A new Similarity Measure for the Context Quantization based on the Statistic Counting Model

Fuyan Wang
Information and Technology College
Kunming University
Kunming, China
e-mail: 4626747@qq.com

Yiping Zhang
Department of Technology and Science
Yunnan Police College
Kunming, China
e-mail: zhangyiping_001@163.com

Min Chen *
Information Security College
Yunnan Police College
Kunming, China
e-mail: minkeychen@sina.cn
* Corresponding Author

Qin Zhao
Information and Technology College
Kunming University
Kunming, China
e-mail: painkiller5230@qq.com

Abstract—In this paper, one new similarity measure which holds better mathematical description is given and discussed in details. The increment of the amazing measure, which denotes the similarity measure between two count vectors is discussed in this paper and its corresponding properties are also explained. We also give the analysis and the proof to explain the efficiency of the proposed similarity measure. The experimental results indicate that when using the proposed similarity measure , the corresponding results for different applications can be optimized.

Keywords-Similarity measure; Context modeling; Amazing measure; description length

I. INTRODUCTION

In a lot of application areas, the statistic model play a important role for estimating the corresponding probability distributions. These probability distributions are then used to analysis the probabilities which some things happen. In some application case, the clustering of the probability distributions is one of efficient method for statistic analysis. However, it is different from data clustering, the merging operation for probability distributions is similar to vector clustering, which implies that the traditional similarity measure, such as Euclidean Distance, is not suitable for vector clustering. Although the the relative entropy [1] which is referred to as K-L distance between two probability distributions can be used to measure the similarity between these two distributions. However, K-L distance is asymmetric, which implies that it is not suit for becoming the similarity measure for clustering. On the other hand, the complexity of one probability distribution is significant for clustering operation. In [2,3], Rissanen proposed the description length to describe the statistic complexity of one probability distribution which is estimated by using its corresponding count vector. In [4,5], Wu gave the proof that the description length is equivalent to the adaptive code length when this distribution is used to drive the arithmetic encoder. However, the description length holds high computation complexity. In [6], we proposed the rapid calculation method for the description length to extend the application of the description length. In [7], the increment of the description length is proposed to the similarity measure and it is used for the image wavelet compression in [8]. Although the increment of the description length perform better estimation efficiency than K-L distance. It also satisfy all features the similarity measure should have.

Actually, in the derivation of the increment of the description length, a novel item which is included in the representation of the increment of the description length attracts our focus. In this paper, it is referred to as amazing measure. It implies that the probability distribution which comes from the statistic estimation is better of not. Meanwhile, we find that amazing measure has the feature that this distribution is away from the uniform distribution. It implies that the amazing measure is one distance measure for the probability distribution which comes from the counting estimation. By them, the increment of the amazing measure should be the similarity measure for two distributions. In this paper, we will discuss this problem in detail. A such derivations will be given in our discussion to indicate that those features which the amazing measure holds are suit for clustering. Meanwhile, in [10,11], the previous works similar with our work are extended. Especially, in [12], the similarity measure is suggested for the compression of the genome sequence, which implies that the context quantization is significant to the biological data compression.

The structure of this paper is sketched as follows: In section 2, the description length is introduced firstly and the increment of the description length is discussed. In

section 3, the amazing measure is proposed and the increment of the amazing measure is discussed in detail. In section 4, an easy application of the increment of the amazing measure in the course estimation is given. The conclusion is given in section 5.

II. DESCRIPTION LENGTH

For a event collection, the scale of its possible solution space is equal to I. For some numerical applications, it implies that the source is $I-ary$. When statistic counting method is proposed to estimate the trend of events, the corresponding counting vector $\mathbf{v}$ can be obtained as (1) given.

