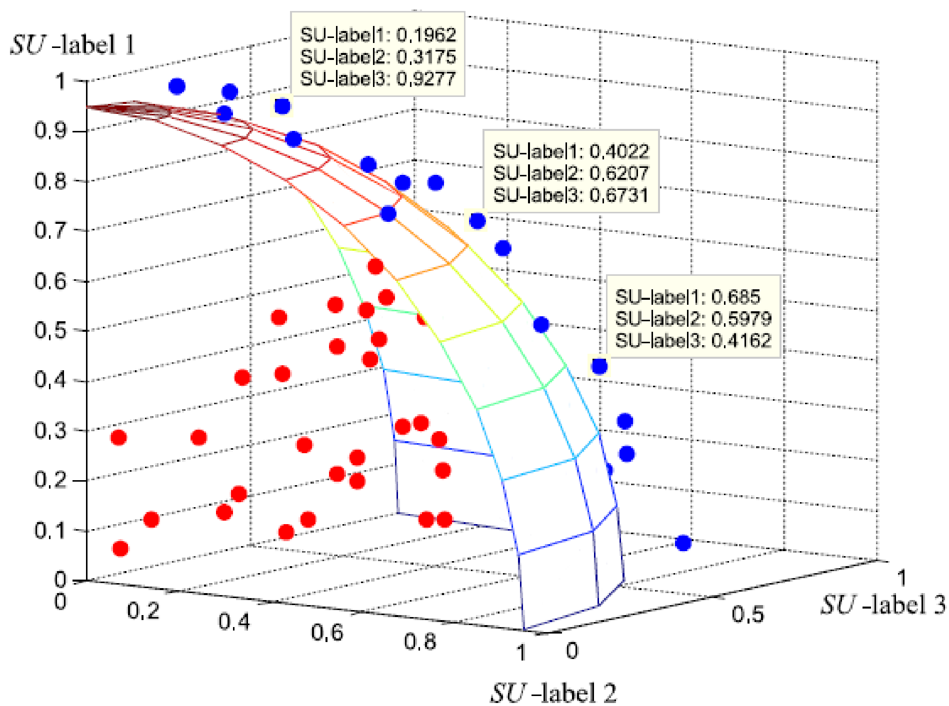
Method 3: Multi-objective Optimization

(Reference) S. Kashef, H. Nezamabadi-pour: A label-specific multi-label feature selection algorithm based on the Pareto dominance concept. *Pattern Recognit.* 88: 654-667 (2019)

Idea: The method transforms label-specific feature learning problem into multi-objective optimization problem. Specially, **objective functions are considered as correlation between each feature and the existing labels**.

Multi-objective Optimization:





