WizWhy White Paper

Abraham Meidan PhD

Many data sets contain valuable information that is not readily obvious. Examples of such information might be:

- Patterns of high-risk companies within financial data;
- Types of customers in a direct mailing list who are most likely to make a purchase;
- The relationships between patients' personal data and their medical diagnosis.

The search for these valuable, yet hidden, patterns and relationships within the data is known as data mining.

Users may be interested in data mining applications for several reasons:

- Some users are interested in data mining for issuing a summary of the data. When the data is too numerous to be reviewed record-by-record, there is a need for a summary, and revealing the main patterns in the data provides a useful summary.
- Other users expect data mining to reveal interesting phenomena in the data. These users wish to ignore trivial cases and concentrate rather on unexpected phenomena.
- Still other users are interested in issuing predictions for new cases. They seek to reveal the patterns in the data in order to use them for issuing predictions for new cases. For example, revealing the patterns of risky customers in financial data enables one to predict the extent to which a new customer is risky.
- Finally, some users may also be interested in using data mining for auditing purposes. The records that deviate from the discovered patterns in the data might be cases of data entry error or fraud.

How does WizWhy work?

Use of WizWhy is applicable to a data set that one wishes to analyze. WizWhy will determine how the values of one field are affected by the values of other fields.

For example, suppose that you maintain a customer data set where each record contains a range of fields relating to a customer, such as: Customer Name, Address, City, State, Field of Business, Salesperson, Amount Purchased, and % Growth Since Last Year. One of these fields, say, % Growth Since Last Year, should be defined as the dependent variable, while the other fields are the independent variables.

In this example, the aim might be to analyze customer retention; that is, to reveal the patterns of those customers where the % Growth Since Last Year is negative. % Growth Since Last Year might then be analyzed as either Boolean or continuous. In a Boolean analysis, the aim might be to reveal the patterns of the customers where the Growth Since Last Year is less than, say, 0%. A continuous analysis is more detailed: it calculates the specific % growth of each customer as a function of the other fields.

On analyzing the data, WizWhy performs the following operations:

- It first reads the data. The user selects the dependent variable (in the above-mentioned example - % Growth Since Last Year) and can fine-tune the analysis by defining parameters such as the minimum probability of the rules, the minimum number of cases in each rule, and the cost of a miss vs. the cost of a false alarm. WizWhy follows these "instructions" when issuing the rules.
- Within a short time, WizWhy lists the rules that relate between the dependent variable and the other fields. The rules are formulated as "if-then" formulas and "if-and-only-if" sentences. On the basis of the discovered rules, WizWhy also points out the main patterns and the unexpected phenomena and cases in the data.
- WizWhy can now make predictions for new cases; for instance, given the data of new customers, WizWhy calculates the expected % Growth for each of customer. These predictions can be either Boolean (for instance, whether or not the % Growth is above 0%) or continuous (for example, the % Growth is 30).

The scientific metaphor

There are many data-mining (and machine learning) algorithms, several of which are based on metaphors. Artificial Neural Networks allegedly imitate the activity of neurons in the brains and Decision Trees allegedly follow the rational method of mangers when they face many options and have to issue decisions. The inspiration for WizWhy's algorithm is the way scientists explain data. Scientists search for theories (namely, simple and accurate generalizations) that explain the data under research. The rules that WizWhy seek out are similar to the theories that scientists research. When WizWhy searches for rules, the target is to find the simplest and most accurate ones. These rules are presented either by formulas, by if-then sentences, by if-and-only-if statements, or by if-then-formula expressions.

What is an if-then rule?

WizWhy starts analyzing the data by revealing all the if-then rules that relate between the Dependent Variable and the other fields. An example of an if-then rule is the following:

If City is New-York
and Amount Purchased is 200 … 300 (average = 250)
and Salesperson is Dave
Then
Growth Since Last Year is less than 0%

Rule's probability: 0.70
The rule exists in 370 records
Significance Level: Error probability < 0.001

This rule says that for 70% of the customers, residing in the City New York, and purchasing between 200 and 300, and the Salesperson is Dave, the Growth Since Last Year is negative. There are 370 such customers.

The term "probability" designates what other data mining tools call "Confidence Level." Obviously, this probability should be significantly higher than the overall frequency of the value under analysis (i.e., the frequency of customers, where the Growth Since Last Year is less than 0%, is much lower than the rule probability, 70%).

