Information-Theoretic Measures for Knowledge Discovery and Data Mining

Y.Y. Yao

Department of Computer Science
University of Regina
Regina, Saskatchewan
Canada S4S 0A2
E-mail: yyao@cs.uregina.ca
URL: http://www.cs.uregina.ca/∼yyao

Abstract. A database may be considered as a statistical population, and an attribute as a statistical variable taking values from its domain. One can carry out statistical and information-theoretic analysis on a database. Based on the attribute values, a database can be partitioned into smaller populations. An attribute is deemed important if it partitions the database such that previously unknown regularities and patterns are observable. Many information-theoretic measures have been proposed and applied to quantify the importance of attributes and relationships between attributes in various fields. In the context of knowledge discovery and data mining (KDD), we present a critical review and analysis of information-theoretic measures of attribute importance and attribute association, with emphasis on their interpretations and connections.

1 Introduction

Information-theoretic measures have been used in many fields for measuring importance of attributes and relationships between attributes [10,14,55], such as pattern recognition [6], multi-attribute decision making [19,67], machine learning [42], information retrieval [45,59,52], and data mining [60,64]. Watanabe [56] suggested that pattern recognition is essentially a conceptual adaptation to the empirical data in order to see a form in them. The form is interpreted as a structure which always entails a small entropy value. Many algorithms in pattern recognition may be characterized as efforts to minimize entropy [54,56]. The philosophy of entropy minimization for pattern recognition can be applied to related fields, such as classification, data analysis, machine learning, and data mining, where one of the tasks is to discover patterns or regularities in a large data set. Regularities and structuredness are characterized by small entropy values, whereas randomness is characterized by large entropy values.

A database consists of a set of objects represented by their values on a set of attributes. Each attribute describes an object by using a value from an associated set called the domain of the attribute [35]. Statistical and information-theoretic analysis for knowledge discovery and data mining (KDD) treats a

database as a statistical population, and an attribute as a statistical variable taking values from its domain [18,35]. Some fundamental tasks of KDD involve the discovery of relationships between attributes. For this purpose, one can immediately use information-theoretic measures. Lee [28] and Malvestuto [35] provided a systematic information-theoretic analysis of databases. They investigated the issues of correlation and interdependency among attributes. The notions such as functional, multi-valued, hierarchical, and join dependencies are stated in terms of various entropy functions. A related partition model of databases was studied by Spyratos [51].

A database can be partitioned into smaller populations based on the values of an attribute. An attribute is deemed important for data mining if regularities are observable in smaller populations, while being unobservable in a larger population. Regularities are expressed by lower entropy values. This suggests that if an attribute is useful for data mining, then the attribute should lead to entropy reduction. The well known ID3 inductive learning algorithm uses exactly such a measure for attribute selection in a learning process [42]. The entropy reduction is the difference between the entropy of the decision attribute and the conditional entropy of the decision attribute given a particular attribute. It is in fact the mutual information between the decision attribute and the given attribute. Other entropy-related measures have also been proposed and studied [25,34,53].

Potential success of information-theoretic analysis for KDD depends on, to a large extent, the interpretations of different information-theoretic measures and their connections. Based on the philosophy of entropy minimization and our earlier investigations on the topic [60,64], we review and examine information-theoretic measures for evaluating attribute importance and attribute association. The objective is to provide a systematic analysis of information-theoretic measures in the setting of KDD. Measures that have been used successfully in related fields, but have not been used in KDD, are discussed. Four classes of measures are identified. They are measures of attribute importance, measures of one-way attribute association, measures of two-way attribute association, and measures of dissimilarity and similarity of populations. Each type of measures captures a particular aspect of attribute importance for KDD. Different types of measures can be combined or used in various stages in a data mining process.

The rest of the article is organized as follows. Section 2 provides an overview of some basic issues of KDD using the notion of information tables [62]. A database is viewed as an information table. Section 3 is a brief review of information-theoretic measures for information tables. We focus on two special forms of entropy related measures. One is expressed in terms of Kullback-Leibler divergence measure [27], and the other is expressed in terms of expected values [10,11,14,48]. They offer useful interpretations of information-theoretic measures for KDD. Section 4 examines and classifies various information-theoretic measures used in KDD and related fields.

2 Analysis of Information Tables

In many information processing systems, a set of objects are typically represented by their values on a finite set of attributes. Such information may be conveniently described in a tabular form. Each column corresponds to an attribute and each row corresponds to an object. A cell, defined by a pair of object and attribute, gives the value of the object on the attribute. Formally, an information table is defined by a quadruple:

$$T = (U, At, \{V_X \mid X \in At\}, \{I_X \mid X \in At\}), \tag{1}$$

where

U is a finite and nonempty set of objects,
At is a finite and nonempty set of attributes,
V_X is a nonempty set of values for each attribute $X \in At$,
$I_X : U \longrightarrow V_X$ is an information function for each attribute $X \in At$.

An information table represents all available information and knowledge about the objects under consideration. Objects are perceived, observed, or measured based on only a finite number of properties. For simplicity, we assume that the domain of each attribute is finite. An information function I_X is a total function that maps an object t of U to one value in V_X. For an object $t \in U$, $I_X(t)$ is the value of t on the attribute X. A database may be viewed as an example of information tables. Additional information and applications of information tables can be found in [38,62,65].

One can extend information functions to subsets of attributes. For $Y \subseteq At$, its domain V_Y is the Cartesian product of the domains of all individual attributes in the set. The symbol $I_Y(t)$ is the value of t on a set of attributes Y, which is a vector of individual attribute values. A single attribute is equivalent to a singleton subset of At. In subsequent discussions, we will use $X, Y, \ldots$ to denote sets of attributes, and $x, y, \ldots$ to denote the values in the domain of $X, Y, \ldots$. We will also use "an attribute" and "a set of attributes" interchangeably.

With respect to the notion of information tables, there are extensive studies on the relationships between values of *different* attributes and relationships between values of the *same* attribute. Studies of the two kinds of relationships correspond to the *horizontal* analysis and the *vertical* analysis of an information table [62].

