Information Entropy Estimation Using Indirect Conditional Information

Nol Premasathian and Watcharee Tantikittipiisut

***Abstract*—Conditional entropy measures information flow from one discrete random to another. Conventional entropy is calculated from conditional or joint probabilities of two variables. When the joint probability is not known, the conditional entropy cannot be calculated. This paper presents an approach to estimate the conditional entropy using indirect conditional information. The result of the estimation is overestimated by less than ten percent of the actual value for normal cases.**

***Index Terms*—Entropy, conditional entropy, information flow, indirect conditional information.**

I. Introduction

Entropy is the statistical mechanics in measurement of uncertainly and is of practical uses in various fields such as thermodynamics, mathematics, information theory computer sciences, social sciences and economics [1]. In information theory, entropy is a measurement of amount of information [2], which is usually measured in bits, nats or bans [3]. The concept of entropy was first introduced by C. E. Shannon in a 1948 paper "Mathematical Theory of Communication [4]. Entropy of information is of an important role in information security. It I used in the determination of the unicity distance of a ciphertext, the prefect secrecy and its conditional value can be measured to see if there is an information flow or leakage from one variable to another [5]. This paper proposes an approach to estimate the value of entropy using indirect conditional information. This means that when the random variable associated with the entropy is unknown, and the direct conditional information is not available, we can estimate the value of the entropy from indirect conditional information using the proposed approach. This paper is organized in five parts as follows. The second part explains related theories, such as the use of the information entropy as well as the calculation of the information entropy and the conditional entropy. The third part gives details of the proposed approach, which estimates the information entropy when only indirect conditional is given. The fourth part is the results of some experiments using the proposed approach. The last part is the conclusions and future works followed by acknowledgment and the references.

Manuscript received August 25, 2012; revised September 26, 2012.

Nol Premasathian is with the Faculty of Information Technology, King Mongkut's Institute of Technology Ladkrabang, Bangkok, Thailand (e-mail: nol@it.kmitl.ac.th).

Watcharee Tantikittipisut is with the Bangkok, Thailand (e-mail: charlieante@gmail.com).

II. Related Theories

Entropy H(M) of a discrete random variable M, with possible values $\{m_1, m_2, \ldots, m_n\}$ in bits can be calculated as in the following formulae.

$$\mathrm{H}(M) = \sum_{i=1}^{n} p(mi) \frac{1}{\log_2 p(mi)} \quad (1)$$

Information entropy involves a number of theories. In information security, information entropy is used to determine the unicity distance and to measure the information flow between variables.

A. Unicity Distance

Unicity distance is a technical term in cryptography representing the size of a ciphertext that is required to break the ciphertext by lessening the number of possible spurious keys to zero [6], [7], [8]. This means that there should be only one key that can decrypt the ciphertext and yield a meaningful result. The unicity distance of a ciphertext depends upon the information entropy of the encryption key as well as the redundancy of the plaintext from which the ciphertext is constructed.

The unicity distance of a ciphrtext can be calculated using the following formulae.

$$Unicity\ Distance = \frac{Entropy\ of\ the\ Key}{Redundancy\ of\ the\ Plaintext} \quad (2)$$

When the plaintext has no redundancy at all, the unicity distance of its ciphertext becomes infinity regardless of the encryption key size. This achieves the perfect secrecy, a crypto system in which the ciphertext gives no information about the plaintext [9].

B. Information Flow Model

Controlling the flow of information is an important part of information security. The flow of information from a discrete random variable to another can be measured in an information flow model. One of the models can be constructed as an entropy based analysis [10] using the conditional entropy [11], [12]. For instance, given two random discrete random variables, X and Y, we can measure if there is any information flow from the variable X to the variable Y by calculating the entropy of Y and the entropy of Y given X. There is no information flow from the variable X to the variable Y if and only of the entropy of Y and the entropy of Y given X are equal. This means that getting any information about X gives no information about Y. The

difference between the entropy and its conditional entropy indicates the level of information flow or leakage from one variable to another. The conditional entropy of Y given X is usually written as $H(Y|X)$ and can be calculated as in the following formulae.

