The Impact of Data Accuracy on System Learning

DANIEL E. O'LEARY

DANIEL E. O'LEARY is an Associate Professor in the School of Business, at the University of Southern California. He received his Ph.D. from Case Western Reserve University, his master's degree from the University of Michigan and a B.S. from Bowling Green University. Professor O'Leary is the editor of the *International Journal of Intelligent Systems in Accounting, Finance and Management*, and is on the editorial boards of a number of journals including *Expert Systems with Applications*, *Accounting, Management and Information Technologies*, and *Advances in Mathematical Programming and Financial Planning*. Dr. O'Leary has published a number of papers in the areas of artificial intelligence and expert systems; decision sciences and operational research; and information systems. In particular, he has published papers in *Decision Sciences*, *European Journal of Operational Research*, *IEEE Expert*, *International Journal of Expert Systems*, *International Journal of Intelligent Systems*, and *International Journal of Man–Machine Studies*, among others.

ABSTRACT: The purpose of this paper is to study the impact of database accuracy on system learning. The paper assumes a basic model of an information system with a database, a rulebase, and an embedded machine learning approach that is used to add rules to the rulebase. The system learns from its database, changes to that database, and the examination of other databases. The results in this paper can be of use in the analysis of the design and behavior of such learning systems. It is found that the information system accuracy impacts the magnitude of a measure of goodness of individual rules. Thus, if only rules of a certain magnitude are kept, then some rules will be discarded because of database inaccuracy, unless that inaccuracy is accounted for. In addition, by accounting for database inaccuracy, the direction of the impact on measure of goodness can be determined. In some cases, the impact on the direction is monotonic. This finding allows us to understand the impact of database inaccuracy, without explicitly taking account of that inaccuracy. Further, information system accuracy can impact the resulting order of importance of rules, within a set of rules. Since only those higher-ranked rules are kept, database accuracy and measure of goodness can impact what rules are retained in the rulebase of the system. As a result, it is important to account for the information system accuracy in learning information systems.

KEY WORDS AND PHRASES: artificial intelligence, decision support systems, expert systems, machine learning.

Acknowledgment: The author would like to thank the anonymous referees for their comments on an earlier version of this paper.

Copyright © M.E. Sharpe, Inc., 1993

1. Introduction

THE PURPOSE OF THIS PAPER IS TO INVESTIGATE THE IMPACT of database accuracy on the learning of an information system. It is assumed that there is an information system that "learns" by analyzing database information using a machine learning approach, embedded in the information system. Learning is actualized by adding new rules to a rulebase, based on that database analyisis.

The resulting learning system is assumed to function in a "real-world" environment, where the database may not be perfectly accurate. It is found that measures of goodness of the derived rules either overestimate or underestimate the appropriate value, in the presence of database inaccuracy. In addition, it is found that the resultant order of the measure of goodness for a set of rules differs if we consider the accuracy of the database. As a result, by not accounting for database accuracy, learning done by the system may be affected.

1.1. Learning Information Systems

If an information system is to be a learning system, it needs to "update the knowledge about the real world and add new knowledge" [7, p. 118]. An information system that learns might learn from its own database, from changes to that database, and from other available external databases. The resulting learning would then be stored in some corporate knowledge repository or knowledge base for future use.

The learning information system is a useful concept for a number of reasons. First, development of a dynamic and learning information system can help an organization respond to the environment. If the information system can learn, then the system may facilitate dynamic adoption to changes in the environment without direct and explicit intervention of human agents.

Second, an information system model that allows learning from the organization's database provides a prototype model of one way that organizations, in general, can learn. In one such sequential model, the firm's database is updated, based on events in the environment. Then the information system learns from the database and updates the organization's knowledge base or rulebase. Then the database is updated, and so on. In this model, organizational learning is dependent on the firm's database and changes to that database. The learning information system simulates the process of an organization learning from data.

Third, the ability of a system to learn from its own database can be viewed as a lower bound to what organizations should be able to learn. In particular, at the very least, organizations should be able to generate knowledge from their own databases. It is a lower bound since other learning may occur from information not captured explicitly in the organization's database, such as that captured by other types of agents for the organization.

In any case, in each of these three models, the database is a critical yet intermediate step in the learning of an information system. Data are captured and then used to develop knowledge. However, the data in the information system may be inaccurate.

Thus, as Simon might say, in a "real-world" situation, we would need to account for information system accuracy in the context of such learning models.

1.2. Information System Accuracy

There have been a number of approaches aimed at automating the learning process (e.g., [5]). However, that research typically ignores the intermediary nature of the database. Generally, it is assumed that the information system database is correct.

Unfortunately, rather than the actual underlying data, the information system database is simply a "report" of the actual underlying data. As a result, there may be errors in the database. In particular, if the underlying value is y, then there is some probability that the reported value is $-y\#$, rather than $y\#$, where # indicates the version of the variable captured in the database and $-y$ is "not y."

There are many reasons for the existence of inaccuracy. First, with the inputting of data there may have been some satisficing, so that the data that were inputted were regarded as "close enough." Unfortunately, if individual errors cascade, then they may not be "close enough." Second, humans make errors and some errors are not found. As a result, databases have errors.

Third, humans in the process of capturing or entering data may have made the errors on purpose. At any rate, the data may be in error.

