Astroinformatics
Proceedings IAU Symposium No. 325, 2016
M. Brescia, S.G. Djorgovski, E. Feigelson,
G. Longo & S. Cavuoti, eds.

doi:10.1017/S1743921316012679

ELM-KNN for photometric redshift estimation of quasars

Yanxia Zhang,[1] Yang Tu,[1,2] Yongheng Zhao[1] and Haijun Tian[2]

[1]Key Laboratory of Optical Astronomy, National Astronomical Observatories, Chinese Academy of Sciences, Beijing, P.R.China
email: zyx@bao.ac.cn

[2]College of Science, China Three Gorges University, Yichang, Hubei, P.R.China

Abstract. We explore photometric redshift estimation of quasars with the SDSS DR12 quasar sample. Firstly the quasar sample is separated into three parts according to different redshift ranges. Then three classifiers based on Extreme Learning Machine (ELM) are created in the three redshift ranges. Finally k-Nearest Neighbor (kNN) approach is applied on the three samples to predict photometric redshifts of quasars with multiwavelength photometric data. We compare the performance with different input patterns by ELM-KNN with that only by kNN. The experimental results show that ELM-KNN is feasible and superior to kNN (e.g. rms is 0.0751 vs. 0.2626 for SDSS sample), in other words, the ensemble method has the potential to increase regressor performance beyond the level reached by an individual regressor alone and will be a good choice when facing much more complex data.

Keywords. methods: data analysis, techniques: photometric, quasars: general, catalogs, surveys

1. Introduction

Photometric redshifts may provide distance information of celestial objects so that it is an important tool to study many significant issues, such as the formation and evolution of galaxies, large-scale structure of cosmology, reionization of the early universe, galaxy clustering, high-redshift quasars and so on. The development and operation of large photometric survey missions (e.g. SDSS, WISE, UKIDSS) creates large opportunities and data testbed for the research of photometric redshift techniques. The studies on this respect are thriving. Taking the photometric redshift estimation of quasars for example, Wu & Jia (2010) explored template-matching; Zhang *et al.* (2013) applied k-Nearest Neighbors algorithm (kNN); Brescia *et al.* (2013) put forward the Multi Layer Perceptron with Quasi Newton Algorithm (MLPQNA); Han *et al.* (2016) presented an integration of KNN and SVM.

2. Data

Our data sets are adopted from the Data Release 12 Quasar catalog (DR12Q) (Paris *et al.* 2016). After removing the records with default values and z warning, we obtain the multiwavelength samples. According to which surveys the samples are from, the data sets include four samples: SDSS sample, SDSS-UKIDSS sample, SDSS-WISE sample, SDSS-UKIDSS-WISE sample. For each sample, we divide it into three subsamples: one with $z_{\rm spec} < 1.3$, one with $1.3 \leqslant z_{\rm spec} < 4.3$ and one with $z_{\rm spec} \geqslant 4.3$, based on the spectroscopic redshift distribution histagram of known quasars.

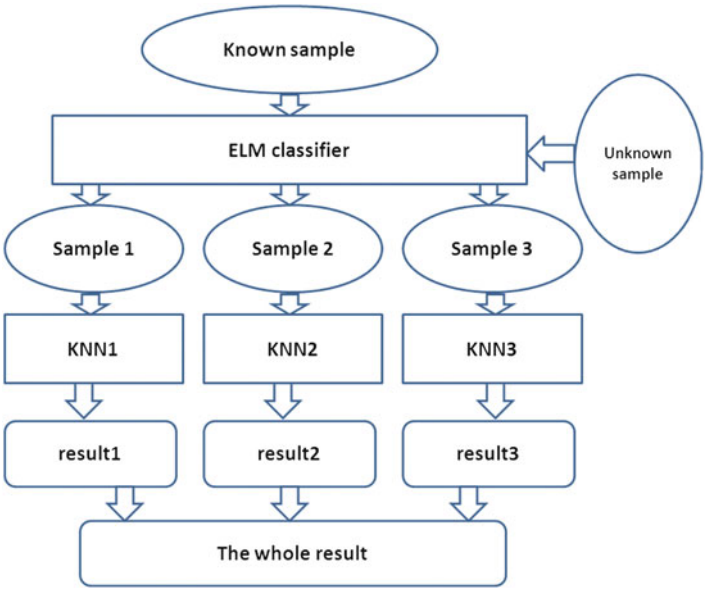

Figure 1. The flowchart of ELM-KNN for Photometric Redshift Estimation

3. Method

A popular learning algorithm called extreme learning machine (ELM) was proposed for both generalized single hidden layer feedforward network and multi-hidden-layer feedforward networks by Huang (2015) and Huang, Zhu & Siew (2006). This algorithm randomly chooses hidden nodes and analytically determines the output weights of feedforward networks. Compared to the other feedforward network learning algorithms, ELM is easily implemented, tends to reach the smallest training error, obtains the smallest norm of weights, provides the good generalization performance at extremely fast learning speed. It is applied for feature learning, clustering, regression and classification. It gradually comes into view due to its superiorities in the big data era.

