Analysis of Uncertain Data: Tools for Representation and Processing

Bin Fu
Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
binf@cs.cmu.edu

Eugene Fink
Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
e.fink@cs.cmu.edu

Jaime G. Carbonell
Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213, USA
jgc@cs.cmu.edu

***Abstract*—We present initial work on a general-purpose system for the analysis of incomplete and uncertain data, integrated with Excel. We explain the representation of main types of uncertainty, and outline tools for the analysis of uncertain data and planning of additional data collection.**

***Index Terms*—Uncertainty, data collection, Excel.**

I. Introduction

The available knowledge about the real world is inherently uncertain, and we usually make decisions based on incomplete and partially inaccurate data. For example, we purchase goods without learning about all their features, sign contracts without fully understanding the fine print, and make investments without full knowledge of the financial markets. Human experts almost never have complete data relevant to their work, and they usually can make reasonable decisions in the face of uncertainty. Furthermore, they are able to estimate the risks involved in their decisions, anticipate major contingencies, and determine which additional data would help to perform their tasks.

The purpose of the described work is to develop tools for the use of uncertain data in reasoning and optimization. We have previously investigated the problem of scheduling based on uncertain information about available resources and scheduling constraints [Bardak *et al*., 2006a; Bardak *et al*., 2006b; Fink *et al*., 2006a; Fink *et al*., 2006b, Bardak, 2007]. We are now working on a general-purpose system for the analysis of uncertainty, which will help human experts to use incomplete and approximate data in specific reasoning tasks. This work involves several challenges, including representation of uncertain knowledge, its use in specific tasks, evaluation of the certainty of resulting conclusions, analysis of contingencies, identification of critical missing data, and planning of additional data collection. We have explored these problems and built an initial system for processing uncertain data, integrated into the Excel software. This work has been part of the RAPID project at Carnegie Mellon University, aimed at building an intelligent system for the analysis of homeland security data.

The manuscript was received on March 16, 2008. The described work has been supported by the Air Force Research Laboratory (AFRL) under Contract No. FA8750-07-2-0137.

We begin with an example of analyzing uncertain facts (Section II). We then explain the representation of uncertain data in the developed system (Section III), and describe tools for the analysis of these data (Section IV).

II. Example

Suppose that we need to analyze data about four companies that may be trying to evade taxes, and decide which of them require further investigation. For each company, we look at its category, number of employees, annual revenue, and amount of last-year taxes (Figure 1).

If some of these data are uncertain, we specify ranges of possible values. For instance, if we do not know the exact revenue of the Sazan Sepahan Café chain, but we know that it is between 100 and 200 million, then we represent it as [100, 200]. For each company, we define the utility of available data, which reflects their certainty and relevance to the given task (see Column L in Figure 1).

If the system is unable to draw definite conclusions based on available data, it estimates the probability of its conclusions. For example, the estimated probability that Sazan Sepahan may be evading taxes is 0.368 (see Column H in Figure 1).

The system also helps to determine which additional data would improve the conclusion certainty. For every uncertain value, we specify the available actions for obtaining more accurate data, called *probes*. The specification of a probe includes its cost and probability of getting a more accurate value. For example, the cost of a probe for the revenue of Sazan Sepahan is 50, and its success probability is 0.3 (see the highlighted cells in Figure 1).

For each probe, the system evaluates the expected utility of obtaining a more accurate value, and computes the probe utility as the difference between its utility impact and cost. For example, the estimated utility of the revenue probe for Sazan Sepahan is 0.958.

The system identifies the probes with positive utility, and requests the user to gather the related data. If the user obtains some of these data, the system uses them to re-evaluate its conclusions. For example, if the user determines that the revenue of Sazan Sepahan is 120 million, the system removes it from the list of suspected tax evaders, and then focuses on probes related to other companies (Figure 2).

