

Automated Quality Assurance of Continuous Data

Mark LAST¹ and Abraham KANDEL²

1. Ben-Gurion University of the Negev, Department of Information Systems Engineering, Beer-Sheva 84105, Israel
2. Department of Computer Science and Engineering, ENB 118, University of South Florida, Tampa, Florida 33620, USA.

Abstract

Most real-world databases contain some amount of inaccurate data. Reliability of critical attributes can be evaluated from the values of other attributes in the same data table. This paper presents a new fuzzy-based measure of data reliability in continuous attributes. We partition the relational schema of a database into a subset of input (predicting) and a subset of target (dependent) attributes. A data mining model, called information-theoretic connectionist network, is constructed for predicting the values of a continuous target attribute. The network calculates the degree of reliability of the actual target values in each record by using their distance from the predicted values. The approach is demonstrated on the voting data from the 2000 Presidential Elections in the US.

Keywords. Data reliability, data quality, information-theoretic networks, data mining, fuzzy databases.

1 Introduction

Modern database systems are designed to store accurate and reliable data only. However, the assumption of zero defect data (ZDD) is far from being true in most real-world databases, especially when their data comes from multiple sources. The issue of data reliability has been in the focus of the recent controversy regarding the results of the Year 2000 presidential elections in the State of Florida. After the elections, the leaders of the Democratic Party questioned the accuracy of the official voting results in certain Florida counties, based on the demographic characteristics of the voters in those counties. Their suspicions have led to a manual re-count of the votes, which was aimed at improving the reliability of the results. Though the accuracy of the punch card counting machines, used in some counties, was known to be limited, a complete manual re-count of all Florida votes was not feasible within a several weeks time frame. Eventually, the courts stopped the manual re-count process and Mr. George W. Bush was declared as the new President of the United States.

In small databases the users have enough time to check manually every record “suspected” of poor data quality and correct data, if necessary. In a large database, like the data on Florida voting results, this approach is certainly impractical. The task of assuring data reliability and data quality, known as “data cleaning”, becomes even more acute in rapidly emerging *Data Warehouses*. Thus, there is a strong need for an efficient automated tool, capable of detecting, filtering, representing and analyzing poor quality data in large databases.

In our previous work (see [18] - [20]), we have introduced an information-theoretic fuzzy method for evaluating reliability of discrete (nominal) attributes. Our methodology

for data quality assurance includes three main stages: modification of the database schema, induction of a data mining model (information-theoretic network), and using the constructed network to calculate reliability degrees of attribute values. In this paper, we present a new fuzzy-based measure for evaluating the reliability of continuous attributes and demonstrate it on a set of real voting data from the US presidential elections.

Our paper is organized as follows. In Section 2 we present an overview of existing approaches to various aspects of data quality and data reliability. Section 3 briefly describes the algorithm for building an information-theoretic connectionist network from relational data. In Section 4, we present the fuzzy-based approach to evaluating data reliability of continuous attributes. In Section 5 we apply the info-fuzzy methodology presented in Sections 3 and 4 to a set of real voting data. Section 6 concludes the paper with summarizing the benefits of our approach to data reliability and representing a number of issues for the future research.

2 Data Quality and Data Reliability

As indicated by Wang *et al.* [32], data reliability is one of *data quality dimensions*. Other data quality dimensions include ([31] - [33]): accuracy, timeliness, relevance, completeness, consistency, precision, etc. Various definitions of these and other dimensions can be found in [33]. Ahituv *et al.* [1] refer to accuracy and relevance as *content attributes* of an information system. According to Wand and Wang [31], the reliability “indicates whether the data can be counted on to convey the right information”. Unreliable (deficient) data represents an inconformity between the state of the information system and the state of the real-world system. The process of mapping a real-world state to a wrong state in an information system is termed by [31] as “garbling”. Two cases of garbling are considered: the mapping to a meaningless state and the mapping to a meaningful, but wrong state. In the first case the user knows that the data is unreliable, while in the second case he relies upon an incorrect data. Wand and Wang suggest to solve the garbling problem by adding data entry controls, like check digits and control totals, methods which are not applicable to qualitative data. The paper follows a “Boolean” approach to data reliability: the information system states are assumed to be either correct or incorrect. No “medium” degree of reliability is provided.

An attribute-based approach to data quality is introduced by Wang *et al.* in [33]. It is based on the entity-relationship (ER) model (see [13]) and assumes that some attributes (called *quality indicators*) provide objective information (metadata) about data quality of other attributes. The data quality is expressed in terms of *quality parameters* (e.g., believability, reliability, and timeliness). Thus, if some sources are less reliable than the others, an attribute *data source* may be an indicator of data reliability. Each quality parameter has one or more quality indicators attached to it via *quality keys*. A quality indicator may have quality indicators of its own, leading easily to an exponential total number of quality indicators. Wang *et al.* [33] suggest integration of quality indicators, to eliminate redundancy and inconsistency, but no methodological approach to this problem (crucial for dimensionality reduction) is presented.

An extended database, storing quality indicators along with data, is defined as a *quality database*. The quality indicator values are stored in quality indicator relations. The quality database is strictly deterministic: once the values of quality indicators are given, the values

of quality parameters are uniquely defined by the database structure. The values of quality parameters are often qualitative and subjective (like “highly reliable” vs. “unreliable”). Wang *et al.* [33] warn that quality parameters and quality indicators are strongly user-dependent and application-dependent. The database structure described by [33] enables an experienced user to infer *manually* from values of quality indicators about the quality of relation attributes, but their work provides no method for automated evaluation of data quality in large databases.

Kandel *et al.* [11] mention unreliable information as one of sources of data uncertainty, other sources including fuzziness of human concepts, incomplete data, contradicting sources of information, and partial matching between facts and events. According to Kandel *et al.* (1996), the main drawback of the probabilistic approaches to uncertainty (e.g., the Bayesian approach) is their limited ability to represent human reasoning, since humans are not Bayesian when reasoning under uncertainty.

