Context Quantization based on Minimum Description Length and Hierarchical Clustering

Hui Chen , Jianhua Chen

Department of Electronic Engineering, Dianchi College, Kunming, China

Abstract. The code length of a source can be reduced effectively by using conditional probability distributions in a context model. However, the larger the size of the context model, the more difficult the estimation of the conditional probability distributions in the model by using the counting statistics from the source symbols. In order to deal with this problem, a hierarchical clustering based context quantization algorithm is used to combine the conditional probability distributions in the context model to minimize the description length. The simulation results show that it is a good method for quantizing the context model. Meanwhile, the initial cluster centers and the number of classes do not need to be determined in advance any more. Thus, it can greatly simplify the quantizer design for the context quantization problem.

1 Introduction

Entropy coding based on the probability distribution of an information source is used in the lossless compression of the source. The set of probability distributions of the current symbol which are conditioned on the past observations is called the context model. Clearly, it is the model that determines the rate at which we can encode the symbol sequence. According to the information theory [1], we know that the conditional entropy is less than or equals to the unconditional entropy and the more conditions may result in the lower conditional entropy, which is described by

$$\mathrm{H}(Z \mid Z_1 Z_2) \le \mathrm{H}(Z \mid Z_1) \le \mathrm{H}(Z) \qquad (1)$$

The reduction of the entropy of a source, which is the lower bound of the average code length, will increase the possibility of the compression of the source. However, too large a modeling context spreads the counting statistics too thin among all possible modeling samples to achieve a good conditional probability estimate. This phenomenon is commonly called "context dilution" [2]. Context quantization is an approach to tackle this problem, where the total number of conditional probability distributions is reduced by combine similar distributions with some kind of clustering algorithm. In [3], the authors designed a vector quantization like context quantization algorithm which is called the minimum conditional entropy context quantization （MCECQ） algorithm and the conditional probability distributions are merged with K-means clustering algorithm. A similar method is proposed in [4] to which the optimization objective is the maximum mutual information (MMI) between the current symbol and the contexts. In these algorithms, K-means clustering is used to implement the context quantization. However, the number of classes and initial cluster centers should be set in advance and the algorithm is easy to be trapped into local optima. In [5], Forchhammer et. proposed the minimum adaptive code length context quantization (MCLCQ) algorithm which is implemented with the dynamic programming algorithm. The most important advantage of this algorithm is that it does not need the predetermined number of classes and the initial cluster centers. Meanwhile, the optimality is ensured by using the dynamic programming. Another way to implement this context quantization technique is to use the shortest path algorithm [6]. However, these two context quantization methods can only be applied to binary source coding.

In the light of the above observation, we want to find a clustering algorithm that does not need to know the optimal number of classes and the initial cluster centers in advance. In fact, the clustering algorithm is divided into partitional clustering algorithms and hierarchical clustering algorithms. K-means algorithm is the most well-known partitional clustering algorithm, where the K initial cluster centers can be determined by using the randomly selected K objects. An iterative procedure is used to assign the objects to their most appropriate classes according to a distance measure between an object and its cluster centers until a given criterion is met. The computational complexity of K-means is moderate and generally, the similarity among the classes is low.

The hierarchical clustering algorithms can be divided into agglomerative hierarchical clustering and split hierarchical clustering algorithms [7]. The most commonly used are the agglomerative hierarchical

clustering algorithms. In this kind of algorithms, each object is treated as a class at first. And then, they are merged one by one, according to some distance measure, until all objects are merged into one class or the algorithm can be stopped by a given termination condition. The agglomerative hierarchical clustering is used in statistical classification with a large data sample set initially. In [8], it is used in the segmentation of image pixels and in [9] it is applied to the fingerprint recognition.

In this paper, a hierarchical clustering based context quantization algorithm is proposed. The optimal number of classes and the initial cluster centers do not need to be determined in advance and good context quantization results can still be obtained.