Analysis of horizontal relationships reveals the similarity, association, and dependency of different attributes [64]. The notion of similarity may be easily explained for binary attributes. Similarities of attributes indicate the closeness of attributes reflected by their values on a set of objects. Two attributes are similar to each other if an arbitrary object is likely to have the same value for both attributes. Associations (dependencies) show the connections

between attributes. They are normally characterized by the problem of determining the values of one set of attributes based on the values of another set of attributes. Associations can be classified into two types. They are one-way and two-way associations [64]. A one-way association reflects that the values of one set of attributes determine the values of another set of attributes, but does not say anything of the reverse. A two-way association is a combination of two one-way associations, representing two different directions of associations. Two levels of associations, referred to as the *local* and *global* associations, may be observed. A local association shows the relationship between *one* specific combination of values on one set of attributes and *one* specific combination of values on another set of attributes. That is, a local association deals with a particular pair of attribute values (x, y). A global association shows the relationships between *all* combinations of values on one set of attributes and *all* combinations of values on another set of attributes. That is, a global association considers a pair of attributes (X, Y) by taking into consideration all pairs of attribute values (x, y)'s.

Finding local one-way association is one of the main tasks of machine learning and data mining [36,38,41,42]. The well known association rules [1], which state the presence of one set of items implies the presence of another set of items, may be considered as a special kind of local one-way associations. Functional dependency in relational databases is a typical example of global one-way association [2,4]. Attribute (data) dependency studied in the theory of rough sets is another example of global one-way association [38]. There are differences between functional dependency in relational database and data dependency in rough set theory. The functional dependency states the semantics constraints on objects in taking their attribute values. The data dependency summarizes the dependency of attributes with respect to a particular information table. Similarity between attributes may be considered as a global two-way association.

Analysis of vertical relationships deals with semantic closeness of values of an attribute. Examples of vertical analysis include the discretization of real-valued attributes, and the use of binary relations, order relations, concept hierarchies, neighborhood systems, fuzzy binary relations, similarity measures or distance functions on attribute values [15,22,61,62,65]. Using the vertical relationships between attribute values, one may study relationships between objects. Objects may be clustered and classified based on their attribute values. The semantic closeness of attribute values also offers a basis for approximate retrieval [62].

The horizontal and vertical analyses of information tables focus on different aspects of an information table. It may be potentially useful to combine the two analyses. One may introduce more flexibility in horizontal analysis by taking into consideration vertical analysis. For example, attribute values can be clustered to obtain more generalized decision rules in machine

learning [31,41]. The use of concept hierarchies in data mining can produce multi-level association rules [15].

Each type of relationships between attributes captures a specific type of knowledge derivable from a data set. Some authors have proposed methods that generalize a particular type of relationships, in order to take into consideration of others [37,49]. There is a need for systematic studies on the characterization, classification, and interpretations of various types of relationships between attributes, as well as their connections to each other [64]. We address these issues from an information-theoretic point of view [64].

3 A Review of Information-Theoretic Measures

For an attribute X, its values divides the set of objects U into a family of disjoint subsets. The subset defined by the value $x \in V_X$ is given by:

$$m(X = x) = m(x) = \{t \in U \mid I_X(t) = x\}. \tag{2}$$

It consists of all objects whose value on X is x. An information table can be viewed as a statistical population and X a statistical variable. We associate X with a probability distribution defined by:

$$P(X = x) = P(x) = |m(x)|/|U|, \qquad x \in V_X, \tag{3}$$

where $|\cdot|$ denotes the cardinality of a set. Other related probability distributions can be similarly defined. In particular, $P(X, Y)$ is the joint probability distribution of X and Y, and $P(X|Y)$ is the conditional probability distribution of X given Y. The set of objects $m(y)$ may be considered as a subpopulation of U. The conditional probability distribution $P(X|y)$ is the probability distribution associated with X in the subpopulation $m(y)$.

Shannon's entropy function H is defined over P as:

$$\begin{aligned} H(P(X)) &= \mathbf{E}_{P(X)}[-\log P(X)] \\ &= -\sum_{x \in V_X} P(x) \log P(x), \end{aligned} \tag{4}$$

where $\mathbf{E}_{P(X)}[\cdot]$ denotes the expected value with respect to the probability distribution of X and $P(x) \log P(x) = 0$ if $P(x) = 0$. We also say that the entropy is over X and write $H(P(X))$ as $H(X)$ when the distribution P over X is understood. The entropy is a nonnegative function, i.e., $H(X) \geq 0$. It may be interpreted as a measure of the information content of, or the uncertainty about, the attribute X. Entropy reaches the maximum value $\log |V_X|$ when P is the *uniform* distribution, i.e., $P(x) = 1/|V_X|$, $x \in V_X$. The minimum value 0 is obtained when the distribution P focuses on a particular value x_0, i.e., $P(x_0) = 1$ and $P(x) = 0, x \neq x_0$. Entropy depends on the probabilities, and does not depend on the actual values taken by attribute X.

One may interpret the entropy value as representing the degree of structuredness or diversity of a probability distribution [43,56]. A lower entropy value indicates a higher degree of structuredness. This may be seen from the notion of relative entropy or Kullback-Leibler divergence measure [27]. Consider two probability distributions $P(X)$ and $Q(X)$. Suppose P is absolutely continuous with respect to Q, namely, $P(x) \to 0$ if $Q(x) \to 0$. The Kullback-Leibler divergence measure $D(P||Q)$, also known as I-divergence measure [20,27], is defined by:

$$\begin{aligned} D(P||Q) &= \mathbf{E}_{P(X)}\left[\frac{P(X)}{Q(X)}\right] \\ &= \sum_{x \in V_X} P(x) \log \frac{P(x)}{Q(x)}. \end{aligned} \tag{5}$$

It measures the degree of deviation of the probability distribution $P(X)$ from another distribution $Q(X)$. The divergence measure is nonnegative, i.e., $D(P||Q) \geq 0$. It becomes minimum 0 if $P(x) = Q(x)$ for all $x \in V_X$. The maximum value of $D(P||Q)$ is realized when $P(x) = 1$ for a particular x for which $Q(x)$ is the smallest [56]. The divergence measure is non-symmetric, i.e., in general, $D(P||Q) \neq D(Q||D)$. A symmetric measure of *mutual* deviation between two distributions $P(X)$ and $Q(X)$ is defined by [56]:

$$J(P, Q) = D(P||Q) + D(Q||P), \tag{6}$$

which is known as the J-divergence measure [6,20,27].

As a special case, one can compute the degree of deviation of a probability distribution P from the uniform distribution $Q(x) = 1/|V_X|$, $x \in V_X$. We obtain [10,56]:

$$\begin{aligned} D(P||Q) &= \sum_{x \in V_X} P(x) \log \frac{P(x)}{1/|V_X|} \\ &= \log |V_X| + \sum_{x \in V_X} P(x) \log P(x) \\ &= \log |V_X| - H(X). \end{aligned} \tag{7}$$

The uniform distribution represents a maximum state of unstructuredness. A larger deviation from the uniform distribution implies a higher degree of structuredness. Thus, entropy may be a good measure of structuredness and evenness.