$$\mathbf{V}=\{n_0, n_1, \ldots, n_{I-1}\} \tag{1}$$

This counting vector will be used to estimate the probability distribution. If $n_i >> n_j,\ j\neq i$, the resulted distribution will be very singular. Then the possibility of one event will become more clearly. However, in practice, this situation is not usually present. More is the distribution represents different form, which is difficult to lead to well estimation. In order to tackle this problem, the description length is proposed. For the counting vector $\mathbf{v}$, its corresponding description length L is calculated by (2)

$$L=\log(N-1)!-\sum_{i=0}^{I-1}\log n_i!-\log(I-1)! \tag{2}$$

Where N denotes the sum of all n_i. The description length actually means the description complexity of the counting model, which also means the singular feature of one respective distribution. The value of L is smaller, the corresponding distribution estimated will be more singular. However, it is difficult to calculate the description length directly and for each counting model, its initial counting number is set to 1. In this case, the calculation will be more complex. Actually, if the Strling form is used, the calculation will be simplified. The resulted form is given in (3)

$$\begin{aligned} L = {} & N\log N-\sum_{i=0}^{I-1}n_i\log n_i-\frac{1}{2}\log\frac{N}{\prod_{i=0}^{I-1}n_i} \\ & -\log(I-1)!-\log 2\pi \end{aligned} \tag{3}$$

Based on this approximation, the description length can be calculated more easily. Meanwhile, its corresponding application can be extended.

In some practical application, some distribution should be merged into one to obtain the estimation more accurate. This merging operation relies on the similarity measure. On the sight of the description length, the accurate estimation comes from the fact that the description length can be shorten after merging. Let L_1 and L_2 denote the description length of two counting vectors $\mathbf{v_1}$ and $\mathbf{v_2}$. $L_{1,2}$ denotes the description length of the counting vector from merging $\mathbf{v_1}$ and $\mathbf{v_2}$. Then the increment of the description length ΔL can be represented by (4)

$$\Delta L=L_{1,2}-(L_1+L_2) \tag{4}$$

After derivation, ΔL can also be represented by (5)

$$\begin{aligned} \Delta L = {} & N_m\sum_{i=1}^{I}\left(\frac{n_i^{(m)}}{N_m}\right)\log\left[\left(\frac{n_i^{(m)}}{N_m}\right)/\left(\frac{n_i^{(mk)}}{N_{mk}}\right)\right]+ \\ & N_k\sum_{i=1}^{I}\left(\frac{n_i^{(k)}}{N_k}\right)\log\left[\left(\frac{n_i^{(k)}}{N_k}\right)/\left(\frac{n_i^{(mk)}}{N_{mk}}\right)\right] \\ & -\frac{I-1}{2}\log\frac{N_mN_k}{N_m+N_k} \\ = {} & N_mD\big(p(x\,|\,c_m)\,\|\,p(x\,|\,c_{mk})\big)+ \\ & N_kD\big(p(x\,|\,c_k)\,\|\,p(x\,|\,c_{mk})\big)- \\ & \frac{I-1}{2}\log\frac{N_mN_k}{N_m+N_k} \end{aligned} \tag{5}$$

It is clearly that the increment of the description length is equivalent to the similarity measure. The first item of the representation (5) is actually the weighting for the relative entropy and the second item is related to the number of data these two counting vectors obtained. Namely, for two counting vectors, their similarity measure is correspond to not only the relative entropy but also the number of data on the current time. In [7,8], the characters of the increment of the description length is discussed in details and it is used as the similarity measure for image wavelet coding. However, the increment of the description length holds not all features that the similarity measure should hold.

Actually, the second item of (5) can be considered as one measure for estimation the probability distribution are "mature" or not. Namely, there is a new measure which is correspond to the number of data obtained on the current time. In next section, we will discuss this and give the definition of the amazing measure.

III. AMAZING MEASURE

From (2), it can also be represented by (6)

$$\begin{aligned} L = {} & N\log N-\sum_{i=0}^{I-1}n_i\log n_i+\frac{1}{2}\log\frac{\prod_{i=0}^{I-1}n_i}{N} \\ & -\log(I-1)!-\log 2\pi \end{aligned} \tag{6}$$

It is obvious that if $n_i >> n_j,\ j\neq i$, the value of the third item is near to 0. In this case, the value of the description length L will be shorten. Oppositely, if $n_i \approx n_j,\ j\neq i$,the value of the third item will not be small and the value of L will also not be shorten. Namely, the third item in (6)is also a measure. We call it amazing measure in this paper. Strictly, for the counting vector (1), its amazing measure ζ is described by (7)

$$\zeta=\log\frac{\prod_{i=0}^{I-1}n_i}{\sum_{i=0}^{I-1}n_i} \tag{7}$$

Then, we will calculate the relative entropy between the distribution from counting vector and the uniform distribution from the vector with whole number 1. Let $D(1\| p)$ denote this relative entropy.

