

A Practical Measure of the Uncertainty Level in a Data Set

Hao Yin, Research Assistant
Elsayed A. Orady, Yubao Chen, Chia-hao Chang
Professors
Department of Industrial and Manufacturing System Engineering
University of Michigan – Dearborn

ABSTRACT

Data mining is a process of knowledge extraction from databases that can be used in decision-making. However, a data mining method that performs well on artificial data sets does not necessarily work adequately on realistic data sets. This is because the uncertainty associated with these data sets usually has a great effect on the test results. When the uncertainty level inherited in a data set is high, most data mining techniques become ineffective. The uncertainty associated with a data set is usually a result of inconsistency, error, incompleteness and redundancy.

Since inconsistency of a data set is the most common source of uncertainty, this paper presents a study on the uncertainty related to the inconsistency of the data. A new practical measure, referred to hereafter as adjCI, is proposed to quantify this kind of uncertainty. The adjCI determines the certainty for a data set. This measure is relatively simple to calculate and interpret. When a data set has no certainty, the measure adjCI is equal to zero (0). On the other hand, when a data set has full certainty, the measure adjCI is equal to one (1). The higher the value of the measure is, the higher the certainty of the data set. These properties of adjCI are proved and application issues are discussed. Several examples are also provided to demonstrate how this measure can be utilized with real data sets.

1. INTRODUCTION

Data mining, known as Knowledge Discovery in Database (KDD), is a process of knowledge extraction from databases that can be used in decision-making. Data classification, one of data mining functionalities, is the process that sorts out the common properties among a set of objects in a database and classifies them into different classes. Many researchers have been working on the data mining methods and data classification. However, it is known that a data mining method that performs well on artificial data sets does not necessarily work on realistic data sets. This is because the uncertainties associated with these data sets. Uncertainty usually has a great effect on the test results. When uncertainty level inherent in a data set is high, related data mining techniques become ineffective. Therefore, there is a need to study the uncertainty of a data set systematically and quantitatively so that the relationships existing in performance of a data mining method and uncertainty of data sets can be determined.

Entropy theory [Shannon, 1948] can be used for uncertainty measurement. This method has gained great success in the effort to develop a practical uncertainty

evaluation procedure [Quinlan, 1986, Quinlan, 1993], [Chen, Y., et al., 1999]. However, they use entropy theory only to determine the quantitative value of the uncertainty carried by each attribute in a data set, but not the whole data set.

This paper presents a measure for uncertainty measurement of a data set caused by inconsistency, which is the most common source of uncertainty in a data set. Hereafter, a practical quantitative measure, referred to as Adjusted Certainty Indicator (adjCI), is proposed to determine the uncertainty of a data set. The uncertainty and inconsistency are defined and a practical algorithm for the measurement is described, followed by some illustrating examples and concluding remarks.

2. UNCERTAINTY BY INCONSISTENCY

Uncertainty can be categorized into two terms: (a) vagueness and (b) ambiguity. In general, vagueness is associated with the difficulty of making sharp or precise distinctions, while ambiguity is associated with more alternative choices. The ambiguity of a data set is usually caused by error, inconsistency, incompleteness and redundancy. The available data in many practical

situations are often inconsistent from the point of view of data mining, and the fact that data have been collected and organized around the needs of organizational activities also causes inconsistency in the data. This paper concentrates on the uncertainty caused by inconsistency. Inconsistency means that in a data set, some observations with the same attribute values have different class labels, which makes it difficult to determine which observation is correct and can be used as a rule in future decision making. Moreover, in this study, the uncertainty denotes the uncertainty of a data set, instead of the uncertainty of attributes.

The Adjusted Certainty Indicator (*adjCI*) is a quantitative measure proposed to determine the certainty of a data set.

The reverse of uncertainty used here is to make the measure of *adjCI* more understandable. When a data set has no certainty, $adjCI = 0$; when a data set has full certainty, $adjCI = 1$. The higher this measure is, the higher the certainty of the data set. The concepts of certainty are illustrated in the following sections.

In order to interpret the concepts of certainty clearly, a small data set in Table 1, adapted from [Quinlan, 1986], is introduced. In this data set, each observation has four attributes and a mutually exclusive class label, which is decided by the values of the attributes. In the example presented in Table 1, *P* means that the Saturday morning, which has certain weather, is suitable for playing golf, and *N* means not suitable for playing golf.

