# Data Mining for Forensic Investigators

Abraham Meidan, Ph.D.

One of the main tasks of forensic investigators, data-quality managers, and auditors is the revealing of errors and potential cases of fraud in data. Effective software tools exist that can help to carry out this task. Investigators can confidently and comfortably conduct thorough analysis of data and reveal inconsistencies and "strange" cases to be investigated on any desired data type and/or size.

The standard method for revealing fraudulent cases and errors uses reports that filter and sort the data (CAAT). For example, one such report may list all the transactions in which the discount percentage is above a certain threshold. There are several tools – mainly ACL and IDEA – that can be used to issue these kinds of reports.

The ideal data mining tool does not compete with these systems, instead it complements them. This is the strength behind some very cost effective data mining software products on the market today (e.g. WizRule or WizWhy).

By definition the previously mentioned reports can only reveal frauds and errors that the report was designed to find. For example, if the investigators suspect that some transactions include fraudulent discounts they may issue the above-mentioned report. If they don't suspect a problem related to the discount, then the reports issued with CAAT software will not discover a fraudulent discount. Since there is an unlimited number of potential frauds, it would be impractical to generate a report for each of them.

To overcome this weakness, data mining software works automatically – the user just selects the data and the software does the analysis, checking all the relationships among the values within the various fields and reporting unexpected and unlikely cases. Fraudulent cases that are missed by standard auditing tools can be revealed by data mining software.

## How do *data mining tools* work?

Data mining programs reveal interesting patterns in data.

Typical data mining tools reveal the patterns in data registered in the past and use these patterns to issue predictions for new cases. For example, a bank may apply a data mining program in order to reveal the patterns of customers that did not pay their loans. Then, when a new customer asks for a loan these patterns are used to calculate the probability of default on this loan. This approach is also used for revealing frauds. For example, credit cards companies use data mining programs in order to discover the patterns of known fraudulent cases and apply these patterns when checking new transactions. However, it is important to note that this approach cannot be used when one searches for fraudulent cases without having previous examples.

Data mining software products based on association rule technology can perform this task. Instead of issuing predictions for new cases, these data mining tools point out cases

deviating from valid patterns. The approach is based on the following assumption: ***In many cases, frauds are exceptions to the rule.*** For example, if in all sales transactions to a certain customer the salesperson is Dan, and there is a single transaction in which the salesperson is someone else who is usually connected with other customers, then this is a suspected case that should be investigated.

In creating a software application that discovers exceptions to the rules, the program first needs to discover all the rules (patterns) in a given data set. In other words, the software should do a "reverse engineering" of the rules that created the data. Driven by a mathematical algorithm that is capable of revealing all the rules governing a data set, data mining analysis generates a list of records that are *unlikely* in reference to the discovered rules. These records are *suspected cases*, or at least *cases to be investigated.*

The process is simple: select the data that you wish to analyze and the software does all the rest. Within a short time the analysis report is displayed on the screen.

During the analysis the following operations are performed:

First, the data is read. Then, you have the opportunity to "fine-tune" the analysis parameters such as "*minimum probability of if-then rules*" and "*minimum number of cases of a rule.*" You can also define exactly which types of rules the product should search for and determine whether or not some fields should be ignored.

All the rules governing the data are revealed and the cases deviating from the discovered rules are listed. For easy reference each deviation is displayed along with the rules from which it deviates.

One such data mining software product is ***WizRule*** which was developed by *WizSoft Inc*. based on the association rule technology.

## What is association rule technology?

Association rule mining finds interesting associations and/or correlation relationships among large set of data items. Association rules show attribute value conditions that occur frequently together in a given dataset.

Association rules provide information in the form of "if-then" statements. These rules are computed from the data and, unlike the if-then rules of logic, association rules determine how the values of one field are affected by the values of other fields.

In addition to the condition (the "if" part) and the result (the "then" part), an association rule has three numbers that express the degree of uncertainty about the rule. An example of an if-then rule is:

If City *is* NYC
*and* Amount Purchased *is* 200 … 300 (average=250)
*and* salesperson *is* Dave
*Then*
Growth Since Last Year is less than 0.1%

*Rule's probability: 0.70*
*The rule exists in 3,700 records*
*Significance Level: Error probability ,0.001*

The first number (probability) is called the **Confidence Level** for the rule. The Probability in if-then rules designates the ratio between the number of records in which the condition(s) and the result hold, and the corresponding number of records in which the condition(s) hold with or without the result.

