Research Article

Metric Divergence Measures and Information Value in Credit Scoring

Guoping Zeng

Think Finance, 4150 International Plaza, Fort Worth, TX 76109, USA

Correspondence should be addressed to Guoping Zeng; guopingtx@yahoo.com

Received 13 August 2013; Accepted 4 September 2013

Academic Editor: Baoding Liu

Copyright © 2013 Guoping Zeng. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Recently, a series of divergence measures have emerged from information theory and statistics and numerous inequalities have been established among them. However, none of them are a metric in topology. In this paper, we propose a class of metric divergence measures, namely, $L_p(P \parallel Q)$, $P \geq 1$, and study their mathematical properties. We then study an important divergence measure widely used in credit scoring, called information value. In particular, we explore the mathematical reasoning of weight of evidence and suggest a better alternative to weight of evidence. Finally, we propose using $L_p(P \parallel Q)$ as alternatives to information value to overcome its disadvantages.

1. Introduction

The information measure is an important concept in Information theory and statistics. It is related to the system of measurement of information or the amount of information based on the probabilities of the events that convey information. Divergence measures are an important type of information measures. They are commonly used to find appropriate distance or difference between two probability distributions.

Let

$$\Delta_n = \left\{ P = (p_1, p_2, \dots, p_n) \mid p_i \geq 0, \sum_{i=1}^n p_i = 1 \right\}, \quad n \geq 2, \quad (1)$$

be the set of finite discrete probability distributions as in [1]. For all $P, Q \in \Delta_n$, the following divergence measures are well known in the literature of information theory and statistics.

Hellinger Discrimination [2].

$$h(P \parallel Q) = \frac{1}{2} \sum_{i=1}^n (\sqrt{p_i} - \sqrt{q_i})^2. \quad (2)$$

Shannon's Entropy [3].

$$H(P) = - \sum_{i=1}^n p_i \log_2(p_i), \quad (3)$$

which is sometimes referred to as *measure of uncertainty*. The entropy $H(P)$ of a discrete random variable is defined in terms of its probability distribution P and is a good measure of randomness or uncertainty.

Note that in the original definition of Shannon's Entropy the log is to the base 2 and entropy is expressed in bits in information theory. The log can be any other bases and the entropy will be a constant factor of the one in base 2 by the change-base formula of the logarithm function. Hence, without loss of generality, we will assume all the logs are natural logarithms.

Kullback and Leibler's Relative Information [4].

$$D(P \parallel Q) = \sum_{i=1}^n p_i \ln \left(\frac{p_i}{q_i} \right). \quad (4)$$

Its symmetric form is the well-known J -divergence.

J-Divergence (Jeffreys [5], Kullback, and Leibler [4]).

$$J(P \parallel Q) = D(P \parallel Q) + D(Q \parallel P) = \sum_{i=1}^n (p_i - q_i) \ln \left(\frac{p_i}{q_i} \right). \quad (5)$$

Triangular Discrimination [6].

$$\Delta(P \parallel Q) = \sum_{i=1}^n \frac{(p_i - q_i)^2}{p_i + q_i}. \quad (6)$$

Symmetric Chi-Square Divergence (Dragomir et al. [7]). One has

$$\begin{aligned} \psi(P \parallel Q) &= \chi^2(P \parallel Q) + \chi^2(Q \parallel P) \\ &= \sum_{i=1}^n \frac{(p_i - q_i)^2 (p_i + q_i)}{p_i q_i}, \end{aligned} \quad (7)$$

where $\chi^2(P \parallel Q) = \sum_{i=1}^n ((p_i - q_i)^2 / p_i)$ is the well-known χ^2 -divergence (Pearson [8]).

Jensen-Shannon Divergence (Sibson [9], Burbea, and Rao [10, 11]).

$$I(P \parallel Q) = \frac{1}{2} \left[\sum_{i=1}^n p_i \ln \left(\frac{2p_i}{p_i + q_i} \right) + \sum_{i=1}^n q_i \ln \left(\frac{2q_i}{p_i + q_i} \right) \right]. \quad (8)$$

Arithmetic-Geometric Divergence (Taneja [12]). Moreover

$$T(P \parallel Q) = \sum_{i=1}^n \frac{p_i + q_i}{2} \ln \left(\frac{p_i + q_i}{2\sqrt{p_i q_i}} \right). \quad (9)$$

Taneja's Divergence (Taneja [12]). One has

$$d(P \parallel Q) = 1 - \sum_{i=1}^n \left(\frac{\sqrt{p_i} + \sqrt{q_i}}{2} \right) \left(\sqrt{\frac{p_i + q_i}{2}} \right). \quad (10)$$

The information measures $J(P \parallel Q)$, $I(P \parallel Q)$ and $T(P \parallel Q)$ can be written as

$$\begin{aligned} J(P \parallel Q) &= 4 [I(P \parallel Q) + T(P \parallel Q)], \\ I(P \parallel Q) &= \frac{1}{2} \left[D\left(P \parallel \frac{P+Q}{2}\right) + D\left(Q \parallel \frac{P+Q}{2}\right) \right], \\ T(P \parallel Q) &= \frac{1}{2} \left[D\left(\frac{P+Q}{2} \parallel P\right) + D\left(\frac{P+Q}{2} \parallel Q\right) \right]. \end{aligned} \quad (11)$$

Relative Information of Type s. Cressie and Read [13] considered the one-parametric generalization of information measure $D(P \parallel Q)$, called the relative information of type s given by

$$D_s(P \parallel Q) = [s(s-1)]^{-1} \left[\sum_{i=1}^n p_i^s q_i^{1-s} - 1 \right], \quad s \neq 0, 1. \quad (12)$$

It also has some special cases:

- (i) $\lim_{s \rightarrow 0} D_s(P \parallel Q) = D(Q \parallel P)$,
- (ii) $\lim_{s \rightarrow 1} D_s(P \parallel Q) = D(P \parallel Q)$,
- (iii) $D_{-1}(P \parallel Q) = (1/2)\psi^2(Q \parallel P)$,
- (iv) $D_{1/2}(P \parallel Q) = 4h(P \parallel Q)$,
- (v) $D_2(P \parallel Q) = (1/2)\psi^2(P \parallel Q)$,
- (vi) $D_2(P \parallel Q) = D_{-1}(Q \parallel P)$,
- (vii) $D_1(P \parallel Q) = D_0(Q \parallel P)$.

It is shown that $D_s(P \parallel Q)$ is nonnegative and convex in P and Q in [1].