Table 1 The Data Set of Play and Don't Play

Observation	Attributes				Class
	Outlook	Temperature	Humidity	Windy	
1	Sunny	Hot	High	False	N
2	Sunny	Hot	High	True	N
3	Overcast	Hot	High	False	P
4	Rain	Mild	High	False	P
5	Rain	Cool	Normal	False	P
6	Rain	Cool	Normal	True	N
7	Overcast	Cool	Normal	True	P
8	Sunny	Mild	High	False	N
9	Sunny	Cool	Normal	False	P
10	Rain	Mild	Normal	False	P
11	Sunny	Mild	Normal	True	P
12	Overcast	Mild	High	True	P
13	Overcast	Hot	Normal	False	P
14	Rain	Mild	High	True	N

This small data set has full certainty, because no inconsistent observations exist. However, for the assumed data set in Table 2, there is inconsistency between two observations because they have same attribute values, {sunny, hot, high and false}, but different class labels. Moreover, no conclusive decision can be reached because

the occurrence probability of class label *N* or *P* is 50% for both under this condition. Therefore, this data set has no certainty. In reality, most data sets fall some where between these two extreme cases. A data set closer to the first case is preferred in data mining, because it possesses more certainty.

Table 2 An Assumed Data Set

Observation	Attributes				Class
	Outlook	Temperature	Humidity	Windy	
1	Sunny	Hot	High	False	N
2	Sunny	Hot	High	False	P

3. STEPS OF DETERMINING THE ADJUSTED CERTAINTY INDICATOR (*adjCI*)

The adjusted certainty indicator (*adjCI*) is a quantitative measure proposed to determine the certainty of a data set caused by inconsistency. The steps of generating *adjCI* are

described below:

- (1) Construct the contingency table for inconsistent observations in a data set, named the contingency table for an inconsistent subset.

- (2) Calculate the *adjCI* of a data set subject to the functions (2) and (3), presented hereafter in Section 3.2.

3.1 Contingency Table for an Inconsistent Subset

Contingency table is proposed to calculate the uncertainty of attributes in EBMV algorithm [Chen, Y., et al., 1990, Chen, Y., et al., 1999]. In this paper, contingency table is also used to evaluate the certainty of a data set, but it is constructed for a data set, instead of attributes. In order to construct a contingency table for a data set, it is necessary to describe the concept of two-way contingency table.

Two-Way Contingency Table

Assume a single population of interest is available, and each individual in the population categorized with respect to two different factors *a* and *b*. There are *I* and *J* categories associated with the two factors *a* and *b*, respectively. When a single sample is collected, the number of individuals belonging to both category *i* of factor *a* and category *j* of factor *b* is entered in the cell in row *i* and *j* column *j* where $i = 1, \dots, I; j = 1, \dots, J$, as shown in Table 3.

Table 3 Two-Way Contingency Table

	b_1	b_2	...	b_j	...	b_J
a_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1J}
:						
a_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{iJ}
:						
a_I	n_{I1}	n_{I2}	...	n_{Ij}	...	n_{IJ}

Where a_i ($i = 1, \dots, I$) represents the categories of factor *a*; b_j ($j = 1, \dots, J$) represents the categories of factor *b*; and n_{ij} denotes the number of individuals in the sample falling in the (*i*, *j*)th cell of the table.

The Contingency Table for an Inconsistent Subset

The observations of a data set are classified into two categories: consistent observations and inconsistent

observations. Inconsistent observations have the same values of attributes but different classes. Assume that N_I inconsistent observations are selected from total *N* observations of a data set to form an inconsistent subset. Then, based on the concept of two-way contingency table, a contingency table for inconsistent subset is constructed as Table 4.