"Error probability" indicates the degree to which the rule can be relied upon as a basis for predictions. Assuming that the data under analysis is a representative sample of an infinite population, the error probability quantifies the chances that the rule does not hold in the entire population and exists accidentally in the file under analysis.

Numeric fields, such as Amount Purchased and % Growth, are automatically segmented into intervals, and these intervals are the values in the if-then rules. For example, in the if-then rule above, the second condition refers to the case where the value in the Amount Purchased field is the interval between 200 and 300. WizWhy employs a unique algorithm for the optimal segmentation of numeric (continuous) fields.

The "association rules" method reveals all the if-then rules. One of the main challenges of such a method is to validate each possible relationship, in a reasonable time-span. For instance, the data might contain:

- 10,000 records
- 20 fields in each record
- An average of 10 possible values for each field

Using conventional means to check every possible relation in such data would require thousands of years. WizWhy employs a sophisticated algorithm that reveals all the rules astonishingly quickly.

What is a formula rule?

When the Dependent Variable is numeric and continuous, WizWhy also looks for formula rules. The formula rules can relate either to all the records (for example, Field A = Field B + Field C) or to a group of records sharing a certain if-clause. An example of

an if-then formula rule is the following:

> If City is New-York
> Then
> A = B*0.4 + 35
> Where:
> A - Growth Since Last Year
> B- Amount Purchased
>
> Accuracy level: 0.95
> The rule exists in 80 records

The accuracy level denotes the percentage of records in which the formula holds (within the formula limits determined when issuing the rule reports) relative to the records regarding which the condition (the if clause) holds.

What is an if-and-only-if rule?

On the basis of if-then rules, WizWhy proceeds to search for if-and-only-if rules. An example of an if-and-only-if rule is the following:

> The following conditions explain when the Growth Since Last Year is less than 0%:
>
> If at least one of these conditions holds, the probability that the Growth Since Last Year is less than 0% is 0.9
>
> If none of these conditions holds, the probability that the Growth Since Last Year is not less than 0% is 0.95
>
> The conditions are:
>
> 1. The Amount Purchased is between 0 … 199 (average = 100)
>
> 2. The Salesperson is Dan
> and the City is Boston

In other words, the Growth Since Last Year is negative if and only if the Amount Purchased is between 0 … 199 or the Salesperson is Dan and the City is Boston.

If-then rules represent sufficient conditions (the "if" condition is a sufficient condition for the result). If-and-only-if rules go one step further: they represent necessary and sufficient conditions. In the previous example, the two conditions, (1) and (2), are necessary and sufficient conditions for the growth being negative. If at least one of them holds, there is a high probability that the Growth is negative; and if the two of them do not hold, there is high probability that the Growth is not negative.

Obviously, such a relation cannot be accidental, and therefore might be relied upon when issuing predictions. Indeed, when WizWhy reveals if- and-only-if rules it takes them into account when issuing predictions for new cases.

Revealing the if-and-only-if rules also helps in pointing out the main patterns in data. One common "complaint" against if-then rules is that they are too numerous. Indeed, in many data sets, WizWhy may discover thousands of if-then rules. All of them are valid and may be used for issuing predictions, but cannot be manually reviewed in any practical way. Revealing if-and-only-if rules solves this problem. Each value of the dependent variable is explained by one or two if-and-only-if-rules. These if-and-only-if rules are optimal in the sense that they cover the maximum number of both positive and negative cases.

How does WizWhy summarize the data?

As already mentioned, some users are interested in data mining in order to issue a data summarization. They are interested in a report that presents the main patterns in the data.

WizWhy meets this objective by listing the relations between all the values in each field and the dependent variable. Consider the Amount Purchased field in the above-mentioned example. WizWhy segments the field into intervals (as mentioned, WizWhy employs a unique algorithm that segments numeric fields optimally) and displays the relation between each interval and the value under analysis (Growth Since Last Year is less than 0%).

For example, when analyzing the Amount Purchased field, WizWhy may reveal the following relations:

IF The Amount Purchased is between:	THEN The probability that the Growth Since Last Year is less than 0% is:
0 - 199	50%
200 - 300	40%
301 - 480	30%
481 - 791	15%

Each line is a one-condition if-then statement. Some or even all of these if-then statements may not be rules, since they may not meet the requirement that the rule probability be significantly higher or lower than the primary frequency. Still, all of them represent basic trends in the data. WizWhy applies such an analysis to each of the fields.

