The Potential of Associative Classifiers

Osmar R. Zaïane

Data Mining Research Group

Database Laboratory

Department of Computing Science
University of Alberta, Canada

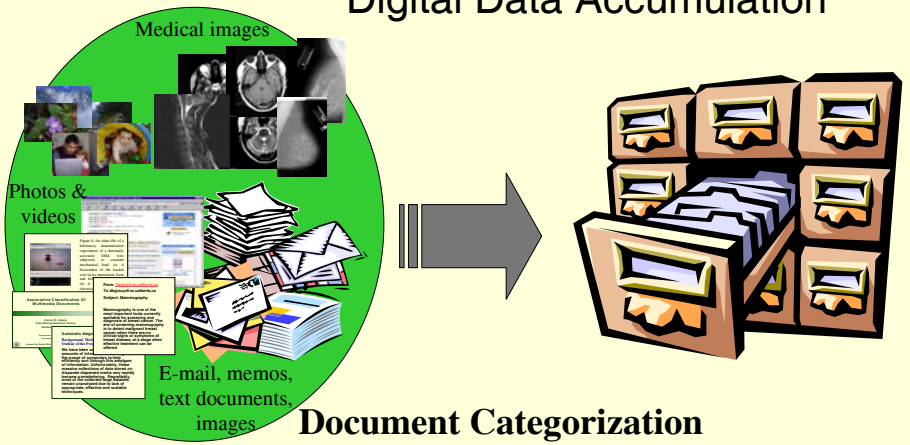
December 1, 2005

School of Information Technology & Electrical Engineering,

The University of Queensland, Brisbane, Australia

Text & Multimedia Mining

Digital Data Accumulation



Document Categorization

Single-label classification and Multi-label classification

Presentation Outline

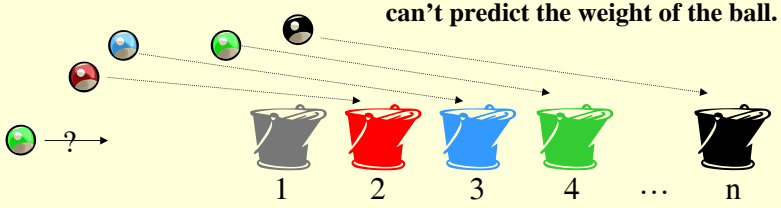
- Typical Machine Learning
- What is Classification?
- What are the Challenges?
- The Associative Classifier
- Breast Cancer Detection
- Other Examples
- Dealing with efficiency & effectiveness

What is Classification?

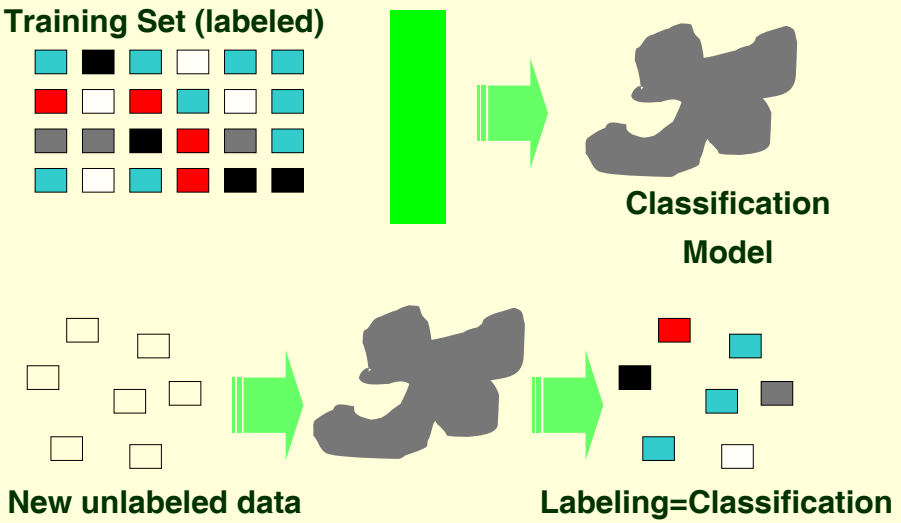
The goal of data classification is to organize and categorize data in distinct classes.