Table 4 The Contingency Table for an Inconsistent Subset

	CO_1	CO_2	...	CO_j	...	CO_{NR}
C_1	n_{11}	n_{12}	...	n_{1j}	...	n_{1NR}
:						
C_i	n_{i1}	n_{i2}	...	n_{ij}	...	n_{iNR}
:						
C_{NC}	n_{NC1}	n_{NC2}	...	n_{NCj}	...	n_{NCNR}
	N_1	N_2	...	N_j	...	N_{NR}

Where CO_j ($j = 1, 2, \dots, NR$) denote combinations of different attribute values and C_i ($i = 1, 2, \dots, NC$) represents all classes of the subset. n_{ij} denotes the number of individuals in the subset falling in the (C_i, CO_j), and

$$N_j = \sum_{i=1}^{NC} n_{ij} \quad (j = 1, 2, \dots, NR).$$

For example, following data set in Table 5 is assumed with only two attributes Humidity and Windy, and five observations.

Table 5 An Assumed Data Set

Observation	Humidity	Windy	Class
1	High	False	N
2	High	False	N
3	Normal	True	P
4	Normal	True	P
5	Normal	True	N

Three inconsistent observations are selected from the data set above and form the subset shown in Table 6.

Table 6 An Inconsistent Subset

Observation	Humidity	Windy	Class
3	Normal	True	P
4	Normal	True	P
5	Normal	True	N

Therefore, the contingency table for this subset is presented in Table 7 as follows:

Table 7 The Contingency Table for an Inconsistent Subset

Class	{Normal, True}
N	1
P	2
	3

When the Humidity is Normal, Windy is True, there are 1 observation belonging to class *N*, while there are 2 observations belonging to class *P*. The total number of observations in the inconsistent subset is 3 with this combination of attribute value.

3.2 Certainty Indicator (CI) and adjusted Certainty Indicator (adjCI)

Based on Table 4, certainty indicator (*CI*) of a data set is defined as:

$$\begin{aligned}
 CI = & \left(\frac{n_{11}}{N} \times \frac{n_{11}}{N_1} + \dots + \frac{n_{i1}}{N} \times \frac{n_{i1}}{N_1} + \dots + \frac{n_{NC1}}{N} \times \frac{n_{NC1}}{N_1} \right) \\
 & + \dots + \left(\frac{n_{1j}}{N} \times \frac{n_{1j}}{N_j} + \dots + \frac{n_{ij}}{N} \times \frac{n_{ij}}{N_j} + \dots + \frac{n_{NCj}}{N} \times \frac{n_{NCj}}{N_j} \right) + \dots \\
 & + \left(\frac{n_{1NR}}{N} \times \frac{n_{1NR}}{N_{NR}} + \dots + \frac{n_{iNR}}{N} \times \frac{n_{iNR}}{N_{NR}} + \dots + \frac{n_{NCNR}}{N} \times \frac{n_{NCNR}}{N_{NR}} \right) \quad (1) \\
 & + \frac{N - N_I}{N}
 \end{aligned}$$

$$CI = \frac{1}{N} \sum_{j=1}^{NR} \left(\sum_{i=1}^{NC} \frac{n_{ij}^2}{N_j} \right) + \frac{N - N_I}{N} \quad (2)$$

Where N = Number of all observations in a data set.

N_I = Number of all inconsistent observations in subset.

The adjusted certainty indicator can then be defined as:

$$adjCI = \frac{CI - \frac{1}{NC}}{1 - \frac{1}{NC}} = \frac{CI \times NC - 1}{NC - 1} \quad (3)$$

Where NC = Number of distinct class labels in contingency table for inconsistent subset.

As an example, if equations (2) and (3) are applied to the data set in Table 5, CI and $adjCI$ are determined as follows:

$$CI = \frac{1}{5} \times \left(\frac{1^2}{3} + \frac{2^2}{3} \right) + \frac{5-3}{5} = 0.7333$$

$$adjCI = \frac{0.7333 - \frac{1}{2}}{1 - \frac{1}{2}} = 0.4666$$

Which shows that this data set has a certainty of 0.4666 $adjCI$

3.3 Explanation of CI

Referred to the data set in Table 5, at most five rules can be derived before redundancies are revealed, because there are

5 cases in the data set. For each rule, the probability of occurrence is $1/5$.