$$\begin{aligned} D(1\| p) &= \sum_{i=0}^{I-1}(\frac{1}{I}\log\frac{\frac{1}{I}}{\frac{n_i}{N}}) \\ &= \log\frac{1}{I} - \frac{I-1}{I}\log N + \\ &\quad \frac{1}{I}\log\frac{\prod_{i=0}^{I-1} n_i}{N} \end{aligned} \tag{8}$$

It is obviously that when the counting vector is determined, the total number of data in this vector is equal to N. Namely, N could be considered as a constant. Based on this discussion, amazing measure can be represented as (9)

$$\begin{aligned} \zeta &= \log\frac{\prod_{i=0}^{I-1} n_i}{N} = I*\log\frac{1}{I} - \\ &\quad (I-1)\log N - I*D(1\| p) \end{aligned} \tag{9}$$

It is a similarity measure for the counting vector between the uniform distribution. Meanwhile, different from the increment of the description length, the amazing measure holds a static reference, the uniform distribution. It implies that the amazing measure of the counting vector is actually the distance away from the uniform distribution. The the distance between two counting vectors can be described by Fig. 1.

From Fig. 1, it means that the distance between two counting vectors can be described by (10)

$$dis = \zeta_1 - \zeta_2 \tag{10}$$

However, in practice, the distance used is actually the difference between the counting vectors and the counting vector merged. Then this ideal can be represented by (11)

$$\Delta\zeta = (\zeta_1 + \zeta_2) - \zeta_{1,2} \tag{11}$$

Then the Fig. 1 is given here.

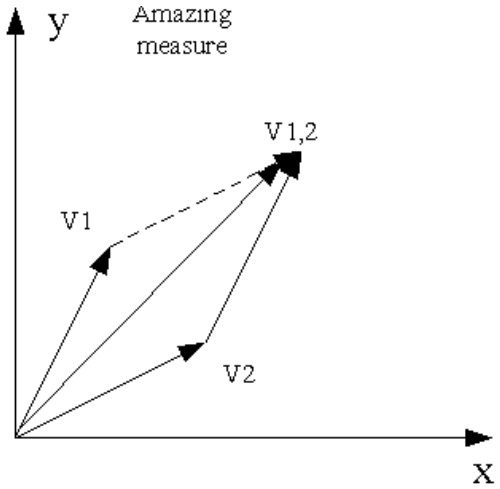

Figure 1. The space-image of the amazing measure

Where $\Delta\zeta$ denotes the increment of the amazing measure when two counting vectors are merged into one and $\zeta_{1,2}$ denotes the amazing measure this merged counting vector holds. The representation (11) is obviously that the value of $\Delta\zeta$ is always positive, which means that the sum of two amazing measure is larger than another one. This is a very important feature for a similarity measure. Based on this analysis, it is the proof that the amazing measure holds every feature one similarity measure should have. Therefore, the amazing measure is actually one similarity measure.

Meanwhile, from (11), if two counting vectors are same with each, namely, every number of data are same, the value of their respective amazing measure are equal. Furthermore, the value of $\zeta_{1,2}$ is equal to the value of $\zeta_1 + \zeta_2$. In this case, the value of the increment of the amazing measure is equal to 0, i.e. It is the smallest value. On the other hand, when two counting vectors are merged into one, if these two vectors are not similar, the resulted amazing measure should be changed from 0. Meanwhile, in practice, we hope that amazing measure can not be changed more. Therefore, the increment of the amazing measure can be considered as one parameter to describe the similarity between two counting vectors.

Consequently, when the merging operation happen, the increment of the amazing measure should be calculated firstly to judge weather these two counting vectors can be merged. In order to testify our new similarity measure, it is used to some designed experiments in next section.

IV. EXPERIMENTS AND RESULTS

In this section, we design some experiments to construct a system to test our proposed similarity measure. The description length is used as the testifying parameter. The theory of these experiments is the minimum description length, which implies that if the description length is small, the merging is considered as a suitable operation. But how to merge is determined by using the proposed increment of the amazing measure as the similarity measure.