- A model is first created based on the data distribution.
- The model is then used to classify new data.
- Given the model, a class can be predicted for new data.

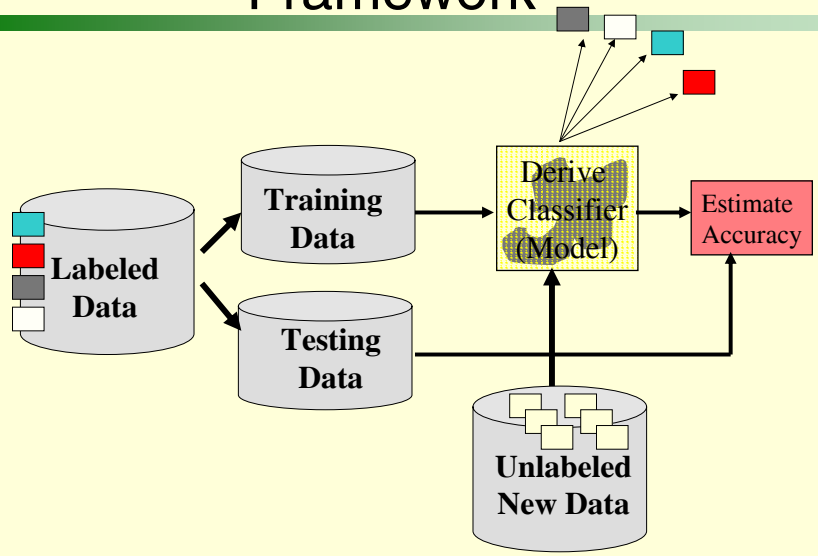
With classification, I can predict in which bucket to put the ball, but I can't predict the weight of the ball.



Classification = Learning a Model

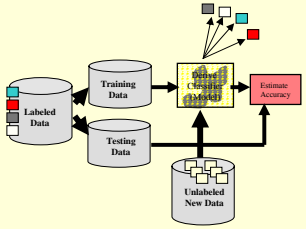


Framework



Classification Methods

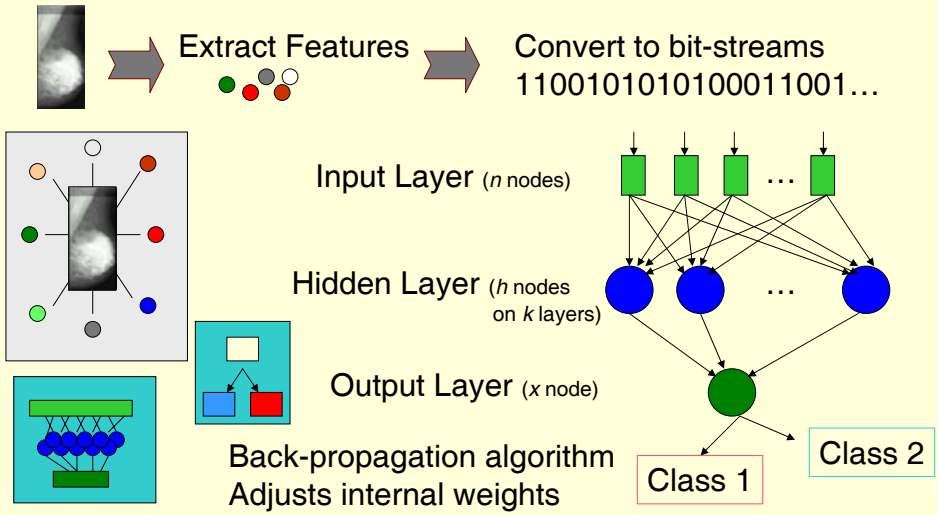
- Decision Tree Induction
- Neural Networks
- Bayesian Classification
- K-Nearest Neighbour
- Support Vector Machines
- Case-Based Reasoning
- Genetic Algorithms
- Rough Set Theory
- Fuzzy Sets
- Etc.