Kurutach [14] discusses three types of data imperfection in databases: *vagueness*, or *fuzziness* (the attribute value is given, but its meaning is not well-defined), *imprecision* (the attribute value is given as a set of possible items), and *uncertainty* (the attribute value is given along with its degree of confidence). All these types of imperfection are defined by users themselves during the data entry process. The author suggests a unified approach, based on fuzzy set theory, to incorporating these aspects of imperfection in an extended relational database containing, primarily, discretely-valued, qualitative data. In addition to imprecision and uncertainty, Motro [22] defines a third kind of imperfect data: erroneous information. Database information is erroneous, when it is different from the true information. Motro [22] follows the binary approach to errors: both “small” and “large” errors in a database should not be tolerated. He also mentions inconsistency as one of the important kinds of erroneous information.

Since, in a general case, data reliability is a *linguistic variable* (the data can be considered “very reliable”, “not so reliable”, “quite unreliable”, etc.), the models of fuzzy databases seem to be helpful for treating reliability of database attributes. As indicated by Zemankova and Kandel [35], the main problem of fuzzy databases is to propagate the level of uncertainty associated with the data (reliability degree in our case) to the level of uncertainty associated with answers or conclusions based on the data. The fuzzy relational algebra proposed by Klir and Yuan [12] enables to check similarity between values of fuzzy attributes by using a similarity relation matrix and a pre-defined threshold level of minimum acceptable similarity degree.

Zemankova and Kandel [35] and Kandel [10] propose a Fuzzy Relational Data-Base (FRDB) model which enables to evaluate fuzzy queries from relational databases. The attribute values in the FRDB can represent membership or possibility distributions defined on the unit interval $[0,1]$. According to this model, a single value of a membership distribution can be used as a value of a fuzzy attribute. Another model of fuzzy querying from regular relational databases (called SQLf) is presented by Bosc and Pivert [2]. The main purpose of this model is to define imprecise answers based on precise data and on fuzzy conditions (which contain fuzzy predicates and fuzzy quantifiers).

The Fuzzy Data model developed by Takahashi [29] assumes that some nonkey attributes may have values defined by fuzzy predicates (e.g., “very reliable”). All key attributes and some other attributes are assumed to have nonfuzzy values only. Any tuple in Takahashi data model has a *truth value* z defined over the unit interval $[0,1]$. The value

of z is interpreted as a degree to which the tuple is true, with two special cases: $z = 0$ when the tuple is completely false and $z = 1$ when the tuple is completely true. This approach treats a tuple as a set of attribute values, all having the same truth-value. The case of different truth-values associated with values of different attributes in the same tuple is not covered by the model of [29]. A similar idea of associating a single truth value (a *weight*) with each tuple is described by Petri [25]. Petri terms such tuples as *weighted tuples* and defines their weight as a membership degree expressing the extent to which a tuple belongs to a fuzzy relation. Three possible meanings of tuple weights are proposed. One of them is “the certainty of information stored in the tuple”, i.e. the reliability of all tuple attributes. The concept of reliability degree associated with every column in a fuzzy spreadsheet table is used by [23]. According to their definition, the degree of reliability can take any continuous value between 0 and 1, but no explicit interpretation of this variable is provided.

All the above-mentioned models assume that both crisp and fuzzy quality dimensions of database attributes are available from the database users. Obviously, this assumption may not be realistic for large and dynamically changing databases. Consequently, there is a need for methods that perform automated assessment of data quality. An information theoretic approach to automated data cleaning is presented by Guyon *et al.* [8]. The paper assumes that erroneous (“garbage”) data has a high information gain. The information gain is defined by [8] as a self-information (logarithm of probability) of predicting the correct data value. This means that the most “surprising” patterns (having the lowest probability to be predicted correctly) are suspicious to be unreliable. The authors propose a computer-aided cleaning method where a human operator must check only those patterns that have the highest information gain and remove from the database patterns, which are truly corrupted, while keeping all the rest. The prediction itself is performed in [8] by using a neural network trained with a “cross-entropy” cost function. One can easily accept the approach of [8] that values having lower probability are more likely to be erroneous. However, the values having the *same* probability (and, accordingly, the same information gain) cannot be treated alike in different databases. Reliability may also depend on the inherent distributions of database attributes and some other, user-related factors. Thus, the approach of [8] should be enhanced to cope with real-world problems of data quality.

In [19], we have presented a fuzzy-based approach to automated evaluation of data reliability. The method of [19] is aimed at detecting unreliable nominal data by integrating objective (information-theoretic) and subjective (user-specific) aspects of data quality. In this paper, we extend the method of [19] to handle partially reliable continuous attributes.

3 Information-Theoretic Connectionist Networks

Uncertainty is an inherent part of our life. Delivery time of manufactured products is not constant, stock prices go up and down, and people vote according to their personal beliefs. Most real-world phenomena cannot be predicted with perfect accuracy. The reasons for that may include limited understanding of the true causes for a given phenomenon (e.g., detailed considerations of each specific voter), as well as missing and erroneous data (e.g., incomplete or inaccurate voting results).

Data mining methods (see [4], [5], [17], [21], [26], and [27]) are aimed at reducing the amount of uncertainty, or gaining *information*, about the data. More information means

higher prediction accuracy for future cases. If a model is useless, it does not provide us with any new information and its prediction accuracy is not higher than just a random guess. On the other hand, the maximum amount of information transferred by a model is limited: in the best case, we have an accurate prediction for every new case. Intuitively, we need more information to predict a multi-valued outcome (e.g., percentage of votes for a certain candidate) than to predict a binary outcome (e.g., customer credibility).

