Markov 그라픽 데이타에 대한 Incremental-Runlength의 확률분포

論 文
29—6—1

Incremental-Runlength Distribution for Markov Graphic Data Source

金 在 均*
(Jae - Kyoon Kim)

Abstract

For Markov graphic source, it is well known that the conditional runlength coding for the runs of correct prediction is optimum for data compression. However, because of the simplicity in counting and the stronger concentration in distribution, the incremental run is possibly a better parameter for coding than the run itself for some cases.

It is shown that the incremental-runlength is also geometrically distributed as the runlength itself. The distribution is explicitly described with the basic parameters defined for a Markov model.

1. Introduction

The graphic data is a binary data generated by scanning the graphic images such as printed documents and drawings. The most realistic statistical model for the graphic data is the two-dimensional Markov model [1]. We assume that all the conditional probabilities of present data value given N surrounding data are known for all possible patterns of data, i.e., for 2^N states. Based on these conditional probabilities, we can predict the present data values, correctly most of the time. Although several prediction schemes are conceivable [2,3] given the Markov assumption, it is most natural to take the maximum-likelihood prediction. For the purpose of transmission of the data source information, it will be sufficient to send only the information that where the prediction is not correct, i.e., the position of prediction error. In fact, it is well known that the conditional runlength coding for the runs of correct predictions is optimum for Markov source [1].

We consider a fourth-order Markov model with sixteen states as shown in Fig. 1 and Table 1. The predicted value $\hat{x}_0$ is the maximum-likehood value for a measured statistics.

The conditional probability of the present value x_0 given the previous values is given by

$$p(x_0|x_1,\ x_2,\cdots)=p(x_0|x_1,x_2,x_3,x_4) \qquad (1)$$

where $\{x_i\}$, $i\geq1$, are the previous pitcure elements used for prediction of x_0.

While we are scanning or receiving the current scan line, we may generate a dual sequence of states and predictions, $\{s_j,e_j\}$, where $e_j=1$ or 0 depending on whether a prediction error occurred or not. From the dual sequence, we may count the separate runs of correct predictions for each states as shown in Fig. 2b [1].

Let $\{r_j\}$ be an arbitrary sequence of runs and let $\{r_{i,j}\}$ be the sequence of runs of state S_i. Let p_i be the probability of prediction error and $q_i=1-p_i$ for each state S_i, $i=0,1,\cdots\cdots,2^N-1$. Then from the Markov property, we get the following results [1,2].

(i) All the conditional runs are independent.

* 正會員：韓國科學院 敎授
接受日字：1980年 3月 5日

$$p_r(\{e_j|s_j\})=\prod_i\prod_j p(r_j|S_i)=\prod_i\prod_j p(r_{i},j) \qquad (2)$$

From this, we see easily that the interleaved runs $\{r_j\}$ can be treated separately for each state.

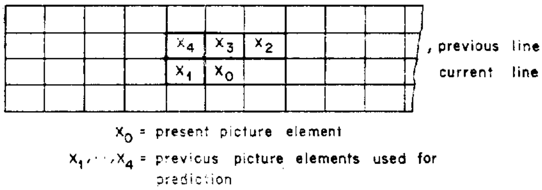

Fig. 1. A 4th order Markov model

Table 1. States and Predicted values

state	x_1	x_2	x_3	x_4	$\hat{x}_0$
S_0	w	w	w	w	w
S_1	b	w	w	w	b
S_2	w	b	w	w	w
S_3	b	b	w	w	b
S_4	w	w	b	w	w
S_5	b	w	b	w	b
S_6	w	b	b	w	b
S_7	b	b	b	w	b
S_8	w	w	w	b	w
S_9	b	w	w	b	w
S_{10}	w	b	w	b	w
S_{11}	b	b	w	b	b
S_{12}	w	w	b	b	w
S_{13}	b	w	b	b	b
S_{14}	w	b	b	b	w
S_{15}	b	b	b	b	b

b=black element　　w=white element

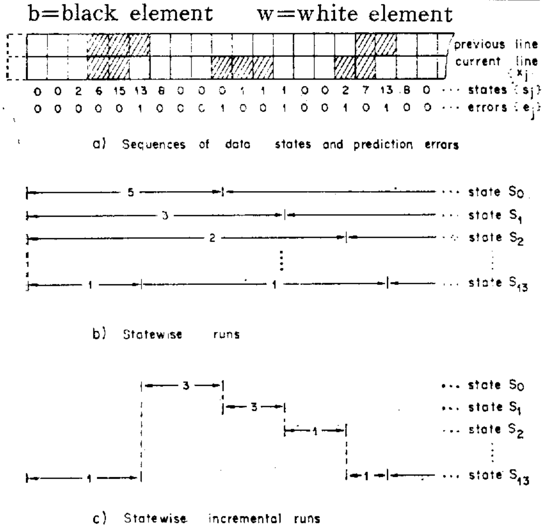

Fig. 2. Conditional runlength countings

(ii) For each state, the runs are geometrically distributed, and the average runlength is the inverse of probability of prediction error.

$$p(r_{i},j)=p_i\, q_i^{\,j-1}, \qquad j\geq 1 \qquad (3)$$

$$E(R|S_i)=p_i^{-1} \qquad \text{for all } i \qquad (4)$$

In fact, the equation (4) holds for any distrubution of empirical probabilities.

