Research on Small Sample Radio Signal Analysis Based on the VSGGTD Algorithm

Kai Zhou[1, a], Kaiyu Qin[2], Yuqi Zeng and Xueling Zhang

[1]The State Radio Monitoring Center, Beijing 100000, China;

[2]University of Electronic Science and Tech, Chengdu, China.

[a]zhoukai@srrc.org.cn

Abstract. The current feature extraction methods for small sample signals focus on improving the performance. How to build a virtual sample set with strong rationality and high accuracy based on the original sample is one of the challenges in realizing small sample signal analysis. In line with the radio service feature in the time domain and the frequency domain of signals, this paper proposes a virtual sample generation method based on an improved Gaussian algorithm, the VSGGTD algorithm, and proves its rationality from mathematics. In the experiment, feature extraction was carried out for the sample set before and after expansion, and the different extraction results were compared. Experimental results show that this algorithm can improve the underfitting phenomenon caused by small sample size.

Keywords: Small sample data; feature extraction; virtual sample; underfitting.

1. Introduction

In the analysis of the application of 5G frequency band, signal sparsity in frequency domain, occurrence time uncertainty in time domain and low interception probability of radio services such as radar, satellite communication and deep space exploration, etc. will give rise to the small number of signal samples. Most data analysis models are built on the premise of large data volume and uniform distribution, and it is difficult to accurately and comprehensively extract the features of small sample signals. Data analysis thoughts of small samples are mainly divided into two types: one is to improve feature extraction performance and enhance feature recognition capability based on prior knowledge; the other is to expand original data and construct a reasonable and applicable virtual sample set.

For the improvement of feature extraction performance, an improved support vector machine algorithm is adopted in reference [1] to extract features. However, it is difficult to determine the parameters needed to construct accurate kernel function for this algorithm. Reference [2] adopts three-layer wavelet transform for face image under small sample condition, and then puts the approximate components of each layer into MMC processing. This algorithm is easy to lose signal features or generate false signals when the signal noise is low. In fact, for the small sample radio signals, the prior knowledge is limited, so improving the feature extraction performance according to the theoretical model is usually accompanied by weak generalization ability.

Under the background of increasingly abundant machine learning resources, constructing virtual samples with strong applicability and meeting the services' requirements has become an important way to improve feature extraction ability. The method of generating virtual sample set based on geometric changes proposed by Poggio has strong applicability, but some singular sample points will be generated when expanding the data set, which makes it difficult to deal with. The sample construction algorithm based on the prior knowledge is generally reasonable in the specific field but has poor adaptability. Other algorithms with a larger range of applicability are usually limited in rationality.

Considering the rationality along with applicability, reference [4] proposed the Gaussian virtual sample generation method to expand the small sample data, so as to improve the accuracy of feature extraction. At first this paper introduced the Gaussian virtual sample generation method, and then improved the algorithm to propose the VSGGTD algorithm and theoretically proved the validity of the algorithm to generate samples. Finally, the feature extraction results of the original samples, Gaussian virtual samples and VSGGTD virtual samples were compared through experiments. Experimental results show that the VSGGTD algorithm is the best way to extract virtual samples.

2. Gaussian Virtual Sample Algorithm

Virtual sample generation method is based on Gaussian distribution, and for each real sample, according to the Gaussian distribution $\mathrm{N}(\mu, \eta^2)$ its random disturbance is increased.

Suppose you have the original sample set $\mathrm{X_N} = \{\mathrm{X_1},\ \mathrm{X_2},\ \mathrm{X_3},\ \mathrm{X_4}, \dots \dots \mathrm{X_n}\}$,regard each element in the original sample X_N as truth value, i.e. $\mu = X_i$. According to the theory put forward in reference [3], the variance is ranked $\eta = {10^{-a*\varepsilon}}/{6}$, where a denotes the precision cut to the decimal place, and the disturbance size can be adjusted through $a * \varepsilon$. On the basis of the original sample, increase $\mathrm{N}(0, \eta^2)$ disturbance in accordance with Gaussian distribution.

Generate M virtual samples for each original sample in the sample sequenceX_N, and then we can get $\mathrm{M} * \mathrm{N}$ virtual samples. Taking two-dimensional data as an example, the expanded sample distribution is shown in the figure below:

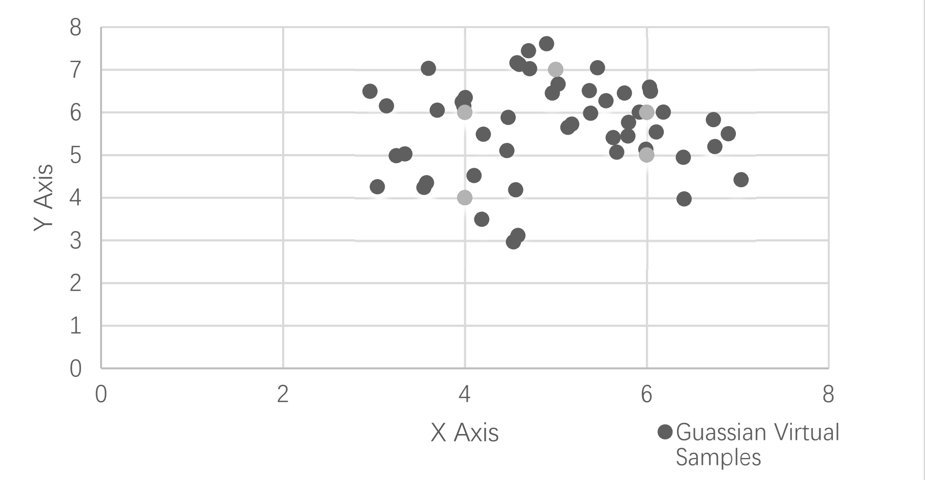

Fig. 1 Example of virtual samples generation based on Gauss distribution

3. VSGGTD Algorithm

3.1 The VSGGTD Algorithm Steps

The virtual samples generated by Gaussian algorithm have high adaptability, but low feature extraction accuracy. On this basis, this paper proposes an improved Gaussian algorithm, that is, the virtual sample generation method based on Guassian time sequence distribution algorithm (the VSGGTD algorithm), based on the prior knowledge in the time domain and frequency domain: it is assumed that each original sample value is true and the variation of feature value in each dimension satisfies smoothness. Considering the continuity of signal feature variation, as for sample sequence $\mathrm{X_N} = \{\mathrm{X_1},\ \mathrm{X_2},\ \mathrm{X_3},\ \mathrm{X_4}, \dots \dots \mathrm{X_n}\}$, in the process from moment t1 to moment t2, assume that the feature varied between the signals neighboring points is continuous.

The nearest adjacent point X_{ti} between the original sample set and X_{tj} is obtained, that is, the minimum Euclidean distance $L = \left|\left|X_{ti} - X_{tj}\right|\right|^2$.

According to X_{ti}, X_{tj}, calculate the parameters of Gaussian distribution $\mathrm{N}(\mu i, \eta^2)$ to generate $\mathrm{Q} = \frac{\omega}{L}$ virtual samples, namely $\mu i = X_{ti} + i * \frac{(X_{tj} - X_{ti})}{Q}$. According to reference [3], the variance was set as $\eta = {10^{-a*\varepsilon}}/{6}$.

According to X_{ti},X_{ti+1}, calculate the parameters of the Gaussian distribution of the virtual sample point to generate P virtual samples. $\mu i = X_{ti} + i * \frac{(X_{tj} - X_{ti})}{P}$ According to reference [3], the variance was set as $\eta = {10^{-a * \varepsilon}}/{6}$, in which $P + Q = M$.

In the above steps, $Q = \frac{\omega}{L}$, L is the minimum Euclidean distance in formula (1), and ω is the empirical constant. The larger ω is, the more numbers of virtual samples will be generated in the neighborhood of the point, and the better it'll be for clustering analysis of different sample points. The smaller ω is, the more virtual samples will be generated during the transition in time domain, and the more beneficial it'll be to the feature classification of sample points in different dimensions.

3.2 Mathematical Proof

Taking two-dimensional data as an example, it is assumed that there is a sample set, which is divided into k categories according to the sample features. The kernel of each category is represented by eigenvectort_k, and each sample feature is represented by vector$y(x, s)$. The distribution quality of the set can be represented by the sum of the Euclidean distances of all samples and the eigenvectors, and then the deviation expected value of the original sample set is:

$$\mathrm{E} = \frac{1}{2}\iint ||y(x,s) - t_k||^2 p(x,s) d_x d_s$$

In the formula, $p(x, s)$ is the probability sample, $||y(x, s) - t_k||^2$ represents the Euclidean distance between the samples and the kernel of each category. To simplify the formula above, the minimum deviation from the expected value is:

$$\mathrm{E} = \frac{1}{2n}\sum_{i=1}^{n} ||y(x_i, s) - t||^2$$

In the formula, n represents the number of samples. For the original data sample set, the formula shows that the smaller n is, the larger the deviation from the expected value will be.

$$E^{\wedge} = K * E_v + E$$

The expected deviation value of the virtual sample set is:

$$E_v = \frac{1}{2}\sum_{i=1}^{k} \iiint ||y(x_i + \tau, s) - t||^2 p(t_k|x)p(x)p(t_k) d_x d_{t_k} d_\tau$$

The formula above can be simplified as follows:

$$E_v = E + E_c$$
$$E^{\wedge} = (1 + K) * E + K * E_c$$
$$E_1^{\wedge} = E + \frac{K}{K+1} * E_c$$

According to this formula, the expected deviation of the expanded virtual sample set is equivalent to adding a regular term$\frac{K}{K+1} * E_c$, on the basis of the original sample set, which is consistent with the regularization method of pattern recognition given in the reference [5][6], thus proving the validity of the virtual sample.