J-Divergence of Type s [14].

$$\begin{aligned} V_s(P \parallel Q) &= D_s(P \parallel Q) + D_s(Q \parallel P) \\ &= [s(s-1)]^{-1} \left[\sum_{i=1}^n (p_i^s q_i^{1-s} + q_i^s p_i^{1-s}) - 2 \right], \\ &\quad s \neq 0, 1. \end{aligned} \quad (13)$$

It admits the following particular cases:

- (i) $\lim_{s \rightarrow 0} V_s(P \parallel Q) = \lim_{s \rightarrow 1} V_s(P \parallel Q) = J(P \parallel Q)$,
- (ii) $\lim_{s \rightarrow 1} V_s(P \parallel Q) = J(P \parallel Q)$,
- (iii) $V_{-1}(P \parallel Q) = V_2(P \parallel Q) = (1/2)\Psi(P \parallel Q)$,
- (iv) $V_0(P \parallel Q) = V_1(P \parallel Q) = J(P \parallel Q)$,
- (v) $V_{1/2}(P \parallel Q) = 8h(P \parallel Q)$.

Unified Generalization of Jensen-Shannon Divergence and Arithmetic-Geometry Mean Divergence [14].

$$W_s(P \parallel Q) = \frac{1}{2} \left[D_s\left(\frac{P+Q}{2} \parallel P\right) + D_s\left(\frac{P+Q}{2} \parallel Q\right) \right]. \quad (14)$$

It admits the following particular cases:

- (i) $W_{-1}(P \parallel Q) = (1/4)\Delta(P \parallel Q)$,
- (ii) $W_0(P \parallel Q) = I(P \parallel Q)$,
- (iii) $W_{1/2}(P \parallel Q) = 4d(P \parallel Q)$,
- (iv) $W_1(P \parallel Q) = T(P \parallel Q)$,
- (v) $W_2(P \parallel Q) = (1/16)\Psi(P \parallel Q)$.

Taneja proved [14] that all the 3 s -type information measures $D_s(P \parallel Q)$, $V_s(P \parallel Q)$, and $W_s(P \parallel Q)$ are nonnegative and convex in the pair (P, Q) . He also obtained inequalities regarding the various divergence measures:

$$\begin{aligned} \frac{1}{4} \Delta(P \parallel Q) &\leq I(P \parallel Q) \leq h(P \parallel Q) \leq 4d(P \parallel Q) \\ &\leq \frac{1}{8} J(P \parallel Q) \leq T(P \parallel Q) \leq \frac{1}{16} \Psi(P \parallel Q). \end{aligned} \quad (15)$$

Here, we observe that $\Delta(P \parallel Q) > 0$. Hence, from the above inequalities we see that $I(P \parallel Q)$, $h(P \parallel Q)$, $d(P \parallel Q)$, $J(P \parallel Q)$, $T(P \parallel Q)$, and $\psi(P \parallel Q)$ are all positive.

We also note that all p_i in the original definition of Δ_n in [1] are required to be positive. Yet, in reality some p_i may be 0. In this case $I(P \parallel Q)$, $T(P \parallel Q)$, $\psi(P \parallel Q)$, and $\Delta(P \parallel Q)$ will be undefined. We have extended the definition of Δ_n to include the cases when $p_i = 0$. We assume that $0 \ln(0) = 0$, which is easily justified by continuity since $x \ln(x) \rightarrow 0$ as $x \rightarrow 0$. For convenience, we also assume $0 \ln(0/0) = 0$ and $t \ln(t/0) = 0$ for $t > 0$.

A problem with the above divergence measures is that none of them are a real distance, that is, a metric, in topology. In this paper, we will study a class of metric divergence measures $L_p(P \parallel Q)$. We then study the underlying mathematics of a special divergence measure called information value, which is widely used in credit scoring. We propose using $L_p(P \parallel Q)$ as alternatives to IV in order to overcome the disadvantages of information value.

The rest of this paper is organized as follows. In Section 2, after reviewing the metric space, we disprove that the above divergence measures are metrics. We then study a class of metric divergence measures $L_p(P \parallel Q)$. Section 3 is concerned with information value. We examine a rule of thumb and weight of evidence and suggest a better alternative to weight of evidence. We then propose using $L_p(P \parallel Q)$ as alternatives of IV to overcome the disadvantages of information value. Section 4 presents some numerical results. Finally, the paper is concluded in Section 5.

2. Metric Divergence Measures

2.1. Review of Metric Space

Definition 1. Suppose a real valued function $d : M \times M \rightarrow R$ and that for all x, y, z of set M

- (1) $d(x, y) \geq 0$ (nonnegative),
- (2) $d(x, x) = 0$ if and only if $x = y$ (identity),
- (3) $d(x, y) = d(y, x)$ (symmetry),
- (4) $d(x, z) \leq d(x, y) + d(y, z)$ (triangle inequality).

Such a “distance function” d is called a metric on M , and the pair (D, M) is called a metric space. If d satisfies (1)–(3) but not necessarily (4), it is called a semimetric.

A metric space is a topological space in a natural manner, and therefore all definitions and theorems about general topological spaces also apply to a metric space. For instance, in a metric space one can define open and closed sets, convergence of sequences of points, compact space, and connected space.

Definition 2. A metric d_1 is said to be upper bounded by another metric d_2 if there exists a positive constant c such that $d_1(x, y) \leq c \times d_2(x, y)$ for all $x, y \in M$. In this case, d_2 is said to be lower bounded by d_1 .

If d_1 is upper bounded by d_2 , then the convergence in the metric space (M, d_2) implies the convergence in the metric space (M, d_1) .

Definition 3. Two metrics d_1 and d_2 are equivalent if there exist 2 positive constants α and β such that $\alpha \times d_2(x, y) \leq d_1(x, y) \leq \beta \times d_2(x, y)$.

If two metrics d_1 and d_2 are equivalent, they will have the same convergence.

2.2. Nonmetric Divergence Measures

Proposition 4. None of the divergence measures $\Delta(P \parallel Q)$, $I(P \parallel Q)$, $h(P \parallel Q)$, $d(P \parallel Q)$, $J(P \parallel Q)$, $T(P \parallel Q)$, and $\psi(P \parallel Q)$ are a metric in topology. Indeed, none of them satisfy the triangle inequality.

Proof. We disprove them either numerically or analytically by constructing counter examples in Δ_2 .