1.3. Outline of This Paper

This paper proceeds as follows. Section 2 provides some background information and a brief discussion about a measure of knowledge (rule) "goodness." Rule goodness is used to choose from among different rules that have been generated from analysis of the data. That measure assumes that the data are correct. Section 3 extends the measure of rule goodness to include the notion that the database is a report of the underlying value, and that the data may not be accurate. An example is used to illustrate that the introduction of the notion of information accuracy can have a substantial impact on the measure of goodness. Section 4 investigates a number of properties of this model, including the situation of perfect database accuracy, and provides an example illustrating the substantial impact of accounting for database accuracy. Section 5 finds that in some situations the direction of movement of the change in the measure of goodness can be anticipated, via a monotonicity property. Section 6 investigates the resulting impact of accuracy on relative magnitude of the measure of goodness for rules. Order changes in the measure of goodness would impact the ultimate ranking and choice of rules. Section 7 provides a brief summary and analysis of the contributions of this paper; it also discusses implementation and some extensions to the paper.

2. Background: Organizational Discovery of Knowledge from Data

THIS SECTION PROVIDES SOME BACKGROUND FOR THE ARGUMENTS that follow in the remainder of the paper. The notation, generation of rules, database accuracy, and measure of rule goodness are discussed.

2.1. Notation

It is assumed that $-x$ is used to represent "not x." Pr (x) is used to represent the probability of x. Pr(x,y) is used to represent the probability of "x and y." Pr$(x|y)$ denotes the probability of x given y.

Throughout this paper, for purposes of presentation, the concern is with dichotomous decisions. However, the results presented here could be extended to other cases of more than two choices of x and not x (x).

2.2. Development of Rules

In this paper the learning mechanism is assumed to be one that generates rules from data. As the rules are generated, a measure of goodness is developed to help choose between the rules to determine which rules should be captured.

Generally, it is assumed that rules are of the form, "If $Y = y$ then $X = x$, with probability p." However, the approach presented in this paper could be extended to more general rules of the form "If $Y_1 = y_1, \ldots ,Y_n = y_n$, then $X_1 = x_1, \ldots ,X_n = x_n$, with probability p."

The underlying values of the conditions and the consequences will be denoted y and x, respectively. The values representing those underlying values that are captured in the information system are represented as $y\#$ and $x\#$, respectively. Thus, in the case of perfect accuracy of both conditions and consequences, $y = y\#$ and $x = x\#$.

2.3. Database Accuracy

It will be assumed that there are two types of errors that can occur in a database. We will consider situations where the database contains x, but should contain $-x$ (or contains $-x$ and should contain x) and where the database contains y, but should contain $-y$ (or contains $-y$ and should contain y).

Two models are presented. In the first model, it is assumed that y can be inaccurate, but that x is perfectly accurate. Such a situation may occur if only the y is in the database and x is generated as part of the learning process. That first model is used to generate the next model. In the second model it is assumed that both the x and the y can be inaccurate. The second model is the primary focus of the paper.

2.4. Measure of Rule Goodness

There are a number of different measures of rule goodness [5]. The measure of goodness is used to determine the relative importance (or order of importance) of the rules generated in the learning process.

Typically, the learning mechanism will use one of two approaches to prune the list of rules. First, a cutoff point on the order of importance may be used to choose which rules are added. If the rules have a value of the rule goodness above a certain quantity, g^*, then those rules are added to the rulebase. Second, only the n rules with the largest goodness measures might be added to the rulebase. Other approaches might be used; however, throughout, the magnitude and the order of the measures of goodness are the critical issue in terms of adding rules to the knowledge base.

The measure of goodness used in this paper was developed by Piatetsky-Shapiro [4]. That measure is based on the incremental contribution of the peice of information y. That measure attributes a larger measure of rule goodness to rules "if y then x," for which $\Pr(x|y)$ is larger than $P(x)$. Thus, learning requires that the conditional probability with the new information is greater than the prior without that information. In particular, that measure is:

$$\Pr(y)\ P(x|y) - \Pr(y)\ \Pr(x). \tag{1}$$

2.5. Rule Goodness and Database Accuracy

Database accuracy is critical since, as will be shown later, accuracy impacts the measure of goodness. By not accounting for the database accuracy (or inaccuracy) the cutoff point may eliminate rules that should be kept or it may lead to keeping rules that should be dropped. Further, by not accounting for database accuracy, the relative magnitudes may be altered, thus influencing which rules are kept when relative order is used as the selection basis for the rules.

3. Inaccuracy and the Measure of Goodness

IN THIS SECTION, DIFFERENT ASSUMPTIONS ABOUT THE ACCURACY of the elements x and y will be made in order to develop a report-based version of $\Pr(y|x)$, in particular, $\Pr(y\#|x\#)$. This section then investigates a number of properties of $\Pr(y\#|x\#)$. In addition, the comparative performance between $\Pr(y|x)$ and $\Pr(y\#|x\#)$ is illustrated in an example.

3.1. Assumption that x Is Perfectly Accurate

Assume that x is always perfectly accurate, that is, $\Pr(x|x\#) = 1$ and $\Pr(\sim x|x\#) = 0$. This might occur in those situations where the consequences are generated at the time of the analysis or contained in a different database than the y values.

From Bayes' theorem we know that $\Pr(y)\Pr(x|y) = \Pr(y|x)\Pr(x)$. Thus, the measure of goodness, equation (1), becomes:

$$\Pr(x)\ \ (\Pr(y|x) - \Pr(y)). \tag{2}$$

Information system accuracy is captured in the variable y. The only part of equation (2) that can consider the accuracy of the information system output y is $\Pr(y|x) - \Pr(y)$.