The above characteristics of the data mining problem resemble the communication task: predicting attributes can be seen as input messages and each value of the system output is an output message. If we have a model with a perfect accuracy, each output value can be predicted correctly from the values of input attributes. In terms of the Information Theory (see [3]), this means that the entropy of the output Y , given the input X is zero, i.e., the mutual information between Y and X is maximal.

The information-theoretic approach to data mining (see [6], [7], [15], [16], [18], [19], and [20]) is a powerful methodology for inducing information patterns from large sets of imperfect data, since it uses meaningful network structure, called *information-theoretic connectionist network*. The measures of information content, expressed by the network connection weights, include mutual information, conditional mutual information, and divergence. The connection weights can incorporate prior knowledge on probability distributions of database values. Information-theoretic connectionist techniques have been successfully applied to the problems of extracting probabilistic rules from pairs of interdependent attributes [6], speech recognition [7], feature selection [15], and rule induction [16]. The procedure for constructing a multi-layer information-theoretic network is briefly described in the next sub-sections. Complete details can be found in [20].

3.1 Extended Relational Model

We use the following formal notation of the relational model [13]:

- $R = (A_1, \dots, A_N)$ - a schema of a relation (data table) containing N attributes
- D_i - the domain of an attribute A_i .
- V_{ij} - the value j in the domain D_i .
- $t_k[A_i]$ - value of an attribute A_i in a tuple k , $t_k[A_i] \in D_i$.

To build an information-theoretic network, we define the following types of attributes in a relation schema:

- 1) A subset $O \subset R$ of *target* (“output”) attributes ($|O| \geq 1$). This is a subset of attributes, which can be predicted by the information-theoretic network. If the values of these attributes are already available, we can evaluate their reliability by using the method of Section 4 below.
- 2) A subset $C \subset R$ of *candidate input* attributes ($|C| \geq 1$). These attributes can be used to predict the values of target attributes.

The following constraints are imposed on the above partition of the relation schema:

- 1) $C \cap O = \emptyset$, i.e. the same attribute cannot be both a candidate input and a target.
- 2) $C \cup O \subseteq R$, i.e. some attributes are allowed to be neither candidate inputs nor targets. Usually, these will be the key (identifying) attributes.

Now we proceed with describing the structure of a connectionist network designed to predict the values of target attributes.

3.2 Connectionist Network Structure

An information-theoretic connectionist network has the following components:

- 1) I - a subset of *input* (predicting) attributes selected by the network construction algorithm from the set C of *candidate input* attributes.
- 2) $|I|$ - total number of *hidden* layers (levels) in a network. Unlike the standard decision tree structure [27], where the nodes of the same tree level are independent of each other, all nodes of a given network layer are labeled by the same input attribute associated with that layer. Consequently, the number of network layers is equal to the number of input attributes. In layers associated with continuous attributes, an information network uses multiple splits, which are identical at all nodes of the corresponding layer. The first layer in the network (Layer 0) includes only the root node and is not associated with any input attribute.
- 3) L_l - a subset of nodes z in a hidden layer l . Each node represents an attribute-based test, similarly to a standard decision tree. If a hidden layer l is associated with a nominal input attribute, each outgoing edge of a non-terminal node corresponds to an attribute distinct value. For continuous features, the outgoing edges represent the intervals obtained from the *discretization* process. If a node has no outgoing edges, it is called a *terminal node*. Otherwise, it is connected by its edges to the nodes of the next layer, which correspond to the same subset of input values.
- 4) K - a subset of target nodes representing distinct values in the domain of the target attribute. For continuous target attributes (e.g., percentage of votes for certain candidate), the target nodes represent the user-specified intervals of the attribute range. The target layer does not exist in the standard decision-tree structure. The connections between terminal nodes and the nodes of the target layer may be used for predicting the values of the target attributes and extracting information-theoretic rules (see [16]).

3.3 The Network Construction Procedure

The network construction algorithm starts with defining the target layer, where each node stands for a distinct target value, and the “root” node representing an empty set of input attributes. The connections between the root node and the target nodes represent unconditional (prior) probabilities of the target values. The network is built only in one direction (top-down). After the construction process is stopped, there is no bottom-up post-pruning of the network branches. The process of *pre-pruning* the network is explained below

A node is split on the values of an input attribute if it provides a statistically significant increase in the *mutual information* of the node and the target attribute. Mutual information, or information gain, is defined as a decrease in the conditional entropy of the target attribute (see [3]). If the tested attribute is nominal, the splits correspond to the attribute values. Splits on continuous attributes represent thresholds, which maximize an increase in mutual information. At each iteration, the algorithm re-computes the best threshold splits of continuously-valued candidate input attributes and chooses an attribute (either discrete, or continuous), which provides the maximum overall increase in mutual information across all nodes of the current final layer.

The maximum increase in mutual information is tested for statistical significance by using the Likelihood-Ratio Test [28]. This is a general-purpose method for testing the null hypothesis H_0 that two discrete random variables are statistically independent. If H_0 is

rejected, a new hidden layer is added to the network and a new attribute is added to the set I of input attributes. The nodes of a new layer are defined for a Cartesian product of split nodes of the previous final layer and the values of a new input attribute. According to the chain rule (see [3]), the mutual information between a set of input attributes and the target (defined as the overall decrease in the conditional entropy) is equal to the sum of drops in conditional entropy at all the layers. If there is no candidate input attribute significantly decreasing the conditional entropy of the target attribute, no more layers are added and the network construction stops.

The main steps of the construction procedure for a single target attribute are summarized in Table 1. If a data table contains several target attributes, a separate network is built, by using the same procedure, for each target attribute. Complete details are provided in [20].