Associative classifiers

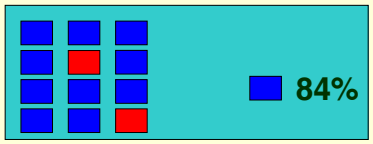
A new Model

Neural Networks



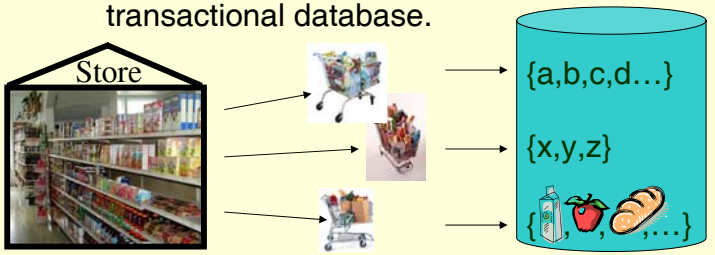
Challenges

- Dealing with high dimensional spaces
- Handling missing data
- Deriving a model that can be interpreted (Transparency leads to trust for some applications)
- Deriving a model that can be edited by human experts (to inject domain knowledge)
- Dealing with very large or evolving training sets
- Allowing multi-label classification
- Dealing with uneven representation of classes in training sets



What Is Association Rule Mining?

- **Association rule mining searches for relationships between items in a dataset:**
  - aims at discovering associations between items in a transactional database.



- Rule form: "**Body → Head [support, confidence]**"

buys(x, "bread") → buys(x, "milk") [0.6%, 65%]
major(x, "CS") ^ takes(x, "DB") → grade(x, "A") [1%, 75%]

Basic Concepts

A transaction is a set of items: $T=\{i_a, i_b,\ldots i_t\}$

$T \subset I$, where I is the set of all possible items $\{i_1, i_2,\ldots i_n\}$

D, the task relevant data, is a set of transactions.

An association rule is of the form:
$P \rightarrow Q$, where $P \subset I$, $Q \subset I$, and $P \cap Q = \varnothing$

$P \rightarrow Q$ holds in D with <u>support</u> s
and
$P \rightarrow Q$ has a <u>confidence</u> c in the transaction set D.

$Support(P \rightarrow Q) = Probability(P \cup Q)$
$Confidence(P \rightarrow Q) = Probability(Q/P)$

Association Rule Mining

Frequent Itemset Mining

FIM (1) abc

Bound by a **support** threshold

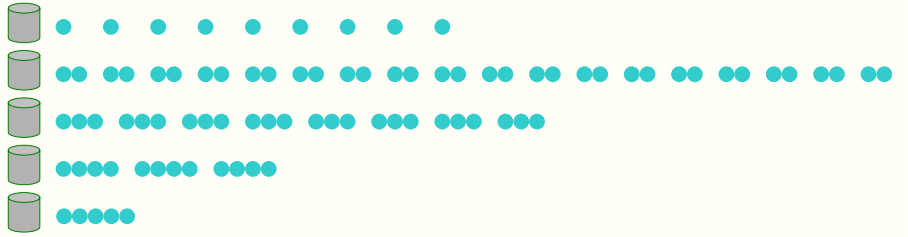
Association Rules Generation

(2) ab→c
b→ac

Bound by a **confidence** threshold

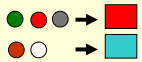
Apriori (Agrawal et al. 1994)

Repetitive I/O scans
Huge Computation to generate candidate items



Associative Classifier

- We want to find associations between extracted features and class labels



- Constrain the association rule mining such that the rules found are of the following form:

$$F_\alpha \wedge F_\beta \wedge F_\gamma \wedge \ldots \wedge F_\delta \rightarrow \textbf{class}$$

- We used a constrained version of frequent itemset mining.
  - The class label has to be part of any frequent itemset — {A, B, C, Class}
  - The class label is a consequent, and all other items are the antecedent of a rule — A, B, C → Class

How do Associative Classifiers Work?

