UDC 004.89

doi.org/10.15407/fmmit2023.36.022

Method for the automatic modulation classification based on linear regression and feature selection

Vasyl Semenov[1]

[1] Dr ph.-m. sc., chief of research and development department, Delta SPE LLC, V. Vasylkivska, 13, 01004, Kyiv; senior researcher, Kyiv Academic University, e-mail: vasyl.delta@gmail.com

The paper considers the task of automatic modulation classification, i.e. blind identification of modulation type of unknown signal before reconstructing its information content. This issue is especially important for the conditions of limited bandwidth of communication channels especially when two or more signals occupy the same frequency bandwidth. The proposed method uses linear logistic regression based on features calculated on the base of higher order cumulants of the received signal. The selection of informative features based on the absolute values of regression coefficients is proposed. The simulation results for the classification of composite BPSK/QPSK signals with various channel parameters and noise levels show the advantage of proposed approach with reduced set of features over the application of linear regression based on normal equation.

Keywords: automatic modulation classification, logistic regression, gradient descent, normal equation.

Introduction. Automatic modulation classification (AMC) is an important step in modern wireless communication systems. It is especially important for so-called non-cooperative communication [1, 2]. Robust recognition of modulation type is important to improve spectrum efficiency. In practical communication systems the radio signals are encoded by different modulation formats from a predefined set where modulation format is selected depending on system specifications and channel conditions, while being unknown to the receiver.

So, AMC is a preliminary stage providing signal type detection for the purpose of subsequent demodulation. From the perspective of machine learning (ML), AMC is a multi-class decision-making task. The most appropriate modulation of an incoming signal is determined by comparing its parameters with a learned ML model.

A lot of different modulation formats are currently used in practice [2]. In this paper we limit ourselves to the two most widespread types: BPSK (Binary Phase Shift Keying) and QPSK (Quaternary Phase Shift Keying).

In this paper we consider the BPSK/QPSK classification tasks by means of feature-based machine learning methods, namely linear regression and logistic linear regression.

1. Traditional approaches

Usually AMC methods are divided into two groups: likelihood-based (probabilistic)

approaches and machine learning frameworks in feature-based approaches. Although the likelihood-based approaches can reach the optimal classification accuracy, they require high computation complexity to estimate model parameters [1].

On the contrary, feature-based approaches are more practical due to their relatively easy implementation and low complexity [1-2]. Such algorithms for modulation recognition mainly include decision tree, the k-nearest neighbor, support vector machine and deep learning architectures [1-2].

2. Signal model

Suppose we have the sum of two signals as the observational signal [3]:

$$z(t) = z_1(t) + z_2(t) + w(t),$$

where $z_p(t),\ p = 1,2$ are the signals from two sources:

$$z_p(t) = a_p e^{j(\phi_p + \omega_p t)} \sum_{n=-\infty}^{\infty} s_p(n) g(t - nT - \tau_p),\ p = 1,2 \qquad (1)$$

and $s_p(n), p = 1,2$ are original sequences to be estimated; T is the symbol period; a_p are the signals' amplitudes; ϕ_p are the phases; τ_p are the time shifts; ω_p are the carrier frequencies; $g(t)$ is a total channel response (assumed to be raised square-root cosine with known roll-off); $w(t)$ is a white Gaussian noise.

The signals $s_p(n), p = 1,2$ are supposed to have the same modulation (either BPSK and BPSK or QPSK and QPSK). The task is to determine the modulation type on the base of received signal (1).

3. Feature extraction

Following the paper [2] in this investigation we use the higher order cumulants (HOC) as input features to the proposed classifier. Generally, HOCs are expressed as functions of the signals' high order moments. For a complex-valued signal z, the (p,q) moment is defined as:

$$M_{pq} = E\{z^{p-q}(z^*)^q\},$$

where z^* is a complex conjugate of z. The exact formulas for calculating M_{pq} can be found in [2].

For example, cumulants C_{42} and C_{63} are calculated as follows:

$$C_{42} = M_{42} - |M_{20}|^2 - 2M_{21}^2,$$

$$C_{63} = M_{63} - 9M_{42}M_{21} + 12M_{21}^3 - 3M_{43}M_{20} - 3M_{41}M_{22} + 18M_{20}M_{21}M_{22}.$$

The example of mutual placement and potential of these parameters for the classification task is presented on the figure below.

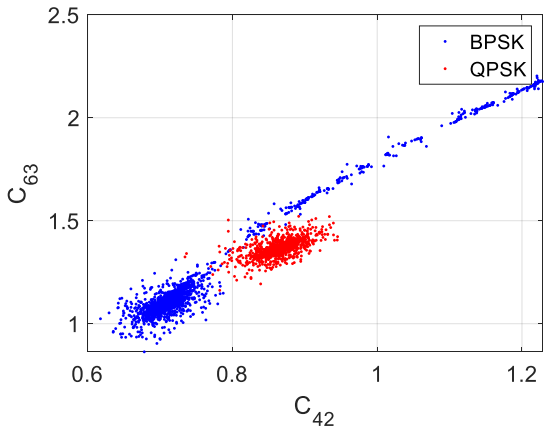

Fig. 1. The example of mutual placement of cumulants C_{42} and C_{63}

4. Application of logistic regression for classification.

As is known, supervised learning methods break down into two major categories: regression and classification. Classification is essentially taking an input and determining the class it belongs to out of the few classes available. Classification is different to regression in the sense that it only deals with a few discrete values while the latter has a continuous scale.

Suppose we have the set of features $X = X_1 \cup X_2$, $X_1 \cap X_2 = \varnothing$ and the task it to find mapping $y(x)$ to map each vector $\mathbf{x}_i \in X$ to one of the classes:

$y_i = 0$ if $\mathbf{x}_i \in X_1$ and $y_i = 1$ if $\mathbf{x}_i \in X_2$.