Table 1 Network Construction Algorithm

<i>Input:</i>	The set of n training instances; the set C of candidate input attributes (discrete and continuous); the target (classification) attribute A_i ; the minimum significance level <i>sign</i> for splitting a network node (default: <i>sign</i> = 0.1%).
<i>Output:</i>	A set I of selected input attributes and an information-theoretic network. Each input attribute has a corresponding hidden layer in the network.
<i>Step 1</i>	Initialize the information-theoretic network (single root node representing all records, no hidden layers, and a target layer for the values of the target attribute).
<i>Step 2</i>	While the number of layers $ I < C $ (number of candidate input attributes) do
<i>Step 2.1</i>	For each candidate input attribute $A_{i'} \notin I$ do If $A_{i'}$ is continuous then Return the best threshold splits of $A_{i'}$. Return the conditional mutual information $cond_MI_{i'}$ between $A_{i'}$ and the target attribute A_i . End Do
<i>Step 2.2</i>	Find the candidate input attribute $A_{i'}^*$ maximizing $cond_MI_{i'}$
<i>Step 2.3</i>	If $cond_MI_{i'} = 0$, then End Do. Else Expand the network by a new hidden layer associated with the attribute $A_{i'}$, and add $A_{i'}$ to the set I of selected input attributes.
<i>Step 2.4</i>	End Do
<i>Step 3</i>	Return the set of selected input attributes I and the network structure

3.4 Predicting Continuous Target Values

Like in decision trees, a predicted target value is assigned to every terminal node of an information-theoretic network. Each record of a training set is associated with one and only one terminal node, which can be found by the procedure described in Table 2 below. The *predicted value* $Pred_{iz}$ of a continuous target attribute A_i at a terminal node z is calculated as the expected value of A_i over all the training records associated with the node z .

Table 2 Associating Record with a Terminal Node

<i>Input:</i>	The set I of selected input attributes; the values of input attributes in a tuple (record) k ; the information-theoretic network
<i>Output:</i>	The ID of a terminal node corresponding to the tuple k : $Node_F_k$
<i>Step 1</i>	Initialize the current node ID: $z = 0$
<i>Step 2</i>	Initialize the layer number: $m = 0$
<i>Step 3</i>	If a node z is terminal, then go to <i>Step 7</i> Else , go to the next step
<i>Step 4</i>	Increment the number of layers: $m = m+1$
<i>Step 5</i>	Find the next hidden node z by following the edge corresponding to the value of the input attribute m in the tuple k
<i>Step 6</i>	Go to <i>Step 3</i>
<i>Step 7</i>	Return $Node_F_k = z$

4 Evaluating Reliability of Target Attributes

4.1 Fuzzy Approach to Data Reliability

The main cause of having unreliable data in a database are the errors committed by an information source, which may be a human user, an automated measuring device, or just another database. In the case of the Year 2000 elections in the State of Florida, the Democrats have argued that the votes were not counted properly. The legal controversy was focused on the so-called “undervotes”, votes not tabulated by the counting machine due to apparent defects in the punch cards. The claim of the Democrats was that the undervotes have biased the results in favor of their opponent, the Republican Candidate George W. Bush. For example, they have questioned the voting results of Palm Beach County, which seemed particularly unreliable based on the demographic characteristics and the voting traditions of people in that specific county.

An expert user examining a familiar database can estimate quickly, and with a high degree of confidence, the reliability of stored information. He, or she, would define some records as “highly reliable”, “not so reliable”, “doubtful”, “absolutely unreliable”, etc. However, what is the exact definition of “data reliability”?

The most common “crisp” approach to data reliability is associated with data validity: some attribute values are valid while others are not. For example, if the valid range of a numeric attribute is $[50,100]$, the value of 100.1 is considered invalid and will be rejected during the data entry process. This is similar to the statistical concept of confidence intervals: any observation outside the interval boundaries is rejected, which means that its statistical validity is zero. The limitations of this approach are obvious: a real validity range may have “soft” boundaries.

It seems reasonable to define the reliability of an attribute value as a mean frequency (or probability) of that particular value, since values of low probability may be assumed less reliable than the most common values. This is similar to the information gain approach of [8]: the most surprising patterns are suspicious as unreliable. However, the information gain approach is not readily applicable to evaluating reliability of continuous

attributes, which can take an infinite number of distinct values, each having a very low probability of occurrence.

Noisy data is not necessarily unreliable data, and vice versa. In some areas, like the stock market, the data may be inherently noisy (having a high variance and a high entropy), because the real-world phenomenon, it represents, depends on many independent and dynamic, mostly unknown, factors. Still, the source of noisy data may be completely reliable. On the other hand, the information on a very stable phenomenon (having a low variance) may be corrupted during the data entry process.

Statistical information, obtained from training data, is certainly not sufficient for distinguishing between reliable and unreliable values. People use their intuition, background knowledge, and short-time memory, rather than any probabilistic criteria, for detecting lowly reliable data. Moreover, as indicated by Kandel *et al* [10], the probabilistic approach seems to be against the nature of human reasoning. Thus, we turn to the fuzzy set theory, which is a well-known approach to catching different aspects of human perception and making use of available prior knowledge.

The fuzzy set theory provides a mathematical tool for representing imprecise, subjective knowledge: the fuzzy membership functions. These functions are used for mapping precise values of numeric variables to vague terms like “low”, “high”, “reliable”, etc. The form of a specific membership function can be adjusted by a set of parameters. For example, a triangular membership function is defined by its prototype, minimum, and maximum values. For modeling human perception of reliability, the non-linear, sigmoid function seems more appropriate, since more probable values are usually perceived as more reliable, though all lowly reliable values are considered unreliable to nearly the same degree. The shape of this membership function depends on user perception of unexpected data, ranging from a “step function” (the crisp approach: only values in a specific range are reliable) to a continuous membership grade, giving a non-zero reliability degree even to very distant and unexpected values.

