A New Measure of Rule Importance Using Hellinger Divergence

Chang-Hwan Lee
Department of Information and Communications
Dongguk University
Seoul, Korea
Email: chlee@dgu.ac.kr

***Abstract*—Many rule induction algorithms generate a large number of rules in data mining problem, which makes it difficult for the user to analyze them. Thus, it is important to establish some numerical importance measure for rules, which can help users to sort the discovered rules. In this paper, we propose a new rule importance measure, called HD measure, using information theory. A number of properties of the new measure are analyzed.**

***Keywords*-Rule Importance; Association; Information Theory;**

I. Introduction

Determining the importance of rules is an important data mining problem since many data mining algorithms produce enormous amounts of rules, making it difficult for the user to analyze them manually.

The large number of rules generated by the algorithms commonly used makes it impossible for users to take all the rules into consideration. A common way of reducing the number of rules is to pre-filter the output of data mining algorithms according to importance measures. By selecting a subset of important rules out of a larger set of ones, users can focus on what should be of interest to them. Therefore, importance measures of rules play a major role within a data mining process.

Thus, it is important to establish some numerical importance measure for rules, which can help users to sort the discovered rules. However, the choice of a measure responding to a user's needs is not easy. Therefore there is no optimal measure, and a way of solving this problem is to try to find good compromises.

There are two kinds of rule importance measures: the subjective ones and the objective ones. Subjective measures take into account the user's domain knowledge [1] [2], whereas objective measures are calculated using only the data [3] [4] [5] [6]. We are interested in objective measures, and this article focuses on the objective aspect of rule importance.

Numerous measures are used for performance evaluation in machine learning and data mining. In classification learning, the most frequently used measure is classification accuracy while other measures include precision and recall in information retrieval. With new tasks being introduced in knowledge discovery, new measures need to be defined.

Among the objective measures of rule importance, the information-theoretic measures are important and useful since they are based on theoretical background. In addition, there is an interesting parallel to draw between the use of information theory to evaluate rules [7]. As for rule, the relation is interesting when the antecedent provides a great deal of information about the consequent. The information-theoretic measures commonly used to evaluate rule importance are the Shannon conditional entropy [8], the Theil uncertainty coefficient [5], the J-measure [7], and the Gini index [3].

The Shannon conditional entropy measures the average amount of information of the rule given that the condition is true [8]. The Theil uncertainty coefficient measures the entropy decrease rate of the consequent due to the antecedent [5]. The J-measure is the part of the average mutual information relative to the truth of the antecedent [7]. Finally, the Gini index is the quadratic entropy decrease [3]. Even if these measures are commonly used to evaluate association rules, they are all better suited to evaluate classification rules.

In this paper, we propose a new measure of rule importance, called HD measure, based on information theory. We employ Hellinger divergence as a tool for calculating the importance of rule. A number of properties of the new measure are analyzed. The proposed HD measure shows a number of important and necessary properties.

II. HD Measure

For the purposes of this paper, a rule is a knowledge representation of the form b $\rightarrow$ a. We restrict the right-hand expression to being a single value assignment expression while the left-hand side may be a conjunction of such expressions.

The basic idea of rule importance starts with the assumption that the value assignments in the left hand side of each rule affects the probability distribution of the right-hand side(target attribute). The target attribute forms its a priori probabilities without presence of any left-hand conditions. It normally represents the class frequencies of the target attribute. However, its probability distribution changes when it is measured under certain conditions usually given as value assignments of other attributes. Therefore, it is a natural definition, in this paper, that the significance of a

Copyright (c) IARIA, 2012. ISBN: 978-1-61208-242-4

rule is interpreted as the degree of dissimilarity between a priori probability distribution and a posteriori probability distribution of the target attribute. The critical part now is how to define or select a proper measure which can correctly measure the instantaneous information.