Thus, consider $\Pr(y\#|x) - \Pr(y\#)$ from (2), where $y\#$ is the report of the value from our information system. $\Pr(y\#|x)$ can be written as:

$$\Pr(y\#|x) = \Pr(y\#|x,y)\Pr(y|x) + \Pr(-y\#|x,y)\Pr(-y|x). \tag{3}$$

Equation (3) can be made simpler in those cases where the accuracy of the report of y is not contingent on the consequence x [6]. In those situations where there is an independence between reporting accuracy and consequence, $\Pr(y\#|x,y) = \Pr(y\#|y)$ and $\Pr(-y\#|x,-y) = \Pr(-y\#|-y)$. Thus in that situation, equation (3) becomes:

$$\Pr(y\#|x) = \Pr(y\#|y)\Pr(y|x) + \Pr(y\#|-y)\Pr(-y|x). \tag{4}$$

As a result, by substituting equation (4) into equation (2) with $y\#$, we have:

$$\Pr(x)\ (\Pr(y\#|y)\Pr(y|x) + \Pr(y\#|\ -y)\Pr(-y|x) - \Pr(y\#)). \tag{5}$$

This is a revised measure of goodness given inaccuracy in the database of the variable y. It still assumes that x is perfectly accurate. In the next section, equation (5) is generalized so that both x and y are inaccurate.

3.2. x and y Can Be Inaccurate

Next assume that both x and y can be inaccurate. Using equation (3), we can substitute $x\#$ for x.

$$\Pr(y\#|x\#) = \Pr(y\#|x\#,y)\Pr(y|x\#) + \Pr(y\#|x\#,-y)\Pr(-y|x\#). \tag{3'}$$

Now, $\Pr(y|x\#) = \Pr(y,x|x\#) + \Pr(y,-x|x\#)$; thus, $\Pr(y|x\#) = \Pr(y|x,x\#)\ \Pr(x|x\#) + \Pr(y|-x,x\#)\ \Pr(-x|x\#)$. As a result, equation (3′) becomes:

$$\begin{aligned}\Pr(y\#|x\#) = {} & \Pr(y\#|x\#,y)\ [\Pr(y|x,x\#)\Pr(x|x\#) + \Pr(y|-x,x\#)\Pr(-x|x\#)] + \\ & \Pr(y\#|x\#,-y)[\Pr(-y|x,x\#)\Pr(x|\ x\#) + \Pr(-y|\ -x,x\#)\Pr(-x|\ x\#)].\end{aligned} \tag{3''}$$

Again, if we assume that the state of the world is such that the report of the consequence is not dependent on the report of the condition [6], then equation (3″) can be rewritten as:

$$\begin{aligned}\Pr(y\#|x\#) = {} & \Pr(y\#|y)[\Pr(x|x\#)\Pr(y|x,x\#) + \Pr(-x\mid x\#)\Pr(y|\ \ -x,x\#)] \\ & + \Pr(y\#\mid -y)[\Pr(x\mid x\#)\Pr(-\ y\mid x,x\#) + \Pr(-x\mid x\#)\Pr(-y\mid -x,x\#).\end{aligned} \tag{6}$$

Further, if we assume that the actual state of the world is such that the condition is not dependent on the report of the consequence, then we have:

$$\begin{aligned}\Pr(y\#\mid x\#) = {} & \Pr(y\#\mid y)[\Pr(x\mid x\#)\Pr(y\mid x) + \Pr(-x\mid x\#)\Pr(y\mid -x)] \\ & + \Pr(y\#\mid -y)[\Pr(x\mid x\#)\Pr(-y\mid x) + \Pr(-x\mid x\#)\Pr(-y\ \mid -x)].\end{aligned} \tag{7}$$

In the remainder of the paper it will be assumed that both x and y can be inaccurate, and equation (7) will be the primary focus. This equation can be used to examine what happens to our measure of goodness when we make the real-world assumption of database inaccuracy.

3.3. Interpretation of the Probabilities

Consider the interpretation of the probabilities in equations (3″) and (7). First, the underlying events are not used, only the reports of events are actually used. Pr(y# | x#) is the probability that is ultimately used in the computation of the measure of goodness.

Second, when the assumption of independence between report of condition and report of consequence is made, Pr(y# | x#,y) becomes Pr(y# | y). This last probability is a measure of the reporting accuracy of the condition database.

Third, Pr(x | x#) also is a measure of the accuracy of the database, only from the perspective of the consequence information. Finally, if the condition is assumed independent of the report of the consequence, then Pr(y | x,x#) becomes Pr(y# | x), the probability that generally most learning algorithms assume they are deriving. Thus, using equation (7), we can compare the underlying conditional probability to the conditional probability of the report of the data.

3.4. Impact of Inaccuracy: Example

The impact of incorporating the accuracy of the information system in the learning algorithms can be substantial, as illustrated by the example in Table 1.

For illustration purposes, it has been assumed that Pr(y# | y) is symmetric, so that Pr(y# |¬y) = 1 – Pr(y# | y). In addition, Pr(y | x) also is assumed to be symmetric. The assumption of symmetry reduces the number of combinations that need to be illustrated. In addition, as noted in the next section, with the assumption of symmetry, we can study the behavior of Pr(y# | x#). This is critical since it permits us to study the impact of consideration of information system inaccuracy.