Logistic regression aims at the estimating coefficients $\mathbf{w}, b$ so that $g((\mathbf{w},\mathbf{x})+b)$ produces probability of belonging feature vector $\mathbf{x}$ either to class X_1 or X_2. Function $g(t)$ is usually called activation function and in the scope of this investigation is supposed to be sigmoid function:

$$g(t) = \frac{1}{1+e^{-t}}.$$

For example, $g((\mathbf{w},\mathbf{x})+b) = 0.8$ means that vector $\mathbf{x}$ belongs to class X_2 with probability 0.8. The coefficients $\mathbf{w}$, b can be estimated by the gradient descent method [5].

So, the essence of the proposed classification method is as folllows.

1. Having the received signal (1), calculate the feature vector $\mathbf{x}$ of first 9 cumulants.

2. Make classification based on the sigmoid function value: $g((\mathbf{w},\mathbf{x})+b)$.

The coefficients $\mathbf{w}$, b have to be preliminary calculated by the gradient descent method [5].

Another possible approach [2] is based on using simple linear regression to find the coefficients $\mathbf{w}$, b so that linear combinations $(\mathbf{w},\mathbf{x})+b$ approximate the binary values y_i in the least-squares sense. In this case the coefficients $\mathbf{w}$, b can be estimated by the gradient descent method or by solving normal equation as in work [2]. In this paper we will refer to such method as "normal equation" approach.

5. Feature selection.

Variable selection algorithms are especially used in situations, where there are many variables suspected to be informative and comparatively few examples are in a training set.

There are different feature selection approaches, e.g. based on Pearson correlation, chi-squared, recursive feature elimination, tree-based etc. However, in this work we use a simple scheme based on the absolute values of the regression coefficients. For example, if we got the vector of weight coefficients $\mathbf{w}=[\text{-}11.5 \quad 8.9 \quad \text{-}0.8]$, the conclusion is made that the first two features are highly informative, while the third is not so important for making classification decision (note that before the experiment all features are normalized to have a similar range).

6. Experimental results.

We verified the effectiveness of the proposed method on the base of 17500 signals, of which 8750 were used for training and 8750 for testing purposes. For each signal some random values of parameters a_p, ϕ_p, τ_p, ω_p ($p=1,2$) were considered. As for the noise $w(t)$, different signal-to-noise ratios (SNR) from 0 to 15 dB were modeled.

Using proposed feature selection approach, out of 9 first cumulants we selected the 3 most informative ones: features 1, 5, 7 (i.e. cumulants C_{20}, C_{42}, C_{61}).

The relative classification errors for the proposed method based on logistic regression and approach based on the solution of normal equation as in paper [2] are given in the table 1.

Table 1

Classification errors for the different learning approaches

	Logistic regression	Normal equation
Features 1,5,7	0.0288	0.0374
Features 1-9	0.0216	0.0279

From the table 1 it follows that the proposed method based on logistic regression provides expected advantage over the "normal equation" approach. Also, the classification based on three most informative features provides the result quite close to that of the full feature set (0.0288 vs 0.0216). It shows the potential for the practical

application of the proposed method. One of the directions for future research might include the inclusion of polynomial features.

Conclusions. In this paper we considered automatic modulation classification method to determine the modulation type (BPSK/QPSK) of unknown composite signal. The proposed method uses linear logistic regression based on features consisting of higher order cumulants. The selection of informative features based on the absolute values of regression coefficients was proposed. The simulation results for the classification of BPSK/QPSK signals have shown the advantage of proposed approach over the application of linear regression based on normal equation. Besides, the classification based on three most informative features provides the result quite close to that of the full feature set (0.0288 vs 0.0216). One of the directions for future research might include the inclusion of polynomial features.

References

[1] *Huynh-The Th. et al.* Automatic Modulation Classification: A Deep Architecture Survey. —IEEE Access, 2021. — 51, P. 142950–142971.
[2] *Abdelmutalab A., Assaleh Kh., El-Tarhuni M.* Automatic Modulation Classification Based on High Order Cumulants and Hierarchical Polynomial Classifiers. — Physical communication, 2016. — 21, P. 10–18.
[3] *Semenov V., Omelchenko P., Kruhlyk O.* Method for the detection of mixed QPSK signals based on the calculation of fourth order cumulants. — Signal and Image Processing, 2019. — 10. — P. 11-20.
[4] *Cramer, J. S.* The origins of logistic regression. — Tinbergen Institute, 2022. — P. 167–178.
[5] *Shalev-Shwartz Sh., Ben-David Sh.* Understanding machine learning. From theory to algorithms. — Cambridge University Press, 2014.

Метод автоматичної класифікації модуляції на основі лінійної регресії та вибору ознак

Василь Семенов

У статті розглядається задача автоматичної класифікації модуляції, тобто сліпого визначення типу модуляції невідомого сигналу перед відновленням його інформаційного контенту. Це питання є особливо важливим в умовах обмеженої смуги пропускання каналів зв'язку, особливо коли два або більше сигналів займають однакову смугу частот. Запропонований метод використовує лінійну логістичну регресію на основі ознак, розрахованих через кумулянти вищих порядків для прийнятого сигналу. Запропоновано процедуру вибору інформативних ознак на основі абсолютних значень коефіцієнтів регресії. Результати моделювання для класифікації композитних сигналів BPSK/QPSK з різними параметрами каналу та рівнями шуму показують перевагу запропонованого підходу зі зменшеним набором ознак над застосуванням лінійної регресії на основі нормального рівняння.

Received 14.03.23