Introduction to Data Mining

Advanced Topics in Multi-Label Classification

Jun Huang

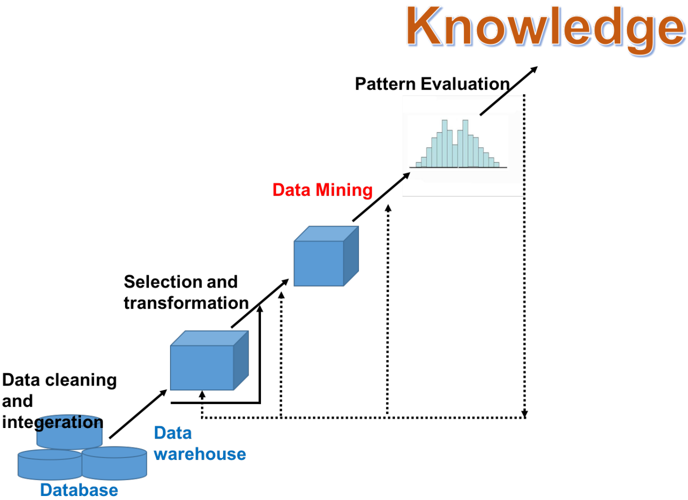
Anhui University of Technology

Spring 2018

huangjun_cs@163.com

KDD Process

Data Mining-Core of Knowledge discovery process



Multi-Label Classification

- **Single-label classification**: Is this a picture of beach?

$$\in \{\text{yes,no}\}$$

- **Multi-label classification**: Which labels are relevant to this picture?

$$\subseteq \{\text{beach, sunset, foliage, field, mountain, urban}\}$$

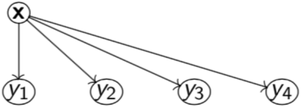
- i.e., each instance can have multiple labels instead of a single one

Advanced Topics in Multi-Label Classification

- Label Correlation
- Missing labels
- Extream Multi-label Classification
- Discover New Labels

Binary Relevance (BR): A Probabilistic View

- BR model: $\mathbf{h} = (h_1, h_2, ..., h_L)$
- each $h_j : \mathcal{X} \to \{0, 1\}$
- for $\widetilde{\mathbf{x}}$, predict:



$$\hat{y}_j = h_j(\widetilde{\mathbf{x}}) \equiv \arg \max_{y_j \in \{0,1\}} p(y_j|\widetilde{\mathbf{x}})$$

- predictions made independently

$$\mathbf{h}(\widetilde{\mathbf{x}}) \equiv [h_1(\widetilde{\mathbf{x}}), h_2(\widetilde{\mathbf{x}}), ..., h_L(\widetilde{\mathbf{x}})]$$

- If labels are independent,...but they are not!

$$p(\mathbf{y}|\mathbf{x}) \neq \Pi_{j=1}^{L} p(y_j|\mathbf{x})$$

Label Correlation

- In multi-label classification, labels often have correlations with each other
- It has been shown that exploiting label correlations between labels can improve the performances of classifiers
- For example, if one image is annotated with label "sailing boat", it has a high probability to be labeled as "sea water"

Label Correlation

Existing strategies to label correlations exploitation could among others be roughly categorized into three families, based on the order of correlations that the learning techniques have considered

- First-order strategy: decomposing the multi-label learning problem into a number of independent binary classification problems (one per label)
- Second-order strategy: considering pairwise relations between labels
- High-order strategy: considering high-order relations among labels such as imposing all other labels' influences on each label

Label Correlation
First-Order

First-order strategy: decomposing the multi-label learning problem into a number of independent binary classification problems (one per label) , e.g.,

- Binary Relevance (BR): M. R. Boutell, J. Luo, X. Shen, and C. M. Brown, Learning multi-label scene classification, Pattern Recognit., vol. 37, no. 9, pp. 1757–1771, 2004.
- ML*k*NN: M. Zhang and Z. Zhou, Ml-knn: A lazy learning approach to multi-label learning, Pattern Recognit., vol. 40, no. 7, pp. 2038–2048, 2007.
- ...

First-order algorithms are simple and efficient, these algorithms could be less effective due to the ignorance of label correlations

Label Correlation
Second-Order

Second-order approaches exploit pairwise relationships between labels

- The most popular way to model pairwise relationship is to exploit the interaction between any pair of labels
- There are several methods for calculating the correlations between class labels, such as
  - Pearson correlation coefficients
  - hamming distance
  - jaccard similarity
  - cosine similarity
  - ...
- Some approaches:
  - CLR[1], LLSF[2], ML-TLLT...