A	B	D	E	F	G	H	I	J	L
	Company Name	Category	Size (Employees)	Revenue (Millions)	Tax (Millions)	Evading Taxes?	Data Utility	Utility Weight	Weighted Utility
Values	Pishgamon	IT	[20, 50]	[100, 200]	40	FALSE	1	8	8
Probe cost	High-Tech	10	20	50	100				
Probe chance		0.9	0.5	0.3	0.2				
Probe utility									
Values	National	Mineral	[100, 250]	40	[0, 10]	{ .842, TRUE}	0.684	6	4.104
Probe cost	Coal Mines	10	20	50	100				
Probe chance		0.9	0.5	0.3	0.2				
Probe utility			0.024		-0.256				
Values	Sazan Sepahan	Food	[15, 20]	[100, 200]	[30, 40]	{ .368, TRUE}	0.264	8	2.112
Probe cost	Cafe	10	20	50	100				
Probe chance		0.9	0.5	0.3	0.2				
Probe utility			-0.076	0.958	-0.964				
Values	Global Tech	unknown	[50, 100]	[400, 1000]	200	{ .474, TRUE}	0.52	10	0.52
Probe cost	Corporation	10	20	50	100				
Probe chance		0.9	0.5	0.3	0.2				
Probe utility		-0.568	0.678	0.422					
									Total Utility
									14.736

Fig. 1: Example scenario, which includes uncertain data about four companies, along with related utilities and probes. For each company, the system evaluates the probability of tax evasion (Column H), as well as the quality of the data used for this evaluation (Column L), and selects probes for gathering critical additional data.

A	B	D	E	F	G	H	I	J	L
	Company Name	Category	Size (Employees)	Revenue (Millions)	Tax (Millions)	Evading Taxes?	Data Utility	Utility Weight	Weighted Utility
Values	Pishgamon	IT	[20, 50]	[100, 200]	40	FALSE	1	8	8
Probe cost	High-Tech	10	20	50	100				
Probe chance		0.9	0.5	0.3	0.2				
Probe utility									
Values	National	Mineral	[100, 250]	40	[0, 10]	{ .79, TRUE}	0.58	6	3.48
Probe cost	Coal Mines	10	20	50	100				
Probe chance		0.9	0.5	0.3	0.2				
Probe utility			-0.046		-0.276				
Values	Sazan Sepahan	Food	[15, 20]	120	[30, 40]	FALSE	1	8	8
Probe cost	Cafe	10	20	50	100				
Probe chance		0.9	0.5	0.3	0.2				
Probe utility									
Values	Global Tech	unknown	[50, 100]	[400, 1000]	200	{ .48, TRUE}	0.04	10	0.4
Probe cost	Corporation	10	20	50	100				
Probe chance		0.9	0.5	0.3	0.2				
Probe utility		-0.508	0.634	0.422					
									Total Utility
									19.88

Fig. 2: Update of the example scenario in Figure 1. The user has found out the exact revenue of Sazan Sepahan, which has enabled the system to exclude Sazan Sepahan from suspected tax evaders (Column H). The system now focuses on probes for National Coal Mines and Global Tech Corporation.

III. Representation

We have implemented the uncertainty-analysis system in Microsoft Excel, thus combining all standard Excel capabilities with the processing of uncertain numeric values, strings, and mathematical functions.

A. Uncertain values

We represent an uncertain number by a *piecewise-uniform distribution*, which is a set of disjoint intervals, with a probability assigned to each interval (Figure 4a). For example, suppose that the revenue of Sazan Sepahan Café is definitely between 100 and 200 million, and that it is between 140 and 160 million with 60% probability; then, we represent it by the piecewise-uniform distribution in Figure 3. Note that we may approximate any probability-density function by a piecewise-uniform distribution; for instance, the distribution in Figure 3 may be an approximation of the normal distribution shown by the grey line.

We also support uncertain strings, which are discrete probability distributions over specific strings, and we allow concatenation and nesting in their representation (Figure 4b). For instance, we may represent typical misspellings of the first word in the name of Sazan Sepahan Café by the following uncertain string:

"S" · (0.8: "a", 0.2: "u") · (0.8: "z", 0.1: "s", 0.1: "zz") · "an",

which specifies the probabilities of six different spellings:

0.64: Sazan	0.16: Suzan
0.08: Sasan	0.02: Susan
0.08: Sazzan	0.02: Suzzan

Furthermore, we support *nominal types*, which are strings with limited sets of possible values. For example, we may represent the company category as a nominal type with four possible values: "IT", "Food", "Mineral", and "Banking". We view Booleans as a special nominal type, whose possible values are TRUE and FALSE.