- Rule 1: IF attribute values are {high, false}, THEN class is N ;
- Rule 2: IF attribute values are {high, false}, THEN class is N ;
- Rule 3: IF attribute values are {normal, true}, THEN class is P ;
- Rule 4: IF attribute values are {normal, true}, THEN class is P ;
- Rule 5: IF attribute values are {normal, true}, THEN class is N ;

Rule 1 has 100% correct rate because according to this rule, when the attribute values are {high, false}, the class must be N without doubt, as well as Rule 2. However, when the attribute values are {normal, true}, class is P according to Rule 3 and Rule 4, or class is N according to

Rule 5. Therefore, Rule 3 and Rule 4 have $2/3$ correct rate, respectively, and Rule 5 has only $1/3$ correct rate.

Thus, the CI of a data set can be defined as:

$$CI = \sum_{i=1}^n (OR_i \times CR_i) \quad i = 1, 2, \dots, n \quad (4)$$

Where OR_i is the probability of occurrence of a rule; CR_i is the correct rate of corresponding rule; and n is the number

of all rules. Therefore, the CI for the data set in Table 5 can be calculated as

$$\begin{aligned} CI &= \sum_{i=1}^5 OR_i \times CR_i \\ &= \left(\frac{1}{5} \times \frac{2}{3} + \frac{1}{5} \times \frac{2}{3} + \frac{1}{5} \times \frac{1}{3} \right) + \left(\frac{1}{5} \times 1 + \frac{1}{5} \times 1 \right) \\ &= \left(\frac{1}{5} \times \frac{2}{3} + \frac{1}{5} \times \frac{2}{3} + \frac{1}{5} \times \frac{1}{3} \right) + \frac{5-3}{5} \\ &= 0.7333 \end{aligned}$$

This result is the same as that obtained using equation (2).

The rules with 100% correct rate come from consistent observations in the data set, so $\sum OR_i \times CR_i$ of this part

$$\text{of the rules} = (N - N_l) \times \frac{1}{N} \times 1 = \frac{N - N_l}{N} .$$

Meanwhile, based on Table 4, OR_i of rule i , which come

from the inconsistent observation i , equals to $\frac{n_{ij}}{N}$ and CR_i

equals to $\frac{n_{ij}}{N_j}$, so $\sum OR_i \times CR_i$ of the rules coming

from inconsistent observations = $\sum \frac{n_{ij}}{N} \times \frac{n_{ij}}{N_j}$. The sum

of these two parts is the same as equation (2). Equation (3) is to scale the output of CI to a fixed range of value [0,1]. Therefore, $adjCI$ is used as a measure to determine the certainty of a data set.

$$CI = \frac{1}{N} \sum_{j=1}^{NR} \left(\sum_{i=1}^{NC} \frac{n_{ij}^2}{N_j} \right) + \frac{N - N_I}{N} = \frac{1}{N} \times 0 + \frac{N - 0}{N} = 1$$

$$adjCI = \frac{CI \times NC - 1}{NC - 1} = 1$$

3.4.2 $adjCI$ of a data set with no certainty

When all observations in a data set are inconsistent, i.e. $N=N_I$, and classes (C_i) in each attribute values combination

3.4 Further Study of $adjCI$

3.4.1 $adjCI$ of a data set with full certainty

When no inconsistent observations exist in a data set, the data set has full certainty and the $adjCI$ is 1. This is because n_{ij} in a contingency table of inconsistent subset is equal to 0 and

(CO_j) are all with uniform distributions, this data set has no certainty and $adjCI$ is 0, because if equation (2) is applied:

$$\begin{aligned} \sum_{i=1}^{NC} \frac{n_{ij}^2}{N_j} &= \frac{n_{1j}}{N} \times \frac{n_{1j}}{N_j} + \dots + \frac{n_{ij}}{N} \times \frac{n_{ij}}{N_j} + \dots + \frac{n_{NCj}}{N} \times \frac{n_{NCj}}{N_j} \\ &= \frac{1}{N \times N_j} (n_{1j}^2 + \dots + n_{ij}^2 + \dots + n_{NCj}^2) \\ &= \frac{NC}{N \times N_j} \left(\frac{n_{1j}^2 + \dots + n_{ij}^2 + \dots + n_{NCj}^2}{NC} \right) \\ &\geq \frac{NC}{N \times N_j} \left(\frac{n_{1j} + \dots + n_{ij} + \dots + n_{NCj}}{NC} \right)^2 \\ &= \frac{NC}{N \times N_j} \times \frac{N_j^2}{NC^2} \\ &= \frac{N_j}{N \times NC} \end{aligned}$$

Where only when $n_{11} = n_{21} = \dots = n_{i1} = \dots = n_{NC1}$, the equal signs come into existence.