$$H(Y|X) = \sum_{x \in X,\ y \in Y} p(x,y) \cdot \log_2 \frac{p(x)}{p(x,y)} \quad (3)$$

The conditional entropy can be calculated when the joint probability of the two discrete random variables are known. In some situations, such as the one having three random variables, X, Y, and Z, with joint probabilities $p(x,y)$ and $p(y,z)$ are all known, it is not possible to find the information flow between X and Z as their joint probabilities are not given. This paper presents the estimation of the conditional probability in this situation as explained in the next section.

III. The Proposed Approach

As mentioned in the next section, our proposed approach aims to estimate the conditional entropy of a discrete random variable when the joint probabilities are absent. This can be achieved by measuring the reduction in the amount of information in one step and proportionally estimate the reduction of the amount of information in the next step. The approach can be applied to a problem of any number of random variables. For simplicity, we use an example of three random variables to demonstrate the approach.

Random variables: X, Y, and Z.

Known probabilities: $p(x)$, $p(y)$, $p(z)$, $p(x, y)$, $p(x, z)$

Find: $H(X|Z)$

The estimation of the conditional entropy $H(X|Z)$ is performed in the following steps.

1) Calculate the entropy $H(X)$ using $p(x)$.
2) Calculate the entropy $H(Y)$ using $p(y)$.
3) Calculate the entropy $H(X|Y)$ using $p(x)$ and $p(x, y)$.
4) Calculate the entropy $H(Y|Z)$ using $p(x)$ and $p(y, z)$.
5) Calculate the entropy remaining rate

$$R = \frac{H(Y|Z)}{H(Y)}$$

6) Calculate the entropy difference

$$D = H(X) - H(X|Y)$$

7) $H(X|Z)$ is estimated as $H(X) - (1\text{-}R).D$

The entropy remaining rate R is the remaining value of the entropy of a random variable Y, given the value of the random variable Z. When the initial entropy of Y is normalized to 1, by getting the information, the entropy of Y is reduced from 1 to 1-R. The entropy difference rate D is the reduced amount of information of X when Y is known. So when Y is only partially known, the entropy of X is partially reduced as described.

The proposed approach is tested against sets of inputs and the result is shown in the next section

IV. Result

The testing of the proposed approached is conducted by arbitrarily assign probabilities to three random variables X, Y, and Z and use the proposed algorithm to estimate the conditional entropy H(X|Z), pretending that the joint probability $p(x, z)$ is not known. In the first test, the probabilities of the three variables are shown in table I.

A. The First Test

TABLE I: Probabilities of Three Variables

Probabilities					
$x = 0$	0.5	$y = 0$	0.9	$z = 0$	0.9
				$z = 1$	0.1
		$y = 1$	0.1	$z = 0$	0.1
				$z = 1$	0.9
$x = 1$	0.5	$y = 0$	0.1	$z = 0$	0.9
				$z = 1$	0.1
		$y = 1$	0.9	$z = 0$	0.1
				$z = 1$	0.9

Using the proposed algorithm, we calculate the relevant values as the results shown below.

$H(X) = 1$
$H(Y) = 1$
$H(X|Y) = 0.468996$
$H(Y|Z) = 0.468996$
$D = 0.531004$
$R = 0.468996$
Estimated $H(X|Z) = 0.718034$

The estimation gives a reasonable entropy value as the actual entropy $H(X|Z)$, calculated when the joint probabilities $p(x, z)$ are used, is 0.680077, or about 5.58% overestimation.

B. The Second Test

The second test is conducted similar to the first one but with adjusted probabilities as in table II shown below.

TABLE II: Probabilities of Three Variables

Probabilities					
$x = 0$	0.4	$y = 0$	0.2	$z = 0$	0.4
				$z = 1$	0.6
		$y = 1$	0.8	$z = 0$	0.3
				$z = 1$	0.7
$x = 1$	0.6	$y = 0$	0.8	$z = 0$	0.7
				$z = 1$	0.3
		$y = 1$	0.2	$z = 0$	0.5
				$z = 1$	0.5

Using the proposed algorithm, we calculate the relevant values as the results shown below.