For all uncertain types, we allow two special values, UNKNOWN and ABSENT. The UNKNOWN value indicates that we have no data about a specific value or related probabilities, whereas ABSENT means that this value is inapplicable. For example, the UNKNOWN amount of tax indicates that we have no information about a company's tax payment, whereas ABSENT tax means that the company is tax-exempt.

B. Uncertain functions

We support three mechanisms for representing uncertain mathematical functions, called *piecewise-linear functions*, *list functions*, and *library functions*.

Piecewise-linear functions: We may represent a dependency between two numeric values by a piecewise-linear function, encoded by the coordinates of its segment endpoints. In Figure 5, we show an example function, which describes a dependency between revenue and taxes.

The representation of an uncertain dependency is based on the combination of a piecewise-linear function with uncertain numeric values. Specifically, we represent it by a function with uncertain coordinates of endpoints, as shown by the grey boxes in Figure 5. We do not allow the use of ABSENT and UNCERTAIN values in the representation of endpoints, and we also do not allow overlaps among possible intervals of different *x*-coordinates.

Furthermore, we allow specifying an uncertain dependency by multiple possible functions and their probabilities. We summarize the representation of uncertain piecewise-linear functions in Figure 6(a).

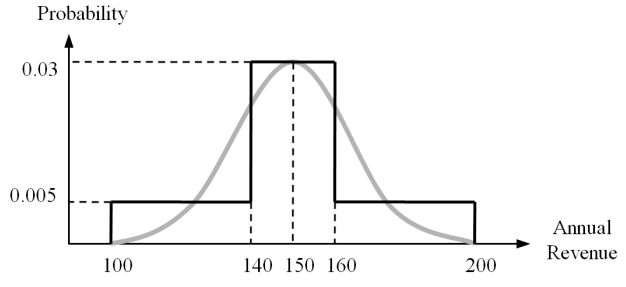

Fig. 3: Piecewise-uniform distribution for an uncertain revenue, which is between 100 and 140 with 20% probability, between 140 and 160 with 60% probability, and between 160 and 200 with 20% probability. This distribution is an approximation of the normal distribution shown by the grey line.

(a) Uncertain numeric value

$prob_1$: from min_1 to max_1
$prob_2$: from min_2 to max_2
…
$prob_m$: from min_m to max_m

We describe an uncertain numeric value by multiple intervals and their probabilities, and we specify each interval by its minimal and maximal value. The intervals do not overlap, and the sum of the probabilities is 1.0, which means that we impose the following constraints:

$min_1 \le max_1 \le min_2 \le max_2 \le \ldots \le min_m \le max_m$
$prob_1 + prob_2 + \ldots + prob_m = 1.0$

(b) Uncertain string

Certain string:
<*uncertain-str*> ::= <*certain-str*>

Possible values with probabilities:
<*uncertain-str*> ::= $prob_1$: <*uncertain-str*$_1$>
$prob_2$: <*uncertain-str*$_2$>
…
$prob_m$: <*uncertain-str*$_m$>

where $prob_1 + prob_2 + \ldots + prob_m = 1.0$

Concatenation:
<*uncertain-str*> ::= <*uncertain-str*$_1$> · … · <*uncertain-str*$_n$>

We construct an uncertain string from specific strings using discrete probability distributions and concatenations, and we allow multiple levels of nested distributions and concatenations.

Fig. 4: Representation of uncertain values.

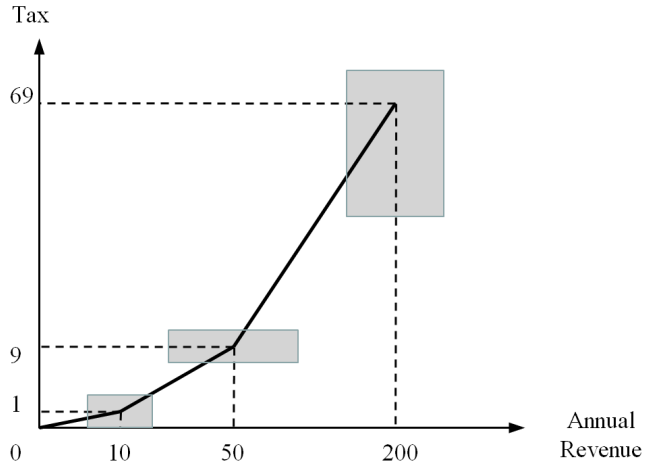

Fig. 5: Example of a piecewise-linear function (solid line), which may include uncertainty in the representation of its endpoints (grey boxes).

