

Using Domain Knowledge in Data Mining

Zohreh Nazeri

703-983-5841 • nazeri@mitre.org

MITRE Sponsored Research

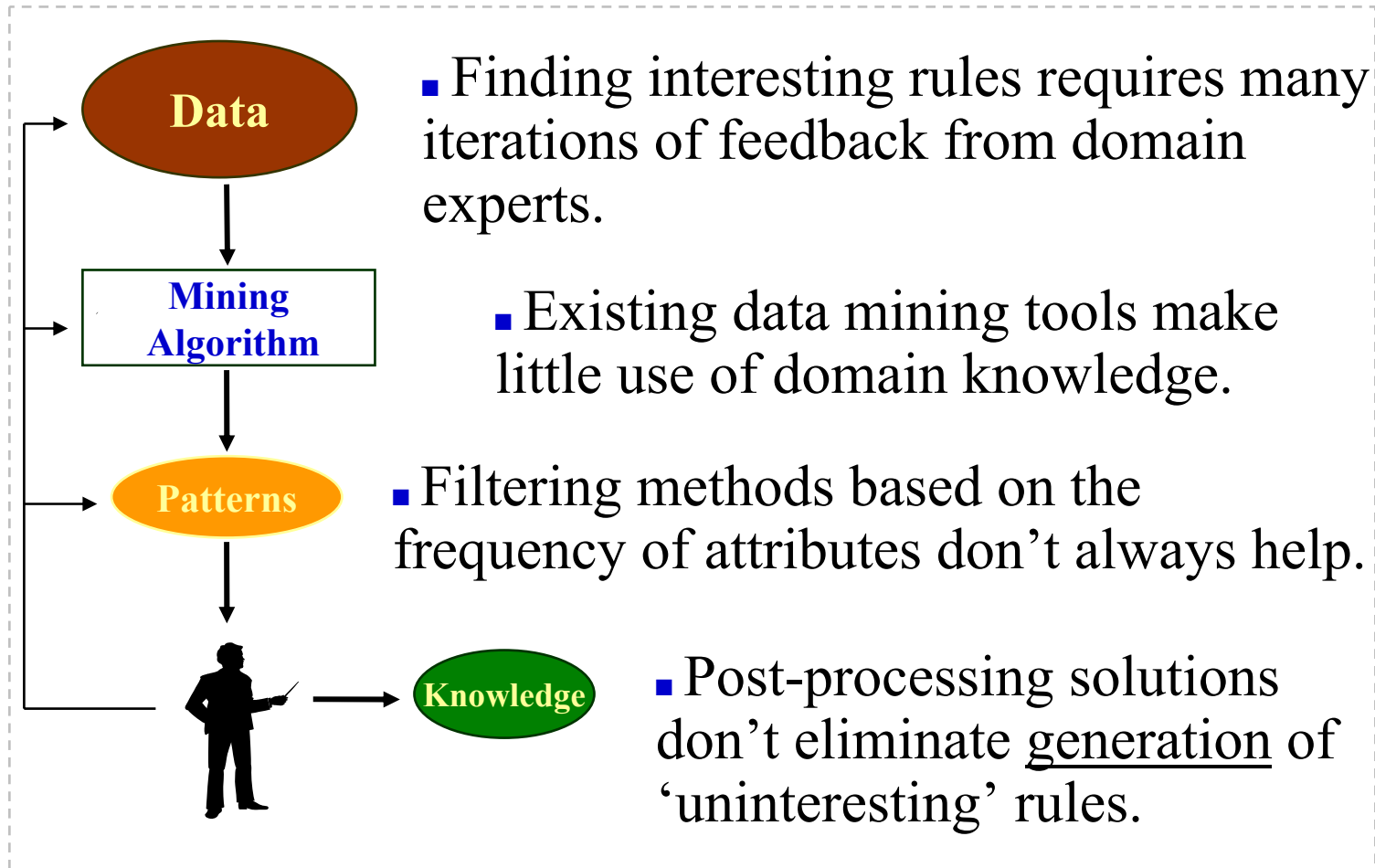
The logo for the MITRE Technology Program, featuring a stylized graphic of stacked blocks in yellow, orange, and blue to the left of the text.

MITRE
Technology
Program

Problem

- Data mining tools generate a large number of patterns, many of which are ‘**uninteresting**’ to domain experts or **irrelevant** to the problem in hand.
- Discovered models are **hard to understand** by domain experts.
- Existing methods to incorporate the experts’ feedback into the mining process are **limited**.