WizWhy also calculates the explanatory power of each field. The explanatory power designates the extent to which the field explains the dependent variable. When sorting the fields by this parameter, the fields appear listed according to their "importance" in explaining which customers are likely to leave for a competitor.

WizWhy illustrates graphically these one-condition relations: each value is represented as a bar, whose height denotes the probability of the customer leaving for a competitor, and whose width signifies the number of cases (i.e., customers) with this value.

This analysis of the basic rules and trends results in a data summarization. These basic rules and trends summarize the data, in the sense that they explain most of the other rules. More accurately, they explain all the rules that have more than one condition, with the exception of the unexpected rules. The idea of unexpected rules will be discussed further in the next section.

How does WizWhy reveal interesting phenomena?

Revealing interesting phenomena relies on the assumption that unexpected phenomena are interesting. For example, an event that is inconsistent with (namely, unexpected by) an accepted theory is an interesting event. Now, each rule can be viewed as an event, and the one-condition rules and trends (discussed in the previous section) can be viewed as the "basic theory" that describes the data. Therefore, calculation of how unlikely each rule is relative to the basic rules and trends can reveal the unexpected rules. These unexpected rules signify the interesting phenomena in the data.

The unexpected rule has at least two conditions. The basic rules have fewer conditions (in many cases they have one condition only); and by definition, the basic trends also have exclusively one condition. Each condition of the unexpected rule appears in the basic rules and trends. The unexpected rule is an unlikely event relative to the basic rules and trends.

The level of unlikelihood is computed in the following manner: consider a data set of 1,000 records, where each record refers to one patient, and contains the information regarding whether the patient shows either symptom A or B and the diagnosis (whether or not the patient suffers from the disease D). Suppose also that 30% of the patients have the disease D, and the following three rules were discovered:

1. If a patient exhibits symptom A, the probability that he or she suffers from disease D is 60%.
2. If the patient exhibits symptom B, the probability that he or she suffers from disease D is 60%.
3. If the patient exhibits both symptom A and symptom B, the probability that he or she suffers from disease D is 10%.

In this example, rules #1 and 2 are the basic rules, and rule #3 is the unexpected rule.

WizWhy calculates what should have been the probability of having the disease D among the patients exhibiting both symptoms, A and B, on the basis of rules 1 and 2, contrary to the actual probability in rule 3. This is the expected probability. To calculate the expected probability, WizWhy measures the dependency between symptom A and B. The difference between the expected probability and the actual probability signifies how unlikely rule 3 is in regard to rules 1 and 2.

WizWhy measures this unlikelihood in an additional way. WizWhy calculates the conditional probability of the event described in rule 3, given the conditions described in rules 1 and 2. In the example under discussion, the conditional probability is almost 0; therefore, the level of unlikelihood, which is 1 minus the

conditional probability, is almost 1. Note that the probability of the unexpected rule may be much higher or much lower than the expected probability. Any significant deviation is unexpected.

How does WizWhy issue predictions?

WizWhy makes use of the rules discovered in the data set in order to issue predictions for new cases. When a new record is entered, WizWhy applies the rules to the values of this record and calculates the expected value of the dependent variable. For example, one can run WizWhy on financial data, where the dependent variable is a field signifying whether the company has gone bankrupt. WizWhy will reveal the rules that relate between the company data and the probability of going bankrupt. When the data of a new company is entered, WizWhy will then apply the relevant rules and calculate the probability of this company going bankrupt. When issuing the predictions, WizWhy can list the rules that entail each prediction. These rules serve as the explanations for the predictions.

The predictions can be Boolean (for instance, whether the company will go bankrupt or not), multi-value (for example, given the patient's symptoms, what the disease is), or continuous (for instance, given the financial data, what the expected rate of growth is).

Still, when issuing the predictions on the basis of rules, one faces the following two problems:

(i) How can rules representing noise or overfitting be avoided?

For any rule discovered in the data, one may ask, what is the reason or the explanation for the existence of this rule? The two possible extreme answers are: (i) the rule exists in the entire population, of which the data under analysis is a representative sample; or (ii) the rule is the result of chance, i.e., it is a case of a noise or overfitting. Obviously, when issuing predictions for new cases, rules existing in the entire population are relevant, while rules representing noise should be ignored.