Analysis of the example yields a number of possible implications that are explored in more detail later in sections 4, 5, and 6. First, the impact of inaccuracy is substantial. If we increase the accuracy from 0.90 to 1.00 (go from observation *b* to *a*), increasing the accuracy of both the condition and consequence by 0.1, the impact is larger than that 0.1 on Pr(y# | x#). In particular, Pr(y# | x#) increases by 0.162, to 0.950 for a 20.6 percent increase. Second, the value of Pr(y# | x#) decreases monotonically from 1 to 0.5 for observations *b* to *f* and *g* to *k*. Third, if any one of Pr(y# | y) and Pr(x | x#) is 0.5 (complete uncertainty of accuracy) then Pr(y# | x#) is 0.5 (complete uncertainty about the impact of x# on y#).

4. Impact of Accuracy on Magnitude

THE MODEL AS GIVEN IN EQUATION (7) (AND [3″]) is explored to understand its behavior in terms of changes in magnitude for individual rules. The first two subsections find that the model has desirable properties in the cases of completely uncertain and completely certain information. The following three subsections focus on other issues, including what happens when either the conclusion or the consequent data are accurate and the other one is not accurate and the duality of Pr(y# | x#).

Table 1 Information System Accuracy: Example*

Item	$\Pr(y\# \mid y)$	$\Pr(y\# \mid -y)$	$\Pr(x \mid x\#)$	$\Pr(y \mid x)$	$\Pr(y \mid -x)$	$\Pr(y\# \mid x\#)$
a	1.0	0.0	1.0	0.95	0.05	0.950
b	0.9	0.1	0.9	0.95	0.05	0.788
c	0.8	0.2	0.8	0.95	0.05	0.662
d	0.7	0.3	0.7	0.95	0.05	0.572
e	0.6	0.4	0.6	0.95	0.05	0.518
f	0.5	0.5	0.5	0.95	0.05	0.500
g	0.9	0.1	1.0	0.95	0.05	0.860
h	0.8	0.2	1.0	0.95	0.05	0.770
i	0.7	0.3	1.0	0.95	0.05	0.680
j	0.6	0.4	1.0	0.95	0.05	0.590
k	0.5	0.5	1.0	0.95	0.05	0.500

* Assumes the following relationship:
(7) $\Pr(y\#|x\#) = \Pr(y\#|y)[\Pr(x|x\#)\Pr(y|x)+\Pr(-x|x\#)\Pr(y| -x)] + \Pr(y\#| -y)[\Pr(x|x\#)\Pr(-y|x) + \Pr(-x|x\#)\Pr(-y|-x)]$.

4.1. Completely Certain Accuracy

In the case of complete certainty of accuracy, $\Pr(y\# \mid x\#)$ has the desired property that it reduces to $\Pr(y \mid x)$. Consider equation (7). In the situation of complete certainty, $\Pr(x \mid x\#) = 1$ and $\Pr(-x \mid x\#) = 0$. In addition, $\Pr(y\# \mid y) = 1$ and $\Pr(y\# \mid -y) = 0$. Thus, $\Pr(y\# \mid x\#) = \Pr(y \mid x)$.

4.2. Completely Uncertain Accuracy of Condition Data

First consider the case where the accuracy of the condition database system is completely uncertain. If the accuracy of the condition evidence is completely uncertain, then $\Pr(y\# \mid y) = \Pr(y\# \mid -y) = 0.5$. If the evidence from which our information system would learn would be completely uncertain, then we would anticipate that it would be better not to attribute different rule goodness (depending on $\Pr(y\#)$ and $\Pr(x\#)$) to derived rules. The finding of theorem 1 is that $\Pr(y\# \mid x\#)$ is the same for all such rules. This is not to say that the measure of goodness is the same, since it is normalized by prior probabilities as in equation (2).

Theorem 1

Assume that $\Pr(y\# \mid y) = \Pr(y\# \mid -y) = 0.5$. Assume that $\Pr(y \mid x)$ is symmetric. $\Pr(y\# \mid x\#) = 0.5$.

Proof: The proofs for theorem 1 and the remainder of the theorems presented in the paper are summarized in the appendix. A similar result can be developed for the case of $Pr(x \mid x\#) = Pr(-x \mid x) = 0.5$ and $Pr(y\# \mid y)$ is symmetric.

4.3. Partial Complete Accuracy

If $Pr(y\# \mid y)$ and $Pr(y / x)$ are symmetric, $Pr(y\# \mid x\#)$ takes the same value whether there is inaccuracy in condition and accuracy in consequence, or the converse. This result is important since it indicates that, in that situation, efforts to ensure accuracy of the condition or consequence data will encounter equal results. This is demonstrated in the theorem 2.

Theorem 2

Assume that $Pr(y / x,x\#)$ is symmetric. If $Pr(x \mid x\#) = Pr(y\# \mid x\#,x)$ and $Pr(-x / x\#) = Pr(y\# \mid x\#,-y)$, then $Pr(y\# \mid x\#$, condition information is accurate) = $Pr(y\# \mid x\#$, consequent information is accurate).

4.4. Duality

$Pr(y\# \mid x\#)$, under consideration of accuracy, has a duality property, when accuracies of both the condition and consequence are the same value, k. In that situation, the value of $Pr(y\# \mid x\#)$ is the same when that accuracy parameter is k or $1 - k$. This duality property is useful since it indicates, in some situations, that we need only consider the information systems with accuracy of condition and consequence greater than or equal to 0.5. Thus, the example in Table 1 only includes the values for $k \geq 0.5$ because the values for $k \leq 0.5$ are the mirror image.

Theorem 3

Consider equation (7). Assume that $Pr(y\# \mid y)$ and $Pr(y \mid x)$ are symmetric. Assume that the accuracy of the condition data and the consequent data are the same and symmetric. In that case, $Pr(y\# \mid x\#, [Pr(y\# \mid y) = Pr(x \mid x\#) = k]) = Pr(y\# \mid x\#, [Pr(y\# \mid y) = Pr(x \mid x\#) = 1-k])$.