- (a) Let $p = (0, 1)$, $q = (1, 0)$, and $r = (0.5, 0.5)$. Then $\Delta(P \parallel Q) = 2$, $\Delta(P \parallel R) = \Delta(R \parallel Q) = 2/3$, and $\Delta(P \parallel R) + \Delta(R \parallel Q) = 4/3 < 2 = \Delta(P \parallel Q)$.
- (b) Let $p = (0.5, 0.5)$, $q = (0.2, 0.8)$, and $r = (0.3, 0.7)$. Then $I(P \parallel Q) = 0.02357767$, $I(R \parallel Q) = 0.00350105$, $I(P \parallel R) = 0.01020550$, and $I(P \parallel R) + I(R \parallel Q) < I(P \parallel Q)$.
- (c) Let $p = (0, 1)$, $q = (1, 0)$, and $r = (0.5, 0.5)$. Then $h(P, Q) = 1$, $h(P \parallel R) = h(R \parallel Q) = (1/2)(2 - \sqrt{2})$, and $h(P \parallel R) + h(R \parallel Q) = 2 - \sqrt{2} < 1 = h(P \parallel Q)$.
- (d) Let $p = (0.2, 0.8)$, $q = (0.4, 0.6)$, and $r = (0.3, 0.7)$. Then

$$\begin{aligned}
 & D(P \parallel R) + D(R \parallel Q) - D(P \parallel Q) \\
 &= 0.2 \times \ln \frac{2}{3} + 0.8 \times \ln \frac{8}{7} + 0.3 \times \ln \frac{3}{4} + 0.7 \\
 &\quad \times \ln \frac{7}{6} - 0.2 \times \ln \frac{1}{2} - 0.8 \times \ln \frac{4}{3} \\
 &= 0.2 \left(\ln \frac{2}{3} - \ln \frac{1}{2} \right) + 0.8 \left(\ln \frac{8}{7} - \ln \frac{4}{3} \right) \\
 &\quad + 0.3 \times \ln \frac{3}{4} + 0.7 \times \ln \frac{7}{6} \\
 &= 0.2 \times \ln \frac{4}{3} + 0.8 \times \ln \frac{6}{7} + 0.3 \times \ln \frac{3}{4} + 0.7 \times \ln \frac{7}{6} \\
 &= 0.2 \times \ln \frac{4}{3} + 0.8 \times \ln \frac{6}{7} - 0.3 \times \ln \frac{4}{3} - 0.7 \times \ln \frac{6}{7} \\
 &= -0.1 \times \ln \frac{4}{3} + 0.1 \times \ln \frac{6}{7} \\
 &= 0.1 \left(\ln \frac{6}{7} - \ln \frac{4}{3} \right) = 0.1 \times \ln \frac{9}{14} < 0.1 \times \ln 1 = 0.
 \end{aligned} \tag{16}$$

Hence, $D(P \parallel R) + D(R \parallel Q) < D(P \parallel Q)$. Indeed, $D(P \parallel Q)$ is not symmetric either (see [15]).

(e) Let $p = (0.2, 0.8)$, $q = (0.4, 0.6)$, $r = (0.3, 0.7)$. Then

$$\begin{aligned} J(P \parallel R) + J(R \parallel Q) &= \left(0.1 \times \ln \frac{3}{2} + 0.1 \times \ln \frac{8}{7}\right) \\ &\quad + \left(0.1 \times \ln \frac{4}{3} + 0.1 \times \ln \frac{7}{6}\right) \\ &= 0.1 \times \ln \frac{8}{3}, \end{aligned} \quad (17)$$

$$J(P \parallel Q) = 0.2 \times \ln 2 + 0.2 \times \ln \frac{4}{3} = 0.2 \times \ln \frac{8}{3}.$$

Hence, $J(P \parallel R) + J(R \parallel Q) < J(P \parallel Q)$.

(f) Let $p = (0.5, 0.5)$, $q = (0.2, 0.8)$, and $r = (0.3, 0.7)$. Then $T(P \parallel Q) = 0.05330024$, $T(P \parallel R) = 0.02135897$, $T(R \parallel Q) = 0.00677313$, and $T(P \parallel R) + T(R \parallel Q) < T(P \parallel Q)$.

(g) Let $p = (0.5, 0.5)$, $q = (0.2, 0.8)$, $r = (0.3, 0.7)$. Then, $\psi(P \parallel Q) = 369/400$, $\psi(Q \parallel R) = 185/1680$, $\psi(P \parallel R) = 368/1050$, and $\psi(P \parallel R) + \psi(R \parallel Q) < \psi(P \parallel Q)$.

□

2.3. A Natural Metric Divergence Measure. If we pick up the common part of $D(P \parallel Q)$ and $J(P \parallel Q)$, we will obtain a metric divergence measure

$$Z(P \parallel Q) = \sum_{i=1}^n |\ln(p_i) - \ln(q_i)| = \sum_{i=1}^n \left| \ln \left(\frac{p_i}{q_i} \right) \right|. \quad (18)$$

Since $p_i \leq 1$ and $|p_i - q_i| \leq 1$ for all $1 \leq i \leq n$, both $D(P \parallel Q)$ and $J(P \parallel Q)$ are upper bounded by $Z(P \parallel Q)$; that is, $D(P \parallel Q) \leq Z(P \parallel Q)$, $J(P \parallel Q) \leq Z(P \parallel Q)$.

2.4. L_p -Divergence. Recall that for a real number $p \geq 1$, the L_p -norm of vector $x = (x_1, x_2, \dots, x_n) \in \mathbf{R}^n$ is defined by

$$\|x\|_p = \left(\sum_{i=1}^n |x_i|^p \right)^{1/p}. \quad (19)$$

We will apply the L_p -metric from the L_p -norm to divergence measures to obtain L_p -divergence. For convenience, we will use the upper case notation.

Definition 5. For two probability distributions P and Q , one defines their L_p -divergence as

$$L_p(P \parallel Q) = \sqrt[p]{\sum_{i=1}^n |p_i - q_i|^p}. \quad (20)$$

Here, $p \geq 1$ is used for superscript, subscript, and radial root. It should not be confused with the vector P .

In particular, when $p = 1$, we have the L_1 distance:

$$L_1(P \parallel Q) = \sum_{i=1}^n |p_i - q_i|. \quad (21)$$

When $p = 2$, we obtain the Euclidean Distance:

$$L_2(P \parallel Q) = \sqrt{\sum_{i=1}^n (p_i - q_i)^2}. \quad (22)$$

When $p \rightarrow \infty$, we have

$$L_\infty(P \parallel Q) = \max_{1 \leq i \leq n} |p_i - q_i|. \quad (23)$$

It is known that L_p -norms are decreasing in p . Moreover, all L_p metrics are equivalent.

Lemma 6. If $r > p \geq 1$, then the L_p -norms in $\mathbf{R}^n$ satisfy

$$\|x\|_r \leq \|x\|_p \leq n^{(1/p-1/r)} \|x\|_r. \quad (24)$$

Corollary 7. If $r > p \geq 1$, then the L_p -divergences satisfy

$$L_r(P \parallel Q) \leq L_p(P \parallel Q) \leq n^{(1/p-1/r)} L_r(P \parallel Q). \quad (25)$$

Theorem 8. L_p -divergences are all bounded by constant 2 for $p \geq 1$; that is, $L_p(P \parallel Q) \leq 2$. In particular, $L_\infty(P \parallel Q) \leq 1$ and $L_2(P \parallel Q) \leq \sqrt{2}$.