2 Hierarchical Clustering

It is known that the agglomerative hierarchical clustering algorithm is a popular method that is mostly used in the hierarchical clustering. The various similarity measures between the classes can be applied in practice, the four mostly used similarity measures are as follows:

(1)The minimum distance

$$d_{\min}(c_i, c_j) = \min_{p \in c_i p' \in c_j} | p - p' | \quad (2)$$

(2)The maximum distance

$$d_{\max}(c_i, c_j) = \max_{p \in c_i p' \in c_j} | p - p' | \quad (3)$$

(3) The average distance

$$d_{mean} = | m_i - m_j | \quad (4)$$

(4) The average weighted distance

$$d_{mg}(c_i, c_j) = \frac{1}{n_i n_j} \sum_{p \in c_i} \sum_{p' \in c_j} | p - p' | \quad (5)$$

Where $| p - p' |$ is the distance between the classes p and p'; m_i is the mean of class c_i, m_j is the mean of class c_j, n_i is the number of objects in class c_i, n_j is the number of objects in class c_j.

With a properly defined distance measure between two classes, the agglomerative hierarchical clustering algorithm which minimizes the given distance measure at each level can be implemented as follows:

- Step1. Initialize each object as a class, calculate the distance between each pair of objects.
- Step2. Merge the two classes into a new class with the minimum distance.
- Step3. Recalculate the distances among the new class and all other classes.
- Step4. If the remaining number of classes is not equal to one, go back to Step 2 for the next iteration.

The most important advantage of the hierarchical clustering is that the number of classes and the corresponding clusters are directly obtained at each level of the tree structured clustering procedure. However, the computational complexity of hierarchical clustering is higher than that of partitional clustering.

A basic hierarchical clustering algorithm will eventually combine all objects into one class. However, if the optimal number of classes is reached, we need to terminate the hierarchical clustering procedure with a reasonable criterion. The description length introduced by Rissanen in [10] can be used as the criterion. Since it reflects not only the complexity of a statistical model, but also the average code length of a source sequence coded based on this model.

3 Description Length

For an I-ary source sequence $x_1, ..., x_t$, $x_t \in \{1,2,..., I\}$, the symbols $x_{t-1}, x_{t-2}, \cdots, x_{t-k}$ $(t > k)$ (with $K = I^k$ possible combinations) observed before the time t are referred to as the context of x_t. Let c_n, $n \in \{1,2,..., K\}$ denote a context event of x_t. It is actually a specific combination of the context symbols $x_{t-1}, x_{t-2}, \cdots, x_{t-k}$. The conditional probability $P(x_t = i \mid c_n)$ can be estimated by

$$P(x_t = i \mid c_n) = \frac{N_i^t + 1}{N^t + I} \qquad N^t = \sum_{i=1}^{I} N_i^t \quad (6)$$

Where N_i^t denotes the number of those symbols, in the sequence $x_1, ..., x_t$, with context event c_n and taking on value i. Apparently, N^t is the sum of the counts in the vector $\{N_1^t, ..., N_I^t\}$ on which the estimate of the conditional probability distribution $P(X_t \mid c_n)$ is based.

According to [5], let L_n denote the description length of $P(X_t \mid c_n)$ which is estimated from the vector of counts $\{N_1^t, ..., N_I^t\}$. Then, L_n can be calculated by the following formula:

$$L_n = \sum_{s=1}^{N^t} \log(s + I - 1) - \sum_{i=1}^{I} \sum_{j=0}^{N_i^t - 1} \log(j + 1) \quad (7)$$

In fact, (7) can be calculated in the form of the factorial operation

$$L_n = \log \frac{(N^t + I - 1)!}{(I-1)! N_1^t! N_2^t! \ldots N_I^t!} \quad (8)$$

However, the computation of (8) is inefficient due to the high cost of the factorial calculation. Instead, the Striling formula (9) is used in our work to calculate the factorials in (8) such that the computational complexity can be alleviated.

$$\ln(x!) \approx (x+0.5)\ln x - x(2x\sin\frac{1}{x})^2 + \ln\sqrt{2\pi} \quad (9)$$

The total description length of the sequence $x_1,...,x_t$ can then be calculated by adding together the description lengths for all possible context events: $L = \sum_{n=1}^{K} L_n$. From [10], the cost for transmitting the specific subsequence in the source sequence $x_1,...,x_t$, in which all symbols have the same context event c_n can be represented equivalently by

$$L_n = N_t H(X \mid c_n) + \frac{I\text{-}1}{2}\log N_t \quad (10)$$

where the conditional entropy $H(X \mid c_n)$ is the average code length for coding each symbol x_t in the subsequence based on the conditional probability distribution $P(X_t \mid c_n)$.

From the above analysis, it is apparent that the cost for encoding a subsequence includes two parts, i.e., the cost for encoding the symbols in the subsequence: $N_t H(X \mid c_n)$ and the cost for encoding the conditional probability distribution $P(X_t \mid c_n)$: $\frac{I\text{-}1}{2}\log N_t$.