{Tid, $Item_1$, $Item_2$, $Item_3$,…$Item_{t1}$}
…
{Tid, $Item_1$,…$Item_{tn}$}

→ Frequent k-itemsets
{$Item_a$, $Item_b$,…$Item_k$}

Rules

{Itemset → Itemset}

Atr1	Atr2	Atr3	…	AtrN	Class Label

Constrained Itemsets

{$Item_1$, $Item_2$, $Item_3$,…$Item_n$, $Class_1$}
…
{$Item_1$, $Item_2$, $Item_3$,…$Item_n$, $Class_n$}

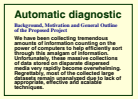
Constrained Association Rules

→ Frequent k-itemsets
{$Item_a$, …$Item_k$, $Class_x$}

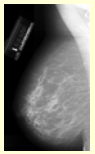
{Itemset → Class}

Modeling documents

{bread, milk, beer,…} → (Bread, milk) → Bread → milk



{term1, term2,…,Ca} → (term2, Ca) → term2 → Ca



{f1, f2,…,Ca} → (f3, f5, Ca) → f3^f5 → Ca

General Approach

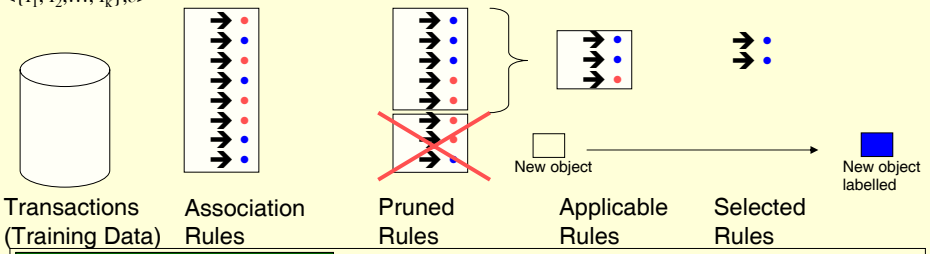
Model input data into transactions

Unlabeled new objects

Also modeled into transactions

Set of transactions → Rule Generation → Set of rules → Rule Pruning → Set of rules → Rule Selection → Labeled objects

<{i_1, i_2,…, i_k},c>



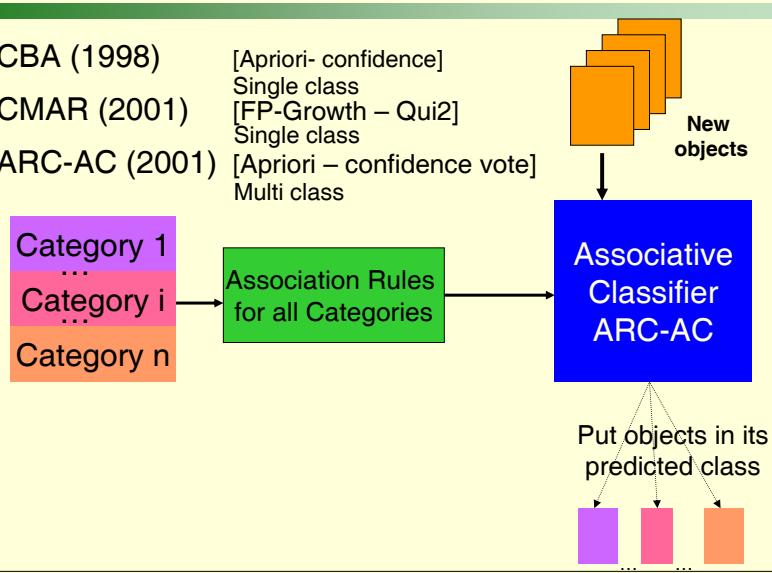
Transactions (Training Data) | Association Rules | Pruned Rules | Applicable Rules | Selected Rules