Proof. We first prove the general case. From Corollary 7, it is sufficient to prove that the L_1 -divergence is bounded by 2. Let $P = (p_1, p_2, \dots, p_n)$ and $Q = (q_1, q_2, \dots, q_n)$ be two probability distributions. Without loss of generality, let us assume that $p_1 \geq q_1$, $p_2 \geq q_2, \dots, p_k \geq q_k$, and $p_{k+1} < q_{k+1}, \dots, p_n < q_n$. Then, we have

$$\begin{aligned} L_1(P \parallel Q) &= \sum_{i=1}^n |p_i - q_i| \\ &= \sum_{i=1}^k (p_i - q_i) + \sum_{i=k+1}^n (q_i - p_i) \\ &\leq \sum_{i=1}^k p_i + \sum_{i=k+1}^n q_i \leq \sum_{i=1}^n p_i + \sum_{i=1}^n q_i = 1 + 1 = 2. \end{aligned} \quad (26)$$

Noting that $|p_i - q_i| \leq 1$ for $1 \leq i \leq n$, we have $(p_i - q_i)^2 \leq |p_i - q_i| \leq 1$. Hence,

$$L_\infty(P \parallel Q) = \max_{1 \leq i \leq n} |p_i - q_i| \leq 1,$$

$$\begin{aligned} L_2(P \parallel Q) &= \sqrt{\sum_{i=1}^n (p_i - q_i)^2} \\ &\leq \sqrt{\sum_{i=1}^n |p_i - q_i|} = \sqrt{L_1(P \parallel Q)} \leq \sqrt{2}. \end{aligned} \quad (27)$$

Therefore, we have proved the 2 particular cases. □

The following result shows that the relative entropy $D(P \parallel Q)$ is lower bounded by the square of the $L_1(P \parallel Q)$. Its proof can be found at [15].

Lemma 9.

$$D(P \parallel Q) \geq \frac{1}{2}(L_1(P \parallel Q))^2. \quad (28)$$

Theorem 10. *The square root of $J(P \parallel Q)$ is lower bounded by $L_1(P \parallel Q)$; that is,*

$$\sqrt{J(P \parallel Q)} \geq L_1(P \parallel Q). \quad (29)$$

Proof. Applying lemma to $D(P \parallel Q)$ and $D(Q \parallel P)$, we obtain

$$\begin{aligned} D(P \parallel Q) + D(Q \parallel P) &\geq \frac{1}{2}(L_1(P \parallel Q))^2 \\ &\quad + \frac{1}{2}(L_1(Q \parallel P))^2 = (L_1(P \parallel Q))^2. \end{aligned} \quad (30)$$

Note that the left hand side is nothing but $J(P \parallel Q)$. The proof is completed by taking the square root on both sides. $\square$

Remark 11. $J(P \parallel Q)$ and hence $\sqrt{J(P \parallel Q)}$ are unbounded by any constants. This can be seen by taking $P = (s, 1-s)$ and $Q = (1-s, s)$, $0 < s < 1$, and taking limit $s \rightarrow 0$. Since L_p -divergences are all bounded by constant 2 for $p \geq 1$, $J(P \parallel Q)$ and hence $\sqrt{J(P \parallel Q)}$ are not equivalent to L_p -divergences.

We now establish the convexity property for L_p -divergence, which is useful in optimization.

Theorem 12. $L_p(P \parallel Q)$ is convex in the pair (p, q) , that is, if (P_1, Q_1) , and (P_2, Q_2) are two pairs of probability distributions, then

$$\begin{aligned} L_p(\lambda P_1 + (1-\lambda)P_2 \parallel \lambda Q_1 + (1-\lambda)Q_2) \\ \leq \lambda L_p(P_1 \parallel Q_1) + (1-\lambda)L_p(P_2 \parallel Q_2) \end{aligned} \quad (31)$$

for all $0 \leq \lambda \leq 1$.

Proof. Let $P_1 = (p_{11}, p_{12}, \dots, p_{1n})$, $P_2 = (p_{21}, p_{22}, \dots, p_{2n})$, $Q_1 = (q_{11}, q_{12}, \dots, q_{1n})$, and $Q_2 = (q_{21}, q_{22}, \dots, q_{2n})$. Then

$$\begin{aligned} L_p(\lambda P_1 + (1-\lambda)P_2 \parallel \lambda Q_1 + (1-\lambda)Q_2) \\ = \left(\sum_{i=1}^n |\lambda p_{1i} + (1-\lambda)p_{2i} - \lambda q_{1i} - (1-\lambda)q_{2i}|^p \right)^{1/p} \\ = \left(\sum_{i=1}^n |\lambda(p_{1i} - q_{1i}) + (1-\lambda)(p_{2i} - q_{2i})|^p \right)^{1/p} \end{aligned}$$

$$\begin{aligned} &\leq \left(\sum_{i=1}^n |\lambda(p_{1i} - q_{1i})|^p \right)^{1/p} \\ &\quad + \left(\sum_{i=1}^n |(1-\lambda)(p_{2i} - q_{2i})|^p \right)^{1/p} \\ &= \left(\lambda \sum_{i=1}^n |p_{1i} - q_{1i}|^p \right)^{1/p} \\ &\quad + (1-\lambda) \left(\sum_{i=1}^n |p_{2i} - q_{2i}|^p \right)^{1/p} \\ &= \lambda L_p(P_1 \parallel Q_1) + (1-\lambda)L_p(P_2 \parallel Q_2). \end{aligned} \quad (32)$$

Here, the first inequality is from the well-known Minkowski's Inequality. $\square$

It follows from the following results that we can generate infinitely many metric divergence measures using the existing ones.

Proposition 13. *If $d_1(P \parallel Q)$ and $d_2(P \parallel Q)$ are two metric divergence measures, so are the following 3 measures:*

- (1) $\alpha \times d_1(P \parallel Q) + \beta \times d_2(P \parallel Q)$ for all $\alpha \geq 0$, $\beta \geq 0$ and $\alpha + \beta > 0$,
- (2) $\max(d_1(P \parallel Q), d_2(P \parallel Q))$,
- (3) $\sqrt{(d_1(P \parallel Q))^2 + (d_2(P \parallel Q))^2}$.

Proof. The proof of (1) and (2) is trivial and hence will be omitted. As for (3), it is sufficient to verify the triangle inequality since nonnegative, identity, and symmetry are all easy to verify.

To begin with, let us first prove an inequality: for any nonnegative a, b, c, d ,

$$\begin{aligned} &\sqrt{a^2 + b^2 + c^2 + d^2 + 2ac + 2bd} \\ &\leq \sqrt{a^2 + b^2} + \sqrt{c^2 + d^2}. \end{aligned} \quad (33)$$

It is easy to see that inequality (33) is equivalent to the following inequality:

$$ac + bd \leq \sqrt{(a^2 + b^2)(c^2 + d^2)}. \quad (34)$$

Inequality (34) is equivalent to the following inequality:

$$2abcd \leq a^2 d^2 + b^2 c^2. \quad (35)$$

Inequality (35) is equivalent to the following inequality:

$$(ad - bc)^2 \geq 0. \quad (36)$$

Since inequality (36) is always true, inequality (34) is true.