The divergence measure may be used to compute the degree of independence of two attributes X and Y. By taking $Q(X, Y) = P(X) \times P(Y)$, i.e.,

the independence distribution formed by the same marginals, we have:

$$\begin{aligned} D(P(X,Y)||Q(X,Y)) &= D(P(X,Y)||P(X) \times P(Y)) \\ &= \mathbf{E}_{P(X,Y)}\left[\log \frac{P(x,y)}{P(x)P(y)}\right] \\ &= \sum_{x \in V_X} \sum_{y \in V_Y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \\ &= I(X;Y). \end{aligned} \tag{8}$$

The measure of deviation of the joint distribution from the independence distribution is in fact the mutual information $I(X;Y)$ between the two attributes X and Y. It is non-negative and symmetric, i.e., $I(X;Y) \geq 0$ and $I(X;Y) = I(Y;X)$. Mutual information can also be expressed in terms of divergence between conditional and marginal probability distributions as follows [50]:

$$\begin{aligned} I(X;Y) &= \sum_{x \in V_X} \sum_{y \in V_Y} P(x,y) \log \frac{P(x,y)}{P(x)P(y)} \\ &= \sum_{y \in V_Y} P(y) \sum_{x \in V_X} P(x|y) \log \frac{P(x|y)}{P(x)} \\ &= \sum_{y \in V_Y} P(y) D(P(X|y)||P(X)) \\ &= \mathbf{E}_{P(Y)}[D(P(X|Y)||P(X)]. \end{aligned} \tag{9}$$

The quantity $D(P(X|y)||P(X))$ shows the degree of deviation of the conditional probability distribution $P(X|y)$ from the unconditional distribution $P(X)$. The distribution $P(X)$ is characterized by the partition of the entire database by values of X, while $P(X|y)$ is characterized by the partition of the subpopulation $m(y)$. A larger divergence implies that the characteristics of subpopulation $m(y)$ is very different from that of the entire population. It may happen that there is a regularity in the subpopulation which may not be present in the entire population. The mutual information is the expectation of divergence.

For two attributes X and Y, their joint entropy is defined by:

$$\begin{aligned} H(X,Y) &= \mathbf{E}_{P(X,Y)}[-\log P(X,Y)] \\ &= -\sum_{x \in V_X} \sum_{y \in V_Y} p(x,y) \log p(x,y). \end{aligned} \tag{10}$$

The conditional entropy $H(X|Y)$ is defined as the expected value of subpopulation entropies $H(X|y)$ with respect to the probability distribution $P(Y)$:

$$H(X|Y) = \sum_{y \in V_Y} P(y) H(X|y)$$

$$
\begin{aligned}
&= -\sum_{y\in V_Y} P(y) \sum_{x\in V_X} P(x|y)\log P(x|y) \\
&= -\sum_{x\in V_X}\sum_{y\in V_Y} P(x,y)\log P(x|y) \\
&= \mathbf{E}_{P(X,Y)}[-\log P(X|Y)]. \qquad (11)
\end{aligned}
$$

Conditional entropy is nonnegative and non-symmetric, namely, $H(X|Y) \geq 0$ and in general $H(X|Y) \neq H(Y|X)$. Conditional entropy can also be expressed by:

$$H(X|Y) = H(X,Y) - H(Y). \qquad (12)$$

It measures the additional amount of information provided by X if Y is already known.

Mutual information can be equivalently expressed by using entropy and conditional entropy:

$$
\begin{aligned}
I(X;Y) &= H(X) - H(X|Y) \\
&= H(Y) - H(Y|X) \\
&= H(X) + H(Y) - H(X,Y). \qquad (13)
\end{aligned}
$$

Accordingly, mutual information measures the decrease of uncertainty about X caused by the knowledge of Y, which is the same as the decrease of uncertainty about Y caused by the knowledge of X. It measures the amount of information about X contained in Y, or the amount of information about Y contained in X. The amount of information contained in X about itself is obviously $H(X)$, namely, $I(X;X) = H(X)$.

Let $P(X)$ and $Q(X)$ be two probability distributions representing information about two related populations. Entropy related functions can be used to measure the similarity of two populations [43,58,66]. Suppose $\lambda_1, \lambda_2 \in [0,1]$ is a pair of real numbers with $\lambda_1 + \lambda_2 = 1$. One may form a composite distribution $\lambda_1 P + \lambda_2 Q$. If P and Q are similar, then both of them are similar to the composite distribution. We would expect a small increase of entropy for the composite distribution. The following entropy difference may be used as a dissimilarity measure of two distributions [30,43,58]:

$$\beta(P,Q:\lambda_1,\lambda_2) = H(\lambda_1 P + \lambda_2 Q) - [\lambda_1 H(P) + \lambda_2 H(Q)]. \qquad (14)$$

The measure β is a nonnegative function, i.e., $\beta(P,Q:\lambda_1,\lambda_2) \geq 0$. The function reaches the minimum value 0 when the two probability distributions P and Q are identical, and reaches the maximum value $H(\lambda) = -(\lambda_1 \log \lambda_1 + \lambda_2 \log \lambda_2)$ when P and Q are totally different, i.e., $P(x) = 0$ whenever $Q(x) \neq 0$ and $Q(x) = 0$ whenever $P(x) \neq 0$.

There exists a close relationship between the divergence measure $D(P||Q)$ and the entropy increase $\beta(P,Q:\lambda_1,\lambda_2)$:

$$\beta(P,Q:\lambda_1,\lambda_2) = \lambda_1 D(P||\lambda_1 P + \lambda_2 Q) + \lambda_2 D(Q||\lambda_1 P + \lambda_2 Q). \qquad (15)$$

The measure β can be viewed as the expected divergence, if (λ_1, λ_2) is considered to be the distribution of a binary random variable. In general, given a set of n populations with probability distributions $P_1(X), \ldots, P_n(X)$ and a set of real numbers $\lambda_1, \ldots, \lambda_n$ with $\sum_{i=1}^{n} \lambda_i = 1$, we have:

$$\begin{aligned}\beta((P_i)_{1\leq i\leq n} : (\lambda_i)_{1\leq i\leq n}) &= H(\sum_{i=1}^{n} \lambda_i P_i) - \sum_{i=1}^{n} \lambda_i H(P_i) \\ &= \sum_{i=1}^{n} \lambda_i D(P_i || \sum_{i=1}^{n} \lambda_i P_i) \end{aligned} \tag{16}$$

Similar to conditional entropy and mutual information, the measure β involves comparisons of probability distributions of various populations. The difference is that β starts with a set of populations and construct a composite population, while conditional entropy and mutual information divide a population into subpopulations based on attribute values.