[1]Multi- label classification via calibrated label ranking,2008
[2]Learning label specific features for multi-label classification,ICDM2015

Label Correlation
Second-Order

Second-order approaches exploit pairwise relationships between labels

- Another way is to incorporate the criterion of ranking loss into the objective function to be optimized when learning the classification models
  - RankSVM[3], BP-MLL[4], RELIAB[5], ...

---

[3] A kernel method for multi-labelled classification,NIPS2001

[4] Multilabel neural networks with applications to functional genomics and text categorization,TKDE2006

[5] Leveraging implicit relative labeling-importance information for effective multi-label learning, ICDM2015

Label Correlation

High-Order

High-order approaches tackle the multi-label learning prob- lem by mining relationship among all the class labels or a subset of class labels

- Problem transformation approaches
  - LP[6], RA*k*EL[7], EPS[8], Classifier Chains (CC)[9]...

---

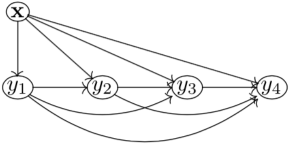
[6] Mining multi-label data,2010

[7] Random *k*-labelsets: An ensemble method for multilabel classification,ECML2007

[8] Multi-label classification using ensembles of pruned sets,ICDM2008

[9] Classifier chains for multi-label classification, ECML2009

- build $\mathbf{h} = (h_1, h_2, ..., h_L)$
- each $h_j : \mathcal{X} \times \{0,1\}^{j-1} \to \{0,1\}$
- and, for any $\widetilde{\mathbf{x}}$, predict



$$\hat{y}_j = h_j(\widetilde{\mathbf{x}}, \hat{y}_1, ..., \hat{y}_{j-1}) \equiv \arg \max_{y_j \in \{0,1\}} p(y_j | \widetilde{\mathbf{x}}, \hat{y}_1, ..., \hat{y}_{j-1})$$

- models label correlations

$$\mathbf{h}(\widetilde{\mathbf{x}}) \equiv [h_1(\widetilde{\mathbf{x}}), h_2(\widetilde{\mathbf{x}}, \hat{y}_1), ..., h_L(\widetilde{\mathbf{x}}, \hat{y}_1, ..., \hat{y}_{L-1})]$$

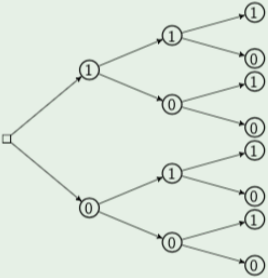
- Inspiration from the chain rule (a greedy approximation):

$$p(\mathbf{y}|\mathbf{x}) = p(y_1|\mathbf{x}) \Pi_{j=2}^{L} p(y_j | \mathbf{x}, y_1, y_2, ..., y_{j-1})$$

Label Correlation

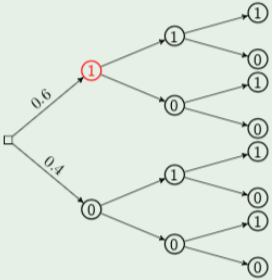
High-Order Approaches: Classifier Chains (CC)

Example



$$\hat{\mathbf{y}} = \mathbf{h}(\tilde{\mathbf{x}}) = [?, ?, ?]$$

Example



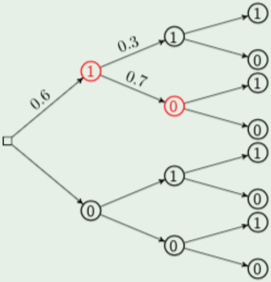
1. $\hat{y}_1 = h_1(\tilde{\mathbf{x}}) = 1$

$\hat{\mathbf{y}} = \mathbf{h}(\tilde{\mathbf{x}}) = [1, ?, ?]$

Label Correlation

High-Order Approaches: Classifier Chains (CC)

Example



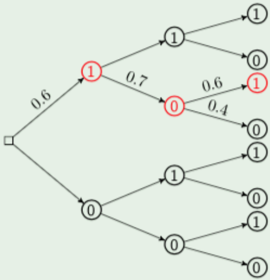
1. $\hat{y}_1 = h_1(\tilde{\mathbf{x}}) = 1$
2. $\hat{y}_2 = h_2(\tilde{\mathbf{x}}, \hat{y}_1) = 0$

$\hat{\mathbf{y}} = \mathbf{h}(\tilde{\mathbf{x}}) = [1, 0, ?]$

Label Correlation

High-Order Approaches: Classifier Chains (CC)