Therefore,

$$CI = \frac{1}{N} \sum_{j=1}^{NR} \left(\sum_{i=1}^{NC} \frac{n_{ij}^2}{N_j} \right) + \frac{N - N_I}{N}$$

$$\begin{aligned}
&\geq \frac{N_1}{N \times NC} + \dots + \frac{N_j}{N \times NC} + \dots + \frac{N_{NR}}{N \times NC} + \frac{N - N}{N} \\
&= \frac{N}{N \times NC} \\
&= \frac{1}{NC}
\end{aligned}$$

Then,

$$adjCI = \frac{\frac{1}{NC} \times NC - 1}{NC - 1} = 0$$

Where, only when $n_{1j} = n_{2j} = \dots = n_{ij} = \dots = n_{NCj}$ ($j = 1, 2, \dots, NR$), the equal signs come to existence. The

contingency table for a data set with no certainty is illustrated below:

Table 8 The Contingency Table for a Data Set with No Certainty

	CO_1	CO_2	...	CO_j	...	CO_{NR}
C_1	n_{11}	n_{12}	...	N_{1j}	...	n_{1NR}
C_2	n_{21}	n_{22}	...	N_{2j}	...	n_{2NR}
:						
C_i	n_{i1}	n_{i2}	...	N_{ij}	...	n_{iNR}
:						
C_l	n_{l1}	n_{l2}	...	N_{lj}	...	n_{lNR}

In this condition, no conclusive decision can be reached, even though the observations are given.

3.4.3 *adjCI* of A Data Set With General Certainty

The degree of certainty for a data set, i.e. the scale of *adjCI*, is related to the framework of contingency table for inconsistent subset. Moreover, two effects determine it: (a) the number of inconsistent observations and (b) their distribution in the contingency table. The effectiveness of these two factors is explained below using simple examples.

(a) The number of inconsistent observations.

Assume that combination $CO_k = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$ is converted to $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$

in the contingency table of an inconsistent subset of a data set while the other conditions does not change, that is, only

the number of observations involved in inconsistent is reduced while the distribution of class labels, number of observations in data set do not change. The certainty level of the data set will increase for the change, i.e. *adjCI* of the data set will rise, because

$$\begin{aligned}
\Delta CI &= CI_1 - CI_0 \\
&= \left(\frac{1}{N} \sum_{i=1}^{NC} \frac{n_{ik1}^2}{N_{k1}} + \frac{N - N_{I1}}{N} \right) - \left(\frac{1}{N} \sum_{i=1}^{NC} \frac{n_{ik}^2}{N_k} + \frac{N - N_I}{N} \right) \\
&= \left(\frac{1^2}{2N} + \frac{1^2}{2N} - \frac{N_I - 2}{N} \right) - \left(\frac{2^2}{4N} + \frac{2^2}{4N} - \frac{N_I}{N} \right) \\
&= -\frac{1}{N} + \frac{2}{N} \\
&= \frac{1}{N} > 0
\end{aligned}$$

$$\Delta adjCI = \frac{CI_1 \times NC - 1}{NC - 1} - \frac{CI_0 \times NC - 1}{NC - 1} = \frac{(CI_1 - CI_0) \times NC}{NC - 1} > 0$$

(b) The distribution of inconsistent observations in the subset.

If the combination $CO_k = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$ is converted to be $\begin{bmatrix} 3 \\ 1 \end{bmatrix}$, that is, the distribution of inconsistent observation is changed

while the other conditions do not change. The certainty level of the data set will increase, i.e. $adjCI$ will rise, because

$$\begin{aligned}
\Delta CI &= CI_1 - CI_0 \\
&= \left(\frac{3^2}{N4} + \frac{1^2}{N4} \right) - \left(\frac{2^2}{N4} + \frac{2^2}{N4} \right) \\
&= \frac{1}{2N} > 0
\end{aligned}$$

$$\Delta adjCI = \frac{CI_1 \times NC - 1}{NC - 1} - \frac{CI_0 \times NC - 1}{NC - 1} = \frac{(CI_1 - CI_0) \times NC}{NC - 1} > 0$$