Now, let us assume P, Q, R are 3 arbitrary probability distributions. Since $d_1(P \parallel Q)$ and $d_2(P \parallel Q)$ satisfy the triangle inequality, we have

$$\begin{aligned}
 & \sqrt{(d_1(P \parallel Q))^2 + (d_2(P \parallel Q))^2} \\
 & \leq ([d_1(P \parallel R) + d_1(R \parallel Q)]^2 \\
 & \quad + [d_2(P \parallel R) + d_2(R \parallel Q)]^2)^{1/2} \\
 & = (d_1^2(P \parallel R) + 2d_1(P \parallel R)d_1(R \parallel Q) \\
 & \quad + d_1^2(R \parallel Q) + d_2^2(P \parallel R) + 2d_2(P \parallel R) \\
 & \quad \times d_2(R \parallel Q) + d_2^2(R \parallel Q))^{1/2} \\
 & = ([d_1^2(P \parallel R) + d_2^2(P \parallel R)] \\
 & \quad + [d_1^2(R \parallel Q) + d_2^2(R \parallel Q)] \\
 & \quad + 2d_1(P \parallel R)d_1(R \parallel Q) \\
 & \quad + 2d_2(P \parallel R)d_2(R \parallel Q))^{1/2} \\
 & \leq \sqrt{d_1^2(P \parallel R) + d_2^2(P \parallel R)} \\
 & \quad + \sqrt{d_1^2(R \parallel Q) + d_2^2(R \parallel Q)}.
 \end{aligned} \tag{37}$$

The last inequality results from inequality (33). $\square$

Remark 14. da Costa and Taneja [16] show that $\sqrt{W_s(P \parallel Q)}$ and $\sqrt{V_s(P \parallel Q)}$ are metrics divergence measures for all s . Since $\Delta(P \parallel Q)$, $I(P \parallel Q)$, $h(P \parallel Q)$, $d(P \parallel Q)$, $J(P \parallel Q)$, $T(P \parallel Q)$, and $\psi(P \parallel Q)$ are all constant factors of special cases of $\sqrt{W_s(P \parallel Q)}$ or $\sqrt{V_s(P \parallel Q)}$, they are all metric divergence measures by Proposition 13. Yet, da Costa and Taneja did not disprove or discuss any applications of these divergence measures.

3. Information Value in Credit Scoring

Information value, or IV in short, is a widely used measure in credit scoring in the financial industry. It is a numerical value to quantify the predictive power of an independent continuous variable x in capturing the binary dependent variable y . Mathematically, it is defined as [17]

$$IV = \sum_{i=1}^n \left(\frac{g_i}{g} - \frac{b_i}{b} \right) \times \ln \left(\frac{g_i/g}{b_i/b} \right), \tag{38}$$

where n is the number of bins or groups of variable x , g_i and b_i are the numbers of good and bad accounts with bin i , and g and b are the total number of good accounts and bad accounts in the population. Hence, g_i/g and b_i/b are distributions of good accounts and bad accounts. Therefore,

$$\sum_{i=1}^n \frac{g_i}{g} = \sum_{i=1}^n \frac{b_i}{b} = 1. \tag{39}$$

Usually, “good” means $y = 0$ and “bad” means $y = 1$. It could be the other way, since IV is symmetric about good and bad. If $g_i/g = b_i/b$ for all $i = 1, \dots, n$, then $IV = 0$; that is, x has no information on y .

IV is mainly used to reduce the number of variables as the initial step in the logistic regression, especially in big data with many variables. IV is based on an analysis of each individual predictor in turn without taking into account the other predictors.

3.1. IV and WOE. One advantage of IV is its close tie with weight of evidence (WOE), defined by $\ln((g_i/g)/(b_i/b))$. WOE measures the strength of each grouped attribute in separating good and bad accounts. According to [17], WOE is the log of odds ratio, which measures odds of being good. Moreover, WOE is monotonic and linear.

Yet, WOE is not an accurate measure in that it is not the log of odds ratio and hence its linearity is not guaranteed. Indeed, g_i/g and b_i/b are from two different probability distributions. They represent the number of good accounts in bin i divided by the total number of good accounts in the population and the number of bad accounts in bin i divided by the total number of bad accounts in the population, respectively. In general, $g_i/g + b_i/b \neq 1$ as can be seen from Exhibit 6.2 in [17].

To make WOE a log of odds, let us change its definition to

$$WOE1 = \ln \left(\frac{g_i/n_i}{b_i/n_i} \right) = \ln \left(\frac{g_i}{b_i} \right) = -\ln \left(\frac{b_i}{g_i} \right) \tag{40}$$

and denote it by WOE1. The cancelled $n_i = g_i + b_i$ is the number of accounts in bin i , and so $g_i/n_i + b_i/n_i = 1$.

As is well known, the logistic regression models the log odds, expressed in conditional probabilities, as a linear function of the independent variable; that is,

$$\ln \left(\frac{P(Y = 1 | x)}{P(Y = 0 | x)} \right) = \beta_0 + \beta_1 x. \tag{41}$$

When x falls into bin i , $\ln(P(Y = 1 | x)/P(Y = 0 | x))$ becomes $\ln(b_i/g_i) = -\ln(g_i/b_i)$. Hence, the WOE1 values are either continuously increasing or continuously decreasing in a linear fashion.

IV and WOE1 can be used together to select independent variables for logistic regression. When a continuous variable x has a large IV, we make it a candidate variable for logistic regression if WOE1 values are linear. It is common to plot the WOE1 values versus the mean values of x at bin i .

3.2. A Rule of Thumb of IV. Intuitively, the larger the IV, the more predictive the independent variable. However, if IV is too large, it should be checked for over predicting. For instance, x may be a postknowledge variable.

To quantify IV, a rule of thumb is proposed in [17, 18]:

- (i) less than 0.02: unproductive,
- (ii) 0.02 to 0.1: weak,
- (iii) 0.1 to 0.3: medium,
- (iv) 0.3+: strong.

In addition, mathematical reasoning of the rule of thumb is given in [18]. In more detail, IV can be expressed as the average of 2 likelihood ratio test statistics $G(P, Q)$ and $G(Q, P)$ of Chi-square distributions with $(n - 1)$ degrees of freedom:

$$2 \times IV = 2 \sum_{i=1}^n p_i \ln \left(\frac{p_i}{q_i} \right) + 2 \sum_{i=1}^n q_i \ln \left(\frac{q_i}{p_i} \right) \quad (42)$$

$$= G(P, Q) + G(Q, P).$$

The close relationship between IV and the likelihood ratio test allows using the Chi-square distribution to assign a significance level.