As we may see from Fig. 2b, it is rather complicated to count the statewise runs in practice. Instead, the incremental runs of Fig. 2C are much simpler to keep trace. The reference point of an incremental runs for a state is the end of the previous run of any state. We notice that the runlengths of the incremental runs are always shorter than or equal to those of the runs themselves. Therefore, the incremental runlength is possibly a better parameter for coding and transmission of the graphic data information[4].

We well find the incremental-runlength probabilities from the given Markov assumption and the probabilities of runs themselves in the following section.

2. Probability of Incremental Runs

By the definition of the incremental runs, the runs of state S_i are transformed to the incremental runs of shorter or equal runlengths of the same state S_i. The change of runlength occurs when some other states in between S_i's are in an error. It does not depend on what other state is in error. Hence, we may account the effect of other states by a complement state S_i^C for state S_i with the averaged probability of prediction error p_i^C.

$$p_i^C=\sum_{j\neq i}\left(\frac{p_j}{\sum_{j\neq i} p_j}\right)p_j \qquad (5)$$

where p_j is the probability of state S_j and p_j is the probability of prediction error in state S_j.

From the independency of conditional runs as shown in (2), we may write the probability of an arbitrary interleaved run including a S_i run of runlength j_0 as following.

$$p(r_{i},j_0)\prod_{1\neq i}\prod_j p(r_1,j)=p(r_i,j_0)\; p(r^C_{i},j) \qquad (6)$$

where r^C_{i},j denotes the run of j complement states, regardless of error, interleaved with (j_0+1)

S_i states. Within the complement state run $r^c_{i,j}$' only the last error state as shown in Table 2 decides the runlength of the incremental run by the definition. Depending on the position of the last complement state in error, a run of runlength j_0 is changed to an incremental run of runlength $1, 2, \cdots\cdots j_0-1$, or remained to be j_0 itself. For example, Table 2 explains how a S_i run of runlength $j_0=3$ is changed to the incremental runs of different runlengths.

We notice that the basic property of Markov model is still valid for the complement states. Let us define a probability $p_{\underline{i}}'$

$$p_{\underline{i}} \triangle p_E(S_{\underline{i}}^c) = \text{probability of prediction error for } S_{\underline{i}}^c \text{ in between two } S_i\text{'s} \qquad (7)$$

where $S_{\underline{i}}^c$ denotes the complement states of arbitrary length in between two adjacent S_i's. This probability will be discussed further later on. For simplicity, we may assume that $p_{\underline{i}}$ is independent of the positions of the two S_i's. Then, the probability of a run of state S_i can be represented as following:

$$\begin{aligned} p(r_{i,j_0}) &= p(r_{i,j_0}) \sum_{k=1}^{j_0} p(r^c_{i,k}) \\ &= p(r_{i,j_0}) \ \{(p_{\underline{i}} + q_{\underline{i}} p_{\underline{i}} + q^2_{\underline{i}} p_{\underline{i}} + \cdots \\ &\quad \cdots + q_{\underline{i}}^{j_0-2} p_{\underline{i}}) + q_{\underline{i}}^{j_0-1}\} \end{aligned} \qquad (8)$$

where

$$q_i = 1 - p_i,$$

and $\{r^c_{i,k}\}$ are the complement state runs sorted by the position of the last state in error as shown in Table 2.

The terms inside the parenthesis are for the incremental runs of length $1, 2, \cdots\cdots j_0-1$, and the last term $q_{\underline{i}}^{j_0-1}$ is for the incremental runlength j_0 itself.

Table 2

runs	possible run patterns	incremental runs
$r_{i,3}$	⊘ ○ ○ ⊘	
$r^c_{i,1}$	• • • ▲	$r'_{i,1}$
$r^c_{i,2}$	• • ▲ △	$r'_{i,2}$
$r^c_{i,3}$	• • △ △	$r'_{i,3}$

NOTE: ⊘, ▲ : States with predictions error

• : Dont care whether s_i^c is in error or not

Similarly, for the run of runlength (j_0+1),

$$p(r_{i,j_0+1}) = p(r_{i,j_0+1})\{(p_{\underline{i}} + q_{\underline{i}} p_{\underline{i}} + \cdots \cdots + q_{\underline{i}}^{j_0-2} p_{\underline{i}} + q_{\underline{i}}^{j_0-1} p_{\underline{i}}) + q_{\underline{i}}^{j_0}\} \qquad (9)$$

and so on.