(a) Uncertain piecewise-linear function

$prob_1$: (x_{11}, y_{11}), (x_{12}, y_{12}), …
$prob_2$: (x_{21}, y_{21}), (x_{22}, y_{22}), …
…
$prob_m$: (x_{m1}, y_{m1}), (x_{m2}, y_{m2}), …

We describe an uncertain function by multiple piecewise-linear functions and their probabilities. The description of each piecewise-linear function is a list of segment endpoints sorted by x-coordinate. The point coordinates may be uncertain; however, they cannot be UNKNOWN or ABSENT. For each piecewise-linear function, the possible intervals of different x-coordinates do not overlap, which means that the following inequalities hold with full certainty:

$$x_{11} < x_{12} < \ldots$$
$$x_{21} < x_{22} < \ldots$$
$$\ldots$$
$$x_{m1} < x_{m2} < \ldots$$

Furthermore, the probabilities of different functions sum to 1.0:

$$prob_1 + prob_2 + \ldots + prob_m = 1.0$$

(b) Uncertain list function

$(set\text{-}x_1, y_1)$, $(set\text{-}x_2, y_2)$, …, $(set\text{-}x_m, y_m)$, (ELSE, y)

We describe a function by a list of pairs, where the *set-x* element of a pair is a nonempty set of arguments, and y is the respective function value. The ELSE pair specifies the function value for all arguments that do not belong to any set.

Fig. 6: Representation of uncertain functions.

List functions: A list function is an alternative mechanism for representing a dependency between two values; its domain includes certain numeric values or strings, whereas its range may be any certain or uncertain type. We encode it by sets of values in its domain and respective function values (Figure 6b).

For example, we may represent the dependency between a company category and typical revenue by a list function, where the domain includes company categories (strings), and the function values are respective revenues (uncertain numbers):

{Food}:	[100, 200]
{IT, Mineral}:	[50, 200]
ELSE:	UNKNOWN

Library functions: A library function is a mathematical function encoded by a Java procedure, which may have multiple arguments. Its arguments may be numbers or strings, whereas its range may be any certain or uncertain type. This mechanism allows the use of any computable functions, and thus it is more powerful than piecewise-linear and list functions. On the negative side, the system treats library functions as black boxes, which limits the effectiveness of reasoning algorithms.

IV. UNCERTAINTY ANALYSIS

We next outline the developed tools for the analysis of uncertainty, which include arithmetic operations on uncertain data, evaluation of data utility, and planning of data collection.

Arithmetic: We may view each uncertain value as a random variable, which allows the use of the Monte-Carlo simulation for approximate computation of standard arithmetic and logical operations. The system generates multiple sets of specific values according to given probabilities, applies given operations to each set of values, and constructs distributions of the results. If a resulting distribution is a set of scattered points, the system converts it into a smoothed piecewise-uniform distribution, by dividing the range of these points into intervals and determining the probability of each interval. In Figure 7, we illustrate the use of this procedure to compute an approximate sum of two uniform distributions.

Utility functions: The system supports the specification of utility functions, which are numeric expressions that represent the utility of available data for completing specific tasks. For each utility, we may include its weight, which is a numeric expression representing the importance of the related task.

The system computes the overall utility of the available data as the weighted sum of all specified utilities. It uses these utilities to keep track of the quality of available data, identify critical uncertainties, and evaluate the available probes for collecting additional data.

For example, the spreadsheet in Figure 1 includes utility functions for evaluating data about each company (Column I), which depend on the certainty of related conclusions about tax evasion. It also includes the weights of these utilities (Column J), which depend on company sizes.

Probes: We represent available actions for gathering additional data by probes, which specify target data, costs of the related actions, and chances of getting the desired data as a result of these actions (see Figure 1). We may specify probe costs and success chances by arithmetic expressions, thus encoding their dependency on other data.