In this paper, we employ Hellinger divergence as the measure of instantaneous information. The Hellinger divergence was originally introduced by Beran [9], and, in this paper, we modified it in order to use it as the information content of rules. The original Hellinger divergence of variable A given the value of b is defined as

$$\left(\sum_i \left(\sqrt{p(a_i)} - \sqrt{p(a_i|b)}\right)^2\right)^{1/2} \tag{1}$$

where a_i denotes the value of variable A. It becomes zero if and only if both a priori and a posteriori distributions are identical, and ranges from 0 to 1. In other words, the Hellinger measure is continuous on every possible combination of a priori and a posteriori values. It can be interpreted as a distance measure where distance corresponds to the amount of divergence between a priori and a posteriori distribution. Therefore, we employ Hellinger measure as a measure of divergence, which will be used as the information amount of rules.

In terms of the probabilistic rules, let us interpret the event $A = a$ as the target concept to be learned and the event (possibly conjunctive) $B = b$ as the hypothesis describing this concept. In this paper, we slightly modify the Hellinger divergence. The information content of a rule (denoted as $IC(b \rightarrow a)$) using Hellinger divergence is defined as

$$\begin{aligned} IC(b \rightarrow a) &= \left(\sqrt{p(a|b)} - \sqrt{p(a)}\right)^2 + \\ &\quad \left(\sqrt{p(\neg a|b)} - \sqrt{p(\neg a)}\right)^2 \\ &= \left(\sqrt{p(a|b)} - \sqrt{p(a)}\right)^2 + \\ &\quad \left(\sqrt{1-p(a|b)} - \sqrt{1-p(a)}\right)^2 \end{aligned} \tag{2}$$

where $p(a|b)$ means the conditional probability of $A = a$ under the condition $B = b$. Notice that Equation (2) has a different form of definition from that of Equation (1). In rule induction, one particular value of the target attribute appears in the right hand side of the pattern, and thus the probabilities for all other values are included in $1 - p(a)$.

In addition, we squared the original form of Hellinger measure because (1) by squaring the original form of Hellinger measure, we could derive a boundary of the Hellinger measure, which allows us to reduce drastically the search space of possible rule rules. (2) the relative information content of each pattern is not affected by the modified Hellinger measure, and (3) the weights between two terms of Hellinger measure provides more reasonable trade-off in terms of their value range.

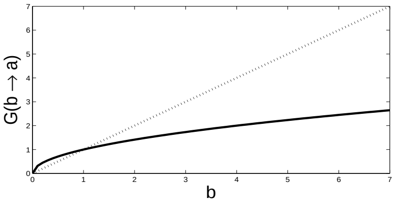

Figure 1. Plot of $\sqrt{p(b)}$ and $p(b)$

Another criteria we have to consider is the *generality* of the rules. The basic idea behind generality is that the more often left-hand side occurs for a rule, the more useful the pattern becomes. The left-hand side must occur relatively often for a pattern to be deemed useful. In this paper, we use

$$G(b \rightarrow a) = \sqrt{p(b)} \tag{3}$$

to represent the probability that the rule will occur and, as such, can be interpreted as the measure of rule generality.

The reason for using the square root form of the original probability is that the square root value can represent the generality of events more correctly. The generality of an event(b) increases rapidly when the event first appears. After that, its importance grows slowly when the event has already happened more than enough. Figure 1 compares the plot of $\sqrt{p(b)}$ with that of straight line, which represents $p(b)$.

As shown by the square root function in Figure 1, the value grows rapidly in early state and, then the observation of the event become less important after the event happens a lot. Meanwhile, the linear function of generality, denoted as $p(a)$, grows proportional to the number of events, which does not match with the characteristics of the generality in real world. Another advantage of using the square root form is that we could also derive some boundaries of H measure, described in Property 8 and 9 of the following section.