Example



1. $\hat{y}_1 = h_1(\tilde{\mathbf{x}}) = 1$
2. $\hat{y}_2 = h_2(\tilde{\mathbf{x}}, \hat{y}_1) = 0$
3. $\hat{y}_3 = h_3(\tilde{\mathbf{x}}, \hat{y}_1, \hat{y}_2) = 1$

$\hat{\mathbf{y}} = \mathbf{h}(\tilde{\mathbf{x}}) = [1, 0, 1]$

Label Correlation
High-Order Approaches: Classifier Chains (CC)

Advantages of CC:

- similar time complexity to BR in practice (if $L < D$)
- better performance than BR
- can improve (a lot) with Bagging Ensembles of CC (ECC):
  - M CC models, each with a random chain and sample of $\mathcal{D}$.

Issues with CC:

- Error Propagation: errors may be propagated down the chain
- It may be inappropriate that each class label is dependent on the previous ones in the chain

Extentions of CC:

- Bayes Optimal Probabilistic Classifier Chains[10] (PCC)
- Monte-Carlo search for Classifier Chains[11] (MCC)
- Conditional Dependency Networks[12] (CDN)
- Prudent[13]
- ...

[10] Bayes optimal multilabel classification via probabilistic classifier chains, ICML2010

[11] Efficient monte carlo methods for multi-dimensional learning with classifier chains, Pattern Recognition, 2014

[12] Multi-Label Classification Using Conditional Dependency Networks, IJCAI2011

[13] Prudent: A pruned and confident stacking approach for multi- label classification,TKDE2015

Label Correlation

High-Order Approaches

In CC, each class label is dependent on the previous ones in the chain, it may be inappropriate

- Exploit the label correlation between class label by mining their dependent structures, e.g.,
  - Tree structure [14]
  - Bayesian Networks [15]
  - ...

[14] Bayesian chain classifiers for multidimensional classification, IJCAI2011

[15] Multi-label learning by exploiting label dependency, KDD2010

Missing/Noisy Label

- Noisy Label: The set of labels assigned to each example may not be fully valid, e.g. some labels may be wrongly assigned due to mistakes of human labellers
- Missing/Weak Label: The absence of some labels do not necessarily mean they are invalid for the example, e.g. only a "partial" set of proper labels is assigned by the human labeller

Missing/Noisy Label

How to solve multi-label classification with Missing/Noisy class labels?

- Complete the **Label Matrix** before learning
- Combine the **learning step** with label matrix completion
- Calculate **classification loss** without considering the Missing values

Combine the learning step with label matrix completion

- Given a training set with n samples $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n]^T$
- The Label Matrix: $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n]^T$
- $y_{ij} = 1$ or 0 means $\mathbf{x}_i$ belongs to y_j or not, and $y_{ij} = ?$ indicates the value is missing

$$J(\Theta) = \sum_{(i,j)} \ell(y_{ij}, f_j(\mathbf{x}_i, \Theta)) + \lambda\mathcal{R}(\Theta)$$

$$s.t. \quad \mathbf{Y} = \mathbf{Y} + \mathbf{E} \text{ or } \mathbf{Y} = \mathbf{YS} + \mathbf{E} \text{ or } \mathbf{Y} = \mathbf{CD} + \mathbf{E}$$

and other constraints

- where Θ indicates the training data with known labels

Missing/Noisy Label

Calculate classification loss without considering the Missing values

- Given a training set with n samples $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n]^T$
- The Label Matrix: $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_n]^T$
- $y_{ij} = 1$ or 0 means $\mathbf{x}_i$ belongs to y_j or not, and $y_{ij} = ?$ indicates the value is missing

$$J(\mathbf{W}) = \sum_{(i,j)\in\Omega} \ell(y_{ij}, f_j(\mathbf{x}_i, \mathbf{W})) + \lambda\mathcal{R}(\mathbf{W})$$

- where Ω indicates the training data with known labels

Missing/Noisy Label

Important Literatures

- Multi-label learning with weak label, AAAI2010
- Large-scale Multi-label Learning with Missing Labels, ICML, 2014
- Learning low-rank label correlations for multi-label classification with missing labels, ICDM2014
- Multilabel classification with label correlations and missing labels, AAAI2014
- Ml-mg: Multi-label learning with missing labels using a mixed graph, ICCV2015
- Semi-supervised multi-label learning with incomplete labels, IJCAI2015
- Improving multi-label learning with missing labels by structured semantic correlations, ECCV2016
- Learning from Semi-Supervised Weak-Label Data, AAAI2018

