Application of Markov Process Model and Entropy Analysis in Data Classification and Information Retrieval

Udayan Ghose
Asst. Professor
University School of IT
GGS Indraprastha University
Delhi, India
g_udayan@lycos.com

C.S.Rai
Assoc. Professor
University School of IT
GGS Indraprastha University
Delhi, India

Yogesh Singh
Professor
University School of IT
GGS Indraprastha University
Delhi, India

***Abstract*— This paper proposes a statistical approach by a modified Markov chain process model and entropy function in the analysis of a large data set. The basic idea is that entropy and conditional entropy are used to measure the information content. In such analysis of large data sets including signal and image processing, unsupervised partitioning of data is required to build similar classes or clusters. The idea behind this is to identify each data item unambiguously as a member of particular class or cluster. The issue of partitioning is viewed as an information theoretic problem and it has been shown that the minimization of partitioning entropy may be used to evaluate the most probable set of data items. The data set considered for the simulation are the scanned OMR application forms of the candidates applying in various courses of a University. Classes are defined and inter dependence is measured on the basis of Markov process model and entropy analysis..**

Keywords- Markov Chain, Clusters, Entropy, Uncertainty, Mutual Information.

I. INTRODUCTION

Shanon [1] defined the entropy of a random variable, which takes on a finite set of arbitrary values by analogy with the physical Boltzmann entropy. If the random variable A takes on a finite number of values A_k (k = 1, 2, .., m) with the probabilities $p_k > 0$, $\sum_{k=1}^{m} p_k = 1$; then the Shanon's entropy of the random variable A is defined as

$$H(A) = -\sum_{k=1}^{m} p_k \log p_k \qquad (1)$$

Let us assume A_k form a complete ordered set / system of events with the associated probability being given by the ordered set $\{ p_1, p_2, \ldots, p_n\}$.
The probabilities satisfy the property:
$1 \geq p_i \geq 0, \ \forall i$

$$\sum_{i=1}^{n} p_i = 1$$

We propose the following quantity as a suitable measure of the uncertainty of the finite schema:

$$E = \prod_{i=1}^{n} p_i^{-p_i}$$

And we always take $p_i^{-pi} = 1$ for $p_i = 0$. Let's call the quantity E as the power entropy. The power entropy is related to the Shanon's entropy

$$H = -\sum_{k=1} p_i \log_a p_i$$

As
$E = a^H$, $a > 0$
Since exponential is a monotonic transformation, maxima of E occurs for the same configuration of probabilities as the configuration for H. That is the maxima of E occurs when all the events are equi-probable, thus Max E occurs when
$p_i = 1 / n, \quad (i \in \{1,n\}$.
For two mutually independent schemes A and B we have
E(AB) = E(A) E(B)
As H(AB) = H(A) + H(B)
$\Rightarrow E(AB) = a^{H(A)}a^{H(B)} = E(A)\,E(B)$ (2)
When the two schemes are not mutually independent, let q_{kl} be the probability that the event B_l of the scheme B occurs, given that the event A_k of the scheme A occured; so that the joint probability is
$r_{kl} = p_k q_{kl}$, $(1 \leq k \leq n; 1 \leq l \leq m)$
Then,

$$E(AB) = \prod_{i} \prod_{k} (r_{ik})^{r_{ik}}$$

$$= \prod_{i=1}^{n} \prod_{k=1}^{m} [\, p_i\, q_{ik}]^{-p_i\, q_{ik}}$$

$$= \prod_{i=1}^{n} \prod_{k=1}^{m} [pi^{-p_i}]^{q_{ik}} \prod_{i=1}^{n} \prod_{k=1}^{m} [q_{ik}^{-q_{ik}}]^{p_i}$$

$$= \prod_{i=1}^{n} [pi^{-p_i}] \sum_{k=1}^{m} q_{ik} \equiv 1 \prod_{i=1}^{n} [E_i(B)]^{p_i}$$

$$= E(A)\, E(B|A) \qquad (3)$$

Where we may regard E_i (B) as the conditional power entropy of the scheme B evaluated on the assumption that the event A_i

of the scheme A. E(B|A) we designate as the conditional power entropy of the scheme B evaluated under the assumption that an event of the scheme A occurs. Equation (3) may be derived also as

$E(AB) = a^{H(AB)} = a^{H(A)+H(B|A)}$)

$=E(A)\ E(B|A)$

It is also apparent that (2) reduces to (1) when the two schemas are mutually independent.