k-Nearest Neighbors algorithm (kNN) is among the simplest of all machine learning algorithms, belongs to a kind of instance-based learning, or lazy learning, and can be used for classification and regression. For classification, the output of kNN is a class membership, which is determined by a majority vote of its neighbors. For regression, the output of kNN is the average of the values of its k nearest neighbors.

In this paper we firstly divide the quasar sample into three different subsamples (one with $z_{\rm spec} < 1.3$ as Sample 1, one with $1.3 \leqslant z_{\rm spec} < 4.3$ as Sample 2 and one with $z_{\rm spec} \geqslant 4.3$ as Sample 3) according to the redshift ranges, then randomly separate each of the three subsamples into two parts: two thirds for training and one-third for testing. By testing, the reliable ELM classifier is created. During the three redshift ranges, different optimal kNN regressors are built. For any unknown-redshift object, the ELM classifier provides its type as Sample 1, Sample 2 or Sample 3, and then the corresponding kNN regressor gives its predicted redshift. The overall flowchart is shown in Fig 1.

4. Results

For each of SDSS, SDSS-UKIDSS, SDSS-WISE and SDSS-UKIDSS-WISE samples, we explore extreme learning machine (ELM) to classify the quasar sample into three different redshift subsamples with different input patterns and model parameters. The

optimal input and model parameters as well as the classification accuracy are shown in Table 1. When arriving at the best performance, different samples correspond to different input patterns and model parameters. Table 1 indicates that the accuracy is more than 84% for all samples, and even reaches 97.51% for the SDSS-UKIDSS-WISE sample. Therefore the following photometric redshift estimation is reliable based on the classified samples obtained by ELM.

Table 1. The performance of ELM classifiers on different samples.

Sample	Input Pattern	No. of Hidden Neurons	Activation Function	Accuracy(%)
SDSS	$4C+i$	494	Triangular basis function	84.03
SDSS-UKIDSS	$8C+r$	682	Radial basis function	93.44
SDSS-WISE	$8C^{'}+r$	562	Sigmoidal function	94.04
SDSS-UKIDSS-WISE	$12C$	1904	Triangular basis function	97.51

Notes: $4C = u-g, g-r, r-i, i-z$; $8C = u-g, g-r, r-i, i-z, z-Y, Y-J, J-H, H-K$;$8C^{'} = u-g, g-r, r-i, i-z, z-W1, W1-W2, W2-W3, W3-W4$; $12C = u-g, g-r, r-i, i-z, z-Y, Y-J, J-H, H-K, K-W1, W1-W2, W2-W3, W3-W4$.

For each subsample of SDSS, SDSS-UKIDSS, SDSS-WISE and SDSS-UKIDSS-WISE samples, we apply k-Nearest Neighbors algorithm (kNN) to predict photometric redshifts of quasars, respectively. The experimental results are shown in Table 2. The subsamples at low redshift (Sample 1), medium redshift (Sample 2) and high redshift (Sample 3) are represented as a, b and c, respectively. As Table 2 indicates, the percents within different $|\Delta z|$ are more than 86.476% and the root mean square (rms) errors are less than 0.0818. The optimal k value for each subsample is also given in Table 2.

Table 2. Photometric redshift estimation on different subsamples by kNN.

Sample	Input Pattern	Subsample	k	$\|\Delta z\|<0.1$(%)	$\|\Delta z\|<0.2$(%)	$\|\Delta z\|<0.3$(%)	rms
SDSS	$4C+i$	a	9	90.112	97.687	99.225	0.0730
		b	12	93.679	99.020	99.816	0.0562
		c	6	100.000	100.000	100.000	0.0229
SDSS-UKIDSS	$8C+r$	a	10	89.430	98.299	99.503	0.0677
		b	13	88.552	96.402	98.789	0.0800
		c	9	100.000	100.000	100.000	0.0188
SDSS-WISE	$8C^{'}+r$	a	9	86.476	96.390	99.046	0.0818
		b	10	93.826	99.081	99.780	0.0561
		c	6	100.000	100.000	100.000	0.0229
SDSS-UKIDSS-WISE	12C	a	6	88.747	97.680	99.304	0.0746
		b	5	93.987	98.376	99.760	0.0568
		c	1	100.000	100.000	100.000	0.0225

Notes: $4C$,$8C$,$8C^{'}$,$12C$ are the same as in Table 1.

