Tao-cheng Hu, Jin-hui Yu, 2016. Max-margin based Bayesian classifier. *Frontiers of Information Technology & Electronic Engineering*, **17**(10): 973-981. http://dx.doi.org/10.1631/FITEE.1601078

Max-margin based Bayesian classifier

Key words: Multi-class learning, Max-margin learning, Online algorithm

Corresponding author: Tao-cheng Hu
E-mail: hutaocheng@gmail.com

ORCID: http://orcid.org/0000-0002-6722-2420

Motivation

- Multi-class classification is a basic problem in machine learning, it surfaces a variety of domains, including object recognition, speech recognition, document categorization and so on.
- Multi-class classification has been subject to intense study, both theoretical and practical.
- Multi-class learning suffers from poor generalization ability or heavy computational overhead.
- Multi-class learning methods usually lack semantics and are difficult to interpret.

Main idea

- Bayesian network make the model easy to understood.
- A carefully designed objective function with max-margin property attaches the model with good generalization performance.
- The convexity of the objective function could derive an online learning algorithm with logarithmic regret.

Method

1. We design a probability model named the Generative Correlation Multi-class Classifier, which correlated the input feature and the label by a probability embedding.

Algorithm 1 Generative correlation multi-class classifier

1: Choose embedding $\boldsymbol{w} \sim \mathcal{N}(\mu = 0, \sigma^2 = 1)$
2: **for** each independent and identically distributed pair $(\boldsymbol{x}, y)$ **do**
3: $\quad \boldsymbol{q} \leftarrow \mathrm{Emb}(\boldsymbol{x}; \boldsymbol{w})$
4: $\quad$ Choose $y \sim \mathrm{Multinomial}(\boldsymbol{q})$
5: **end for**

Method (Cont'd)

Here are the log-likelihood objective function

$$\max_{\boldsymbol{w}} \mathcal{L}(\boldsymbol{w}; \{\boldsymbol{x}_t, y_t\}_{t=1}^T)$$
$$= -\frac{1}{2}\|\boldsymbol{w}\|_{\mathcal{F}}^2 + \sum_{t=1}^{T} \langle \boldsymbol{e}_{y_t}, \log \boldsymbol{\phi}(\boldsymbol{w}\boldsymbol{x}_t) \rangle,$$

- We prove the objective function had max-margin property and is convex (see the paper for more details).
- Also note that the regularization term is shared by the data term. This is why we aggregates the dualities in the online algorithm [Duality Aggregation].

Major results

- Our model has appreciated generalization performance.