Association Rules - Classification for all Categories

CBA (1998) [Apriori- confidence] Single class

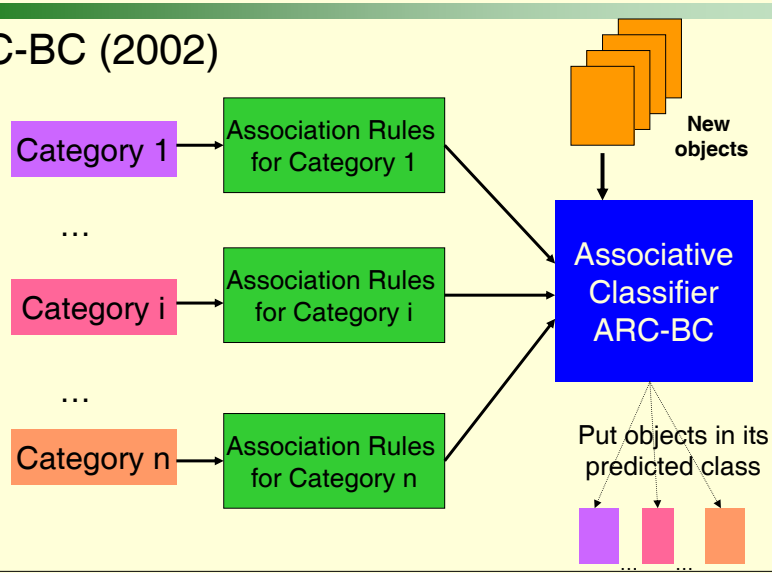
CMAR (2001) [FP-Growth – Qui2] Single class

ARC-AC (2001) [Apriori – confidence vote] Multi class



Association Rules - Classification by Category

ARC-BC (2002)



Association Rules: Advantages & Issues

- AR are well studied – fast – scalable
- No independence assumption btw. attributes
- Attributes: – large number – variable number, can handle missing values
- Transparency

- AC are in an early stage of development
  - use simple rules
  - naïve selection function
- AC models consist of a large number of rules
  - harder selection
  - redundant, uninteresting rules
  - longer classification time
  - difficult to manually revisit rules

Solution: Pruning Techniques

Pruning Rules

Large number of rules ⇨ Noisy information

Large number of rules ⇨ Long classification time

Solution: Pruning Techniques

- Removing low ranked specialized rules;

$R_1 : F_1 \Rightarrow C$ Confidence 90%

$R_2 : F_1 \wedge F_2 \Rightarrow C$ Confidence 80%

$\Rightarrow R_1$

- Eliminate conflicting rules (for single-class classification);

$F_1 \Rightarrow C_1 \wedge F_1 \Rightarrow C_2$

- Database coverage;

Classification Stage

Let S be the classification system

A new object O <f1; f3; f4; f7; f9 >

f1 ➔ C1 confidence 0.9
f3 & f4 ➔ C2 confidence 0.85
f4 ➔ C2 confidence 0.8
f7 ➔ C1 confidence 0.6
f9 ➔ C3 confidence 0.5

C2 0.825
C1 0.75
C3 0.5

Using the *dominance factor* we chose the winning categories. If δ=100% C2 is winning. If δ=80% O is predicted to fall in C2 and C1.

Presentation Outline

- **Motivation**
- **What is Classification?**
- **What are the Challenges?**
- **The Associative Classifier**
- **Breast Cancer Detection**
- **Other Examples**
- **Dealing with efficiency & effectiveness**