As a result, by multiplying the generality ($G(b \rightarrow a)$) with the information content ($IC(b \rightarrow a)$) of the rule, the importance of rule, denoted as $HD(b \rightarrow a)$, is given as the following term

$$\begin{aligned} HD(b \rightarrow a) &= G(b \rightarrow a) \cdot IC(b \rightarrow a) \\ &= \sqrt{p(b)}\left[\left(\sqrt{p(a|b)} - \sqrt{p(a)}\right)^2 + \right. \\ &\quad \left. \left(\sqrt{1-p(a|b)} - \sqrt{1-p(a)}\right)^2\right] \qquad (4) \\ &= 2\sqrt{p(b)}\left[1 - \sqrt{p(a)p(a|b)} - \right. \\ &\quad \left. \sqrt{(1-p(a))(1-p(a|b))}\right] \qquad (5) \\ &= 2\left[\sqrt{p(b)} - \sqrt{p(a)p(ab)} - \right. \\ &\quad \left. \sqrt{(1-p(a))(p(b)-p(ab))}\right] \qquad (6) \end{aligned}$$

which possesses a direct interpretation as a multiplicative measure of the generality and information content of a given rule.

III. Properties of HD Measure

This section describes the properties of the proposed measure in this paper. Assuming we have such a rule as $b \rightarrow a$, the proposed HD measure has the following properties.

Property 1 : $HD(b \rightarrow a) \geq 0$.

The proof of this property is trivial from the definition of the HD measure given in Equation (4). This property is one of the fundamental properties of rule importance measure since negative importance simply does not make sense in rule mining.

Property 2 : *If a and b are independent, then $HD(b \rightarrow a) = 0$.*

If values a and b are independent with each other, it is known that $p(ab) = p(a)p(b)$. Therefore, from Equation (4), it is clear that $HD(b \rightarrow a) = 0$. In case antecedent attribute and consequent attribute are independent, the resulting importance of the rule ought to be zero. In this sense, this property is an important property.

Property 3 : $HD(b \rightarrow a) \neq HD(a \rightarrow b)$.

With respect to the information content of each rule, $IC(b \rightarrow a) = IC(a \rightarrow b)$. However, $G(b \rightarrow a) \neq G(a \rightarrow b)$. Therefore, $HD(b \rightarrow a) \neq HD(a \rightarrow b)$. Rule $b \rightarrow a$ means there is cause-result relationship between b and a, respectively. This rule does not necessarily mean $a \rightarrow b$.

Property 4 : *Suppose the values of $p(a)$ and $p(b)$ are fixed. When the value of $p(ab)$ increases, the HD measure behaves as follows*

$$HD(b \rightarrow a) = \begin{cases} \searrow & \text{if } \ p(ab) < p(a)p(b) \\ 0 & \text{if } \ p(ab) = p(a)p(b) \\ \nearrow & \text{otherwise} \end{cases}$$

The $\searrow$ and $\nearrow$ symbols represent the value of HD measure monotonically increase and monotonically decreases, respectively. From Equation (4),

$$\begin{aligned} \frac{\partial HD(b \rightarrow a)}{\partial p(ab)} &= -2\sqrt{p(a)}\left(\frac{1}{2}\right)\left(\frac{1}{\sqrt{p(ab)}}\right) - \\ & \quad 2\sqrt{1-p(a)}\left(\frac{-1}{2}\right)\left(\frac{1}{\sqrt{p(b)-p(ab)}}\right) \\ &= \sqrt{\frac{1-p(a)}{p(b)-p(ab)}} - \sqrt{\frac{p(a)}{p(ab)}} \qquad (7) \end{aligned}$$

Suppose

$$D = \frac{1-p(a)}{p(b)-p(ab))} - \frac{p(a)}{p(ab)} = \frac{p(ab)-p(a)p(b)}{((p(b)-p(ab))p(ab)}$$

(i) If $p(ab) < p(a)p(b)$, then $D < 0$. Therefore, $\frac{\partial HD(b \rightarrow a)}{\partial p(ab)} < 0$.