With respect to an information table, the measure β is the same as mutual information. Let X, Y be two attributes. Based on the values of Y, one can divide a population into $|V_Y|$ subpopulations. Let $\lambda_y = P(y)$, $y \in V_Y$ and $P_y(X) = P(X|y)$. It follows:

$$P(X) = \sum_{y\in V_Y} P(y)P(X|y) = \sum_{y\in V_Y} \lambda_y P_y(X). \tag{17}$$

We have:

$$\begin{aligned}\beta((P_y)_{y\in V_Y} : (\lambda_y)_{y\in V_Y}) &= \sum_{y\in V_Y} P(y) D(P(X|y)||P(X)) \\ &= I(X;Y). \end{aligned} \tag{18}$$

This provides another interpretation of mutual information. One would expect a large mutual information between X and Y, if Y divides the universe into very different subpopulations as expressed in terms of the values of X.

Two important features of the information-theoretic measures need to be emphasized. All measures are related to the divergence D. If a pattern or a regularity is interpreted as the deviation from some standard probability distribution, those measures are potentially useful. All measures can be expressed in a form of expectation, they thus measure global association by considering some kind of average.

4 Information-theoretic Measures of Attribute Importance

Some tasks of KDD are to find important pattern, regularity, and relationship or association, between attributes. In statistical terms, two attributes are associated if they are not independent [29]. Two attributes are independent if

the changes in the value of one do not affect the values of the other. From this standpoint, information-theoretic measures may be used to evaluate the importance of attributes. The structuredness induced by an attribute may be measured by the entropy of the attribute. One-way and two-way associations of two attributes may be measured by conditional entropy and mutual information.

4.1 Measures of structuredness

For an attribute X, its entropy $H(X)$ is related to the deviation of the probability distribution of X from the uniform distribution. A lower entropy suggests that the distribution is uneven, and consequently one may have a better prediction using the value of X. The attribute entropy $H(X)$ serves as a measure of diversity or unstructuredness. It is determined by the probability distribution of the attribute in the entire population, and does not depend on any other attributes.

An attribute with a larger domain normally divides the database into more smaller classes than an attribute with a smaller domain, and hence may have a higher entropy value. In fact, the maximum value of attribute entropy is $\log|V_X|$, which depends on the size of V_X. A regularity found in a very small portion of database may not necessarily be useful. On the other hand, an attribute with smaller domain, i.e., a lower entropy value, usually divides the database into a few larger classes. One may not be able to find regularities in such large subsets of the database. Attribute entropy values may be used to control the selection of attributes. It is expected that an attribute with middle range entropy values may be useful. Similar ideas have been used successfully in information retrieval [45,59]. A high frequency term tends to have a large entropy value, and a low frequency term tends to have a small entropy value. Both may not be a good index term. The middle frequency terms are useful in describing documents in a collection.

The divergence between probability distribution $P(X)$ and the uniform distribution as defined by equation (7) immediately offers a measure of structuredness, namely,

$$W_1(X) = \log|V_X| - H(X). \tag{19}$$

A normalized measure is given by [59,66]:

$$W_2(X) = 1 - \frac{H(X)}{\log|V_X|}, \tag{20}$$

which satisfies the condition $0 \leq W_2(X) \leq 1$. The ratio $H(X)/\log|V_X|$ is referred to as the *relative entropy* by Shannon [46]. A measure similar to W_2 was used in information theory to estimate the *redundancy* of a language or an information source [14,46]. Such a measure was also used to assess the usefulness of an attribute in multi-attributes decision making [19,67], information retrieval [59], and data mining [66]. Instead of using $\log|V_X|$,

one may use the maximum value of all attribute entropies. Let $H_{\max} = \max\{H(X) \mid X \in At\}$, we have:

$$W_3(X) = H_{\max} - H(X), \tag{21}$$

$$W_4(X) = 1 - \frac{H(X)}{H_{\max}}. \tag{22}$$

It may be interpreted as a relative measure by comparing the attribute X and an attribute with maximum entropy value. For an attribute with a smaller domain, we may have $\log |V_X| < H_{\max}$. Measures W_3 and W_4 may favor an attribute with smaller domain in comparison with W_1 and W_2.

Measures W_1 and W_2 reach the minimum value 0 if the distribution $P(X)$ is a uniform distribution, while W_3 and W_4 may not reach 0. In the context of KDD, an attribute with uniform distribution may not necessarily be unimportant. Thus, measures W_3 and W_4 seem to be reasonable, as they take into consideration of entropy values of other attributes. All four measures reach their maximum values when the distribution $P(X)$ focuses on a particular value of V_X, namely, all objects have the same value on X. Although no uncertainty is involved with the attribute, it is not necessarily a useful attribute. One may use measures of structuredness to control the selection of attributes, in the same manner that attribute entropy is used. More specifically, attributes with middle range degrees of structuredness may be potentially useful.

4.2 Measures of one-way association

The notion of association rules has been proposed and studied extensively in mining transaction data [1]. The interpretation of association rules is essentially is the same as that of decision rules studied in machine learning [36]. Association rules concern the relationships between particular combinations of attribute values [1]. For a pair of values x and y of two attributes X and Y, an association rule, $x \Leftarrow y$, states that the occurrence of y warrants the occurrence of x. The confidence of an association rule is defined by:

$$conf(x \Leftarrow y) = P(x|y). \tag{23}$$

It measures the local one-way association of x on y, and does not say anything about x supports y. Many different measures have also been proposed and studied. A review and analysis of commonly used measures can be found in a recent paper by Yao and Zhong [63].

The negative logarithm of $P(x|y)$, i.e., $-\log P(x|y)$, is a monotonic decreasing transformation of $P(x|y)$. Conditional entropy $H(X|Y)$ is the expected value of $-\log P(x|y)$. It may be viewed as an *inverse* measure of global one-way association of two attributes, namely,

$$IC_1(X \Leftarrow Y) = H(X|Y). \tag{24}$$

A normalized version is given by [39]:

$$IC_2(X \Leftarrow Y) = \frac{H(X|Y)}{\log |V_X|}. \tag{25}$$

Conditional entropy $H(X|Y)$ is non-symmetric. The measures of one-way association are also non-symmetric, which is consistent with the interpretation one-way association.