Thus, adopting the fuzzy logic theory and looking at the reliability degree as a fuzzy measure seems an appropriate approach to automating the human perception of data reliability. In [19], we have proposed the following definition for the degree of data reliability:

Definition 1. *Degree of Reliability of an attribute A in a tuple k is defined on a unit interval $[0,1]$ as the degree of certainty that the value of attribute A stored in a tuple k is correct from user’s point of view.*

This definition is consistent with the definition of fuzzy measures in Klir and Yuan [12], since a set of correct attribute values can be viewed as a “crisp” set, and we are concerned with the certainty that a particular attribute belongs to that set. It is also related to the fuzzy concept of “usuality” [34], where the fuzzy set of normal (or regular) values is considered the complement of a set of exceptions. Two special cases of Definition 1 are: degree of reliability = 0 (the data is clearly erroneous) and degree of reliability = 1 (the data is completely reliable, which is the implicit assumption of most database systems).

According to Definition 1, the degree of reliability is an attribute-dependent, tuple-dependent and user-dependent measure. It may vary for different attributes of the same tuple, for the same attribute in different tuples and for different users who have distinct views and purposes with respect to the same data. The subjectiveness of data reliability was best demonstrated in the 2000 election controversy. While the Democrats complained

about the unreliable voting results, the same numbers seemed perfectly accurate to their political opponents.

Data correctness does not imply precision. It just means that if a user could know the exact state of the real-world system, his or her decision, based on that data, would not be changed. After the 2000 elections, the real controversy was not about the exact number of votes for each candidate, which could be determined only by a tedious hand count. Both parties were just eager to know who won the *majority* of votes in the State of Florida.

4.2 Calculating Degree of Reliability

After finding a *predicted* value of the target attribute A_i in a tuple k , we compute the degree of reliability of the *actual* target value by the following formula:

$$t_k[R_i] = \frac{2}{1 + e^{\alpha \cdot d_{ik}}} \quad (1)$$

Where

α - exponential coefficient expressing the user perception of “unexpected” data. Low values of α (about 1) make it a sigmoid function providing a gradual change of reliability degree between 0 and 1 within the attribute range. Higher values of α (like 10 or 20) make it a step function assigning a reliability degree of zero to any value, which is different from the expected one.

d_{ik} - a measure of distance between the actual value $t_k[A_i]$ and the predicted value $Pred_{iz^*}$ ($z^* = Node_F_k$) of a target attribute A_i in a tuple k . For continuous target attributes, the distance measure is calculated by:

$$d_{ik} = \frac{abs(t_k[A_i] - Pred_{iz^*})}{Range_i} \quad (2)$$

Where $Range_i$ is the difference between the maximum and the minimum values of the attribute A_i . According to Equation 2, d_{ik} is a linear measure of the difference between predicted and actual values, which is normalized to the $[0, 1]$ range. The reliability degree in Equation 1 is defined on the same range, but it represents the *non-linearity* of reliability perception as a function of data deviation from the expected value, which can be predicted from the information-theoretic network.

In Figure 1, we show the reliability degree $t_k[R_i]$ as a function of the distance d_{ik} for two different values of α : $\alpha = 1$ and $\alpha = 5$. Equation 1 satisfies the four requirements of a fuzzy measure (see [12], p. 178): boundary conditions, monotonicity, continuity from below and continuity from above. The way to verify that is to look at the proximity to the predicted value as a reciprocal of the distance d_{ik} . Then the reliability of the empty set (zero proximity, or infinite distance) is zero and the reliability of the complete set (infinite proximity, or zero distance) is one. Reliability degree is a continuous monotonic function of proximity by its mathematical definition in Equation 1.