5. Impact on Magnitude of Measure of Goodness

A PRIORI, IT IS UNCLEAR HOW ACCOUNTING FOR DATABASE ACCURACY will impact the magnitude of the measure of goodness (7). The purpose of this section is to study some special cases in which the direction of the change of magnitude can be predicted, when accuracy of the database is considered. This is done by examining the behavior of the $Pr(y \mid x)$ as compared with $Pr(y\# \mid x\#)$ under selected conditions.

5.1. Monotonic Increasing

In some cases equation (7) is monotonically decreasing or increasing in the accuracy of the information system. This is important since it indicates that by not accounting for the quality of the information system in the learning approach, the "measure of goodness" of discovered rules will be overemphasized or underestimated. Given the underlying probabilities, the magnitude for the measure of goodness actually computed will be too large or too small. Theorems 4 and 5 investigate such results.

Theorem 4

Assume that Pr(y# | y) = Pr(x | x#) are symmetric and greater than or equal to 0.5. Assume that Pr(y | x) ≥ 0.5, is symmetric. Let k_1 and k_2 be two different values of Pr(y# | y), such that $k_1 \geq k_2$. Pr(y# | x#, Pr(y# | y) = Pr(x | x#) = k_1) ≥ Pr(y# | x#, Pr(y# | y) = Pr(x | x#) = k_2).

A similar theorem, for the monotonicity of the Pr(y# | x#) can be developed for the case of Pr(y | x) ≤ 0.5. As might be anticipated from the duality property, instead of being monotonically increasing, it is monotonically decreasing in the accuracy.

Theorem 5

Assume that Pr(y# | y) = Pr(x | x#) are symmetric and greater than or equal to 0.5. Assume that Pr(y | x) ≥ 0.5, is symmetric. Let k_1 and k_2 be two different values of Pr(y# | y), such that $k_1 \geq k_2$. Pr(y# | x#, Pr(y# | y) = Pr(x | x#) = k_1) ≥ Pr(y# | x#, Pr(y# | y) = Pr(x | x#) = k_2).

Other monotonicity results can be developed for other sets of assumptions.

5.2. Implications

The results developed in this section indicate that by not accounting for the accuracy of the information system, Pr(y | x) overestimates or underestimates (in a predictable manner) the value of Pr(y# | x#). As a result, if a cutoff point is used to determine which generated rules are included in the knowledge base, then the measure of goodness either overestimates or underestimates the value of the rules that are gathered. As a result, rules are either included in the knowledge base when they should not be, or they are excluded when they should be in the knowledge base. Database accuracy impacts magnitude which impacts which rules are kept in the knowledge base.

6. Impact of Accuracy on Magnitude

THE PREVIOUS SECTION CONSIDERED THE IMPACT OF ACCURACY on the magnitude of the measure of goodness for single rules. This section considers the relative impact of measure of goodness on the set of rules generated through learning. Consider the development of multiple rules *i* and *j*. This section finds that, in general, by accounting for accuracy of the information system, the relative magnitude of the measure of goodness (7) of those two rules can be affected. As a result, in some situations, if accuracy is accounted for, then the measure of goodness of rule *i* may exceed the measure of goodness for rule *j*. However, if accuracy is not accounted for, then the measure of goodness of rule *j* may exceed the measure of goodness for rule *i*.

6.1. The General Case

In the case of developing multiple rules, *i* and *j*, the information system would be used to generate $\Pr(y_i\#| x_i\#)$ and $\Pr(y_j\# \mid x_j\#)$. If the reporting system accuracy did not make a relative "ordering" difference in the measure of goodness, then if $\Pr(y_i \mid x_i) \geq \Pr(y_j \mid x_j)$, then $\Pr(y_i\# \mid x_i\#) \geq \Pr(y_j\# \mid x_j\#)$. Unfortunately, there is no general reason to assume that the ordering without consideration of accuracy would be the same as the ordering with consideration, except in some special circumstances.

6.2. A Situation Where Order Does Not Change

There is at least one situation where the relative order of the measure of goodness for rules does not change when we consider the impact of the information system accuracy. Consider equation (7). It may be reasonable to assume that the accuracy of both the condition and consequent information is the same in the generation of different rules. In that situation, it would not be unreasonable to expect that order of measure of goodness would be preserved between different rules. That is the case in theorem 6.

Theorem 6 (Relative Order Preservation)

Suppose that $\Pr(y_i \mid x_i) \geq 0.5$ is symmetric for all *i*. Further suppose that $\Pr(y\# \mid y) = \Pr(x \mid x\#)$ is symmetric. If $\Pr(y_j \mid x_j) \geq \Pr(y_k \mid x_k)$, then $\Pr(y_j\# \mid x_j\#) \geq \Pr(y_k\# \mid x_k\#)$.

In the same sense that there is order preservation for $\Pr(y \mid x) \geq 0.5$, there is also order preservation for $\Pr(y \mid x) \leq 0.5$.

Theorem 7 (Relative Order Preservation)

Suppose that $\Pr(y_i \mid x_i) \leq 0.5$ is symmetric for all *i*. Further suppose that $\Pr(y\# \mid y) = \Pr(x \mid x\#)$ is symmetric. If $\Pr(y_j \mid x_j) \leq \Pr(y_k \mid x_k)$, then $\Pr(y_j\# \mid x_j\#) \leq \Pr(y_k\# \mid x_k\#)$.