Here we put forward a method of firstly classifying the sample into different subsamples by ELM and then estimating photometric redshifts by kNN, called as ELM-KNN. In order to check the effectiveness of ELM-KNN, we compare the performance of ELM-KNN with that of kNN, as shown in Table 3. As far as the percents in different $|\Delta z|$ ranges and rms error, ELM-KNN shows better performance than kNN for any sample. The performance rank of samples from top to bottom is SDSS-UKIDSS-WISE sample, SDSS-WISE sample, SDSS sample and SDSS-UKIDSS sample.

Table 3. Comparison of photometric redshift estimation of ELM-KNN and kNN.

Sample	Input Pattern	Method	$\|\Delta z\| < 0.1(\%)$	$\|\Delta z\| < 0.2(\%)$	$\|\Delta z\| < 0.3(\%)$	rms
SDSS	$4C + i$	ELM-KNN	89.440±0.138	97.320±0.064	99.030±0.031	0.0751±0.0005
		kNN(k = 22)	61.140±0.143	78.520±0.100	85.040±0.098	0.2626±0.0010
SDSS-UKIDSS	$8C + r$	ELM-KNN	89.121±0.116	96.892±0.059	98.973±0.052	0.0756±0.0005
		kNN(k = 6)	74.470±0.262	87.050±0.140	92.130±0.109	0.1939±0.0018
SDSS-WISE	$8C^{'} + r$	ELM-KNN	90.015±0.107	97.565±0.102	99.189±0.097	0.0718±0.0010
		kNN(k = 14)	82.970±0.292	92.330±0.173	95.770±0.116	0.1367±0.0025
SDSS-UKIDSS-WISE	$12C$	ELM-KNN	92.293±0.703	98.266±0.324	99.571±0.126	0.0614±0.0030
		kNN(k = 6)	85.610±0.742	92.960±0.411	96.800±0.336	0.1290±0.0074

Notes: $4C$,$8C$,$8C^{'}$,$12C$ are the same as in Table 1.

5. Conclusion

We apply ELM-KNN and kNN for photometric redshift estimation of quasars with different samples and different input patterns. The experimental results indicate that the performances of ELM-KNN are all superior to kNN for the four different samples with different input patterns. It is obvious that when the sample is divided into subgroups according to redshift distribution, the redshift prediction accuracy improves. In general, the more information from more bands are given, algorithms have better performance. But when adding parameters from more bands, the size decrease of a sample sometimes leads to accuracy reduction (e.g. SDSS-UKIDSS sample), at this time, we should pay more attention on whether it is necessary to cross-identify the sample with other band catalogues. Facing a concrete problem, we need pursue a balance between sample size and sample information. Moreover for a special algorithm, the choice of optimal model parameters is necessary in practice, and when samples are set, the input pattern is very important for accuracy improvement. In addition, with more and more data collected, we consider not only how to improve the accuracy of photometric redshift estimation, but also how to increase the speed of photometric redshift estimation. In a word, each data processing step, data mining algorithm choice as well as algorithm application environments directly influence the effectiveness and efficiency of photometric redshift estimation of quasars. New upcoming survey projects (e.g. LSST) will provide more opportunities and challenges for us at this issue.

Acknowledgements

This paper is funded by 973 Program 2014CB845700 and the National Natural Science Foundation of China under grant Nos.11178021, 11033001. We acknowledge the SDSS, UKIDSS and WISE databases.

References

Brescia, M., Cavuoti, S., D'Abrusco, R., Longo, G., & Mercurio, A. 2013, *ApJ*, 772(2), 140
Han, B., Ding, H.-P., Zhang, Y.-X., & Zhao, Y.-H. 2016, *RAA*, 16(5), 74
Huang, G.-B., 2015, *Cogn Comput* 7, 263
Huang, G.-B., Zhu, Q.-Y., & Siew, C.-K. 2006, *Neurocomputing*, 70, 489
Paris, I., Petitjean, P., Ross, N. P., *et al.* 2016, *arXiv e-print*, arXiv:1608.06483
Wu, X.-B. & Jia, Z.-D., 2010, *MNRAS* 406, 1583
Zhang, Y., Ma, H., Peng, N., Zhao, Y., & Wu, X. 2013, *AJ*, 146, 22