The Features of the Statistic Counting Model for Its Application to Estimate the Probability

Baiyun Yang

Information security College, Yunnan Police College, Yunnan Kunming, China, 650223

18459423@qq.com

Keywords: Statistic counting model; Description length; Probability estimation

Abstract. The statistic method is widely used to analyze the statistic features hiding behind the various data. Some famous algorithms such as Ls, LMS are based on the statistic models. Actually on the basis of the description length theory, the statistic model can be considered as the judgment system which should hold different description complexity. It implies that the corresponding features of the statistic model under description length theory are different from these characters under the numerical area. In this paper, three basic features of statistic model are discussed under the minimum description length criterion. They are description complexity, similarity measure and the adding criterion. The corresponding derivations and the discussions are also given. These features can help increasing the efficiency of the estimation of the probability distributions.

Introduction

The statistic method is widely used to analyze the statistic features hiding behind the various data. Some famous algorithms such as Ls, LMS are based on the statistic models. In most of times, the statistic models are obtained by counting the number of data respectively which are corresponding to the different values of states that one thing holds. Then the resulted counting vectors are used to estimate the probability distributions. In this case, these probability distributions can be used to analyze or discuss some events. For math, the probability distribution implies the possibility that one event happen. Namely, one of utilization of the probability distribution is to predict the possibility whether one event expected can occur. The accuracy of this possibility is the criterion to testify the performance of the corresponding probability distribution. Actually, for each probability distribution, its entropy is used to describe the average information content that the corresponding events happened. If the value of the entropy is large, the accuracy will be lower, i, e, the difficulty of the prediction will be increased. Thus, the entropy implies the complexity of prediction. However, the entropy will be obtained difficultly in practice. The true probability distribution is not known, oppositely, the distribution is estimated which relies on the sufficient scale of data. Actually, during the counting procedure, the calculated real entropy is changing along the increment of the number of data. Namely, the complexity of this counting vector is also changed. Then the problem is that how to obtain the current possibility and to achieve the prediction with high accuracy. The predicting process is a system to describe some statistic features of the probability distribution or counting vectors. Based on the description length, the real-time complexity can be represented by using its corresponding description length which is equivalent to the total information contents contained in the current counting vectors or events. The value of the description length is smaller, the predicting accuracy will be higher. Therefore, the description length can be used to testify the performance of one probability distribution which will be estimated by one counting vector. It is worth to notice that the description length is the parameter of counting vector but not for the probability distribution.

However, in practice, the description length is not easy to use it directly. In this paper, we discuss some special derivations of the description length which aims to simplify the analysis procedure based on the statistic model.

Basic Theory

Considering a counting vector $\mathbf{v}$, it is consisted of some counting numbers as follows:

$$\mathbf{V} = \begin{pmatrix} a_1, & a_2,\ldots,a_m \\ n_1, & n_2,\ldots,n_m \end{pmatrix} \tag{1}$$

Where a_i denote the possible value of one event and n_i denote the corresponding number of counted data with event value a_i. Let N denote the total number of data the counting vector contain. Then the description length of this counting vector can be calculated by (2)

$$L = \sum_{w=0}^{N} \log(w+m) - \sum_{j=0}^{m-1}\sum_{i=0}^{n_i} \log(i+1) \tag{2}$$

It is difficult to resolve this equation directly. In its calculation, the factor operation will lead to high computational complexity. When Strilin fomula (3)

$$n! = n^{(n+\frac{1}{2})} e^{-n} \sqrt{2\pi} \tag{3}$$

Is suggested to simplify the calculation, the representation of (2) can hold the form (4)

$$L = N\log N - \sum_{i=1}^{m} n_i \log n_i + \delta \tag{4}$$

Where δ is the no-linear parameter which is included during the derivations. But it have some mathematical meaning which will be discussed in the later section in this paper.

Actually, when the fomula (4) is obtained, some inferences can be achieved. We will discuss them respectively.

Inferences

In practice applications, the description length can perform little efficiency. But some inferences based it is more serviceable. Considering two probability distributions P_1 and P_2 are combined with each other to analyze some problems. For this application, the probability distribution P resulted is obtained by (5)

$$P = P_1 \times P_2 \tag{5}$$

But for counting vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ which these two probability distributions are corresponding to, one easy way to estimate the average probability distribution of P is to merge these two counting vectors. Let N and M denote the total counting number of data each counting vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ obtained. Let n_i and s_i denote the counting number of events holding the event value i of two counting vectors respectively. Then after merging, a new counting vector $\mathbf{v}$ is obtained. The key point we pay more attention to is its corresponding description length. Let L_1 and L_2 denote the description lengths of counting vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ and L denote the description length of $\mathbf{v}$. New parameter can be defined.

Definition 1: The increment of the description length ΔL can be calculated by:

$$\Delta L = L - (L_1 + L_2) \tag{6}$$

It is obviously that the value of ΔL can be negative. We will give its some properties later. Now, we give the representation of the relative entropy which is used to describe the similar measure between two probability distributions. Let $D(P\|Q)$ denote the relative entropy between probability distributions P and Q, it is calculated by (7)

$$D(P\|Q)=\sum_{i=0}^{n}p_i\log\frac{p_i}{q_i} \tag{7}$$

When representations (4) and (6) are used to derive, the increment of the description length can be transformed as the form of (8) approximatively and its derivation is given in Appendix.

$$\Delta L\approx N*D(P_1\|P_2)+M*D(P_2\|P_1) \tag{8}$$

As we know, the relative entropy $D(P\|Q)$ does not satisfy the symmetry, i.e. $D(P\|Q)\neq D(Q\|P)$. But the increment of the description length ΔL is equivalent to the weighting of two relative entropy. Obviously, it is symmetric, which implies that the increment of the description length can be used to testify the similarity between two probability distributions.