Background



Objective

- Using domain knowledge to improve the **quality** of data mining output and the overall efficiency of the knowledge discovery process
- Reducing the **time** experts have to spend on identifying interesting findings and interpreting them

Activities

- **Obtaining data and acquiring domain knowledge**
 - **Aviation safety domain**
 - **Intrusion detection domain**
- **Modifying data mining algorithms to use domain knowledge internally**
 - **A priori algorithm to find association rules**
 - **C4.5 algorithm to build decision trees**
- **Initial evaluation to determine effectiveness of the proposed approach on the quality of output rules, running time, and accuracy**

Highlight

Association Rules Example

Data: 2K records from NTSB database

Options: Min support = 30%

Min confidence = 50%

Rules

1: {WX_COND = VMC} => {VIS_SM = 10.00}	s= 89%	c= 72%
2: {WX_COND = VMC} => {INJURY = NONE}	s= 89%	c= 56%
3: {WX_COND = VMC} => {TYPE_FLY = PERS}	s= 89%	c= 56%
4: {TYPE_FLY = PERS} => {WX_COND = VMC}	s= 55%	c= 92%
5: {INJURY = NONE} => {WX_COND = VMC}	s= 53%	c= 94%
6: {INJURY = FATL} => {WX_COND = VMC}	s= 30%	c= 74%

The expert is not interested in rules 2 and 5; raising support/confidence thresholds will cause loss of other useful rules. Adding the following domain knowledge, however, eliminates generation of unwanted rules without the loss of useful ones.

DK: “-:INJURY=NONE;” (not interested in no-injury incident rules)

Rules (with the use of DK)

1: {WX_COND = VMC} => {VIS_SM = 10.00}	s= 89%	c= 72%
3: {WX_COND = VMC} => {TYPE_FLY = PERS}	s= 89%	c= 56%
4: {TYPE_FLY = PERS} => {WX_COND = VMC}	s= 55%	c= 92%
6: {INJURY = FATL} => {WX_COND = VMC}	s= 30%	c= 74%

Highlight

Decision Tree Example

Which alerts are false?

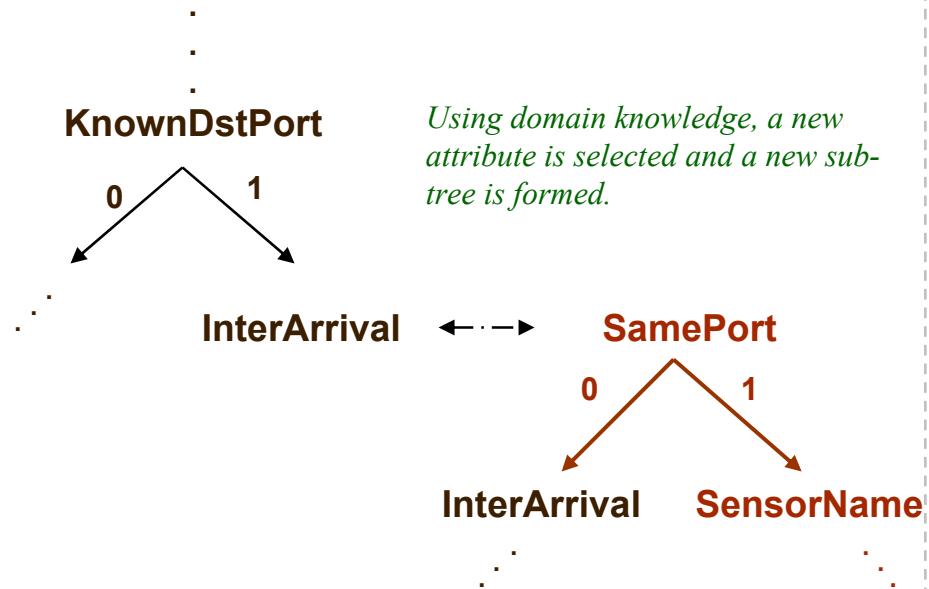
Class Labels: Alert
False Alert

Domain:

...
KnownDstPort 0,1
SamePort 0,1
SensorName a,b,c,d
...

Domain Knowledge:

+ : KnownDstPort =1, SamePort =*;
+ : SamePort=1, KnownDstPort =*;
(denote encouraged paths)



34898 examples in training sample

Using DK	avg tree size	test set errors
no	280	39
yes	296	38

Impacts

- Being used by ongoing data mining projects within **MITRE**
- Transferring the technology to commercial **tool vendors**
- Upgrading Data Mining Workbench (produced in earlier MSR), which is currently being transferred to commercial vendors and air carriers participating in **FAA GAIN Committee**
- Preparing paper submission for **ICTAI 2002**

Future Plans