For an attribute X, conditional entropy can be used to select important attributes for discovering one-way association $X \Leftarrow Y$. Measures IC_1 and IC_2 can be used to rank attributes in increasing order. If one prefers to rank attributes in decreasing order, the following corresponding direct measures of one-way association can be used:

$$C_1(X \Leftarrow Y) = \log |V_X| - H(X|Y), \tag{26}$$

$$C_2(X \Leftarrow Y) = 1 - \frac{H(X|Y)}{\log |V_X|}. \tag{27}$$

In these measures, attribute entropy $H(X)$ may be used in place of $\log |V_X|$. We obtain the following measures [26,35]:

$$C_3(X \Leftarrow Y) = H(X) - H(X|Y) = I(X;Y), \tag{28}$$

$$C_4(X \Leftarrow Y) = 1 - \frac{H(X|Y)}{H(X)} = \frac{I(X;Y)}{H(X)}. \tag{29}$$

Measure C_3 is in fact the mutual information between X and Y. It is commonly referred to as information gain and is widely used in machine learning [42].

For a fixed X, measures of one-way association $X \Leftarrow Y$ show the relative importance of Y. An attribute with a larger domain may possibly divide a database into many small populations. Within a small population, there are not many choices for the values of X, and hence the conditional entropy value $H(X|y)$ might be low. Such an attribute may be perceived to be important based on the entropy related measures discussed so far. A measure that corrects such a bias is given by [42]:

$$C_5(X \Leftarrow Y) = \frac{C_3(X \Leftarrow Y)}{H(Y)} = \frac{H(X) - H(X|Y)}{H(Y)} = \frac{I(X;Y)}{H(Y)}. \tag{30}$$

Similarly, one may use $\log |V_Y|$ to replace $H(Y)$ and obtain the measure [25]:

$$C_6(X \Leftarrow Y) = \frac{C_3(X \Leftarrow Y)}{\log |V_Y|} = \frac{I(X;Y)}{\log |V_Y|}. \tag{31}$$

The discussion on measures of structuredness is also relevant to mining one-way association. One may first use the attribute entropy to select a subset of

attributes with middle range entropy values without considering their relationships to X. Measures of one-way association, concerning dependency of X on other attributes, may then be used to fine tune the mining process.

In pattern recognition, a special attribute X may be viewed a label of patterns, and other attributes are features used for describing patterns. The process of feature selection may be viewed as mining one-way association, namely, the association of patterns on various features. Information-theoretic measures such as IC_1 and C_3 have been used for feature selection. A discussion on this topic and many relevant references can be found in the book by Chen [6] and a recent book by Liu and Motoda [33].

By examining two extreme cases of associations, one may provide further support for conditional entropy and mutual information as measures of one-way association. A functional dependency $Y \rightarrow X$ of a relational database holds if the value of Y determines the value of X, namely, $P(x|y)$ is either 1 or 0 for all $x \in V_X$ and $y \in V_Y$. If $Y \rightarrow X$, the partition of the database by X and Y is the same as the one produced by Y alone. In other words, the partition produced by Y is finer than the partition produced by X in the sense that for every $y \in V_Y$ there is a value $x \in V_X$ such that $m(y) \subseteq m(x)$. In terms of information-theoretic measures, $Y \rightarrow X$ holds if and only if the following equivalent conditions hold [28,35]:

$$\begin{aligned} &\text{(i1)} \quad H(X|Y) = 0, \\ &\text{(i2)} \quad H(X,Y) = H(Y), \\ &\text{(i3)} \quad I(X;Y) = H(X). \end{aligned}$$

Functional dependency may by considered as the strongest one-way association. Conditional entropy obtains the minimum value 0 when X functionally depends on Y. The mutual information $I(X;Y) = H(X)$ reaches its maximum value, provided that X is fixed. If X and Y are probabilistically independent, we cannot use the value of Y to predict the value of X, and vice versa. In other words, knowing the values of Y does not reduce our uncertainty about X, and vice versa. In this case, we have the following equivalent conditions:

$$\begin{aligned} &\text{(ii1)} \quad H(X|Y) = H(X), \\ &\text{(ii2)} \quad H(Y|X) = H(Y), \\ &\text{(ii3)} \quad H(X,Y) = H(X) + H(Y), \\ &\text{(ii4)} \quad I(X;Y) = 0. \end{aligned}$$

Two attributes are associated if they are not independent [29]. Independence of two attributes may be viewed as the weakest one-way (or two-way) association. In this case, conditional entropy $H(X|Y)$ reaches the maximum value and mutual information reaches the minimum value. Condition (ii3) states that if X and Y are independent, the uncertainty about (X,Y) is the sum of uncertainties about X and Y. This implies that X and Y do not have any correlations.

4.3 Measures of two-way association

In data mining, the quantity:

$$i(x,y) = \frac{P(x,y)}{P(x)P(y)} \tag{32}$$

has been widely used as a measure of local two-way association [3,7,13,49,63]. The logarithm of $i(x,y)$ is the mutual information of x and y, $I(x;y) = \log[P(x,y)/(P(x)P(y))]$, which is a monotonic transformation of $i(x,y)$. The quantity $I(x;y)$ is also a measure of local two-way association of x and y. Mutual information $I(X;Y)$ is the expected value of such local associations for all attribute value pairs. We obtain a measure of global two-way association:

$$M_1(X \Leftrightarrow Y) = I(X;Y). \tag{33}$$

From $I(X;Y) \leq \min(H(X), H(Y)) \leq H(X,Y)$, we obtain the normalized versions [26,35]:

$$M_2(X \Leftrightarrow Y) = \frac{I(X;Y)}{\min(H(X), H(Y))}, \tag{34}$$

$$M_3(X \Leftrightarrow Y) = \frac{I(X;Y)}{H(X,Y)}. \tag{35}$$

Two-way association as measured by mutual information is the degree of deviation of a joint distribution from the independence distribution. With a fixed X, the use of $I(X;Y)$ for finding a two-way association is in fact the same as using $H(X|Y)$ for finding a one-way association [34,53].

Mutual information has been used in pattern recognition and information retrieval for finding association between attributes [6,52]. A dependence tree consisting of pairs of most dependent attributes can be constructed by using mutual information as a measure of dependency between two attributes [8]. Mutual information and related dependence trees and generalized dependence graphs have been used in probabilistic networks and expert systems [9,40].