6.3. Implications

The finding that in general the relative order of the measure of goodness for two rules does not stay the same is a critical issue. If rules are choosen by their relative measures of goodness, then unless accuracy is accounted for there is no guaranttee that the order is correct. This is critical since in some cases rules are added to the system knowledge on the basis of their relative measure of goodness, for example, only the rules with the *n* largest measures of magnitude would be added to the knowledge base. This section presented one result where that relative ordering of measures of goodness was not impacted by not accounting for the accuracy of the database. If the particular problem under consideration meets the assumptions of that result, then relative orderings of measures of goodness are maintained even if we do not directly account for database accuracy.

7. Summary, Contributions, Implementation, and Extensions

A LEARNING THEORY MODEL WAS DEVELOPED to account for database accuracy. That model has some desirable characteristics. First, it reduces to the model that assumes away the accuracy in the situation when there is perfect accuracy. Second, where there is complete uncertainty of the accuracy of the data, Pr(y# | x#) is equal to 0.5. Finally, it was shown with an example that the model that accounts for accuracy of the data was found to differ substantially from the model that did not include a model of accuracy.

Additional analysis of the model that incorporates database accuracy revealed two important special cases, given a symmetry assumption. First, Pr(y# | x#) is monotonic in the accuracy parameter. Second, in a special case, Pr(y# | x#) preserves the relative magnitude.

7.1. Contributions

This paper has investigated embedding the impact of the quality of the information system into machine learning approaches. It was found that both the magnitude and the relative order were affected by introducing the accuracy of the information system into the model. These findings indicate that by not accounting for accuracy, inappropriate knowledge may be added to the knowledge base, while appropriate knowledge is left out of the knowledge base. Pr(y# | x#) was found to be monotonic in the accuracy of the database. Thus, by not considering the information system accuracy, the results are likely to be either overstated or understated. Further, it was shown that in one case accounting for the accuracy does not change the order between the measures of goodness for two rules i and j. However, in general, the relative magnitude is not preserved.

7.2. Implementation of the Models

The implementation of the models that account for the accuracy of the information system may be difficult but it should not be overwhelming. The primary difficulty would be in the development of the probabilities. If we assume the form of equation (7), then at least two of the sets of probabilities can be developed by analyzing the accuracy of the database, Pr(y# | y) and Pr(x | x#). In addition, the probabilities Pr(y | x) can be developed from databases that have been thoroughly tested and examined.

The primary results presented in this paper have dealt with the assumption of symmetric probabilities. Empirical tests of the model could be used to determine if a symmetric model is appropriate. Generally, the symmetric model is theoretically appealing, since it suggests that there is symmetry in the errors. In addition, some closed form results can be developed using the symmetric model.

7.3. Extensions

The results in this paper can be extended to other approaches used for machine learning. For example, the approach could be used to investigate the algorithms used by Cheeseman et al. [1, 2] or Liang [3]. The paper examined only rules of the form "if y then x," with a single condition and consequence. The results of this paper could be extended to include either multiple conditions or multiple consequences or both. Primary attention was given to the symmetric model. Other results might be developed for more general forms of accuracy. Finally, this paper focused on the discovery of rules from data sets. Alternative approaches might focus on other forms of knowledge representation, for example, cases or other approaches.

REFERENCES

1. Cheeseman, P.; Kelly, J.; Self, M.; and Stutz, J. Automatic Bayesian induction of classes. *Proceedings of the Seventh National Conference on Artificial Intelligence.* Menlo Park, CA: American Association for Artificial Intelligence, 1988, pp. 607–611.
2. Cheeseman, P.; Kelly, J.; Self, M.; Stutz, J.; Taylor, W.; and Freeman, D. AutoClass: a Bayesian classification system. 54–64, *Proceedings of the Fifth International Conference on Machine Learning*. San Mateo, CA: Morgan Kaufman, 1988, pp. 54–64.
3. Liang, T.P. A composite approach to inducing knowledge for expert systems design. *Management Science, 38,* 1 (January 1992), 1–17.
4. Piatetsky-Shapiro, G. Discovery, analysis and presentation of strong rules. In G. Piatetsky-Shapiro and W. Frawley, *Knowledge Discovery in Databases.* Cambridge, MA: MIT Press, 1991, pp. 121–135.
5. Piatetsky-Shapiro, G., and Frawley, W. *Knowledge Discovery in Databases.* Cambridge, MA: MIT Press, 1991.
6. Schum, D., and De Charme, W. Comments on the relationship between the impact and the reliability of evidence. *Organizational Behavior and Human Performance, 6* (1971), 111–131.
7. Simon, H. *The Sciences of the Artificial,* 2d ed. Cambridge, MA: MIT Press, 1981.

APPENDIX: Theorem Proofs

ALL PROOFS USE THE GENERAL FORM OF EQUATION (7), (3″).