However, this is doubtful. On the one hand, $G(P, Q)$ and $G(Q, P)$ are not necessarily independent. On the other hand, even if they are independent, it is not enough. Let us assume that $2 \times IV$ follows a Chi-square distribution with $2(n - 1)$ degrees of freedom. Yet, the critical values of the Chi-square distribution are too large compared with the values in the rule of thumb, as can be seen from the Chi-square table in many books about Probability, say [19].

We only list the first several rows of Table 1. When Table 1 grows as DF increases, the values in each column will increase. One may use the Excel function $\text{CHIINV}(p, df)$ or its newer and more accurate version $\text{CHISQ.INV.RT}(p, df)$ to build Table 1, which returns the inverse of the right-tailed probability $1 - p$ of the Chi-square distribution with df degrees of freedom.

The critical values are as small as the values in the rule of thumb only when the degrees of freedom are as small as 6. For instance, there is a probability of $1 - 0.005 = 0.995$ that a Chi-square distribution with 6 degrees of freedom will be larger than or equal to 0.68, that is,

$$\text{CHIINV}(0.995, 6) = 0.68. \quad (43)$$

Yet, there is a probability of 0.995 that a Chi-square random variable with 10 degrees of freedom will be larger than or equal to 2.16. There is a probability of 0.995 that a Chi-square random variable with 18 degrees of freedom will be larger than or equal to 6.26. On the basis of the above, the rule of thumb is more or less empirical.

3.3. Calculation of IV. The calculation of IV is simple once binning is done. In this sense, IV is a subjective measure. It depends on how the binning is done and how many bins are used. Different binning methods may result in different IV values, whereas the logistic regression in the later stages will not use the information of these bins.

In practice, 10 or 20 bins are used. The more the bins, the better the chance the good accounts will be separated from the bad accounts. Yet, we cannot divide the values of x indefinitely since we may not avoid 0 good account or 0 bad account in some bins. To overcome the limitation of the logarithm function in the J -divergence, the binning should avoid 0 good account or 0 bad account in any bins.

The idea of binning is to assign values of x with similar behaviors to the same group or bin. In particular, the same values of x must fall into the same bin. A natural way of

binning is to sort the data first and then divide them into n bins with an equal number of observations (the last bin may have less number of observations). This works well if x has no repeating values at all. In reality, x often has repeating values (called the tied values in statistics), which may cause problems when the tied values of x fall into different bins.

Proc Rank in SAS serves, a good candidate for binning (as opposed to function cut in R). When there are no tied values in x , it simply divides the values of x into n bins. When there are tied values in x , it treats the tied values by its option TIES.

Proc Rank begins with sorting the values of x within a BY group. It then assigns each nonmissing value an ordinal number that indicates its rank or position in the sequence. In case of ties, option TIES will be used. Depending on whether TIES = LOW, HIGH, or MEAN (default one), the lowest rank, highest rank, or the average rank will be assigned to all the tied values. Next, the following formula is used to calculate the binning value of each nonmissing value of x :

$$\left\lfloor \frac{\text{rank} \times n}{m + 1} \right\rfloor, \quad (44)$$

where $\lfloor \cdot \rfloor$ is the floor function, rank the value's rank, n the number of bins, and m the number of nonmissing observations. Note that the range of the binning values is from 0 to $n - 1$. Finally, all the values of x are binned according to their binning values. In case one bin has less than 5% of the population, we may combine this bin with its neighboring bin.

To illustrate the use of Proc Rank with $n = 10$ and TIES = MEAN, let us look at an imaginary dataset with one variable age and 100 observations. Assume this dataset has been sorted and has fifty observations with a value of 10, thirty observations with a value of 20, ten observations with a value of 30, nine observations with a value of 40, and one observation with a value of 50.

The first 50 observations have a tied value of 10. Each of them will be assigned an average rank of $k = 25.5$ and hence a binning value of $\lfloor (25.5 \times 10)/101 \rfloor = 2$. The next 30 observations have a tied value of 20. Each of them will be assigned an average rank of $(51 + 52 + \dots + 80)/30 = 65.5$ and hence a binning value of $\lfloor (65.5 \times 10)/101 \rfloor = 6$. The next 10 observations have a tied value of 30. Each of them will be assigned an average rank of $(81 + 82 + \dots + 90)/10 = 85.5$ and hence a binning value of $\lfloor (85.5 \times 10)/101 \rfloor = 8$. The next 9 observations have a tied value of 40. Each of them will be assigned an average rank of $(91 + 92 + \dots + 99)/9 = 95$ and hence a binning value of $\lfloor (95 \times 10)/101 \rfloor = 9$. The last observation has a rank of 100 and hence will be assigned a binning value of $\lfloor (100 \times 10)/101 \rfloor = 9$. In summary, the 100 observations are divided into 4 bins: the first 50 observations, the next 30 observations, the next 10 observations, and the last 10 observations.

Remark 15. Missing values are not ranked and are left missing in Proc Rank. Yet, they may be kept in a separate bin by means of Proc Summary or Proc Means in the calculation of IV.

Remark 16. If x has less than k different values, the number of bins by Proc Rank will be less than k .

TABLE 1: Chi-square table.

DF	$P = 0.005$	0.01	0.025	0.05	0.25	0.5	0.75	0.9	0.95	0.975	0.99
1	0	0	0.001	0.004	0.1	0.45	1.32	2.71	3.84	5.02	6.64
2	0.01	0.02	0.051	0.1	0.58	1.39	2.77	4.61	5.99	7.38	9.21
3	0.072	0.11	0.22	0.35	1.21	2.37	4.11	6.25	7.81	9.35	11.3
4	0.21	0.3	0.48	0.71	1.92	3.36	5.39	7.78	9.49	11.1	13.3
5	0.41	0.55	0.83	1.15	2.67	4.35	6.63	9.24	11.1	12.8	15.1
6	0.68	0.87	1.24	1.64	3.45	5.35	7.84	10.6	12.6	14.4	16.8
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
10	2.16	2.56	3.25	3.94	6.74	9.34	12.5	16.0	18.3	20.5	23.2
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
18	6.26	7.01	8.23	9.39	13.7	17.3	21.6	26.0	28.9	31.5	34.8

After binning is done for x , a simple SAS program can be written to calculate IV. Meanwhile, WOE1 are calculated per bin as for WOE in [17]. If IV is less than 0.02, we will throw this independent variable. If IV is large than 0.3, over predicting will be checked. If IV is between 0.02 and 0.3 and WOE1 are linear, we will include this independent variable as a candidate variable in logistic regression. If IV is between 0.02 and 0.3 but WOE1 are not linear, we may make transformations of the independent variable to make WOE1 more linear. If a transformation can preserve the rank of the original independent variable, the binning by Proc Rank will be preserved. Therefore, we have obtained the following result.