Conditional entropy $H(X|Y)$ is an inverse measure of the one-way association in one direction, and $H(Y|X)$ the one-way association in the other direction. Inverse measures of two-way association can be obtained by combining two one-way associations [34,44,53]:

$$\begin{aligned} IM_1(X \Leftrightarrow Y) &= H(X|Y) + H(Y|X) \\ &= 2H(X,Y) - [H(X) + H(Y)] \\ &= H(X) + H(Y) - 2I(X;Y) \\ &= H(X,Y) - I(X;Y), \end{aligned} \tag{36}$$

$$IM_2(X \Leftrightarrow Y) = \frac{IM_1(X \Leftrightarrow Y)}{H(X,Y)} = 2 - \frac{H(X)+H(Y)}{H(X,Y)} = 1 - \frac{I(X:Y)}{H(X,Y)}. \quad (37)$$

where $IM_2(X \Leftrightarrow Y) = 0$ if $H(X,Y) = 0$. From the various forms of these measures, one may associate different information-theoretic interpretations. Measures IM_1 and IM_2 are pseudo-metrics between two random variables of the two attributes [11,16,44,47]. They have been used as measures of correlation and applied to machine learning [34,53]. A more generalized measure may be defined by [53]:

$$IM_4(X \Leftrightarrow Y) = \lambda_1 H(X|Y) + \lambda_2 H(Y|X), \quad (38)$$

where $\lambda_1 + \lambda_2 = 1$. It is a non-symmetric measure unless $\lambda_1 = \lambda_2 = 1/2$.

4.4 Measures of similarity of populations

In some data mining problems, one may be interested in similarity or dissimilarity of different populations [66]. Similarity is closely related to two-way association [64]. For example, one may analyze local two-way association of a pair of attribute value x and y by examining the similarity of two subpopulations $m(x)$ and $m(y)$ with respect to another attribute Z. Divergence measure can be used for such a purpose.

Let $P_1(X)$ and $P_2(X)$ be probability distributions of X in two populations. A non-symmetric dissimilarity measure of the two populations is given by the Kullback-Leibler divergence measure $D(P_1||P_2)$. A symmetric dissimilarity measure is given by $D(P_1||P_2) + D(P_2||P_1)$. A difficulty with such measures is the requirement that one distribution must be absolutely continuous with respect to the other. The related measure $\beta(P_1, P_2 : \lambda_1, \lambda_2)$ does not suffer from this problem. A similarity measure corresponding to β is defined by [58]:

$$S(P_1, P_2 : \lambda_1, \lambda_2) = 1 - \frac{\beta(P_1, P_2 : \lambda_1, \lambda_2)}{H(\lambda)}, \quad (39)$$

where $H(\lambda) = -(\lambda_1 \log \lambda_1 + \lambda_2 \log \lambda_2)$. The values of λ_1 and λ_2 may be interpreted as the importance associated with P_1 and P_2, or the sizes of the two populations.

Measures β and S have been used in pattern recognition [57] and information retrieval [58]. It is recently used for mining market value functions for targeted marketing [66].

4.5 Discussions

Attribute entropy shows the structuredness induced by the attribute, and hence can be used to design measures of attribute importance. Conditional entropy and mutual information serve as the basic quantities for measuring attribute association. By combination and normalization, one can obtain various information-theoretic measures of attribute importance and attribute association.

Table 1 is a summary of the well known measures. Some references are also given, where more information or applications about the measure can be found. The first group consists of measures of structuredness induced by an attribute. The middle two groups are measures of attribute association. Measures of one-way association are non-symmetric. They can be expressed, in a general form, as different normalizations of conditional entropy. Measures of two-way association are symmetric. Two subclasses can be observed, one class consists of different normalizations of mutual information [26], the other class consists of the combination of two conditional entropies. For a fixed X, some measures of one-way and two-way associations produce the same result, if they are used to rank other attributes Y's. They may be viewed as measuring the relative importance of other attributes with reference to X. The last group consists of measures of dissimilarity of populations. From the relationship between entropy, conditional entropy and mutual information, a measure can be expressed in many different forms.

Entropy and mutual information can be explained in terms of Kullback-Leibler divergence measure. Entropy shows the divergence from the uniform distribution, while mutual information shows the divergence from the independence distribution. Uniform distribution and independence distribution are perceived as uninteresting. Application of information-theoretic measures for KDD is therefore intended to discover regularities and patterns revealing large divergence from unimportant or uninteresting distributions.

All measures are based on some kind of average which is suitable for global association. In some situations, the best average might not be a good choice. For example, Cendrowska [5] presented a learning algorithm that is different from ID3. Instead of using every attribute value of an attribute to decide if the attribute should be selected, only certain values are considered. Populations constrained by some values reveal stronger regularities, although on average populations by all attribute values reveal weaker regularities.

In studying main problem types for KDD, Klösgen [23] discussed the following two types of problems. The *classification and predication* problem deals with the discovery of a set of rules or similar patterns for predicting the values of a dependent variable. The ID3 algorithm [42] and the mining of associate rules [1] are examples for solving this type of problems. The *summary and description* problem deals with the discovery of dominant structure that derives a dependency. Kamber and Shinghal [21] referred to these problems as the discovery of discriminant and characteristic rules, respectively. Differ-

References	Measures
Measures of structuredness of an attribute X:	
Shannon [46], Watanabe [56]	$H(X)$
Hwang and Yoon [19], Shannon [46], Wong and Yao [59], Yao and Zhong [66], Zeleny [67]	$1 - \frac{H(X)}{\log \lvert V_X \rvert}$
Measures of one-way association $X \Leftarrow Y$:	
Lee [28], Malvestuto [35], Pawlak *et al.* [39]	$H(X\lvert Y)$
Kvålseth [26], Malvestuto [35], Quinlan [42]	$\frac{I(X;Y)}{H(Y)}$
Measures of two-way association $X \Leftrightarrow Y$:	
Knobbe and Adriaans [24], Linfoot [32], Quinlan [42]	$I(X;Y)$
Malvestuto [35]	$\frac{I(X;Y)}{H(X,Y)}$
Horibe [17], Kvålseth [26]	$\frac{I(X;Y)}{\max(H(X),H(Y))}$
Kvålseth [26]	$\frac{I(X;Y)}{\min(H(X),H(Y))}$
Kvålseth [26]	$\frac{2I(X;Y)}{H(X)+H(Y)}$
López de Mántaras [34], Shannon [47], Wan and Wong [53]	$H(X\lvert Y) + H(Y\lvert X)$
López de Mántaras [34], Rajski [44]	$\frac{H(X\lvert Y)+H(Y\lvert X)}{H(X,Y)}$
Measures of dissimilarity of populations P_1 and P_2	
Chen [6], Kullback and Leibler [27]	$D(P_1 \lvert\lvert P_2)$
Kullback and Leibler [27], Watanabe [56]	$D(P_1 \lvert\lvert P_2) + D(P_2 \lvert\lvert P_1)$
Lin and Wong [30], Rao [43], Wong and Yao [58], Wong and You [57]	$H(\lambda_1 P_1 + \lambda_2 P_2) - [\lambda_1 H(P_1) + \lambda_2 H(P_2)]$

Table 1. Information-theoretic Measures

ent measures should be used for selecting attributes for distinct problems. A non-symmetric measure of one-way association may be suitable for the first type, while a symmetric measure of two-way association may be appropriate for the second type.