Therefore, $adjCI$ can represent the change of certainty level of a data set under the influence of the two effects: the number of inconsistent observations and their distribution in inconsistent subset. When less inconsistent observations are involved in a data set, $adjCI$ is higher, i.e. the data set has higher certainty level. When the distribution of inconsistent observations is closer to uniform distribution, $adjCI$ is lower, i.e. the data set has lower certainty level.
More Examples

In order to demonstrate how $adjCI$ can be utilized to quantify the uncertainty of a data set, several test data sets are created from a data set in industry. The data set, named car evaluation data set, is from University of California, Irvine KDD Database Repository, the most popular site for data sets used for research in machine learning and knowledge discovery. Car evaluation data set derived from a hierarchical decision model originally developed for the demonstration of DEX (M. Bohanec, V. Rajkovic, 1990). Cars are evaluated according to the following concept structure:

- CAR
 - PRICE overall price
 - Buying buying price
 - Maint price of the maintenance
 - TECH technical characteristics
 - COMFORT comfort
 - Doors number of doors
 - Persons capacity in terms of persons to carry
 - Lug_boot the size of luggage boot
 - Safety estimated safety of the car

That is, the value of a car is determined by six attributes: Buying, Maint, Doors, Persons, Lug_boot and Safety. The attributes' values are presented in Table 9:

Table 9 The Values of Attributes

Attributes	Values
Buying	V-high, high, med, low
Maint	V-high, high, med, low
Doors	2, 3, 4, 5-more
Persons	2, 4, more
Lug_boot	Small, med, big
Safety	Low, med, high

and the distribution of a class, car value, (number of cases per class) is like:

Table 10 The Distribution of Class Labels

Class	N	N(%)
Unacceptable	1210	(70.023%)
Acceptable	384	(22.222%)
Good	69	(3.993%)
Very good	65	(3.762%)

In this database, there are 1728 observations, and none missing attribute value. Assume that 20 observations from the data set is selected to form the test data set 1 as the following:

Table 11 Test Data Set 1

Obs	Attributes						Class
	Price	Maint	Doors	Persons	Lug_boot	Safety	
1	low	med	5more	4	small	med	acc
2	low	med	5more	4	small	high	good
3	low	med	5more	4	med	low	unacc
4	low	med	5more	4	med	med	good
5	low	med	5more	4	med	high	vgood
6	low	med	5more	4	big	low	unacc
7	low	med	5more	4	big	med	good
8	low	med	5more	4	big	high	vgood
9	low	med	5more	more	small	low	unacc

10	low	med	5more	more	small	med	acc
11	low	med	5more	more	small	high	good
12	low	med	5more	more	med	low	unacc
13	low	med	5more	more	med	med	good
14	low	med	5more	more	med	high	vgood
15	low	med	5more	more	big	low	unacc
16	low	med	5more	more	big	med	good
17	low	med	5more	more	big	high	vgood
18	low	low	2	2	small	low	unacc
19	low	low	2	2	small	med	unacc
20	low	low	2	2	small	high	unacc
21	low	low	2	2	med	low	unacc
22	low	low	2	2	med	med	unacc
23	low	low	2	2	med	high	unacc
24	low	low	2	2	big	low	unacc
25	low	low	2	2	big	med	unacc

Since there are no inconsistent observations in this data set, the *adjCI* of this data set is 1, i.e., test data set 1 has full certainty. If an observation as the following is added to the

test data set 1 to form test data set 2, which has 26 observations and 2 of them are inconsistent with each other.

Obs	Attributes						Class
	Price	Maint	Doors	Persons	Lug_boot	Safety	
26	low	low	2	2	big	med	acc

Then the contingency table for inconsistent observations of the test data set 2 is as below.

Table 12 The Contingency Table for an Inconsistent Subset

Class	{low, low, 2, 2, big, med}
acc	1
unacc	1
	2

Applying function (2) and (3), then

$$CI = \frac{1}{26} \times \left(\frac{1^2}{2} + \frac{1^2}{2} \right) + \frac{26-2}{26} = 0.9615$$

and

$$adjCI = \frac{0.9615 - \frac{1}{4}}{1 - \frac{1}{4}} = 0.9487$$

Which imply that the test data set 2 has certainty with 0.9487 *adjCI*.