Proposition 17. *IV, when binning by Proc Rank, is invariant under any strictly monotonic transformations.*

3.4. Mathematical Properties of IV. IV is the information statistic for the difference between the information in the good accounts and the information in the bad accounts. Indeed, IV is the J -divergence with distributions of good accounts and bad accounts. Thus, IV is lower bounded by the square of $L_1(P \parallel Q)$ by Theorem 10.

Property 1. $IV \geq (L_1(P \parallel Q))^2$.

Property 2. IV satisfies the inequalities (15); that is,

$$\begin{aligned} \frac{1}{4} \Delta(P \parallel Q) &\leq I(P \parallel Q) \leq h(P \parallel Q) \\ &\leq 4d(P \parallel Q) \leq \frac{IV}{8} \leq T(P \parallel Q) \\ &\leq \frac{1}{16} \psi(P \parallel Q). \end{aligned} \quad (45)$$

In particular, IV is upper bounded by $\psi(P \parallel Q)$:

$$IV \leq \frac{1}{2} \sum_{i=1}^n \frac{(p_i - q_i)^2 (p_i + q_i)}{p_i q_i}. \quad (46)$$

Note that there is a direct proof of the above inequality, which is much easier than that in [14]. Let us assume without loss of

generality that $p_1 \geq q_1, p_2 \geq q_2, \dots, p_k \geq q_k$, and $p_{k+1} < q_{k+1}, \dots, p_n < q_n$.

Proof. Using the identity $p_i/q_i = 1 + ((p_i - q_i)/q_i)$ and making Taylor's expansion of function $\ln(p_i/q_i)$ around 1 for $i = 1, 2, \dots, k$, we obtain

$$\begin{aligned} \ln\left(\frac{p_i}{q_i}\right) &= \ln(1) + \frac{1}{2} \frac{p_i - q_i}{q_i} \\ &\quad - \frac{1}{2(1 + \xi)^2} \left(\frac{p_i - q_i}{q_i}\right)^2 \leq \frac{1}{2} \frac{p_i - q_i}{q_i}. \end{aligned} \quad (47)$$

Multiplying $p_i - q_i \geq 0$ and summing up from 1 to k , we obtain

$$\begin{aligned} \sum_{i=1}^k (p_i - q_i) \ln\left(\frac{p_i}{q_i}\right) &\leq \frac{1}{2} \sum_{i=1}^k \frac{(p_i - q_i)^2}{q_i} \\ &\leq \frac{1}{2} \sum_{i=1}^n \frac{(p_i - q_i)^2}{q_i}. \end{aligned} \quad (48)$$

Similarly, $\ln(q_i/p_i) \leq (1/2)((q_i - p_i)/p_i)$. Multiplying $q_i - p_i \geq 0$ and summing up from k to n , we obtain

$$\begin{aligned} \sum_{i=k+1}^n (q_i - p_i) \ln\left(\frac{q_i}{p_i}\right) &\leq \frac{1}{2} \sum_{i=k+1}^n \frac{(q_i - p_i)^2}{p_i} \\ &\leq \frac{1}{2} \sum_{i=1}^n \frac{(p_i - q_i)^2}{p_i}. \end{aligned} \quad (49)$$

The proof is completed by noting that

$$\begin{aligned} &\sum_{i=1}^n (p_i - q_i) \ln\left(\frac{p_i}{q_i}\right) \\ &= \sum_{i=1}^k (p_i - q_i) \ln\left(\frac{p_i}{q_i}\right) + \sum_{i=k+1}^n (p_i - q_i) \ln\left(\frac{p_i}{q_i}\right) \\ &= \sum_{i=1}^k (p_i - q_i) \ln\left(\frac{p_i}{q_i}\right) + \sum_{i=k+1}^n (q_i - p_i) \ln\left(\frac{q_i}{p_i}\right). \end{aligned} \quad (50)$$

□

TABLE 2: Outlier domination to IV.

Bin	Count	Goods	Distr. good	Bads	Distr. bad	IV contr.
1	9580	7600	99.74%	1980	83.19%	0.03001
2	420	20	0.26%	400	16.81%	0.68814
Total	10000	7620	100.00%	2380	100.00%	0.71815

Theorem 18. *IV is convex in the pair (p, q) .*

Proof. Applying Theorem 2.7.2 from [15] to both $D(P \parallel Q)$ and $D(Q \parallel P)$, we obtain

$$\begin{aligned}
& J(\lambda P_1 + (1 - \lambda) P_2 \parallel \lambda Q_1 + (1 - \lambda) Q_2) \\
&= D(\lambda P_1 + (1 - \lambda) P_2 \parallel \lambda Q_1 + (1 - \lambda) Q_2) \\
&\quad + D(\lambda Q_1 + (1 - \lambda) Q_2 \parallel \lambda P_1 + (1 - \lambda) P_2) \\
&\leq \lambda D(P_1 \parallel Q_1) + (1 - \lambda) D(P_2 \parallel Q_2) \\
&\quad + \lambda D(Q_1 \parallel P_1) + (1 - \lambda) D(Q_2 \parallel P_2) \\
&= \lambda (D(P_1 \parallel Q_1) + D(Q_1 \parallel P_1)) \\
&\quad + (1 - \lambda) (D(P_2 \parallel Q_2) + D(Q_2 \parallel P_2)) \\
&= \lambda J(P_1 \parallel Q_1) + (1 - \lambda) J(P_2 \parallel Q_2).
\end{aligned} \tag{51}$$

□

Remark 19. $\sqrt{J(P \parallel Q)}$ is not convex albeit a metric.

Property 3. If more than 95% of population of x have the same value, then $IV = 0$. In particular, if x has just one value, then $IV = 0$.

Proof. Assume more than 95% of population of x have the same value x_0 . Then, all the population with value x_0 will fall into the same bin, called the majority bin. The rest of population whose values are different from x_0 will be combined into the majority bin. Thus, there will be only one bin for all the values of x . Therefore $g_1 = g, b_1 = b$, and hence $IV = 0$. □

Remark 20. If the population whose values are not x_0 are not combined into the majority bin, then IV could be larger than 0.02. As shown in Table 2, x has 10000 observations, where 95.8% or 9580 observations have the same value, say 2, and the rest of 420 observations have another value, say 4. Both bins contribute a value larger than 0.02 to IV . Statistically, 4.2% of the population are outliers and can be neglected. Hence, it is more meaningful to say x has no information to y .

3.5. Alternatives to IV. As we have seen above, IV has 3 shortcomings: (1) it is not a metric; (2) no groups are allowed to have 0 bad accounts or 0 good accounts; and (3) its range is too broad, from 0 to ∞ .

