Function Approximation for Adaptive Learning of Label Distributions

Miao Cheng

Email: mcheng@mailbox.gxnu.edu.cn
School of Computer Science and Information Engineering
Guangxi Normal University
Guilin, Guangxi, China

ICSIP, Oct 2021

Outline

Outline

Background Knowledge

Label Distribution Learning

- Traditional classification methods focus on absolute label learning.
- Label Distribution Learning (LDL) aims to estimate the probabilistic labels of instances, which outputs the range **probabilities** of categories with the sums of unity.
- As a consequence, **more** category and distribution information can be provided.

Outline

Keypoints

Keypoints

- Estimation of Label Probabilities: only feature information of instances are available.
- Bridge the **connection** between informative patterns of samples and categories.
- The **adaptive learning** mechanism is hardly to be devised to match with LDL problems, while ubiquitous computing is desired.

Outline

Main Ideas

Main Ideas

- In theory, the distributions of instances are quite **related** with the features of themselves, of which the calculation and estimation are **feasible** for the **rough distirbutions** of instances.
- The estimated values **fails** to accurately depict the true distributions of each instances, but further steps of **adjustments** are available for improvements.
- The probabilistic labels of **out-of-class** samples can be **predicated** by following the learned **models**, while **enhanced** learning is proposed to obtain the ideal estimations with **adaptive learning of iterations**.

Outline

Dimensionality Reduction

- Rather than the state-of-the-art methods, **random** projections are referred in this work, and consequently, calculation convenience and efficiency can be preserved, e.g.,

$$F(x_i) = \theta^T x_i, \tag{1}$$

- Note that, it actually makes **tiny** impact on learning results in this work.

Adherent Affinity of Probabilistic Labels

Adherent Affinity for Transformed Samples

- Obviously, the probabilistic labels of **supervised** instances can be adhered naturally, but dimensionality reduction is preferred **simultaneously**.
- In terms of this, **randomly** selected normalized samples are referred as projection directions, while it resorts to label affinity also.
- As a consequence, the affinity of probabilistic labels can be adopted to pre-processing in a **single** one step, and **class-specific** samples black are generated as

$$a_{i,j} = f_j(x_i) = d_{i,j}F(x_i). \tag{2}$$

Outline

Categorical Distributions of Instances

Gaussian Approximation

- Without loss of generality, the **Gaussian** normal distributions of each classes are calculate, by assuming the generated samples of each category are common with the **Gaussian** distributions.
- More specifically, **the class mean** μ and **covariance** σ are calculated, and the distribution of each sample can be estimated by

$$g_{\mu,\sigma^2}\left(a_{i,j}\right) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{\left(a_{i,j}-\mu\right)^T \Gamma \left(a_{i,j}-\mu\right)}{2\sigma^2}\right). \tag{3}$$

Outline

Function Adjustment

Function Adjustment

- The obtained distribution functions are deduced from the **class-specific** feature data, and fail to match with the corresponding **ground truth** of label distributions normally.
- In terms of this, a further step of **adjustment** is adopted to enhance the connection between the learned distributions and ground truth, e.g.,

$$d_{i,j} = \rho\left(x_i\right) = h_j\left(g_j\left(f_j\left(x_i\right)\right)\right) \tag{4}$$

and

$$a_{i,j} \rightarrow g\left(\cdot\right) \rightarrow h\left(\cdot\right) \sim d_{i,j}. \tag{5}$$

Function Adjustment

Function Ajustment

- There are many candidate solutions to such problem, while **nonnegative learning** is adopted for convenience.
- Without loss of generality, the **optimization** problem and **approximate** solution are respectively,

$$D_i \approx Wg\left(a_i\right) \quad s.t. \quad W \geq 0 \tag{6}$$

and

$$W_{r,c} \leftarrow W_{r,c} \frac{\left(D_i g(a_i)^T\right)_{r,c}}{\left(Wg\left(a_i\right) g(a_i)^T\right)_{r,c}} \tag{7}$$

The Proposed Algorithm of Learning Model

Algorithm 1

Input: Given data pairs of training instances $T_r = \{(x_1, D_1), (x_2, D_2), \cdots, (x_n, D_{n_r})\}$.

Output: The weighted feature function $f(\cdot)$, the distribution function $g(\cdot)$, and the approximate function $h(\cdot)$.