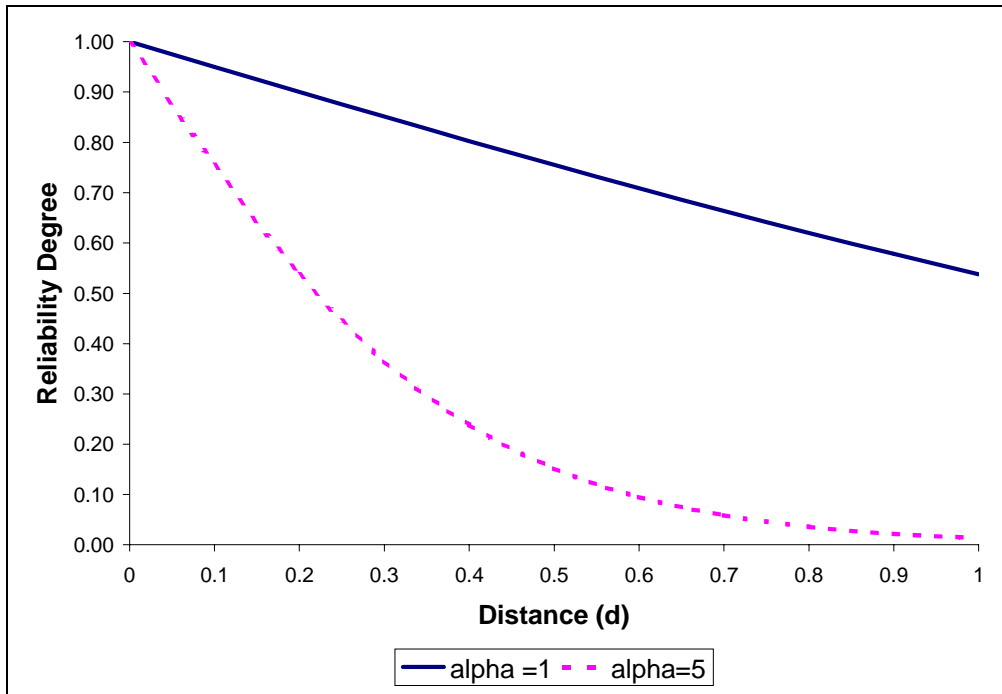


Figure 1 Reliability perceptions for different values of alpha

5 Case Study: Palm Beach Election Data

We have applied the information-theoretic fuzzy approach to the precinct-level voting data of the 2000 Presidential Election in Palm Beach County, Florida. The results of the initial count and the demographic data on each precinct (including voter registration information) have been downloaded from the web page of Dr. Bruce E. Hansen [9] in November 2000. The original source for the data was Palm Beach County web page. The list of attributes in the Palm Beach dataset is presented in Table 3 below. The raw data included absolute numbers (number of votes and number of voters). We have normalized these numbers to the percentage out of the total number of votes / voters in the corresponding precinct. Since there is a strong dependency between the percentages of votes for each major candidate in the same precinct, we have arbitrarily chosen the percentage of votes for Bush as the target attribute. After normalization, the values of the target attribute have been discretized to five intervals of approximately equal frequency. The Palm Beach dataset includes 14 candidate-input attributes, representing the percentage of registered voters in each party and the distribution of the voting population across several age groups. The dataset has 494 records referring to all the voting precincts of Palm Beach County. The 106 absentee precincts were excluded from the analysis due to the lack of demographic information.

Table 3 Palm Beach Dataset - List of Attributes

Ser No	Attribute Name	Meaning	Type	Use in Network
1	Precinct	Precinct No	Nominal	None
2	Bush	Percentage of Votes	Continuous	Target
3	Gore	Percentage of Votes	Continuous	None
4	Nader	Percentage of Votes	Continuous	None
5	Buchanan	Percentage of Votes	Continuous	None
6	Total_Vote	Percentage of Votes	Continuous	None
7	McCollum	Percentage of Votes	Continuous	None
8	Nelson	Percentage of Votes	Continuous	None
9	DEM_PTY	Percentage of Registered Voters	Continuous	Candidate Input
10	REP_PTY	Percentage of Registered Voters	Continuous	Candidate Input
11	OTHER_PTY	Percentage of Registered Voters	Continuous	Candidate Input
12	WHITE	Percentage of Registered Voters	Continuous	Candidate Input
13	BLACK	Percentage of Registered Voters	Continuous	Candidate Input
14	HISPANIC	Percentage of Registered Voters	Continuous	Candidate Input
15	OTHER_RACE	Percentage of Registered Voters	Continuous	Candidate Input
16	MALE	Percentage of Registered Voters	Continuous	Candidate Input
17	FEMALE	Percentage of Registered Voters	Continuous	Candidate Input
18	AGE_18-20	Percentage of Registered Voters	Continuous	Candidate Input
19	AGE_21-29	Percentage of Registered Voters	Continuous	Candidate Input
20	AGE_30-55	Percentage of Registered Voters	Continuous	Candidate Input
21	AGE_56-64	Percentage of Registered Voters	Continuous	Candidate Input
22	AGE_65&UP	Percentage of Registered Voters	Continuous	Candidate Input

The results of applying the information-theoretic procedure of sub-section 3.3 above to the Palm Beach Dataset are shown in Table 4. Only three out of 14 candidate input attributes (*REP_PTY*, *WHITE*, and *DEM_PTY*) have been identified as statistically significant and included in the Information-Theoretic Network. The column “Conditional MI” in Table 4 shows the net decrease in the entropy of the target attribute “*Bush*” due to adding each input attribute. The first input attribute (*REP_PTY*) alone contributes nearly 90% of the overall mutual information (1.435 bits). This attribute is shown in bold. The next two input attributes (*WHITE* and *DEM_PTY*) contribute about 8% and 2% respectively. The first and the third input attributes are not surprising, since people tend to vote by their political association. The input attribute No. 2 (White) is an indicator of some weak relationship between the racial origin of the voters and their votes.

Table 4 Palm Beach Dataset - Summary of Results

Iteration	Attribute Name	Mutual Information	Conditional MI	Percentage Of MI	Conditional Entropy
0	REP_PTY	1.282	1.282	89.3%	1.04
1	WHITE	1.4	0.118	8.2%	0.922
2	DEM_PTY	1.435	0.035	2.4%	0.887

The constructed information-theoretic network has been used for evaluating the reliability of the target attribute (percentage of Bush votes in each precinct) by the fuzzy-based method of Section 4 above. We have calculated the degrees of reliability with $\alpha = 1.00$. The resulting reliability degrees range between 0.617 and 1.000. As indicated above, these reliability degrees refer to the initial voting results certified by the Palm Beach County after the Election Day. During the following weeks, these results were in the center of a legal controversy until the US Supreme Court halted the vote recount on December 12, 2000. However, the public was still interested to know the “ground truth”: who would be the actual winner of the Election in Florida, if the hand recount could be continued to its completion? For this reason, the Miami Herald and other media organizations have conducted a complete review of the “undervote” ballots in all Florida counties. The precinct-level results have been posted on the Miami Herald web site [30]. To evaluate the usefulness of the data reliability calculations, we have examined the number of undervotes and the resulting change in the gap between the candidates for the precincts having the highest and the lowest reliability degrees (see Tables 5 and 6 below).

The total number of undervotes in 20 precincts having the lowest reliability degrees (Table 5) is much larger than the number of undervotes in 20 precincts with highest reliability (Table 6). In other words, starting the count of undervotes in low reliability precincts would help to detect significant gaps, like the one in Precinct No. 191, as early as possible. From a close look at the data of this precinct, one can see that the predicted percentage of Bush votes is high (51.5%) due to high percentages of Republicans and whites among the voters. However, Mr. Bush has got only 37.2% of votes in this precinct. The low reliability of this result (0.838) has been confirmed by the count of undervotes, which has added the net amount of 28 votes to Bush.