The second number is called the **Support** of the rule. It is simply the number of records that include all items in the condition and result parts of the rule.

The third number **Significance Level** indicates the degree of the rule's validity. It is equal to 1 minus the **error probability** which qualifies the probability that the rule exists accidentally in the data under analysis.

## What kind of rules does WizRule reveal?

Three types of rules are revealed and used for the analysis:

- Formula rules
- If-then rules
- Spelling rules

As mentioned, the user does *not* enter the rules. Rather all the rules are discovered automatically.

An example of a *formula* rule is:

**A = B * C**

*Where:* **A = Total**
**B = Quantity**
**C = Unit Price**

*Rule's Accuracy Level*: **0.99**
*The rule exists in* **1890** *records*

The "*Accuracy Level*" in formula rules indicates the ratio between the number of cases in which the formula holds and the total number of relevant cases. The cases in which the formula holds are those cases where the formula matches the data exactly except for deviations that may result from a rounding.

Arithmetical formulas with up to 5 variables that hold in the data are revealed. Formulas where **A** is 0 or 1 are ignored.

Obviously if a formula rule holds for all the records in the data except for just a few records, then these deviating records should be investigated.

An example of an *if-then* rule is:

> *If* **Customer** is **Summit**
> *and* **Item** is **Computer type A**
> T*hen*
> **Price** = **765**
>
> *Rule's probability:* **0.998**
> *The rule exists in* **1002** *records*
> *Significance level: error probability < 0.001*

All the if-then rules with any number of conditions are revealed.

Once again, a deviation from a highly valid rule might point to a fraud.

An example of a *spelling* rule is:

> *The value* **Summit** *appears* **2080** *times*
> *in the* **Customer** *field.*
> *There are* **2** *case(s) containing similar value(s)*

These rules are presented mainly in order to reveal cases of misspelled names. A name is suspected as misspelled if (a) it is similar to another name in this field, and (b) the frequency of the first name is very low, while the frequency of the second name is very high. For example, if the name **Zummit** appears only one time (in the Customer field) it will be presented as a deviation to be examined.

## Avoiding False Alarms:

Following the discovery of the rules that govern the data, any deviations from these rules are checked. However, not every deviation from a rule is a case to be examined. Suppose the following *if-then* rule is revealed:

> *If* Customer is Summit
> Then
> Salesperson is Dan
> *Rule's probability:* 0.98
> *The rule exists in* 1003 *records*
> Significance level: error probability < 0.001

Since the rule's probability is 0.98 and the rule exists in 1003 records, then there are about 20 records in which the salesperson deviates from this rule. Reviewing each of these 20 records is quite tedious and many of these deviations might be false alarms.

To avoid such false alarms an operation is performed to check whether the deviation is explainable by another rule that holds in the data. If the answer is positive the case is not

a suspected one. For example, it may be the case that if the **item** is **computer** then the **salesperson** is **John**, and this rule may explain some of the above-mentioned deviations. Since these deviations are explainable they are not considered as cases to be investigated.

Another operation also checks the frequency of the *then* value in the deviated case. Its frequency under the rule conditions should be lower than its overall frequency in the data. If it does not, then once again the case is not considered as a case to be investigated. For example, if the **salesperson** in two of the deviated records is **Frank** and these are *the only two cases* in the entire data where Frank is the salesperson, then these cases are not considered as cases to be investigated. However, if Frank is usually the salesperson of other customers, then the above-mentioned deviations are *indeed suspected cases.*

When the *then* field is numeric, false alarms are further reduced by only displaying deviations where the *then* value deviates from the expected value by at least one standard deviation. Smaller deviations are ignored.

When viewing the suspected cases you can sort the cases by the *then* field, and sort the values within this field. This sorting is mainly relevant when the report lists many deviations. By sorting the deviations you can concentrate on the most interesting cases.

Applying all these methods nearly all the false alarms are avoided so that only the cases that need to be investigated are pursued.

* * * * * *