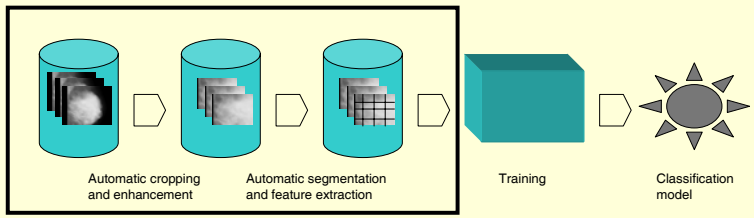
Risks of Errors

- Mammograms can be categorized as:
  - Normal
  - Benign
  - Malignant

  } Abnormal
- False positive ➔ unnecessary biopsy
- False negative ➔ risk of death
- Need to identify preliminary signs of masses and calcification clusters, but in early stages of cancer these are very subtle.
- Some work was done for automatic classification of medical images to medical specialists in detection but low accuracy

Digital Mammograms

- Mammograms are difficult to read even by specialists due to low contrast and different types of tissue.
- In order to extract visual features Image enhancement is necessary

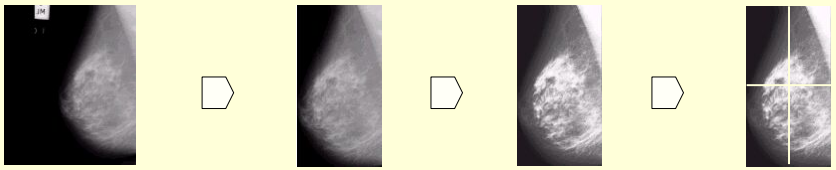


Improving the Quality of Images

- Digitization introduces noise
- Inconsistent illumination conditions
- Inconsistent sizes and distributions

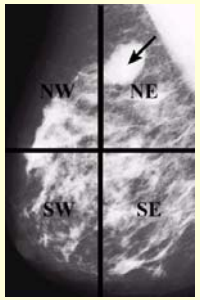
Automatic Cropping: Removes unwanted parts and artifacts.

Enhancement: Diminishes the effect of over brightness and over darkness. Histogram equalization to increase contrast range.



Original mammogram | Cropping | Histogram Equalization | Partitioning & orientation

Feature Extraction



- Mean
- Variance
- Skewness
- Kurtosis

$$\bar{x} = \frac{\sum x}{N}$$

$$\sigma^2 = \frac{\sum (x-\bar{x})^2}{N}$$

$$Sk = \frac{1}{N}\left(\frac{(x-\bar{x})}{\sigma}\right)^3$$

$$Kurt = \frac{1}{N}\left(\frac{(x-\bar{x})}{\sigma}\right)^4 - 3$$

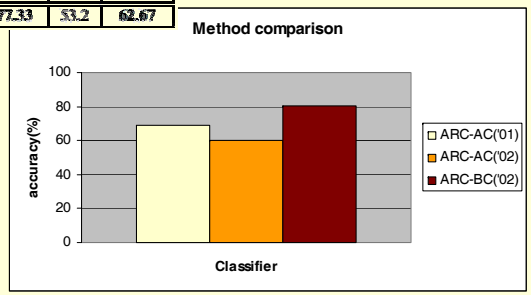
Image Transaction (ImageID, F_{1NE},F_{2NE},F_{3NE},F_{4NE})

Image Transaction (ImageID, F_{1NW},F_{2NW},F_{3NW},…,F_{4NE})

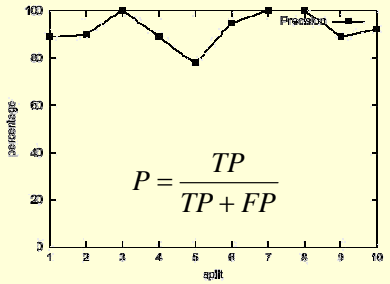
Experimental Results

Split	1st rule		ordered		cut rules		remove specific	
	#rules	accuracy	#rules	accuracy	#rules	accuracy	#rules	accuracy
1	22	76.67	1121	80.00	856	76.67	51	60.00
2	18	86.67	974	93.33	755	90.00	48	86.67
3	22	83.33	823	86.67	656	86.67	50	76.67
4	22	63.33	1101	76.67	842	66.67	51	53.33
5	33	56.67	1893	70.00	1235	70.00	63	50.00
6	16	66.67	1180	76.67	958	73.33	51	63.33
7	30	66.67	1372	83.33	1055	73.33	58	53.33
8	26	66.67	1386	76.67	1089	80.00	57	46.67
9	20	66.67	1353	76.67	1130	76.67	52	60.00
10	18	76.67	895	83.33	702	80.00	51	76.67
avg(%)	22.7	71.02	1209.8	80.33	927.8	77.33	53.2	62.67