In the study of association of random variables using statistical measures, Liebetrau [29] pointed out that many symmetric measures do not tell us anything about causality. When two attributes are shown to be correlated, it is very tempting to infer a cause-and-effect relationship between them. It is very important to realize that the mere identification of association does not provide grounds to establish causality. Garner and McGill [12] showed that information-theoretic analysis is very similar to analysis of variance. One may extend the argument of Liebetrau [29] to information-theoretic measures. In order to establish causality, we need additional techniques in data mining.

5 Conclusion

Many different forms of knowledge and information can be derived from a large data set. Relationships between attributes represent an important class. An analysis of possible relationships between attributes and their connections may play an important role in data mining. Starting with the Shannon entropy function and the Kullback-Leibler divergence measure, we present an overview and analysis of information-theoretic measures of attribute importance and attribute association in the setting of KDD. Four classes of measures are discussed. Attribute entropy shows the structuredness induced by the attribute, and is used to design measures of attribute importance. Conditional entropy is used to define non-symmetric measures of one-way association. Conditional entropy and mutual information are used to define symmetric measures of two-way association. They can be used to measure the relative importance of other attributes with respect to a fixed attribute. Measures of dissimilarity and similarity of populations are also discussed.

This article is mainly a critical analysis of existing results in using information theory in KDD and related fields. Our preliminary study shows that information theory might be used to establish a formal theory for KDD. The systematic analysis of information-theoretic measures may serve as a starting point for further studies on this topic.

References

1. Agrawal, R., Imielinski, T. and Swami, A. Mining association rules between sets of items in large databases, *Proceedings of the ACM SIGMOD International Conference on the Management of Data*, 207-216, 1993.
2. Bell, A. Discovery and maintenance of functional dependencies by independencies, *Proceedings of KDD-95*, 27-32, 1995.

3. Büchter, O. and Wirth R. Discovery of association rules over ordinal data: a new and faster algorithm and its application to basket analysis, in: *Research and Development in Knowledge Discovery and Data Mining*, Wu, X., Kotagiri, R. and Bork, K.B. (Eds.), Springer, Berlin, 36-47, 1998.
4. Butz, C.J., Wong, S.K.M. and Yao, Y.Y. On data and probabilistic dependencies, *Proceedings of the 1999 IEEE Canadian Conference on Electrical and Computer Engineering*, 1692-1697, 1999.
5. Cendrowska, J. PRISM: an algorithm for inducing modular rules, *International Journal of Man-Machine Studies*, **27**, 349-370, 1987.
6. Chen, C. *Statistical Pattern Recognition*, Hayden Book Company, Inc., New Jersey, 1973.
7. Chen, M., Han, J. and Yu, P.S. Data mining, an overview from a database perspective, *IEEE Transactions on Knowledge and Data Engineering*, **8**, 866-883, 1996.
8. Chow, C. and Liu, C. Approximating discrete probability distributions with dependence trees, *IEEE Transactions on Information Theory*, **IT-14**, 462-467, 1968.
9. Cowell, R.G., Dawid, A.P., Lauritzen, S.L. and Spiegelhalter, D.J. *Probabilistic Networks and Expert Systems*, Springer, New York, 1999.
10. Cover, T. and Thomas, J. *Elements of Information Theory*, John Wiley & Sons, Toronto, 1991.
11. Csiszár, I. and Körner, J. *Information Theory: Coding Theorems for Discrete Memoryless Systems*, Academic Press, New York, 1981.
12. Garner, W.R. and McGill, W.J. Relation between information and variance analyses, *Psychometrika*, **21**, 219-228, 1956.
13. Gray, B. and Orlowska, M.E. CCAIIA: clustering categorical attributes into interesting association rules, in: *Research and Development in Knowledge Discovery and Data Mining*, Wu, X., Kotagiri, R. and Bork, K.B. (Eds.), Springer, Berlin, 132-143, 1998.
14. Guiasu, S. *Information Theory with Applications*, McGraw-Hill, New York, 1977.
15. Han, J., Cai, Y. and Cercone, N. Data-driven discovery of quantitative rules in databases, *IEEE Transactions on Knowledge and Data Engineering*, **5**, 29-40, 1993.
16. Horibe, Y. A note on entropy metrics, *Information and Control*, **22**, 403-404, 1973.
17. Horibe, Y. Entropy and correlation, *IEEE Transactions on Systems, Man, and Cybernetics*, **SMC-15**, 641-642, 1985.
18. Hou, W. Extraction and applications of statistical relationships in relational databases, *IEEE Transactions on Knowledge and Data Engineering*, **8**, 939-945, 1996.
19. Hwang, C.L. and Yoon, K. *Multiple Attribute Decision Making, Methods and Applications*, Springer-Verlag, Berlin, 1981.
20. Kazakos, D. and Cotsidas, T. A decision approach to the approximation of discrete probability densities, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-2**, 61-67, 1980.
21. Kamber, M. and Shinghal, R. Evaluating the interestingness of characteristic rules, *Proceedings of KDD-96*, 263-266, 1996.
22. Klir, G.J. and Yuan, B. *Fuzzy Sets and Fuzzy Logic, Theory and Applications*, Prentice Hall, New Jersey, 1995.