Table 5 Low Reliability Precincts

Precinct	Dem	Rep	White	Pred. Vote	Act. Vote	Reliability	Under votes- Bush	Under votes- Gore	Total Under votes	Abs. Gain
154C	4.12	84.02	97.94	0.515	0.866	0.617	4	0	4	4
33	13.17	77.72	98.37	0.515	0.846	0.637	0	1	1	1
154B	13.18	74.00	99.63	0.515	0.797	0.687	2	1	3	1
167	15.25	69.50	97.34	0.515	0.761	0.724	1	0	1	1
162I	8.07	78.60	97.54	0.515	0.758	0.727	0	0	0	0
37	52.20	37.11	59.21	0.295	0.519	0.749	0	5	5	5
122A	49.66	24.16	73.83	0.251	0.452	0.773	0	1	1	1
001A	15.79	67.64	97.68	0.515	0.679	0.814	6	2	8	4
36	46.83	38.03	78.52	0.404	0.562	0.820	0	0	0	0
148E	37.43	45.99	76.47	0.515	0.667	0.827	0	1	1	1
151	17.68	60.10	91.92	0.515	0.659	0.836	0	1	1	1
163	30.87	48.23	94.86	0.515	0.372	0.837	0	1	1	1
191	29.73	50.23	96.77	0.515	0.372	0.838	84	56	140	28
121A	36.14	42.57	89.11	0.398	0.53	0.850	0	0	0	0
158	35.34	48.54	98.25	0.515	0.394	0.863	2	1	3	1
225	32.62	46.34	92.34	0.515	0.398	0.867	23	21	44	2
90	17.31	64.24	96.24	0.515	0.631	0.867	1	1	2	0
045A	33.19	46.36	93.51	0.515	0.399	0.868	5	11	16	6
49	35.22	45.34	97.03	0.333	0.448	0.869	0	4	4	4
093A	39.78	40.37	94.50	0.398	0.511	0.871	1	0	1	1
Total									236	62

Table 6 High Reliability Precincts

Precinct	Dem	Rep	White	Pred. Vote	Act. Vote	Reliability	Under votes- Bush	Under votes- Gore	Total Under votes	Abs. Gain
156C	41.76	31.87	79.12	0.326	0.322	0.996	0	0	0	0
38	82.69	8.20	2.96	0.072	0.069	0.997	0	0	0	0
78	42.18	40.47	83.97	0.404	0.401	0.997	0	1	1	1
110	43.39	39.31	91.10	0.398	0.4	0.997	6	9	15	3
128G	43.26	33.80	80.26	0.326	0.328	0.997	1	1	2	0
159J	36.03	40.24	84.63	0.404	0.406	0.997	2	1	3	1
201	35.63	43.19	90.48	0.398	0.396	0.997	3	0	3	3
003B	32.08	42.64	95.00	0.398	0.4	0.998	2	2	4	0
114	68.21	19.44	26.54	0.253	0.255	0.998	0	2	2	2
119	40.12	41.93	95.18	0.398	0.397	0.998	2	3	5	1
120	45.00	27.50	84.53	0.251	0.249	0.998	0	3	3	3
144E	44.19	34.99	75.35	0.326	0.324	0.998	33	43	76	10
162A	80.42	9.88	98.19	0.072	0.074	0.998	0	2	2	2
205E	41.51	35.82	92.55	0.28	0.279	0.998	4	1	5	3
007A	31.71	60.98	95.12	0.515	0.515	0.999	0	0	0	0
018J	50.45	33.32	93.99	0.28	0.281	0.999	1	7	8	6
88	37.09	42.29	92.81	0.398	0.397	0.999	5	1	6	4
115	45.53	41.30	93.12	0.398	0.397	0.999	4	5	9	1
203	29.88	52.74	96.34	0.515	0.515	1.000	0	0	0	0
Total									144	40

6 Conclusion

In this paper, we have presented a novel fuzzy-based approach to evaluating reliability of continuous attributes in a relational database. The approach includes partition of a data table into input and target attributes, induction of a data mining model (information-theoretic network) from a set of training data, and calculation of reliability degrees for target values based on their distance from the values predicted by the network.

The proposed approach combines objective information about the data, which is represented by an information-theoretic network, with a subjective, user-specific perception of data quality. In our case study, we have shown that the method can be an efficient tool for detection of inaccurate information in a real-world database.

Related issues, to be further studied, include: integrating the method with other data mining models, evaluating reliability of input attributes, and detecting unreliable information in non-relational data.

Acknowledgment. This work was supported in part by the National Institute for Systems Test and Productivity at USF under the USA Space and Naval Warfare Systems Command Grant No. N00039-01-1-2248 and in part by the USF Center for Software Testing under grant No. 2108-004-00.