From (8), (9), we notice that the probability of the incremental run of runlength j is constributed by all the probabilities $p(r_{i,1})$, $1 \geq j$. Combining all these terms, we obtain the probability of incremental runlength j.

$$\begin{aligned} p(r'_{ij}) &= p(r_{\underline{i}j}) \ q_i^{j-1} + p(r_{i,j+1}) q_{\underline{i}}^{j-1} p_{\underline{i}} \\ &\quad + p(r_{i,j+2}) \ q_{\underline{i}}^{j-1} p_{\underline{i}} + \cdots = p(r_{ij}) q_{\underline{i}}^{j-1} \\ &\quad + p_{\underline{i}} \ q_{\underline{i}}^{j-1}\{1 - \sum_{1=1}^{j} p(r_{i,1})\} \end{aligned} \qquad (10)$$

Using the property of geometrical distribution $p(r_{ij})$, this is simplified to another geometrical distribution as following.

$$p(r'_{ij}) = p_i' \ q_i'^{j-1} \qquad (11)$$

where

$$\begin{aligned} p_i' &= 1 - q_i' \\ q_i' &= q_i \ q_{\underline{i}} \leq q_i \\ q_{\underline{i}} &= 1 - p_{\underline{i}} \end{aligned} \qquad (12)$$

We notice here that as the runlength increases, the incremental-runlength distribution is decreasing more rapidly than the runlength distribution of (3). Hence the distribution is more concentrated to result in a smaller entropy.

Now we return to $p_{\underline{i}} = p_E(S_{\underline{i}}^c)$ of (7). Before considering $p_{\underline{i}}$ itself, we discuss first the probability to have a certain number of S_i^c's in between two adjacent S_i's. We notice that the N^{th} order Markov source $\{x_n\}$, the state sequence $\{S_n\}$ becomes a first order Markov. Hence, the conditional probability of a state sequence given the initial state of state S_i is

$$p(S_n, S_{n+1}, \cdots | S_i) = p(S_n | S_i) p(S_{n+1} | S_n) \cdots \qquad (13)$$

Given a state sequence of m S_i^c's, i.e., $S_{\underline{i}_m}^c$ in between two S_i's, the pro-bability of S_i^c in any error is

$$p_E(S_{\underline{i}}^c | m) = 1 - q_i^{c^m}, \ m \geq 0 \qquad (14)$$

where

$$q_i^c = 1 - p_i^c$$

and p_i^c is defined in (5).

Combining (13) and (14), we obtain $p_E(S_{\underline{i}}^c)$.

$$p_E(S_i^c) = \sum_{m=1}^{\infty} p(S_{\underline{i}_m}^c, \ S_i | S_i) \cdot p_E(S_{\underline{i}}^c | m)$$

$$=p(S_i^c|S_i)\ p(S_i|S_i^c)\sum_{m=1}^{\infty}p^{m-1}(S_i^c|S_i^c)\,(1-q_i^{c^m}) \quad (15)$$

Expanding further, we get the following result.

$$p_E(S_i^c)=\frac{p(S_i^c|S_i)}{1-q_i^c\cdot p(S_i^c|S_i^c)}\cdot p_i^c$$

$$=\frac{p(S_i^c|S_i)}{1+\left(\frac{q_i^c}{p_i^c}\right)p(S_i|S_i^c)} \quad (16)$$

where $p(S_i^c|S_i)$ and $p(S_i|S_i^c)$ can be obtained from the given p_i's and the patterns of states.

3. Conclusion

We have shown that for the Markov graphic data, the incremental-runlength is also geometrically distributed as the runs themselves. The distribution is explicitly represented by the parameters defined for the Markov model. As expected, it is less spreaded out than the runlength distribution.

ACKNOWLEDGMENT

The author wishes to express his gratitude to Prof. Hans G. Musmann of Hannover University and to the Alexander von Humboldt Foundation of West Germany for their kind support for this work.

References

1. D. Preuss; "Comparison of two-dimensional facsimile coding scheme," Record of the 1975 International Conference on Communications, pp. 7.12～7.16, June 1975.
2. T.S. Huang; "Coding of two-tone images," IEEE Trans. on Communications, Vol. Com-25, pp. 1406～1424, November 1977.
3. H. Kobayashi and L.R. Bahl; "Image data compression by predictive coding," IBM Journal of Res. and Dev., pp. 164～179, 1974.
4. J. Kim and P. Segin; "Conditional Incremental-Runlength code based on two-dimensional Markov model," To be published.