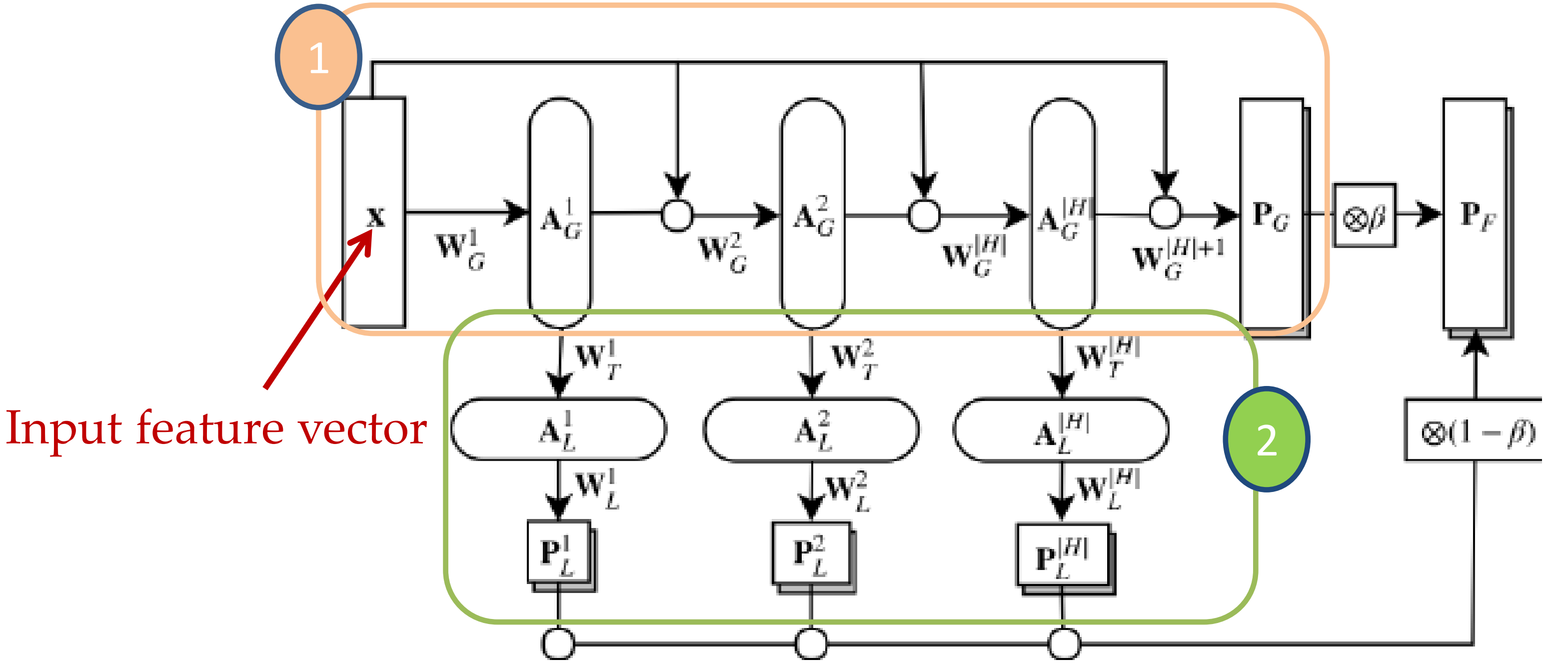
Contents

1 Multi-label Feature Selection

2 Label-specific Feature Learning

3 Learning Deep Features for MLC

Deep Fully-connected Neural Network



β is the parameter to balance the influence of the local and the global.

1. The global flow is designed with a fully-connected neural network to learn deep features for distinguishing all labels.

2. The local flow is designed to learn local features for predicting the set of classes from each level.

J. Wehrmann, R. Cerri, Rodrigo C. Barros: Hierarchical Multi-Label Classification Networks. *ICML* 2018: 5225-5234

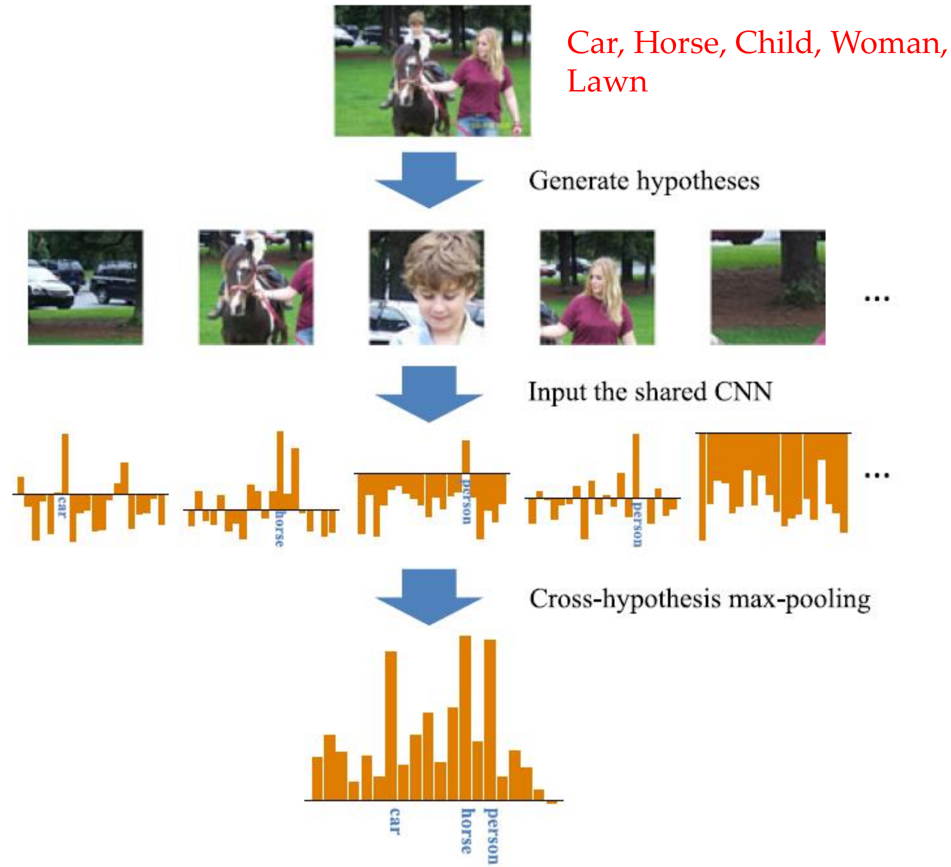
Deep Convolutional Neural Network

HCP framework:

Step 1: Hypotheses Extraction: objectness detection, normalized cut;

Step 2: Initialization. the parameters of CNN pre-trained on ImageNet, the parameters of the final fully-connected layer (the number of outputs equals to the number of class labels) pre-trained on target image data set;

Step 3: Hypotheses-fine-tuning is carried out based on the proposed framework.



Y. Wei, W. Xia, M. Lin, J. Huang, B. Ni, J. Dong, Y. Zhao, S. Yan: HCP: A Flexible CNN Framework for Multi-Label Image Classification. *IEEE Trans. Pattern Anal. Mach. Intell.* 38(9): 1901-1907 (2016)

THANK YOU FOR YOUR ATTENTION