4. Experiment and Verification

In order to verify the feasibility of the algorithm, this paper uses the data of radio monitoring signals for testing. Each sample is two-dimensional data of amplitude-frequency characteristics (i.e. frequency and corresponding strength), and there are 30 samples in total. Now, 10 signal samples are recorded as small sample original data (I). The two methods are used to generate the virtual data. The training data were recorded as follows: Gaussian virtual sample set (II), VSGTD virtual sample set (III), where (II), (III) class data according to 1:M generated virtual data, according to the sample expansion method given in the reference[5], set M = 5, a total of 60 samples (including the original sample).

The above three kinds of sample data are used to train the modulation parameter extraction model based on neural network given in the reference[7], and then examine the learned model with testing data. The output of the model mainly includes the basic attributes of signals, including the following dimensions: center frequency(IF), peak power(Pow), center frequency bandwidth(IFBW), periodic time interval(TI) and so on. In the experiment, the remaining 20 signal samples were used as testing data to compare the performance of each method. The flow chart is as follows:

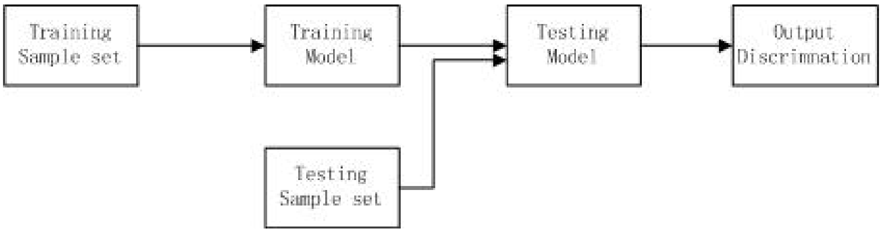

Fig. 2 Experiment Flow Chart

First, the relative error of each signal testing sample is calculated.

$$\sigma_i = \frac{X_t - X_s}{X_s}$$

Then the standard deviation is calculated according to the relative error of the previous measurements. In the formula N=20.

$$\sigma_f = \sqrt{\frac{1}{N}\sum_{1}^{N}\sigma_i^2 - (\frac{1}{N}\sum_{i=1}^{N}\sigma_i)^2}$$

Table 1. Results of different samples

Training Sample Type	Num. of sample set	standard deviation of IF	standard deviation of Pow	standard deviation of IFBW	standard deviation of TI
Original sample set	10	0.26	0.29	0.49	0.18
Gaussian virtual sample set	60	0.21	0.23	0.27	0.08
VSGGTD virtual sample set	60	0.2	0.16	0.18	0.07

According to the test results in Table 1.

(I) The number of original samples is too small, so the training results should go with an expansion of samples, after which the accuracy of feature extraction in different dimensions is improved.

(II) In the same model, the training result of VSGGTD virtual sample set is more accurate than that of the unimproved Gaussian virtual sample set. The difference is especially noticeable in the identification of IF bandwidth, for certain kinds of signals.

Experiments demonstrate that the virtual sample signal generated by VSGGTD algorithm can make a learning model with better reasonability, stronger generalization ability, and higher recognition accuracy.

5. Summary

At present, the relevant theories of artificial intelligence and machine learning are thriving. However, when the number of samples is small and the distribution is uneven, direct learning and training can easily lead to poor applicability of mathematical models and inaccurate feature extraction. The expansion algorithm for small sample data sets is one of the methods to solve this problem. The VSGGTD algorithm proposed in this paper provides a new thought for feature extraction of irregular signals and intercepted signals with low probability, which solves an important problem of machine learning in radio signal recognition. At the same time, this method has strong referential significance for the preprocessing of unbalanced data samples, covariance migration data and other low-quality data sets.

Acknowledgements

This work is supported by the National Science and Technology Major Project under Grants No. 2015ZX03002008.

References

[1]. WeiLu, JianpingCai, JuanLiu. Feature Extraction of Hyperspectral Image from Small Sample classification[J]. Journal of Institute of Surveying and Mapping. 2005 (2) Vol.22 Jun. 116-118.

[2]. HaisenZhong, Research on Face Feature Extraction from Small Sample Size[D]. Xinjiang University. Master's thesis.2014 (7).

[3]. ShaopingLu. Research on Gabor Transform For Face Feature [D]. Sichuan University. Master's thesis.2005 (7).

[4]. Xuyu. Research on training set construction method in pattern classification[D]. Harbin Engineering University. Master's thesis.2012 (7).

[5]. Li D C, Fang Y H. A non-linearly virtual sample generation technique using group discovery and parametric equations of hypersphere. Expert Systems with Applications, 2009, 36(1): 844-851.

[6]. Barron A R. Complexity regularization with application to artificial neural networks. In G. Roussas (Ed.), Nonparametric functional estimation and related topics, 1991, 561-576.

[7]. QinglongWang. Research on automatic modulation recognition of digital communication signals [D]. North China University of Technology. Master's thesis.2012 (7).