4 Context Quantization based on the Minimum Description Length

As stated in section II, a suitable similarity measure between classes should be defined to enable the hierarchical clustering. Here, we propose a description length based similarity measure to fulfill this requirement.

If L_m and L_n are the description lengths of two conditional probability distributions $P(X_t \mid c_m)$ and $P(X_t \mid c_n)$, let L_{mn} denote the description length of the merged probability distributions estimated from the counts vector obtained by adding together the two counts vectors for estimating $P(X_t \mid c_m)$ and $P(X_t \mid c_n)$, an increment of the description length can be observed when these two conditional probability distributions are merged:

$$\Delta L_{mn} = L_{mn} - (L_m + L_n) \quad (11)$$

In fact, this increment reflects the difference of the description length before and after the merging of the two probability distributions. According to (10), this increment can be represented by

$$\begin{aligned} \Delta L_{mn} &= N_t D[P(X_t \mid c_n) \mid P(X_t \mid c_{mn})] \\ &+ M_t D[P(X_t \mid c_m) \mid P(X_t \mid c_{mn})] \\ &- \frac{I\text{-}1}{2}\log\frac{N_t M_t}{N_t + M_t} \end{aligned} \quad (12)$$

where $D[\bullet \mid \bullet]$ is the relative entropy between two probability distributions, M_t is the number of symbols in the sequence $x_1,...,x_t$ with the same context event c_m , $P(X_t \mid c_{mn})$ is the probability distribution estimated by merging the two vectors of counts on which $P(X_t \mid c_m)$ and $P(X_t \mid c_n)$ are estimated.

Some of the interesting properties of this increment can be summarized as follows: Firstly, it is symmetric since $\Delta L_{mn} = \Delta L_{nm}$. Secondly, it also reflects a kind of weighted average distance between $P(X_t \mid c_m)$ and $P(X_t \mid c_n)$ since $D[\bullet \mid \bullet]$ can be used as some kind of distance between two probability distributions.

These two characteristics of ΔL_{mn} means that it can be used as the similarity measure between two conditional probability distributions and can be applied in our hierarchical clustering algorithm to evaluate the similarity between two classes.

Thirdly, it may take a negative value. This feature means that in the context quantization procedure for a given source sequence, the optimal number of classes can be obtained by minimizing the total description length for the sequence since during the evaluation of the fitness of the merging of two probability distributions, a negative increment means the merging will reduce the total description length while a positive one will not.

In this work, the context quantization is actually implemented by merging the conditional probability distributions $P(X_t \mid c_n)$ (for all possible c_n) by the hierarchical clustering algorithm and every conditional probability distribution can be viewed as a class at the beginning of the clustering. Thus, we use the above mentioned increment ΔL_{mn} as the distance measure between two conditional probability distributions in the hierarchical clustering operations. During the clustering procedure, a probability distribution $P(X_t \mid c_n)$ is tested against all other probability distributions $P(X_t \mid c_m)$ for the possible merging and is finally merged with the one that can minimize ΔL_{mn}. In this way, the clustering operation will be stopped when no merging of two distributions can result in a negative ΔL_{mn} . The total description length of the remaining conditional probability distributions will reach the minimum at this time. These remaining classes actually characterize the final clustering results and the optimal number of classes is found accordingly.

The context quantization based on hierarchical clustering and the minimum description length is listed as follows：

- Step1. Initialize the conditions probability distributions by counting from the observed source sequence and treat each of them as an initial class.
- Step2. Calculate ΔL_{mn} for each pair of classes, if $\Delta L_{mn} < 0$, the two classes are viewed as a candidate pair of classes for merging at this level.
- Step3. Find out all candidate pairs, the one that minimize ΔL_{mn}, i.e., $\Delta L_{m'n'} = \min_{m,n} \{\Delta L_{mn}\}$ (corresponding to $\mathrm{P}(X_t \mid c_{m'})$ and $\mathrm{P}(X_t \mid c_{n'})$) is merged, the number of classes is reduced by one.

Step4. If no candidate pair is found, stop the algorithm and output the quantization result. Otherwise, go back to Step 2 for the next iteration.