Extream Multi-label Classification

- Most existing MLL algorithms will fail when the label space is large, e.g. $L > 50$, especially for the second-order and high-order approaches.
- The number of class labels is so high in many applications
- http://manikvarma.org/downloads/XC/XMLRepository.html
- The labeling sparsity and structures should be exploited

Extream Multi-label Classification

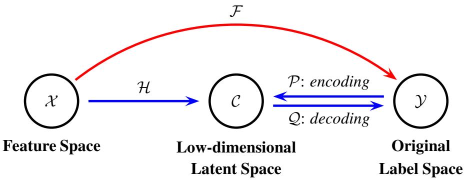
Dataset	Download	Feature Dimensionality	Label Dimensionality	Number of Train Points	Number of Test Points	Avg. Points per Label	Avg. Labels per Point
Mediamill	Download	120	101	30993	12914	1902.15	4.38
Bibtex	Download	1836	159	4880	2515	111.71	2.40
Delicious	Download	500	983	12920	3185	311.61	19.03
RCV1–2K	Download	47236	2456	623847	155962	1218.56	4.79
EURLex–4K	Download	5000	3993	15539	3809	25.73	5.31
AmazonCat–13K	Download Dataset Download Feature & Label Meta–data Download Raw Text for Deep Learning	203882	13330	1186239	306782	448.57	5.04
AmazonCat–14K	Download Dataset Download Feature & Label Meta–data Download Raw Text for Deep Learning	597540	14588	4398050	1099725	1330.1	3.53
Wiki10–31K	Download Dataset Download Feature & Label Meta–data Download Raw Text for Deep Learning	101938	30938	14146	6616	8.52	18.64
Delicious–200K	Download	782585	205443	196606	100095	72.29	75.54
WikiLSHTC–325K	Download	1617899	325056	1778351	587084	17.46	3.19
Wikipedia–500K	Download Dataest Download Feature & Label Meta–data	2381304	501070	1813391	783743	24.75	4.77
Amazon–670K	Download Dataset Download Feature & Label Meta–data Download Raw Text for Deep Learning	135909	670091	490449	153025	3.99	5.45
Ads–1M	–	164592	1082898	3917928	1563137	7.07	1.95
Amazon–3M	Download Dataset Download Feature & Label Meta–data Download Raw Text for Deep Learning	337067	2812281	1717899	742507	31.64	36.17
Ads–9M	–	2082698	8838461	70455530	22629136	14.32	1.79

Extream Multi-label Classification

How to solve multi-label classification with extream number of class labels?

- LSDR: Label Space Dimension Reduction
- Exploit the Structures of Class Labels
- Parallel and Distributed Computing

- To tackle a multi-label classification problem with many classes, recently label space dimension reduction (LSDR) is proposed.
- It encodes the original label space to a low-dimensional latent space by $\mathcal{P}$
- And uses a decoding process $\mathcal{Q}$ for recovery



Important Literatures

- Multi-Label Learning with Millions of Labels: Recommending Advertiser Bid Phrases for Web Pages, WWW, 2013.
- FastXML: A Fast, Accurate and Stable Tree-classifier for eXtreme Multi-label Learning, KDD, 2014
- Multi-label Classification via Feature-aware Implicit Label Space Encoding, ICML, 2014
- Sparse Local Embeddings for Extreme Multi-label Classification, NIPS, 2015
- PD-Sparse: A Primal and Dual Sparse Approach to Extreme Multiclass and Multilabel Classification, ICML, 2016
- PPDSparse: A Parallel Primal-Dual Sparse Method for Extreme Classification, ICML, 2017
- DiSMEC-Distributed Sparse Machines for Extreme Multi-label Classification, WSDM, 2017

Discover New Class Labels

- Previous works on multi-label classification focused on a fixed set of class labels
- In many application, the environment is open and new concepts may emerge with previously unseen instances
- For example, in birdsong recognition, experts label long audio intervals with a fix set of bird species. Other categories of sound such as rain or car sound are not included in the labeling process. Yet, such sounds are present in the data
- Another example is image annotation, the annotator considers only a fixed set of tags and ignores "grass"as it is not included in the tag set.

Discover New Class Labels

How to deal with multi-label data with new class labels?