We also have,

$H(B|A) \leq H(B)$

$=> a^{H(B|A)} \leq a^{H(B)}$

$=> E(B|A) \leq E(B)$

Consistent with the view that the conditional entropy [3] is less than or equal to entropy of the scheme.

II. ENTROPY OF MARKOV CHAIN

A Markov chain can be regarded as an iterative process for getting the information for each data classes [4, 5]. This consists of a number of class states and transition probabilities. Information about the observations is exchanged from one state to another state during the iteration process Based on this information obtained, the observation for each data class are updated after each iteration and finally converges to a single output. Let the Markov chain has finite number of states A_1, A_2, …, A_n and with the transition probability matrix p_{ik} (I, k,= 1,2,…, n). We denote by P_k the probability of the sate A_k ($1 \leq k \leq n$), so that in particular

$$\sum_{k=1} P_k P_{kl} = P_l \qquad (l=1, 2, \ldots, n)$$

If the system is in state A_i, then its transitions to the different states A_k (k=1, 2, …, n) form a finite scheme

$$\begin{pmatrix} A_1 & A_2 & \cdots & A_n \\ p_{i1} & p_{i2} & \cdots & p_{in} \end{pmatrix}$$

The entropy of which is

$$E_i = \prod_{k=1}^{n} [p_{ik}]^{-p_{ik}} \qquad (4)$$

It depends on i and can be regarded as a measure of the amount of information obtained when the Markov chain moves one step ahead starting from the initial state A_i.

From here two directions are possible:

1. The mathematical expectation of E_i over all initial states is to be regarded as the measure of the average amount of information obtained when the given Markov chain moves one step ahead.

   $$E = \sum_{i=1}^{n} P_i E_i = \sum_{i=1}^{n} \prod_{k=1}^{n} P_i\ p_{ik}^{-p_{ik}}$$

2. Arbitrarily let us define

$E = \prod_{i=1}^{n} E_i^{P_i}$ as the entropy of the chain. E_i may be taken as the conditional power entropy of the chain evaluated on the assumption that the event A_i had already occurred. In a same a one step more may be considered as the product event space A^2 event. Then similar to conditional entropy E(B|A) we may define the one step entropy of the chain as:

$$E(A|A) = \prod_{i=1}^{n} E_i^{P_i} = \prod_{i} \prod_{k} p_{ik}^{-p_{ik} P_i} \qquad (5)$$

Similar to the one step entropy we may define the r step entropy as

$$E(r) = \prod_{i=1}^{n} [E_i^{(r)}]^{P_i} \qquad (6)$$

It seems reasonable to demand that

$E^{(r+s)} = E^{(r)}\ E^{(s)}$

Or $E^{(r)} = E^r$

Where E is the one step entropy of the chain.

The relation is trivially true for r = 1. Assume that it is true for some $r \geq 1$; we need to show that $E^{(r+1)} = E^{r+1}$

Let the system be in the state A_i; the finite scheme which describes the fate of the system in the next r+1 trials, can be then regarded as the product of two dependent schemes:

(A) The scheme corresponding to the immediately following trial with the entropy E_i and

(B) The scheme describing the fate of the system in the next r trials; the entropy of this scheme is E_k^r, if the outcome of scheme A was A_k. As