23. Klösgen, W. Explora: a multipattern and multistrategy discovery assistant, in: *Advances in Knowledge Discovery and Data Mining*, Fayyad, U.M, Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, R. (Eds.), AAAI/MIT Press, California, 249-271, 1996.
24. Knobbe, A.J. and Adriaans P.W. Analysis of binary association, *Proceedings of KDD-96*, 311-314, 1996.
25. Kohavi, R. and Li, C. Oblivious decision trees, graphs and top-down pruning, *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*, 1071–1077, 1995.
26. Kvålseth, T.O. Entropy and correlation: some comments, *IEEE Transactions on Systems, Man, and Cybernetics*, **SMC-17**, 517-519, 1987.
27. Kullback, S. and Leibler, R.A. On information and sufficiency, *Annals of Mathematical Statistics*, **22**, 79-86, 1951.
28. Lee, T.T. An information-theoretic analysis of relational databases – part I: data dependencies and information metric, *IEEE Transactions on Software Engineering*, **SE-13**, 1049-1061, 1987.
29. Liebetrau, A.M. *Measures of Association*, Sage University Paper Series on Quantitative Application in the Social Sciences, 07-032, Sage Publications, Beverly Hills, 1983.
30. Lin, J. and Wong, S.K.M. A new directed divergence measure and its characterization, *International Journal of General Systems*, **17**, 73-81, 1991,
31. Lin, T.Y. and Cercone, N. (Eds.), *Rough Sets and Data Mining: Analysis for Imprecise Data*, Kluwer Academic Publishers, Boston, 1997.
32. Linfoot, E.H. An informational measure of correlation, *Information and Control*, **1**, 85-87, 1957.
33. Liu, H. and Motoda, H. *Feature Selection for Knowledge Discovery and Data Mining*, Kluwer Academic Publishers, Boston, 1998.
34. López de Mántaras, R. ID3 revisited: a distance-based criterion for attribute selection, in: *Methodologies for Intelligent Systems, 4*, Ras, Z.W. (Ed.), North-Holland, New York, 342-350, 1989.
35. Malvestuto, F.M. Statistical treatment of the information content of a database, *Information Systems*, **11**, 211-223, 1986.
36. Michalski, R.S., Carbonell, J.G. and Mitchell, T.M. (Eds.), *Machine Learning*, Tioga, 1983.
37. Pfahringer, B. and Kramer, S. Compression-based evaluation of partial determinations, *Proceedings of KDD-95*, 234-239, 1995.
38. Pawlak, Z. *Rough Sets, Theoretical Aspects of Reasoning about Data*, Kluwer Academic Publishers, Boston, 1991.
39. Pawlak, Z., Wong, S.K.M. and Ziarko, W. Rough sets: probabilistic versus deterministic approach, *International Journal of Man-Machine Studies*, **29**, 81-95, 1988.
40. Pearl, J. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*, Morgan Kaufmann Publishers, San Francisco, 1988.
41. Polkowski, L. and Skowron, A. (Eds.), *Rough Sets in Knowledge Discovery 1,2*, Physica-Verlag, Heidelberg, 1998.
42. Quinlan, J.R. Induction of decision trees, *Machine Learning*, **1**, 81-106, 1986.
43. Rao, C.R. Diversity and dissimilarity coefficients: a unified approach, *Theoretical Population Biology*, **21**, 24-43, 1982.
44. Rajski, C. A metric space of discrete probability distributions, *Information and Control*, **4**, 373-377, 1961.

45. Salton, G. and McGill, M.H. *Introduction to Modern Information Retrieval*, McGraw-Hill, New York, 1983.
46. Shannon, C.E. A mathematical theory of communication, *Bell System and Technical Journal*, **27**, 379-423, 623-656, 1948.
47. Shannon, C.E. Some topics in information theory, *Proceedings of International Congress of Mathematics*, **2**, 262, 1950.
48. Sheridan, T.B. and Ferrell, W.R. *Man-Machine Systems: Information, Control, and Decision Models of Human Performance*, The MIT Press, Cambridge, 1974.
49. Silverstein, C., Brin, S. and Motwani, R. Beyond market baskets: generalizing association rules to dependence rules, *Data Mining and Knowledge Discovery*, **2**, 39-68, 1998.
50. Smyth, P. and Goodman, R.M. Rule induction using information theory, in: *Knowledge Discovery in Databases*, Piatetsky-Shapiro, G. and Frawley, W.J. (Eds.), AAAI/MIT Press, 159-176, 1991.
51. Spyratos, N. The partition model: a deductive database model, *ACM Transactions on Database Systems*, **12**, 1-37, 1987.
52. van Rijsbergen, C.J. *Information Retrieval*, Butterworth, London, 1979.
53. Wan, S.J. and Wong, S.K.M. A measure for attribute dissimilarity and its applications in machine learning, in: *Computing and Information*, Janicki, R. and Koczkodaj, W.W. (Eds.), North-Holland, Amsterdam, 267-273, 1989.
54. Wang, Q.R. and Suen, C.Y. Analysis and design of a decision tree based on entropy reduction and its application to large character set recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **PAMI-6**, 406-417, 1984.
55. Watanabe, S. *Knowing and Guessing*, Wiley, New York, 1969.
56. Watanabe, S. Pattern recognition as a quest for minimum entropy, *Pattern Recognition*, **13**, 381-387, 1981.
57. Wong, A.K.C. and You, M. Entropy and distance of random graphs with application to structural pattern recognition, *IEEE Transactions on Pattern Analysis And Machine Intelligence*, **PAMI-7**, 599-609, 1985.
58. Wong, S.K.M. and Yao, Y.Y. A probability distribution model for information retrieval, *Information Processing and Management*, **25**, 39-53, 1989.
59. Wong, S.K.M. and Yao, Y.Y. An information-theoretic measure of term specificity, *Journal of the American Society for Information Science*, **43**, 54-61, 1992.
60. Yao, Y.Y., Wong, S.K.M. and Butz, C.J. On information-theoretic measures of attribute importance, *Proceedings of PAKDD'99*, 133-137, 1999.
61. Yao, Y.Y., Wong, S.K.M. and Lin, T.Y. A review of rough set models, in: *Rough Sets and Data Mining: Analysis for Imprecise Data*, Lin, T.Y. and Cercone, N. (Eds.), Academic Publishers, Boston, 47-75, 1997.
62. Yao, Y.Y. Information tables with neighborhood semantics, in: *Data Mining and Knowledge Discovery: Theory, Tools, and Technology II*, Dasarathy, B.V. (Ed.), The International Society for Optical Engineering, Bellingham, Washington, 108-116, 2000.
63. Yao, Y.Y. and Zhong, N. An analysis of quantitative measures associated with rules, *Proceedings of PAKDD'99*, 479-488, 1999.
64. Yao, Y.Y. and Zhong, N. On association, similarity and dependency of attributes, *Proceedings of PAKDD'00*, 2000.
65. Yao, Y.Y. and Zhong, N. Granular computing using information tables, manuscript, 2000.

66. Yao, Y.Y. and Zhong, N. Mining market value function for targeted marketing, manuscript, 2000.
67. Zeleny, M. *Linear multiobjective programming*, Springer-Verlag, New York, 1974.