Theoretically, any divergence measures of the difference or distance between good distributions and bad distributions

FIGURE 1: Logical Trend of WOE1 for Variable Age.

can be alternatives to IV . In particular, $L_p(P \parallel Q)$ ($p \geq 1$) are good alternatives to IV . They overcome all the 3 shortcomings of IV : (1) $L_p(P \parallel Q)$ are all metrics; (2) they allow bins to have 0 bad accounts or 0 good accounts; and (3) They all have a much narrow range, from 0 to 2.

While $L_p(P \parallel Q)$ do not seem to have a tie with weight of evidence, they can be as quantifiable as IV . For instance, we may adopt a rule of thumb for $L_p(P \parallel Q)$:

- (i) weak: 6% of its upper bound,
- (ii) medium: 6% to 30% of its upper bound,
- (iii) strong: Larger than 30% of its upper bound.

In particular,

- (i) weak: 0.12 for $L_1(P \parallel Q)$, 0.085 for $L_2(P \parallel Q)$, and 0.06 for $L_\infty(P \parallel Q)$,
- (ii) medium: 0.12 to 0.60 for $L_1(P \parallel Q)$, 0.085 to 0.424 for $L_2(P \parallel Q)$, and 0.06 to 0.30 for $L_\infty(P \parallel Q)$,
- (iii) strong: 0.60+ for $L_1(P \parallel Q)$, 0.424+ for $L_2(P \parallel Q)$, 0.3+ for $L_\infty(P \parallel Q)$.

Remark 21. If $L_1(P \parallel Q) > 0.12$, then $IV > 0.0144$.

The lower bound 0.12 of $L_1(P \parallel Q)$ can be adjusted as needed. It can also be combined with IV to enhance the accuracy. For instance, for the number of independent variables is large enough, we may select only those which satisfy both lower bounds of $L_1(P \parallel Q)$ and IV .

4. Numerical Results

To illustrate our results, we use Exhibit 6.2 in [17] but add one column for WOE1. We use the real WOE, not its “more user-friendly” form—100 times WOE.

From Table 3, we see that $L_\infty < L_2 < L_1 < \sqrt{IV}$. From Figure 1, we also see that the WOE1 for nonmissing values has a linear trend for variable age.

5. Conclusions

In this paper, we have proposed a class of metric divergence measures, namely, $L_p(P \parallel Q)$, $p \geq 1$, and studied their mathematical properties. We studied information value, an

TABLE 3: Calculation of IV and WOE.

Age	Count	Tot. distr.	Goods	Distr. good	Bads	Distr. bad	WOE	WOE1
Missing	1000	2.5%	860	2.38%	140	3.65%	-0.42719	4.117875
18-22	4000	10%	3040	8.41%	960	25%	-0.10898	1.15268
23-26	6000	15%	4920	13.61%	1080	28.13	-0.72613	1.516347
27-29	9000	22.5%	8100	22.4%	900	23.44%	-0.04526	2.197225
30-35	10000	25%	9500	26.27%	500	13.02%	0.70196	2.944439
35-44	7000	17.5%	6800	18.81%	200	5.21%	1.28388	3.526361
44+	3000	7.5%	2940	8.13%	60	1.56%	1.64934	3.89182
Total	40000	100%	36160	100%	3840	100%		

IV = 0.6681, and $\sqrt{IV} = 0.8174$.

$L_1 = 0.6684$, $L_2 = 0.2987$, and $L_{\infty} = 0.1659$.

important divergence measure widely used in credit scoring. After exploring the mathematical reasoning of a rule of thumb and weight of evidence, we suggested an alternative to weight of evidence. Finally, we proposed using $L_p(P \parallel Q)$ as alternatives to information value to overcome its disadvantages.

References

- [1] I. J. Taneja, "Generalized relative information and information inequalities," *Journal of Inequalities in Pure and Applied Mathematics*, vol. 5, no. 1, pp. 1-19, 2004.
- [2] E. Hellinger, "Neue begründung der theorie der quadratischen formen von unendlich vielen veränderlichen," *Journal Für Die Reine und Angewandte Mathematik*, vol. 136, pp. 210-271, 1909.
- [3] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379-423, 1948.
- [4] S. Kullback and R. A. Leibler, "On information and sufficiency," *Annals of Mathematical Statistics*, vol. 22, pp. 79-86, 1951.
- [5] H. Jeffreys, "An invariant form for the prior probability in estimation problems," *Proceedings of the Royal Society*, vol. 186, pp. 453-461, 1946.
- [6] F. Topsøe, "Some inequalities for information divergence and related measures of discrimination," *Institute of Electrical and Electronics Engineers*, vol. 46, no. 4, pp. 1602-1609, 2000.
- [7] S. S. Dragomir, J. Šunde, and C. Buşe, "New inequalities for Jeffreys divergence measure," *Tamsui Oxford Journal of Mathematical Sciences*, vol. 16, no. 2, pp. 295-309, 2000.
- [8] K. Pearson, "On the criterion that a given system of deviations from the probable in the case of correlated system of variables is such that it can be reasonable supposed to have arisen from random sampling," *Philosophical Magazine*, vol. 50, pp. 157-172, 1992.
- [9] R. Sibson, "Information radius," *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, vol. 14, no. 2, pp. 149-160, 1969.
- [10] J. Burbea and C. R. Rao, "Entropy differential metric, distance and divergence measures in probability spaces: a unified approach," *Journal of Multivariate Analysis*, vol. 12, no. 4, pp. 575-596, 1982.
- [11] J. Burbea and C. R. Rao, "On the convexity of some divergence measures based on entropy functions," *Institute of Electrical and Electronics Engineers*, vol. 28, no. 3, pp. 489-495, 1982.
- [12] I. J. Taneja, "New developments in generalized information measures," in *Advances in Imaging and Electron Physics*, P. W. Hawkes, Ed., vol. 91, pp. 37-136, 1995.
- [13] N. Cressie and T. R. C. Read, "Multinomial goodness-of-fit tests," *Journal of the Royal Statistical Society B*, vol. 46, no. 3, pp. 440-464, 1984.
- [14] I. J. Taneja, "Generalized Symmetric Divergence Measures and Inequalities," *RGMA Research Report Collection*, vol. 7, no. 4, 2004.
- [15] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, John Wiley & Sons, New York, NY, USA, 1991.
- [16] G. A. T. F. da Costa and I. J. Taneja, *Generalized Symmetric Divergence Measures and Metric Spaces*, Computing Research Repository, 2011.
- [17] N. Siddiqi, *Credit Risk Scorecards—Developing and Implementing Intelligent Credit Scoring*, John Wiley & Sons, 2006.
- [18] N. Siddiqi, *Credit Risk Scorecards—Development and Implementation Using SAS*, LULU, 2011.
- [19] D. Downing and J. Clark, *Barron's E-Z Statistics*, Barron's Educational Series, 2009.