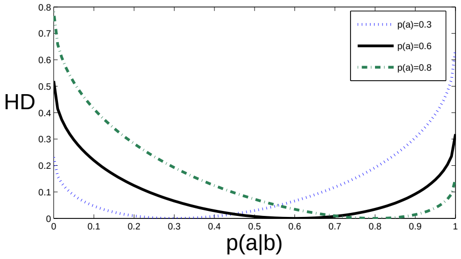

Figure 2. HD values by changing $p(a|b)$

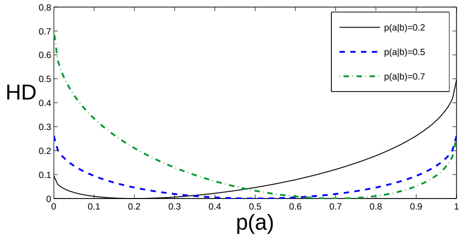

Figure 3. HD values by changing $p(a)$

(ii) If $p(ab) = p(a)p(b)$, then $D = 0$. Therefore, $HD(b \rightarrow a) = 0$.

(iii) If $p(ab) > p(a)p(b)$, then $D > 0$. Therefore, $\frac{\partial HD(b \rightarrow a)}{\partial p(ab)} > 0$. Q.E.D.

This property shows an important characteristic of the HD measure. The HD measure monotonically increases as the degree of deviation from independence between variable of a and b increases.

Property 5 : *Suppose the values of p(a) and p(b) are fixed. When the value of $p(a|b)$ increases, the HD-measure behaves as follows*

$$HD(b \rightarrow a) = \begin{cases} \searrow & \text{if } \ p(a|b) < p(a) \\ 0 & \text{if } \ p(a|b) = p(a) \\ \nearrow & \text{otherwise} \end{cases}$$

The proof of this property is straightforward since we get the same results by dividing the probabilities in Property 4 by $p(b)$. Figure 2 show the relationship between HD measure and $p(a|b)$. For simplicity, in Figure 2, the value of $p(b)$ is given as 0.5. The probability $p(a|b)$ can be interpreted as the accuracy of the rule.

Property 6 : *Suppose the values of $p(a|b)$ and $p(b)$ are fixed. When the value of $p(a)$ increases, $HD(b \rightarrow a)$ behaves as follows*

$$HD(b \rightarrow a) = \begin{cases} \searrow & \text{if } \ p(a) < p(a|b) \\ 0 & \text{if } \ p(a) = p(a|b) \\ \nearrow & \text{otherwise} \end{cases}$$

Figure 3 show the relationship between HD value and $p(a)$. For simplicity, in Figure 3, the value of $p(b)$ is given as 0.2.

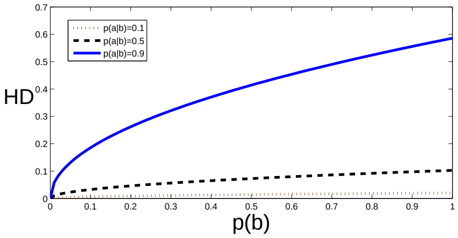

Figure 4. HD values by changing $p(b)$

Property 7 : *HD increases monotonically as the value of p(b) increases.*

This property is true based on Equation 5. Figure 4 shows the relationship between the HD values and $p(b)$ values.

Figure 4 show the relationship between HD value and $p(b)$. For simplicity, in Figure 4, the value of $p(a)$ is given as 0.2.

Property 8 : *In case we add some additional conditions in the rule such as $b \wedge c \to a$ where C means a set of value assignments. The HD measure of this rule, denoted as HD_2, is bounded as follows.*

$$HD_2 \leq \max\{\sqrt{p(a|b)}\sqrt{p(b)}\left[2\sqrt{m} - 2\sqrt{p(a)}\right],$$
$$2\sqrt{p(b)} - \sqrt{1-p(a|b)}\sqrt{p(b)}\left[2\sqrt{p(a)} + 2\sqrt{1-p(a)}\right]\}$$

where m represents the number of class in the target variable.