- Calculate the artificially generated samples of each instance x_i corresponding to class-specific weighted random projections, and obtain $a_{i,1}, a_{i,2}, \cdots, a_{i,c}$.
- Calculate the mean vector μ and variance σ associated with the generated samples of each feature function, and obtain the gaussian distribution function $g_j(\cdot),\ j = 1, 2, \cdots, c$.
- Solve the problem defined in (6) by updating W as (7), and obtain the approximate function $h(\cdot)$.

Outline

Adaptive Label Predication

Label Predication with Learned Model

- With the learned functions, it is still unable to make predication for new instances, owing to the fact that their label affinity are **unknown** beforehand.
- To conquer this problem, the alternating optimization idea is adopted, which is derived from **iterative optimization** of expectation-maximization (EM) algorithm.

Adaptive Label Predication

Label Predication with Learned Model

- The initial probabilities of predicated instances are **averagely** set to be **equivalent** distributions for each class.
- The objective values are calculated with optimized parameters, and then the parameters are updated based on predicated objective values, till **stable** results are obtained.
- During the iterations, the probabilistic labels are necessary to be further normalized, namely

$$p\left(l_i \left| y\right.\right) = \frac{\rho_i\left(y_i\right)}{\sum\limits_{i=1}^{c} \rho_i\left(y_i\right)}. \tag{8}$$

The Predication Procedure for A New Instance y

Algorithm 2

Input: Given the pre-learned approximate functions $g(\cdot)$ and $f(\cdot)$, the feature data of y, the maximum number of iteration t.
Output: The predicated labels $p(l_i|y)$.

- Initial probabilistic labels of instance $p(l_i|y)$ with average label distributions.
- **While** maximum iteration t **or** stable predication $p(l_i|y)$ has never been reached
  - Calculate the generated samples of instance y with the weighted feature function $f(\cdot)$.
  - Calculate the values of approximate distribution function $g(\cdot)$ corresponding to the generated samples, and obtain transformed nonnegative distributions $\rho_i(y)$ with adjusted function $h(\cdot)$.
  - Normalize and obtain the predicated labels $p(l_i|y)$ with Eq. (8).
- Obtain the final $p(l_i|y)$ as output.

Outline

Experimental Setting

Data Sets

- The benchmark LDL data sets consisting of 15 artificial data sets are employed in this experiment.

Hardwares

- 2.8 GHz CPU
- 16 GB Memory

Experimental Setting

Setting

- Each data set is randomly splitted into **five** folds, and **four** folds of them are used for training while the remaining **one** fold data is for testing.
- The original data are reduced down to ten features with random sampling directions.
- 5,000 iterations are mostly performed in the stage of nonnegative learning.

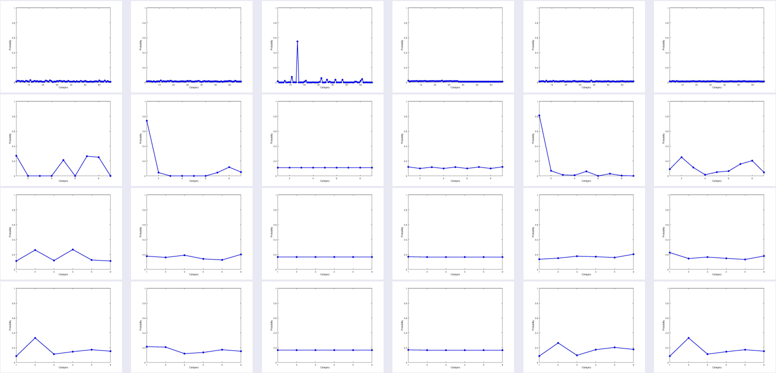
Experimental Results

Experimental Results

- As shown in Tab. I-VI, the proposed FA-LDL method is able to achieve **comparable** results associated with the median ranking ones in most cases.
- **Stable** results are obtained on diverse data sets with different measure distances.

Experimental Results

Some Obtained Label Distributions



The predicated label distributions of different algorithms on diverse data sets. The subfigures from the first to the sixth columns respectively denote the label distributions from original, AA-kNN, PT-Bayes, SA-IIS, SA-BFGS, and FA-LDL.

Conclusion

- By calculating the *Gaussian* distribution functions of each class, the label distributions of each sample is estimated accordingly.
- To improve the estimated results, a further **adjustment** step is proposed with **nonnegative learning**.
- Since the label affinity of testing instances are **unknown** beforehand, an **adaptive label predication** is adopted with **average** initialization.
- Experimental results on diverse artificial data sets demonstrate that, the proposed adaptive learning method is able to give **stable** performance with comparable results.

Thank you for your participation !