- Offline Learning
  - Assume their are k number of new labels hidden in the data
  - Construct the model to discover these labels
- Online Learning
  - Detect new class labels
  - Update old classfication models
  - Learn new models for new class labels

Discover New Class Labels

Important Literatures

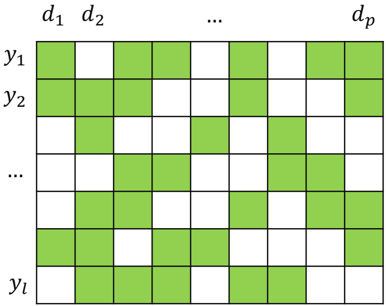
- Online multi-label active annotation: Towards large-scale content-based video search, ACM MM, 2008
- From n to n+1: Multiclass transfer incremental learning, CVPR2013
- Multi-instance multi-label learning in the presence of novel class instances, ICML2015
- Multi-label learning with emerging new labels, ICDM2016
- Discovermultiplenovellabelsinmulti-instancemulti-label learning, AAAI2017

Label-Specific Feature Learning

- Existing approaches learn from multi-label data by manipulating with identical feature set
- Each class label might be determined by some specific characteristics of its own
  - For example, in text categorization, features related to word terms such as government, national security and presidential election would be informative in discriminating political and non-political documents
  - While features related to word terms such as GDP, tax reduction and stock markets would be informative in discriminating economic and non-economic documents

Label-Specific Feature Learning

How to learn label-specific features of it own?



- Feature transformation
- Feature selection
- Label correlation between labels should be considered

LIFT: Multi-label learning with label-specific features

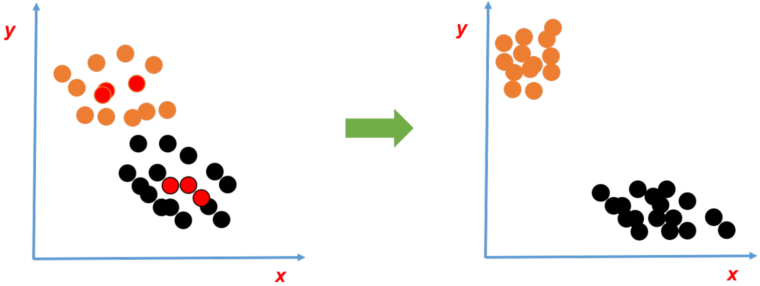
- Given a training data set $\mathcal{D} = \{(\mathbf{x}_i, Y_i)\}_{i=1}^{m}$
- For each label l_k, define the positive training instances $\mathcal{P}_k$ and negative training instances $\mathcal{N}_k$

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

- Clustering the positive and negative training instances into m_k^+ and m_k^- disjoint clusters respectively
- The centers are denoted as $\{\mathbf{p}_1^k, \mathbf{p}_2^k, \ldots, \mathbf{p}_{m_k^+}^k\}$ and $\{\mathbf{n}_1^k, \mathbf{n}_2^k, \ldots, \mathbf{n}_{m_k^-}^k\}$

Label-Specific Feature Learning

- LIFT set $m_k^+ = m_k^- = m_k$, and $m_k = \lceil r \cdot \min(|\mathcal{P}_k|, |\mathcal{N}_k|) \rceil$
- where parameter $r \in [0, 1]$
- Mapping function
  $\phi_k(\mathbf{x}) = [d(\mathbf{x}, \mathbf{p}_1^k), ..., d(\mathbf{x}, \mathbf{p}_{m_k}^k), d(\mathbf{x}, \mathbf{n}_1^k), ..., d(\mathbf{x}, \mathbf{n}_{m_k}^k)]$



Label-Specific Feature Learning

- A new binary training set $\mathcal{B}_k$ with m examples is created from the original multi-label trainingg set $\mathcal{D}$ by the mapping function ϕ_k

$$\mathcal{B}_k = \{(\phi_k(\mathbf{x}_i),\ Y_i(k))|(\mathbf{x}_i,\ Y_i) \in \mathcal{D}\}$$

- where $Y_i(k) = +1$ if $l_k \in Y_i$; otherwise, $Y_i(k) = -1$
- Based on $\mathcal{B}_k$, any binary learner $\Im$ can be applied to induce a classification model: $g_k : \mathcal{Z}_k \to \mathbb{R}$ for l_k
- Given an unseen example $\mathbf{x}_t$, its can be predicted as

$$Y = \{l_k|g_k(\phi_k(\mathbf{x}_t)) > 0, 1 \leq k \leq q\}$$

Important Literatures

- Lift: Multi-label learning with label-specific features, TPAMI2015
- Multi-label learning with label-specific features, ICDM2015
- A dirty model for multi-task learning, NIPS2010
- A multivariate regression approach to association analysis of a quantitative trait network,Bioinformatics2009
- Task sensitive feature exploration and learning for multitask graph classification,TCYB,2016.