WizWhy assesses the probability that a rule is the result of chance by considering two parameters:

- The error probability of the rule: The lower the error probability, the lower the probability that the rule is the result of mere chance, and therefore the higher the probability that the rule exists in the entire population.
- The level of unlikelihood of the rule (in case the rule is unexpected): The higher the level of unlikelihood, the lower the probability that the rule is an accidental phenomenon, and therefore the more the rule can be relied upon in issuing predictions.

Note, however, that above-mentioned problem does not relate to the if-and-only-if rules and to formula rules.

(ii) How can cases of rule inconsistency be solved?

In using if-then rules to issue predictions, one may face cases where some of the rules predict one outcome while others predict another (for example, some of the

rules predict that a certain company tends to go bankrupt, while other rules predict that it does not). WizWhy solves this problem by using the error probability and the level of unlikelihood. Since these two parameters signify the extent of reliability of the rule, WizWhy weighs the rules according to these parameters and calculates the prediction in cases of inconsistency.

Once again, note that the problem does not apply to the case where the predictions are based on the if-and-only-if rules. The if-and-only-if rules are always consistent.

As discussed below, when you apply WizWhy on your data, WizWhy predictions usually suffer from a low rate of over-fitting, even when the number of cases is small, and when there are many missing values.

How does WizWhy explain the predictions?

Contrary to some other data-mining algorithms (such as Artificial Neural Networks), WizWhy issues an easy to understand model that explains the predictions. When the predictions are based on if-then rules, the relevant rules are revealed; and when the predictions are based on an if-and-only-if rule, the rule is usually short and easy to comprehend.

How does WizWhy point out unexpected cases?

In addition to issuing predictions for new cases, the rules can be used to point out unexpected cases in the data. WizWhy issues predictions regarding the records of the data set under analysis and identifies cases where the dependent variable's value deviates from the value anticipated according to the rules. Such a deviation may be the result of noise, but it can also indicate a data entry error, another type of error, or fraud. WizWhy lists those cases, displaying the expected value alongside the relevant rules. The user can review the cases in auditing the data.

Who can benefit from WizWhy?

Data mining in general, and WizWhy in particular, has many practical applications. You can use WizWhy in most cases where simple or complex data analysis and predictions are required. For instance:

- WizWhy can assist professionals in the fields of medicine and social sciences to enhance their diagnostic and research efforts.
- Banks and financial institutions can use WizWhy to identify risky customers.
- Corporations implementing direct marketing will find that WizWhy is ideal tool for increasing the success rate of direct customer appeals.

Scientific research: WizWhy can be applied for inferring rules in a wide range of scientific fields, including medicine, economics, psychology, sociology, and

geology. For example:

- In medical research, WizWhy can reveal laws relating symptoms and diseases.
- In geological research, WizWhy can reveal rules that show the relationships between soil data and mineral location.
- Researchers in social sciences can use WizWhy to discern patterns and characteristics that designate students' aptitude for academic success.

Generally, in research entailing large quantities of data, WizWhy can save significant time and effort by assuming the scientist's burden of revealing rules. Data mining in general, and WizWhy in particular, may revolutionize traditional methods of conducting research in scientific fields.

Banks, credit unions, and insurance companies: These organizations can use WizWhy to discern financially unstable customers and to predict the degree of financial risk of a new customer. To do so, WizWhy reads the customer data file and reveals the patterns of financially risky customers. Rules describing these patterns are saved to the file. Later, company personnel can analyze a new customer by entering the customer's data and having WizWhy calculate the extent of that customer's financial risk.

Market research: Market researchers can use WizWhy to improve the success rate of direct marketing campaigns. Prior to mailing a large quantity of marketing literature, a small batch can be sent to a representative sample of the population, and WizWhy can then discover the patterns of those customers who are and less inclined to purchase, relative to the average purchasing rate. For example, if the purchasing rate is 2%, WizWhy can be instructed to identify those customers whose purchasing rate is 4% and above, or 1% and below. The result can then be applied to the master file, in order to define the optimal direct mailing list.

Explanations about the mathematics behind WizWhy can be found in:
https://www.wizsoft.com/index.php/products/wizwhy/wizwhy-technology/