However, it is not the easiest way to represent the similarity. We give the second definition about amazing measure.

Definition 2: The amazing measure is used to describe the amazing performance of one probability distribution estimated by its counting vector. The increment of the amazing measure is used to describe the similarity between two probability distributions estimated by using their corresponding counting vectors.

The amazing measure Γ of one probability distribution estimated by counting vector v_1 can be calculated by (9)

$$\Gamma=\log\frac{\prod_{i=1}^{m}n_i}{\sum_{i=1}^{m}n_i} \tag{9}$$

And the increment of amazing measure $\Delta\Gamma$ can be calculated by:

$$\Delta\Gamma=\Gamma'-(\Gamma_1+\Gamma_2) \tag{10}$$

Where Γ' denotes the amazing measure of the counting vector $\mathbf{v}$ obtained by merging two counting vectors v_1 and v_2.

It is worth to notice that although the amazing measure comes from the derivation by using the counting vectors. It is not only the property of the counting vector but also the feature of the corresponding probability distribution. It implies that the analysis about the probability distributions can be translated to analyze the property of corresponding counting vector by using amazing measure.

This is a meaningful inference about the description length. Meanwhile, the amazing measure is the part of δ of (4). Namely, the representation (4) holds the true form as (11)

$$L=N\log N-\sum_{i=1}^{m}n_i\log n_i+\frac{1}{2}\log\frac{\prod_{i=1}^{m}n_i}{\sum_{i=1}^{m}n_i} \tag{11}$$

Based on these inferences, the basic properties of the description length is discussed. The third property of the description length is more complex. It is referred to as the linear property. It is widely used to some linear analysis such as regression analysis based on statistic models of the economic applications. We firstly give the explanation of the statistic linear calculation

The linear calculation of the probability distributions is to achieve one new distribution by using the linear transform such as least square algorithm. Namely, by weighting. It can be represented as:

$$P = \sum_{i=1}^{n} w_i * P_i \tag{12}$$

One purpose of this operation is to optimize the corresponding weights w_i. However, it is difficult to optimize these weights directly. When the description length theory is used, the optimization operation will be simplified. We give the inferences as follows:

Inference1: The linear transform for the probability distributions can be translated to the linear transform of the description length.

Inference2: The optimization of weights for the linear transform of the probability distribution can be implemented by optimizing the weights for the weighting of the description length.

These two inferences can be simplified by (13)

$$opm\{w_i \mid \sum_{i=1}^{n} w_i * P_i\} \Leftrightarrow opm\{w_i \mid \sum_{i=1}^{n} w_i * L_i\} \tag{13}$$

Actually, it implies that the weighting for the probability distributions is equivalent to the weighting for their corresponding counting vectors. Thus, the description length of the resulted counting vector can be obtained by weighting description lengths of these counting vectors participated into linear transform respectively as shown (14)

$$L = \sum_{i=1}^{n} w_i * L_i \tag{14}$$

Meanwhile, $\forall wi, \exists w_i \to \min\{L\}$. It means that those optimization algorithms such as least square, searching methods can be used to implement the optimization for weights.

Conclusion

In this paper, three basic features of statistic model are discussed under the minimum description length criterion. They are description complexity, similarity measure and the adding criterion. The corresponding derivations and the discussions are also given. These features can help increasing the efficiency of the estimation of the probability distributions.

References

[1] Jianhua Chen, Yufeng Zhang, Xinling Shi, Image coding based on wavelet transform and uniform scalar dead zone quantizer [J], Signal Processing: Image Communication, vol.21, pp.562-572, 2006.

[2] X.Wu, P.A.Chou, X. Xue, Minimum conditional entropy context quantization [C], in: Proc. of IEEE Inter. Symposiumon Information Theory, Sorrento, Italy, p.43, June 2000.

[3] Marco Cagnazzoa, Marc Antonini, Michel Barlaud, Mutual information-based context quantization [J], Signal Processing Image Communication, vol.25, pp. 64－74, 2010.

[4] S.Forchhammer, X. Wu, J.D. Andersen, Optimal context quantization in lossless compression of image data sequences [J],IEEE Transactions on Image Processing 13(4), pp.509－517, Apr. 2004.

[5] S. Forchhammer, X. Wu, Context quantization by minimum adaptive code length [C], in: Proc. of IEEE Inter. Symposium on Information Theory, Nice, France, pp.246－250, June 2007.

[6] J. Rissanen, Stochastic complexity and modeling [J], The Annals of Statistics, Vol.14, No.3, pp.1080-1100, 1986.

Appendix

Derivation: Let $\Delta L_{m,k}$ denote the increment of the description length of two counting vectors $\mathbf{v_m}$ and $\mathbf{v_k}$. It’s derivation is given as follows:

$$\Delta L_{mk} = L_{mk} - (L_m + L_k)$$

$$= n_m \sum_{i=1}^{I} (\frac{n_{i,m}}{n_m}) \log[(\frac{n_{i,m}}{n_m})/(\frac{n_{i,mk}}{n_{mk}})] + n_k \sum_{i=1}^{I} (\frac{n_{i,k}}{n_k}) \log[(\frac{n_{i,k}}{n_k})/(\frac{n_{i,mk}}{n_{mk}})] - \frac{I-1}{2} \log \frac{n_m n_k}{n_m + n_k}$$

$$= n_m D(P(x \mid c_m) \parallel P(x \mid c_{mk})) + n_k D(P(x \mid c_k) \parallel P(x \mid c_{mk})) - \frac{I-1}{2} \log \frac{n_m n_k}{n_m + n_k}$$