Theorem 1

$$0.5\,[\Pr(y \mid x,x\#)\Pr(x \mid x\#) + \Pr(y \mid -x,x\#)\Pr(-x \mid x\#)$$
$$+ \Pr(-y \mid x,x\#)\Pr(x \mid x\#) + \Pr(-y \mid x,x\#)\Pr(-x \mid x\#)] =$$

$$0.5\,[\Pr(x \mid x\#)](\Pr(y \mid x,x\#)+\Pr(-y \mid x,x\#)]$$
$$+[\Pr(-x \mid x\#)][\Pr(y \mid -x,x\#)+\Pr(-y \mid -x,x\#)] = 0.5.$$

Theorem 2

Consider equation (3″), where:

$$\Pr(y\# \mid x\#) = \Pr(y\# \mid x\#,y)\ [\Pr(y \mid x,x\#)\Pr(x \mid x\#) + \Pr(y \mid -x,x\#)\Pr(-x \mid x\#)]$$
$$+ \Pr(y\# \mid x\#,-y)[\Pr(-y \mid x,x\#)\Pr(x \mid x\#) + \Pr(-y \mid -x,x\#)\Pr(-x \mid x\#)].$$

If the condition information is perfectly accurate, then,

$$\Pr(y\# \mid x\#) = \Pr(x \mid x\#) \qquad \Pr(y \mid x,x\#) + \Pr(-x \mid x\#) \qquad \Pr(y \mid -x,x\#).$$

If the consequence information is perfectly accurate, then,

$$\Pr(y\# \mid x\#) = \Pr(y\# \mid x\#,y)\ \Pr(y \mid x,x\#) + \Pr(y\# \mid x\#,-y)\ \Pr(-y \mid x,x\#).$$

But since $\Pr(y \mid x,x\#)$ is assumed to be symmetric, they are equal for those situations where $\Pr(x \mid x\#) = \Pr(y\# \mid x\#,y)$ and $\Pr(-x \mid x\#) = \Pr(y\# \mid x\#,y)$.

Theorem 3

Using (3″),

$$\Pr(y\# \mid x\#, \Pr(y \mid x,x\#) = \Pr(x \mid x\#) = k) = k\ [\Pr(y \mid x,x\#)\ k$$
$$+ \Pr(y \mid -x,x\#)\ (1-k)] + (1-k)[\Pr(-y \mid x,x\#)\ k + \Pr(-y \mid -x,x\#)\ (1-k)].$$

Similarly, using (3″),

$$\Pr(y\# \mid x\#, \Pr(y\# \mid x\#,x) = \Pr(x \mid x\#) = 1-k) = (1-k)\ [\Pr(y \mid x,x\#)\ (1-k)$$
$$+ \Pr(y \mid -x,x\#)\ k] + k\ [\Pr(-y \mid x,x\#)\ (1-k) + \Pr(-y \mid -x,x\#)\ k].$$

Thus,

$$\Pr(y\# \mid x\#, \Pr(y\# \mid x\#,x) = \Pr(x \mid x\#) = k) = \Pr(y\# \mid x\#, \Pr(y\# \mid x\#,x)$$
$$= \Pr(x \mid x\#) = 1-k).$$

Theorem 4

$$\Pr(y\# \mid x\#, \Pr(y\# \mid x\#,y) = \Pr(x \mid x\#) = k_1) = k_1\ [\Pr(y \mid x,x\#)\ k_1$$
$$+ \Pr(y \mid -x,x\#)\ (1-k_1)] + (1-k_1)[\Pr(-y \mid x,x\#)k_1 + \Pr(-y \mid -x,x\#)(1-k_1)]$$

$$= \Pr(y \mid x,x\#)\ k_1{}^2 + \Pr(y \mid -x,x\#)\ k_1 - \Pr(y \mid -x,x\#)\ k_1{}^2 +$$
$$- \Pr(-y \mid x,x\#)\ k_1{}^2 + \Pr(-y \mid x,x\#)\ k_1 + \Pr(-y \mid -x,x\#)\ k_1{}^2$$
$$\Pr(-y \mid -x,x\#) - \Pr(-y \mid -x,x\#)\ 2\mathrm{k}_1.$$

$$\Pr(y\# \mid x\#, \Pr(y\# \mid x\#,y) = \Pr(x \mid x\#) = k)$$
$$= k_2 \quad [\Pr(y \mid x,x\#)\ k_2 + \Pr(y \mid -x,x\#)\ (1-k_2)]$$
$$+ (1-k_2)[\Pr(-y \mid x,x\#)k_2 + \Pr(-y \mid -x,x\#)(1-k_2)]$$
$$= \Pr(y \mid x,x\#)\ k_2{}^2 + \Pr(y \mid -x,x\#)\ k_2 - \Pr(y \mid -x,x\#)\ k_2{}^2$$
$$+ - \Pr(-y \mid x,x\#)\ k_2{}^2 + \Pr(-y \mid x,x\#)\ k_2 + \Pr(-y \mid -x,x\#)\ k_2{}^2$$
$$\Pr(-y \mid -x,x\#) - \Pr(-y \mid -x,x\#)\ 2k_2.$$

Assume:

$$\Pr(y\# \mid x\#, \Pr(y\# \mid x\#,y) = \Pr(x \mid x\#) = k_1)$$
$$< \Pr(y\# \mid x\#, \Pr(y\# \mid x\#,y) = \Pr(x \mid x\#) = k_2).$$

That would imply:

$$\begin{aligned}
&(k_1{}^2 - k_2{}^2)\,(\Pr(y \mid x,x\#) + \Pr(-y \mid -x,x\#) - \Pr(y \mid -x,x\#) - \Pr(-y \mid x,x\#))\\
&< \ (k_1 - k_2)\,(\Pr(-y \mid -x,x\#) + \Pr(-y \mid -x,x\#) - \Pr(y \mid -x,x\#) - \Pr(-y \mid x,x\#)).
\end{aligned}$$

However, since $(k_1{}^2 - k_2{}^2) = (k_1 - k_2)\,(k_1 + k_2)$, $k_i \geq 0.5$ and $\Pr(y \mid x,x\#)$ is symmetric and greater than or equal to 0.5, there is a contradiction.

