Rare Association Rule Mining

Part I. An Introduction

Based on

- PAKDD’04 Tutorial Data Mining for Analysis of Rare Events: A Case Study in Security, Financial and Medical Application, Aleksandar Lazarevic, Jaideep Srivastava, Vipin Kumar
- Rare Association Rule Mining: An Overview, Yun Sing Koh & Nathan Rountree (in the book)

Mining rare events

- outliers
- anomalies
- exceptions
- rare events (in temporal data)

- in other kind of data, in data where the structure is important, e.g. spatial, spatiotemporal, graph

Motivation

- Despite the enormous amount of data, particular events of interest are still quite rare
- Rare events are events that occur very infrequently, i.e. their frequency ranges from 0.1% to less than 10%
- However, when they do occur, their consequences can be quite dramatic and quite often in a negative sense

(PAKDD’04 Tutoria)

“Mining needle in a haystack.
So much hay and so little time”

Examples

- Network intrusion detection number of intrusions on the network is typically a very small fraction of the total network traffic
- Credit card fraud detection millions of regular transactions are stored, while only a very small percentage corresponds to fraud
- Medical diagnostics when classifying the pixels in mammogram images, cancerous pixels represent only a very small fraction of the entire image

Examples

- Insurance Risk Modeling Claims are rare but very costly
- Web mining Less than 3% of all people visiting Amazon.com make a purchase
- Targeted Marketing Response is typically rare but can be profitable
- Churn Analysis Churn is typically rare but quite costly
- Hardware Fault Detection Faults are rare but very costly
- Airline No-show Prediction Disease is typically rare but can be deadly

Limitations of standard techniques

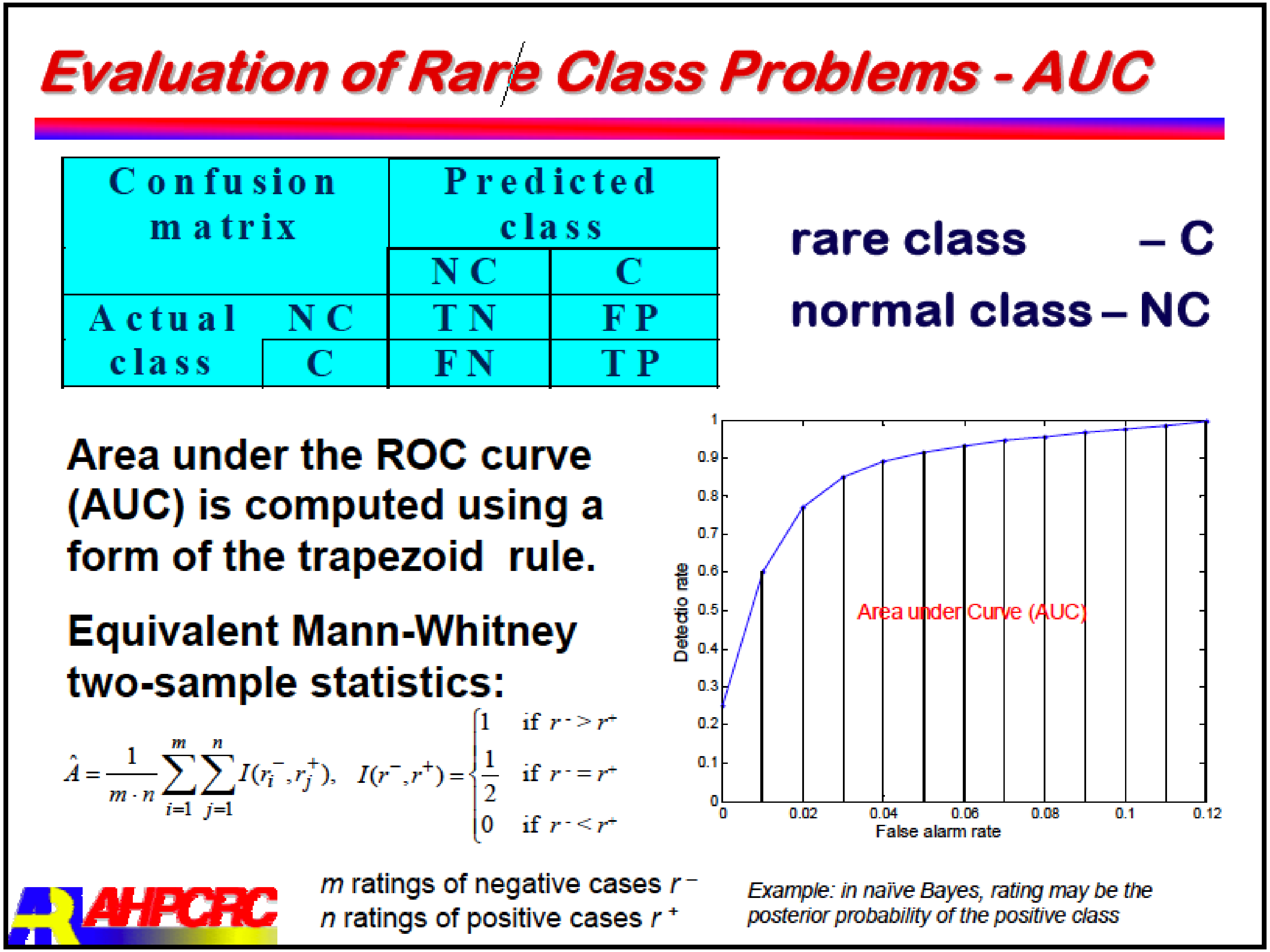
- While most normal events are similar to each other, rare events may be quite different from one another regular credit transaction are fairly standard, while fraudulent ones vary from the standard ones in many different ways
- Metrics: Accuracy is not appropriate for evaluating methods for rare event detection precision, recall,
- On many applications data keeps arriving in an ongoing stream