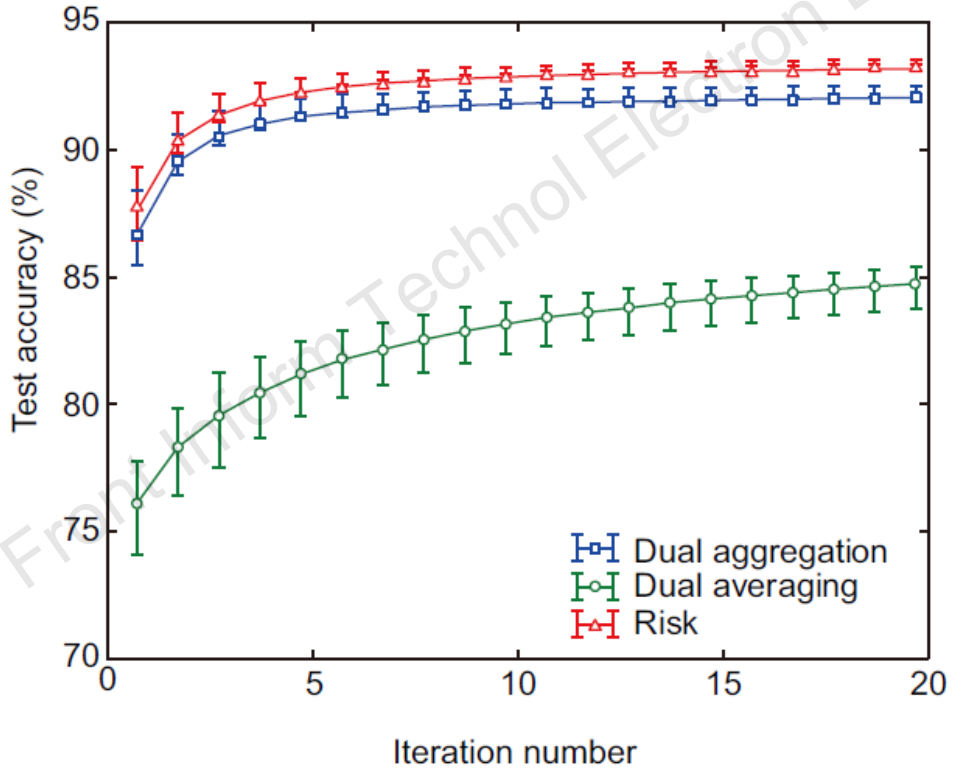

Fig. 2 Evolutions of classification accuracy versus the iteration number on the MNIST dataset

Major results (Cont'd)

- Compared with many classification frameworks, our model had better performance.

Table 1 Performances of various multi-classification schemes on the MNIST dataset with different data representations

Data set	Scheme	Accuracy (%)	Time (s)	
			Training	Prediction
MNIST (raw)	one-vs-one	91.35 ± 0.35	**11.65**	4.09
	one-vs-rest	86.57 ± 1.26	15.30	0.47
	ECOC3	84.57 ± 0.61	48.90	1.30
	ECOC4	85.14 ± 1.00	63.27	1.72
	DAve	85.53 ± 0.62	12.87	0.14
	DAgg	$\mathbf{92.55 \pm 0.12}$	12.87	**0.14**
MNIST (PCA+RBF)	one-vs-one	97.07 ± 0.13	16.40	6.65
	one-vs-rest	95.53 ± 0.08	23.07	0.75
	ECOC3	94.66 ± 0.15	66.93	2.08
	ECOC4	94.89 ± 0.20	97.37	2.08
	DAve	87.58 ± 0.67	16.93	0.15
	DAgg	$\mathbf{97.09 \pm 0.09}$	**12.08**	**0.11**

Bold numbers denote the best performances of the related indicators. ECOCn: ECOC whose code size equals n; DAve: duality averaging; DAgg: duality aggregation

Major results (Cont'd)

- Our model works better when the class number grows.

Table 2 Performances of various multi-classification schemes on the COIL dataset with different class numbers

Data set	Scheme	Accuracy (%)	Time (s)	
			Training	Prediction
COIL-20	one-vs-one	98.89 ± 0.50	0.72	0.18
	one-vs-rest	98.08 ± 2.02	0.59	0.01
	ECOC4	97.97 ± 0.97	2.83	0.04
	ECOC5	97.74 ± 0.91	1.16	0.05
	DAve	88.85 ± 2.29	8.54	0.03
	DAgg	$\mathbf{98.00 \pm 0.66}$	**0.68**	**0.01**
COIL-100	one-vs-one	93.33 ± 0.80	41.42	79.89
	one-vs-rest	87.67 ± 1.77	34.87	0.80
	ECOC6	78.02 ± 1.19	352.31	4.92
	ECOC7	77.86 ± 0.97	410.98	5.58
	DAve	55.28 ± 1.77	**14.31**	**0.03**
	DAgg	$\mathbf{91.62 \pm 0.59}$	14.92	**0.03**

Bold numbers denote the best performances of the related indicators. ECOCn: ECOC whose code size equals n; DAve: duality averaging; DAgg: duality aggregation

Conclusions

- We propose a Bayesian model named the Generative Correlation Multi-Classifier for classification, where Bayesian network and classification are well integrated as we assigned each element with probabilistic semantics.
- The GCMC mind the gap between generalization performance and computational overhead

+ max-margin property

+ convexity property

+ local variables are eliminated, the optimization problem is much easier

+ Duality Aggregation VS Duality Averaging