In the first experiment, the proposed similarity measure is used to help the context quantization for the image context modeling compression. The context quantization

is significant to the compression efficiency. There are a lot of method to implement the context quantization. In order to use our similarity measure, the clustering algorithm, K-means, is used for implementing, which is proposed in [9]. For comparison, both the results obtained by using the increment of the description length and the results by using the proposed measure are also listed in table 1. Meanwhile, in this experiment, the three images (lena, barb and crowd, 512512, 8 bits for each pixel) are used to train the context model respectively. Then those counting vectors contained in the context model are clustered. The total description length of these three images are given in Table 1.

TABLE I. THE COMPARISON OF THE DESCRIPTION LENGTH OF THREE IMAGES

images	**Description length (bits)**	
	Increment of the description length	***Increment of amazing measure***
lena	1551892	1550324
barb	1598449	1596833
crowd	1534821	1533270

It is obviously that the results from the increment of amazing measure is better than the results from the increment of the description length. It also means that the increment of the amazing measure is more suitable than the increment of the description length to become the similarity measure.

Then in experiment 2, The proposed similarity measure is suggested to analysis the efficiency of the setting of the police officer training course. 29 count vectors for the course "Information security" from 29 areas are used as the test data. we testify the efficiency of the increment of description length. It easy to understand that if the clustering results are reasonable, the total description length of these count vectors should be shorter. They are listed in table 2.

TABLE II. THE COMPARISON OF DESCRIPTION LENGTH BASED ON TWO SIMILARITY MEASURE

Count vectors	**Description length (bit)**	
	Proposed measure	***Relative entropy***
Total these 29 count vectors	13,395,432	13,668,519

It is also implies that the amazing measure is better than the relative entropy to become the similarity measure. Above all, the proposed increment of the amazing measure is actually one efficient similarity measure than any other measures previous.

V. CONCLUSIONS

In this paper, one new similarity measure which holds better mathematical description is given. The increment of the amazing measure, which denotes the similarity measure two count vectors is discussed in this paper and its corresponding properties are also explained. The experimental results indicate that when using the proposed similarity measure , the corresponding results for different applications can be optimized.

ACKNOWLEDGMENT

This work is supported by the Foundation of Science of Yunnan under GRANT 2013FD042 and the Foundation of Science of Yunnan under GRANT 2014FD037.

Meanwhile, authors want to thank all reviewers for their constructing suggestions.

REFERENCES

[1] J. Rissanen, A universal data compression system, IEEE Trans. Inform. Theory, 1983, 29: 656−664.

[2] J. Rissanen and G. Langdon, "Universal modeling and coding," IEEE Trans. Inf. Theory, 1981, 27(1): 12-23.

[3] X. Wu, G. Zhai, Adaptive Sequential Prediction of Multidimensional Signals with Applicat -ions to Lossless Image Coding, IEEE Trans. Image Processing, 2011, 20(1):36-42.

[4] X. Wu, Lossless compression of continuous-tone images via context selection and quantization, IEEE Trans. on Image Proc, 1996, 6(5):656−664.

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

[6] Jianhua Chen, Context modeling based on Context quantization with Application in Wavelet Image Coding, IEEE Trans. Image Processing, 2004, 13(1):26-32.

[7] X. Wu. Context quantization with fisher discriminant for adaptive embedded wavelet image coding, Proc. of 1999 Data Compression Conference, pp. 102-111, Mar. 1999.

[8] Min Chen, Jianhua Chen, Context quantization based on the modified genetic algorithm with K-means, proceeding of 9th International Conference on Natural Computation, 424-428, Shengyang China, 2013.7.

[9] Min Chen, Jianhua Chen, Affinity propagation for the Context quantization, Advanced Materials Research, 2013, 791:1533-1536.

[10] B. Huang, Yuanyuan Wang, Jianhua Chen, ECG compression using the context modeling arithmetic coding with dynamic learning vector-scalar quantization, Biomedical Signal Processing and Control, 2013, 8: 59-65.

[11] Di Luo, Jianhua Chen, Qingqing Wang, Ran Hou, Context Weighting Based on the Shortest Code Length, Advanced Materials Research, 2014, 1030:1688-1691.

[12] Taysir H.A.Soliman, A lossless compression algorithm for DNA sequence, Int. J. Bioinformatics Research and Application, 2009, 5(6):593-602.