$H(X) = 0.970951$
$H(Y) = 0.989588$
$H(X|Y) = 0.703291$
$H(Y|Z) = 0.923412$
$D = 0.267659$
$R = 0.933129$
Estimated $H(X|Z) = 0.953052$

The estimation gives a reasonable entropy value as the actual entropy $H(X|Z)$, calculated when the joint probabilities $p(x, z)$ are used, is 0.889257, or about 7.17% overestimation.

From both tests, it can be seen that the algorithm can estimate the entropy value of a random variable using only indirect conditional information. The results are less than 10% over the actual value.

C. Testing with Varying Probabilities of Two Variables

When the probability of each value of X is fixed at 0.5,and the probability of each value of Y given X and Z given Y vary from 0.6-0.4 to 0.95-0.05, we found that the value of the entropy becomes more and more overestimated as in table III shown below.

TABLE III: OVERESTIMATION WITH VARYING PROBABILITIES

Probabilities (Y given X and Z given Y)	Estimated H(X|Z)	Actual H(X|Z)	Overestimation (percent)
0.6 – 0.4	0.999156	0.998846	0.031096
0.7 – 0.3	0.985908	0.981454	0.453842
0.8 – 0.2	0.922676	0.904381	2.022881
0.9 – 0.1	0.718034	0.680077	5.58132
0.95 – 0.05	0.490771	0.452943	8.351644

D. Problems with More Than Three Random Variables

This approach can be applied to a problem with more than three variables by calculating the reduction of the entropy proportionally. However, the accuracy of the estimation tend to decrease as the number of random variables grows.

V. CONCLUSIONS AND FUTURE WORKS

We have proposed an approach to estimate the entropy of a discrete random variable when no direct conditional information is given. The estimated result is close to the actual result that is calculated when the required joint probability value is given.

Our future work is to adjust the approach by introducing some co-efficient to certain value in some steps of the algorithm, as from our experiment, the algorithm seems to always overestimate the entropy value.

ACKNOWLEDGMENT

We would like to express our gratitude to the Faculty of Information Technology, King Mongkut's Institute of Technology Ladkrabang for its continual support.

REFERENCES

[1] A. Valero, L. Serra, and J. Uche, "Fundamentals of energy cost accounting and thermoeconomics. Part I: Theory," *Journal of Energy Resources Technology*, vol. 128, no. 1, pp. 1-8, 2006.
[2] S. Ihara, "Information theory for continuous systems," World Scientific Publishing, 1993.
[3] L. Brillouin, "Science and information theory," Dover Publications, 2004.
[4] C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal*, vol. 27, no.3, pp. 379-423, Jul/Oct 1948.
[5] G. Smith, "Quantifying information flow using min-entropy," *8th International Conference on Quantitative Evaluation of Systems*, pp. 159-167, Aachen, Germany, 2011.
[6] M. E. Hellman, "An extension of the shannon theory approach to cryptography," *IEEE Transaction on Information Theory*, vol. 23, no. 3, pp. 289-294, May 1977.
[7] P. Beauchemin and G. Brassard, "A generalization of hellman's extension to shannon's approach to cryptography," *Journal of Cryptology*, vol. 1, no. 2, pp. 129-132, 1988.
[8] C. E. Shannon, "Communication theory of secrecy systems," *Bell System Technical Journal*, vol. 28, no. 4, pp. 656-715, 1949.
[9] B. Schneier, "Applied cryptography," Wiley, 1996.
[10] M. Bishop, "Computer security," Addison-Wesley, 2003.
[11] G. A. Korn and T. M. Korn, "Mathematical handbook for scientists and engineers: Definitions, theorems, and formulas for reference and review," Dover Publications, 1967.
[12] C. Arndt, "Information measures: Information and its description in Science and Engineering," Springer, 2001.