5 Simulation

In our simulations, the original 256 by 256 gray scale images with 8 bits per pixel are quantized to images of the same size but with 3 bits per pixel (8 gray levels) and we quantize 7 gray scale images (8 bits per pixel) into 8 gray level per pixel images as the testing source sequence. Using the quantized images as the source sequences is to simplify the simulation. Among the 7 quantized images, two images (Girl and Barb) are used as the training sequences to estimate the 2-order conditional probabilities $p(x_i \mid x_{i-1}x_{i-2})$, where $x_{i-1}x_{i-2}$ have 64 possible combinations.

Context quantization is implemented using the proposed algorithm to merge these 64 conditional probability distributions. After context quantization, we can obtain the mapping scheme in which each conditional probability distribution is mapped into its corresponding cluster. We use this mapping scheme to help the coding of the test images. Three images (Lena, Woman and Baby) are used as the test sequences and the conditional probability distributions in the quantized context model are applied to drive an arithmetic encoder to implement the compression. The resulting total numbers of bits in all test images are listed in Table I. For comparison, the coding results based on context quantization implemented by the K-means clustering algorithm (MCECQ) are listed in Table II.

Table 1. The Results For The Test Images By Hierarchical Quantization (bits)

Number of classes	Lena	Woman	Baby
27	117376	170259	185982

Table 2. The Results By K-Means Quantization On Test Images (bits)

Number of classes	Lena	Woman	Baby
5	164601	201254	234216
10	143754	216534	227311
27	118928	173287	195431
30	153864	216543	217512
40	175489	223876	223754
45	176236	230874	237095
50	183765	236892	248541

From Table Ⅰ, The optimal number of classes of the context quantizer by the proposed algorithm is 27 based on the given training sequences. In Table Ⅱ, various numbers of classes are tested and the related context quantizers are designed accordingly by using the MCECQ algorithm. It can be found that the best coding results for all test images are also obtained when the number of classes is set to 27 for the same training sequences. The coding results for the three test images are almost the same as those in Table Ⅰ. However, due to the fact that initial cluster centers are randomly selected, the context quantizers are designed by multiple executions of the algorithm for each given number of classes and the one with the best coding results is chosen. For the proposed algorithm, only one execution is needed although its computational complexity is higher than that of the MCECQ algorithm. That means it takes a longer time for the MCECQ algorithm to design a good context quantizer than the proposed algorithm. Therefore, the good coding results indicate that the design objective is achieved by the proposed hierarchical clustering based context quantization algorithm.

6 Conclusion

Context quantization in high order entropy coding is applied more and more widely. With the employment of hierarchical clustering and the description length, the proposed context quantization algorithm successfully address the problems that the optimal number of classes and the initial cluster centers have to be given in advance, which are faced in the MCECQ algorithm. In addition, simulation results show that better coding results can be obtained with the context quantizer designed with the proposed algorithm.

Acknowledgment

This work is supported by a grant from the National Natural Science Foundation of China under Grant 61062005.

References

1. T. M. Cover, J. A. Thomas, Elements of Information Theory, Wiley, (2006)
2. M. J. Weinberger, J J. Rissanen, R. B. Arps, Applications of universal context modelling to lossless compression of gray-scale images, IEEE Transactions on Image Processing, 5 (4), (1996), 575-586.
3. J. Chen, Context modelling based on context quantization with application in wavelet image coding, IEEE Transactions on Image Processing, 13 (1), (2004), 26-32.

4. M. Cagnazzo, Mutual information-based context quantization, Signal Processing Image Communication, 25, (2010), 64-74.
5. S. Forchhammer, X.Wu, Optimal context quantization in lossless compression of image data sequences, IEEE Trans on Image Processing, 13(4) , (2004), 509-517.
6. S. Forchhammer, Context quantization by minimum adaptive code length, in: Proceedings of IEEE International Symposium on Information Theory, Nice, France, June (2007), pp.246-251.
7. A. K. Jain, R. C. Dubes, Algorithms for Clustering Data, Prentice Hall, (1998).
8. X. Hu, The image segmentation based on hierarchical clustering of LS-SVM, Computer and Digital Engineering, 38 (1), 2010, 143-147.
9. Anil Jain, Lin Hong and Ruud Bolle, On-Line Fingerprint Verification, IEEE Trans on Pattern Analysis and Machine Intelligence, 19 (4), 1997, 302-313.
10. J. Rissannen, Universal coding, information, prediction and estimation, IEEE Trans. Inform. Theory, 30, 1984, 629-636.