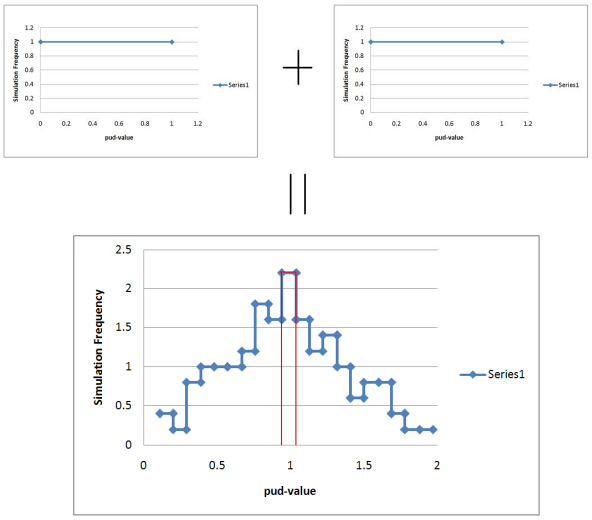

Fig. 7: The result of applying the Monte-Carlo procedure to compute the sum of two uniform distributions (top). The system constructs a piecewise-linear distribution (bottom), which is an approximation of the target sum.

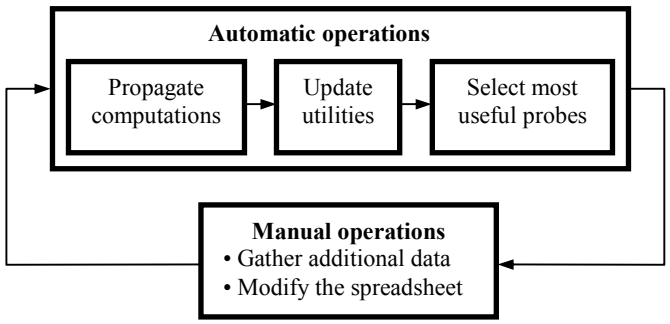

Fig. 8: Data-collection cycle, which includes the selection of probes, gathering of related data, and update of these data in the spreadsheet.

The utility of a probe is the expected impact of the related actions on the overall utility, which is the difference between the expected utility increase due to new data and the expected probe cost. For each probe, the system keeps track of its utility, and re-computes it after each change to the spreadsheet. Furthermore, the system identifies probes with positive utilities, sorts them in the decreasing order of utilities, and presents this sorted list to the user, which serves as data-collection advice. In Figure 8, we show the typical data-gathering cycle, which includes selecting the most useful probes, gathering the related data, and adding these data to the spreadsheet.

V. UNCERTAINTY ANALYSIS

We have described an initial version of general-purpose tools for working with incomplete and partially unknown data, which allow representing uncertainty, applying standard arithmetic operations to uncertain values, evaluating the utility of available data, and planning of data collection. We have integrated them with Excel, thus enabling users without programming experience to build task-specific Excel spreadsheets for uncertainty analysis. We are now working on more advanced tools, which will allow tracking the sources and reliability of available data, analyzing contingencies, and planning of complex data-collection strategies.

AKNOWLEDGMENT

We are grateful to Anatole Gershman for his detailed comments and suggestions, which have helped to extend and focus the presentation. We thank Diwakar Punjani and Andrew Yeager for their work on the development and testing of the described system. We also thank Nancy Roberts for her comments, and Mehrbod Sharifi for his help with formatting.

REFERENCES

[Bardak, 2007] Ulas Bardak. Information elicitation in scheduling problems. Ph.D. Thesis, Language Technologies Institute, Carnegie Mellon University, 2007.

[Bardak *et al*., 2006a] Ulas Bardak, Eugene Fink, and Jaime G. Carbonell. Scheduling with uncertain resources: Representation and utility function. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics,* pages 1486–1492, 2006.

[Bardak *et al*., 2006b] Ulas Bardak, Eugene Fink, Chris R. Martens, and Jaime G. Carbonell. Scheduling with uncertain resources: Elicitation of additional data. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics,* pages 1493–1498, 2006.

[Fink *et al*., 2006a] Eugene Fink, Ulas Bardak, Brandon Rothrock, and Jaime G. Carbonell. Scheduling with uncertain resources: Collaboration with the user. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics,* pages 11–17, 2006.

[Fink *et al*., 2006b] Eugene Fink, P. Matthew Jennings, Ulas Bardak, Jean Oh, Stephen F. Smith, and Jaime G. Carbonell. Scheduling with uncertain resources: Search for a near-optimal solution. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics,* pages 137–144, 2006.