Confusion matrix

		Predicted class		
		NC	C	
Actual	NC	TN	FP	normal class – NC
class	C	FN	TP	rare class – C

- Recall (R) = TP/(TP + FN)
- Precision (P) = TP/(TP + FP)
- F – measure = 2*R*P/(R+P)

AUC



Evaluation of Rare Class Problems - AUC

Confusion matrix		Predicted class	
		NC	C
Actual class	NC	TN	FP
	C	FN	TP

rare class – C

normal class – NC

Area under the ROC curve (AUC) is computed using a form of the trapezoid rule.

Equivalent Mann-Whitney two-sample statistics:

$$\hat{A} = \frac{1}{m \cdot n}\sum_{i=1}^{m}\sum_{j=1}^{n} I(r_i^-, r_j^+), \quad I(r^-, r^+) = \begin{cases} 1 & \text{if } r^- > r^+ \\ \frac{1}{2} & \text{if } r^- = r^+ \\ 0 & \text{if } r^- < r^+ \end{cases}$$

m ratings of negative cases r^-

n ratings of positive cases r^+

Example: in naïve Bayes, rating may be the posterior probability of the positive class

Association rule mining

Apriori algorithm

Rare association rule mining

- why *Apriori* etc. are not appropriate
- approaches
  - a variable support thershold
  - mining without support threshold
  - constraint-based mining
  - structure-based mining

Emerging patterns*

- For 2 classes C1 and C2 and support $s_i(X)$ of itemset X in class C_i, growth rate is defined as $s_2(X)/s_1(X)$. Given a threshold p >1, itemset X is p-emerging pattern (EP) if growth rate ≥ p
- Find EPs in rare class
- Find values that have the highest growth rate from the major class to the rare class
- Replace different attribute values in the original rare class EPs with the highest growth rate values
- New generated EPs have a strong power to discriminate rare class from the major class

*** H. Alhammady, K. Rao, The Application of Emerging Patterns for Improving the Quality of Rare-class Classification, PAKDD 2004, Springer LNAI 3056**.