References

- [1] N. Ahituv, S. Neumann, H.N. Riley, Principles of Information Systems for Management, B & E Tech, Dubuque, Iowa, 1994.
- [2] P. Bosc and O. Pivert, SQLf: A Relational Database Language for Fuzzy Querying, IEEE Transactions on Fuzzy Systems, vol. 3, no. 1, pp. 1-17, 1995.
- [3] T. M. Cover, Elements of Information Theory, Wiley, New York, 1991.
- [4] J.F. Elder IV and D. Pregibon, A Statistical Perspective on Knowledge Discovery in Databases. In Advances in Knowledge Discovery and Data Mining, U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds. AAAI/MIT Press, Menlo Park, CA, pp. 83-113, 1996.
- [5] U. Fayyad, G. Piatetsky-Shapiro, and P. Smith, From Data Mining to Knowledge Discovery in Databases, AI Magazine, vol. 3, pp. 37-54, 1996.
- [6] R. M. Goodman, J. W. Miller, P. Smyth, An Information Theoretic Approach to Rule-Based Connectionist Expert Systems. In Advances in Neural Information Processing Systems, D.S. Touretzky, Ed., Morgan Kaufmann, San Mateo, CA, pp. 256-263, 1988.
- [7] A.L. Gorin, S.E. Levinson, A.N. Gertner and E. Goldman, Adaptive Acquisition of Language, Computer Speech and Language, vol. 5, no. 2, pp. 101-132, 1991.
- [8] I. Guyon, N. Matic, and V. Vapnik, Discovering Informative Patterns and Data Cleaning. In Advances in Knowledge Discovery and Data Mining, U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, and R. Uthurusamy, Eds. AAAI/MIT Press, Menlo Park, CA, pp. 181-203, 1996.
- [9] B. E. Hansen, Florida Data Page [<http://www.ssc.wisc.edu/~bhansen/vote/data.html>]
- [10] A. Kandel, Fuzzy Mathematical Techniques with Applications, Addison-Wesley, Reading, MA, 1986.
- [11] A. Kandel, R. Pacheco, A. Martins, and S. Khator, The Foundations of Rule-Based Computations in Fuzzy Models. In Fuzzy Modelling, Paradigms and Practice, W. Pedrycz, Ed., Kluwer, Boston, pp. 231-263, 1996.
- [12] G. J. Klir and B. Yuan, Fuzzy Sets and Fuzzy Logic: Theory and Applications, Prentice-Hall Inc., Upper Saddle River, CA, 1995.
- [13] H.F. Korth and A. Silberschatz, Database System Concepts, McGraw-Hill, Inc., New York, 1991.
- [14] W. Kurutach, Managing Different Aspects of Imperfect Data in Databases, Proc. 1995 IEEE Int'l Conf. on SMC, Vancouver, BC, Canada, pp. 2812-2817, 1995.
- [15] M. Last, A. Kandel, O. Maimon, Information-Theoretic Algorithm for Feature Selection, Pattern Recognition Letters, Vol. 22, No. 6, pp. 799-811, 2001.
- [16] M. Last, Y. Klein, A. Kandel, Knowledge Discovery in Time Series Databases, IEEE Transactions on Systems, Man, and Cybernetics, Vol. 31, Part B, No. 1, pp. 160-169, 2001.
- [17] H. Lu, R. Setiono, and H. Liu, Effective Data Mining Using Neural Networks, IEEE Transactions on Knowledge and Data Engineering, vol. 8, no. 6, pp. 957-961, 1996.

- [18] O. Maimon, A. Kandel, and M. Last, Information-Theoretic Fuzzy Approach to Knowledge Discovery in Databases. In *Advances in Soft Computing - Engineering Design and Manufacturing*, R. Roy, T. Furuhashi and P.K. Chawdhry, Eds. Springer-Verlag, London, 1999.
- [19] O. Maimon, A. Kandel, and M. Last, Fuzzy Approach to Data Reliability. In *Knowledge Management in Fuzzy Databases*, O. Pons, A. Vila, and J. Kacprzyk, Eds., Physica-Verlag, pp. 89-101, 2000.
- [20] O. Maimon and M. Last, *Knowledge Discovery and Data Mining, The Info-Fuzzy Network (IFN) Methodology*, Kluwer, Boston, 2000.
- [21] T.M. Mitchell, *Machine Learning*, McGraw-Hill, New York, 1997.
- [22] A. Motro, Imprecision and Uncertainty in Database Systems. In *Fuzziness in Database Management Systems*, P. Bosc and J. Kacprzyk, Eds., Springer-Verlag, pp. 3-22, 1995.
- [23] H. Nakajima and Y. Senoh, A Spreadsheet-Based Fuzzy Retrieval System, *International Journal of Intelligent Systems*, vol. 11, no. 0, pp. 661-670, 1996.
- [24] Palm Beach County Elections - Election Results
[<http://www.pbcelections.org/eresults.htm>]
- [25] F. E. Petry, *Fuzzy Databases, Principles and Applications*, Kluwer, Boston, MA, 1996.
- [26] J.R. Quinlan, Induction of Decision Trees, *Machine Learning*, vol. 1, no. 1, pp. 81-106, 1986.
- [27] J. R. Quinlan, *C4.5: Programs for Machine Learning*, Morgan Kaufmann, San Mateo, CA, 1993.
- [28] C.R. Rao and H. Toutenburg, *Linear Models: Least Squares and Alternatives*, Springer-Verlag, 1995.
- [29] Y. Takahashi, Fuzzy Database Query Languages and Their Relational Completeness Theorem, *IEEE Transactions on Knowledge and Data Engineering*, vol. 5, no. 1, pp. 122-125, 1993.
- [30] The Miami Herald Ballot Review
[http://advapps.herald.com/election_results/default.asp]
- [31] Y. Wand and R.Y. Wang, Anchoring Data Quality Dimensions in Ontological Foundations, *Communications of the ACM*, vol. 39, no. 11, pp. 86-95, 1996.
- [32] R.Y. Wang, V.C. Storey, and C.P. Firth, A Framework for Analysis of Data Quality Research, *IEEE Transactions on Knowledge and Data Engineering*, vol. 7, no. 4, pp. 623-639, 1995.
- [33] R.Y. Wang, M.P. Reddy, H.B. Kon, Toward Quality Data: An Attribute-based Approach, *Decision Support Systems*, vol. 13, pp. 349-372, 1995.
- [34] L.A. Zadeh, Syllogistic Reasoning in Fuzzy Logic and its Application to Usuality and Reasoning with Dispositions, *IEEE Transactions on Systems, Man, and Cybernetics*, vol. 15, no. 6, pp. 754-763, 1985.
- [35] M. Zemankova-Leech and A. Kandel, *Fuzzy Relational Databases - a Key to Expert Systems*, Verlag TUV, Koln, Germany, 1984.