Thus,

$$\begin{aligned}
&\Pr(y\# \mid x\#, \Pr(y\# \mid x\#,y) = \Pr(x \mid x\#) = k_1 \geq \Pr(y\# \mid x\#, \Pr(y\# \mid x\#,y)\\
&= \Pr(x \mid x\#) = k_2.
\end{aligned}$$

Theorem 5

The proof to theorem 5 is similar to that for theorem 4.

Theorem 6 (Relative Order of Measure of Goodness Preservation)

$$\begin{aligned}
&\Pr(y\# \mid x\#) = \Pr(y\# \mid y)[\Pr(x \mid x\#)\Pr(y \mid x) + \Pr(-x \mid x\#)\Pr(y \mid -x)]\\
&+ \Pr(y\# \mid -y)[\Pr(x \mid x\#)\Pr(-y \mid x) + \Pr(-x \mid x\#)\Pr(-y \mid -x)].
\end{aligned}$$

Using equation (7), we have:

$$\begin{aligned}
&\Pr(y_j\# \mid x_j\#)\\
&= k_1\ [k_2 \Pr(y_j \mid x_j) + (1 - k_2)\Pr(y_j \mid -x_j)]\\
&+ (1 - k_1)\ [k_2 \Pr(-y_j \mid x_j) + (1 - k_2) \Pr(-y_j \mid -x_j)];
\end{aligned}$$

$$\begin{aligned}
&\Pr(y_k\# \mid x_k\#)\\
&= k_1\ [k_2 \Pr(y_k \mid x_k) + (1 - k_2)\Pr(y_k \mid x_k)]\\
&+ (1 - k_1)\ [k_2 \Pr(-y_k \mid x_k) + (1 - k_2) \Pr(-y_k \mid -x_k)].
\end{aligned}$$

Assume that $\Pr(y_j\# \mid x_j\#) < \Pr(y_k\# \mid x_k\#)$. Then,

$$\begin{aligned}
&k_1\, k_2 \Pr(y_j \mid x_j) + k_1 \Pr(y_j \mid -x_j)\\
&- k_1\, k_2 \Pr(y_j \mid -x_j) + k_2 \Pr(-y_j \mid x_j)\\
&- k_1\, k_2 \Pr(-y_j \mid x_j) + \Pr(-y_j \mid -x_j)\\
&- k_1 \Pr(-y_j \mid -x_j) - k_2 \Pr(-y_j \mid -x_j)\\
&+ k_1\, k_2 \Pr(-y_j \mid -x_j)\\
&< \ k_1\, k_2 \Pr(y_k \mid x_k) + k_1 \Pr(y_k \mid x_k)\\
&- k_1\, k_2 \Pr(y_k \mid -x_k) + k_2 \Pr(-y_k \mid x_k)\\
&- k_1\, k_2 \Pr(-y_k \mid x_k) + \Pr(-y_k \mid -x_k)\\
&- k_1 \Pr(-y_k \mid -x_k) - k_2 \Pr(-y_k \mid -x_k)\\
&+ k_1\, k_2 \Pr(-y_k \mid -x_k).
\end{aligned}$$

Thus,

$$\begin{aligned}
&2\, k_1\, k_2 \Pr(y_j \mid x_j) + (k_1 + k_2) \Pr(y_j \mid -x_j)\\
&- 2\, k_1\, k_2 \Pr(y_j \mid -x_j) - (k_1 + k_2) \Pr(y_j \mid x_j) + \Pr(-y_j \mid -x_j)
\end{aligned}$$

$$
\begin{aligned}
&< 2\,k_1\,k_2\,\Pr(y_k \mid x_k) + (k_1 + k_2)\Pr(y_k \mid -x_k) \\
&- 2\,k_1\,k_2\,\Pr(y_k \mid -x_k) - (k_1 + k_2)\Pr(y_k \mid x_j) + \Pr(-y_k \mid -x_k).
\end{aligned}
$$

As a result,

$$
\begin{aligned}
&(1 + 2\,k_1\,k_2 - k_1 + k_2)\Pr(y_j \mid x_j) \\
&(k_1 + k_2 - 2\,k_1\,k_2)\,(\Pr(y_j \mid -x_j)) \\
&< (1 + 2\,k_1\,k_2 - k_1 + k_2)\Pr(y_k \mid x_k) \\
&(k_1 + k_2 - 2\,k_1\,k_2)\,(\Pr(y_k \mid -x_k))\,.
\end{aligned}
$$

Thus,

$$
\begin{aligned}
&(1 + 2\,k_1\,k_2 - k_1 + k_2)\,(\Pr(y_j - x_j) - \Pr(y_k \mid x_k)) \\
&<\ (k_1 + k_2 - 2\,k_1\,k_2)\,(\Pr(y_k \mid -x_k) - \Pr(y_j \mid -x_j))\,.
\end{aligned}
$$

But $(2\,k_1\,k_2 - k_1 - k_2) < 0$ for all $k_i < 1$ and $(\Pr(y_k \mid -x_k) - \Pr(y_j \mid -x_j))$. In addition, $(1 + 2\,k_1\,k_2 - k_1 + k_2) > 0.5$ for all $k_i < 1$ and $(\Pr(y_j \mid x_j) - \Pr(y_k \mid x_k))$ is greater than 0. Thus, this indicates that a positive quantity is less than 0 and there is a contradiction.

Theorem 7

The proof for theorem 7 is similar to that for theorem 4.