With this property, we are able to estimate the boundary of HD measure value without knowing any information about c. Using Property 8, we can predict in advance whether adding conditions in current rule can increase the HD measure. This property is very useful when we generate rules using the proposed HD measure since we can significantly reduce the search space of rule generation.

Property 9 : *Suppose the HD measure of $b \to a$ and $b \wedge c \to a$ are HD_1 and HD_2, respectively. In case $p(a|b) = 1$, then $HD_1 \geq HD_2$.*

From $p(b) = p(ab) + p(\neg ab)$ and $p(a|b) = \frac{p(ab)}{p(b)} = 1$,

$$p(\neg ab) = p(b) - p(ab) = 0$$

Therefore,

$$\begin{aligned} p(a|bc) = \frac{p(abc)}{p(bc)} &= \frac{p(abc)}{p(abc) + p(\neg abc)} \\ &= \frac{p(abc)}{p(abc) + p(c|\neg ab)p(\neg ab)} = 1 \end{aligned} \quad (8)$$

From Equation (5) and $p(a|b) = 1$,

$$HD_1 = \sqrt{p(b)}\left(2 - 2\sqrt{p(a)}\right)$$

From Equation (5) and (8),

$$HD_2 = \sqrt{p(bc)}\left(2 - 2\sqrt{p(a)}\right)$$

Since $p(bc) \leq p(b)$, $\quad HD_2 \leq HD_1$. Q.E.D.

This property is also useful when we generate rules using the proposed HD measure. Like Property 8, we can significantly reduce the search space of rule generation using Property 9.

IV. Conclusions and Future Work

In this paper, we proposed a new information theoretic measure of rule importance, called HD measure. Specifically, we employed Hellinger divergence as the measure of information content of rules, and combined it with the generality of rule. The proposed rule importance show a number of important and interesting characteristics.

The future work of this paper is as follows.

- More analysis of the characteristics of the HD measure
- Apply the HD measure in a rule generation algorithms using real datasets.
- Use the measure as a tool for classification learning.
- Compare the classification performance with current importance measures.

V. Acknowledgement

This work was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) (Grant number: 2011- 0023296).

References

[1] B. Liu, W. Hsu, S. Chen, and Y. Ma, “Analyzing the subjective interestingness of association rules,” *IEEE Intelligent Systems*, vol. 15, no. 5, pp. 47–55, 2000.

[2] B. Padmanabhan and A. Tuzhilin, “Unexpectedness as a measure of interestingness in knowledge discovery,” *Decision Support Systems*, vol. 27, no. 3, pp. 303–318, 1999.

[3] R. J. Bayardo and R. Agrawal, “Mining the most interesting rules,” in *KDD*, 1999, pp. 145–154.

[4] X.-H. Huynh, F. Guillet, and H. Briand, “Arqat: An exploratory analysis tool for interestingness measures,” in *the 11th international symposium on Applied Stochastic Models and Data Analysis ASMDA-2005*, 334-344, p. 2005.

[5] V. K. P.-N. Tan and J. Srivastava, “Selecting the right objective measure for association analysis,” *Information Systems*, vol. 29, no. 4, pp. 293–313, 2004.

[6] P. L. B. Vaillant and S. Lallich, “A clustering of interestingness measures,” in *Proceedings of the 7th International Conference on Discovery Science*, 2004, pp. 290–297.

[7] P. Smyth and R. M. Goodman, “An information theoretic approach to rule induction from databases,” *IEEE Transactions on Knowledge and Data Engineering*, vol. 4, no. 4, pp. 301–316, 1992.

[8] P. Clark and T. Niblett, “The cn2 induction algorithm,” *Machine Learning*, vol. 3, no. 4, pp. 261–283, 1989.

[9] R. J. Beran, “Minimum hellinger distances for parametric models,” *Ann. Statistics*, vol. 5, pp. 445–463, 1977.