$E(AB) = E(A)\ E(B|A)$

We have,

$$E_i^{(r+1)} = E_i^{(1)} \prod_{k=1}^{n} [E_k^{(r)}]^{p_{ik}} \qquad (7)$$

Thus,

$$H^{(r+1)} = \prod_{i=1}^{n} [E_i^{(r+1)}]^{P_i}$$

$$= \prod_{i=1}^{n} (E_i^{P_i}) \prod_{i=1}^{n} \prod_{k=1}^{n} [E_k^{(r)}]^{P_i p_{ik}}$$

$$= E \prod_{k=1}^{n} [E_k^{(r)}] \sum_{i=1}^{n} P_i p_{ik}$$

But, $\sum_{i=1}^{n} P_i\ p_{ik} = P_k$

$$H^{(r+1)} = E \prod_{k=1}^{n} [E_k^{(r)}]^{P_k}$$

$$= E\ E^{(r)} = E\ E^r = E^{r+1} \qquad (8)$$

Equation (4) represents the entropy of the system after r+1 trials. For assigning appropriate weights based on self and conditional entropy let us consider a state transition from A_i to A_j with associated weight w_{ij}. If this transition is treated as an information flow the the state A_j gains information from state A_i. State A_j can in turn compute its conditional entropy $E(A_j|A_i)$ based on state A_i's observation. A greater weight should be assigned to this transition if the calculated

conditional entropy $E(A_j|A_i)$ is small. This implies that the weight (or transition probability) is inversely proportional to conditional entropy. The same argument applies to self-entropy. The larger the self-entropy, the smaller the corresponding weight. If $E(A_j|A_i)$ is smaller than $E(A_i|A_i)$, then w_{ij} is larger than w_{ii} . This relationship can be written as:

$w_{ij} = \eta / E(A_j|A_i)$ Where η is a constant and i, j = 1, 2, …, m.

Since,

$\sum_{j=1}^{m} w_{ij} = 1$, it follows that the weight is given by

$$w_{ij} = 1 / (E(A_j|A_i) \sum_{j=1}^{m} E(A_j|A_i)) \quad (9)$$

III. PROBABILITY ESTIMATION

Let the chain sized s be Ergodic, then

$P \{ | (m_i / s) – P_i | > \delta \} < Є$

Each possible result of the series of s consecutive trials of the given Markov chain can be written as the sequence A_{k1}, A_{k2}, …, A_{ks}

The probability of realizing the sequence *c* does not depend on the part of the chain where the series of trials begins (because of stationarity) and is obviously equal to

$p(c) = P_{k1} p_{k1\,k2} p_{k2\,k3} \cdots p_{ks-1\,ks}$

Let i and l be two arbitrary numbers from l – n; and let m_{il} be the number of pairs of the form $k_r k_{r+1}$ $(1 \leq r \leq s-1)$ in which $k_r = i$, $k_{r+1} = l$. Then,

$$p(c) = P_{k1} \prod_{i=1}^{n} \prod_{l=1}^{n} (p_{il})^{m_{il}} \quad (10)$$

Let's divide the sequence *c* in two groups as 1st group sequences satisfy the following properties

1. It is a possible outcome $p(c) > 0$
2. For any i, l $(i, l \in \{1, n\})$ the inequality $| m_{il} – sP_i P_{il} | < s \delta$
   Where δ is any small positive number and s is sufficiently large.

Thus for the members of this group

$m_{il} = sP_i P_{il} + s \delta \theta_{il}$, $| \theta_{il} | < 1$

Thus,

$$p(c) = P_{k1} \prod_{i} \prod_{l} (p_{il})^{sP_i P_{il} + s \delta \theta_{il}}$$

$$\Rightarrow p(c) = [\prod_{il} p_{il}^{-P_i P_{il}}]^{-s} P_{k1} \prod_{il} p_{il}^{s \delta \theta_{il}}$$

$$\Rightarrow p(c)\, E^{s} = P_{k1} \prod_{il} p_{il}^{s \delta \theta_{il}}$$

$$\Rightarrow [p(c)]^{1/s} E = (p_{k1})^{1/s} \prod_{il} p_{il}^{\delta \theta_{il}}$$

$$\Rightarrow 1 – [p(c)]^{1/s} E = 1 - (p_{k1})^{1/s} \prod_{il} p_{il}^{\delta \theta_{il}}$$

$$\Rightarrow 1 – [p(c)]^{1/s} E \leq \eta \quad (11)$$

for sufficiently large s and δ.

And similar to the original proof [2] we may prove the sum of the probability of all sequences (classes) that do not belong to the first group is less than Є.