If 3 more observations as the following are added into the test data set 2 to form test data set 3, which has 29 observations and 6 of them are inconsistent.

Obs	Attributes						Class
	Price	Maint	Doors	Persons	Lug_boot	Safety	
27	low	med	5more	more	big	med	vgood
28	low	med	5more	more	big	med	vgood
29	low	med	5more	more	big	med	acc

Then the contingency table for inconsistent observations of the data set is

Table 13 The Contingency Table for an Inconsistent Subset

Class	{low, low, 2, 2, big, med}	{low, low, 2, 2, big, med}
acc	1	1
unacc	1	
vgood		2
good		1
	2	4

Applying function (2) and (3), then

$$CI = \frac{1}{29} \times \left(\left(\frac{1^2}{2} + \frac{1^2}{2} \right) + \left(\frac{1^2}{4} + \frac{2^2}{4} + \frac{1^2}{4} \right) \right) + \frac{29-6}{29} = 0.8793$$

and

$$adjCI = \frac{0.8793 - \frac{1}{4}}{1 - \frac{1}{4}} = 0.8091$$

Which means that the test data set 3 has a certainty of 0.8391 *adjCI*. Thus, it is clear that as the number of inconsistent data increases *adjCI* decreases.

4. CONCLUSIONS

In this paper, the uncertainty of a data set caused by inconsistency is defined and quantified by a practical measure, referred to as adjusted Certainty Indicator (*adjCI*). When a data set has no certainty, *adjCI* is equal to zero (0); and when a data set has full certainty, *adjCI* is equal to (1). The higher this measure is, the more certainty the data set has.

This technique is applied to a practical data set. The results showed that *adjCI* represents mainly the change of certainty level of a data set under the influences of the two effects: the number of inconsistent observations and their distribution in the data set. If less inconsistent observations

are involved in a data set, *adjCI* is higher, i.e. the data set has a higher certainty level. In addition, if the distribution of inconsistent observations is closer to the uniform distribution, *adjCI* is lower, i.e. the data set has a lower certainty level. The simplicity of the proposed measure makes it very practical to evaluate the uncertainty level of a data set before further data processing is made, which, therefore, would enhance the data mining results.

5. REFERENCES

- Chen, M.-S., Han, J., and Yu, P. S., 1996, "Data Mining: An Overview from Database Perspective," *IEEE Trans. on Knowledge and Data Engineering*, Vol. 8, No. 6, pp. 866-883.
- Chen, Y., Li, X., and Orady, E., 1996, "Integrated Diagnosis using Information-Gain-Weighted Radial Basis Function Neural Networks," *Computers in Industrial Engineering*, Vol. 30, No. 2, pp. 243-255.

Chen, Y., Orady, E., 1999, "An Entropy-Based Index Evaluation Scheme for Multiple Sensor Fusion in Classification Process," *ASME Transactions, Journal of Manufacturing Science and Engineering*, Vol. 121, pp 727-732.

Chen, Y., Sha, J. L., and Wu, S. M., 1990, "Diagnosis of the Tapping Process by Information Measure and Probability Voting Approach," *ASME Transactions, Journal of Engineering for Industry*, Vol. 112, pp. 319-324.

Fu, L. M., 1988, "Truth Maintenance under Uncertainty," *In Process 4th Workshop. Uncertainty in Artificial Intelligent*, Minneapolis, MN: AAAI, pp. 119-126.

Michalski, R. S., 1987, "Learning Strategies and Automated Knowledge Acquisition: An Overview," *Computational Models and Learning*, Lonard Bolc Editor, Springer-Verlag, pp. 1-19.

Quinlan, J. R., 1986, "Induction of Decision Trees," *Machine Learning*, Vol. 1, pp. 81-106.

Quinlan, J. R., 1993, *C4.5: Programs for Machine Learning*, Morgan Kaufmann.

Quinlan, J. R., 1994, "Comparing Connectionist and Symbolic Learning Methods," *Computational Learning Theory and Natural Learning Systems, Volume I: constraints and prospects*, pp. 445-456.

Shannon, C. E., 1948, "A Mathematical Theory of Communication," *Bell System Tch. J.*, Vol. 27, pp. 379-423.