IV. EXPERIMENTATION

The information content [7, 8]of the system consisting of the classes created from candidates data who are applying in a University. The various attributes of the data set are: *Application_no, Family_Income, Religion, 10th_Marks, 12th_marks, Gender, Pincode, Total* and *Rank* Classes have been created on the basis of the family income of the candidates. Based on the information the data set is divided in four broad classes viz:

C1: 0 > *Family_Income* ≥ 1 Lac

C2: 1 Lac > *Family_Income* ≥ 3 Lac

C3: 3 Lac > *Family_Income* ≥ 5 Lac

C4: *Family_Income* > 5 Lac

The data set considered are for two consequtive years 2008 and 2009. Based on the information Table 1 has been derived for probability distribution of various classes in consideration.

CLASSES / YEAR	2008	2009
C1	NO	NO
C2	YES	YES
C2	NO	NO
C3	YES	--
C3	YES	YES
C3	NO	YES
C3	NO	YES
C4	YES	YES
C4	--	NO

Table 1

The self and conditional entropy calculated for various classes are shown in Table 2.

	2008	2008	2009	2009
	$E(A_I\|A_I)$	$E(A_J\|A_I)$	$E(A_I\|A_I)$	$E(A_J\|A_I)$
C1	0.125	0	0.125	0
C2	0.25	1	0.25	1
C3	0.5	1	0.66	0
C4	0.125	0	0.5	1

Table 2

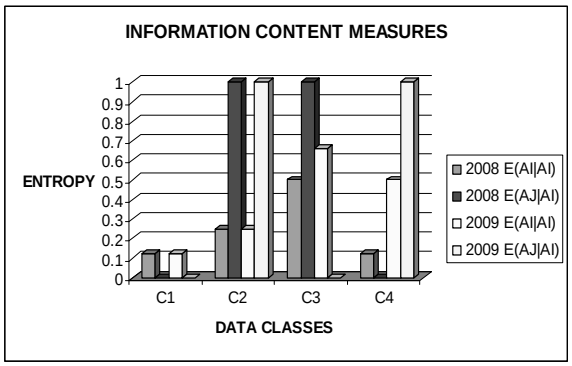

Figure 1

Figure 1 shows the information content of individual classes. It has been observed that for the class C1 conditional entropy is zero for both the years, where as for other classes it has different values.

V. CONCLUSION

The main goal of this work was to investigate the applicability of Markov process model and entropy concept in the analysis of information derived from large data sets. A theoretical approach is taken for the derivation of entropy in Markov chains and then the concept is used in the analysis of data. The experimental results illustrate the usefulness of entropy measures in the derivation of information content in the classes of similar data values.

REFERENCES

[1] C.E.Shanon, " The mathematical theory of communication", Bell System Technical Journal, vol 27, pp 379 – 423, 1948.

[2] J.N.Kapur, " On the concept of useful information", Jour. Org. Behav. Stat. 2 (3, 4), 147 – 162, 1985.

[3] Padhryac Smyth, " An information theoretic approach to rule induction from databases", IEEE Transactions on Knowledge and Data engineering, vol 4, no. 4, Aug 1992.

[4] A.C.S. Chung, H.C.Shen, " Entropy based Markov chains for multisensor fusion", Journal of Intelligent and Robotic Systems, 29: 161 – 189, 2000.

[5] Ram C. Tewari, D.P.Singh, Z.A.Khan, " Application of Markov chain and entropy analysis to lithologic succession – 1n example from early Permian Barakar formation, Bellampalli coalfield, Andhra pradesh, India", Journal of Earth System Science, 118, no. 5, 583 – 596, Oct 2009.

[6] Xilin Zhao, Jianzhong Zhou, Bo Fo, Hui Liu, " Application of entropy based Markov chains data fusion technique in fault diagnosis", IEEE International Conference on Computer Science and Software Engineering, 2008.

[7] Alfred Renyi, "On measures of Entropy and Information", Mathematical Institute, Hungarian Academy of Sciences.

[8] A Fienstein, " Foundations of Information Theory, New York, Mc Graw Hill, 1958.