Minimum support: 25%;
Minimum confidence: 50%;

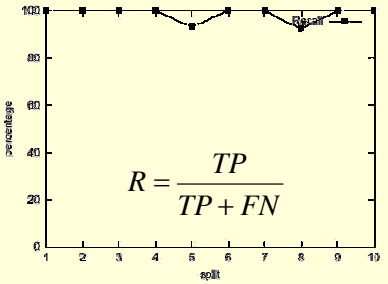


Experimental Results



$$P = \frac{TP}{TP + FP}$$

Precision over 10 splits



$$R = \frac{TP}{TP + FN}$$

Recall over 10 splits

False positives and false negatives tend to zero

Experimental Results with Reuters

Reuters collection: ModApte version: 12,202 documents consisting of 9,603 training documents and 3,299 testing documents.

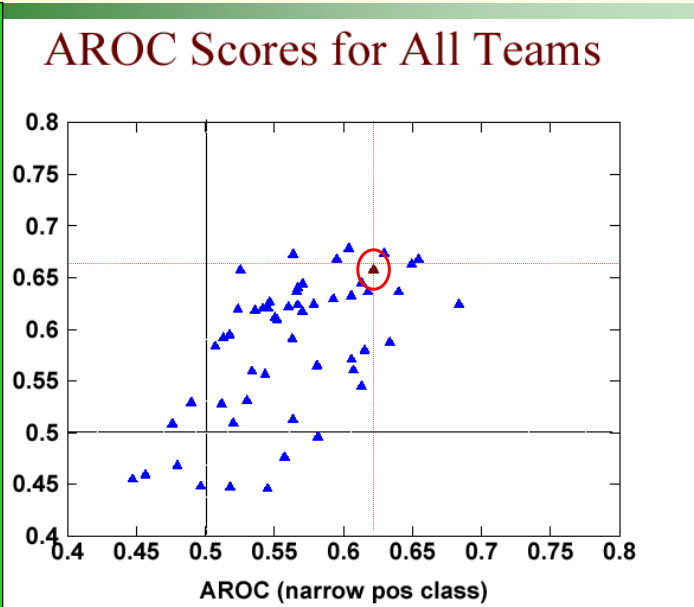
BEP	ARC-BC with δ=50 10%	15%	20%	Bayes	Rocchio	C4.5	k-NN	bigrams	SVM (poly)	SVM (rbf)
acq	**90.9**	89.9	87.8	91.5	92.1	85.3	92.0	73.2	94.5	95.2
corn	69.6	**82.3**	70.9	47.3	62.2	87.7	77.9	60.1	85.4	85.2
crude	77.9	77.0	**80.7**	81.0	81.5	75.5	85.7	79.6	87.7	88.7
earn	**92.8**	89.2	86.6	95.9	96.1	96.1	97.3	83.7	98.3	98.4
grain	68.8	72.1	**73.1**	72.5	79.5	89.1	82.2	78.2	91.6	91.8
interest	70.5	70.1	**75.3**	58.0	72.5	49.1	74.0	69.6	70.0	75.4
money-fx	70.5	**72.4**	70.5	62.9	67.6	69.4	78.2	64.2	73.1	75.4
ship	**73.6**	73.2	63.0	78.7	83.1	80.9	79.2	69.2	85.1	86.6
trade	68.0	69.7	**69.8**	50.0	77.4	59.2	77.4	51.9	75.1	77.3
wheat	84.8	**86.5**	85.3	60.6	79.4	85.5	76.6	69.9	84.5	85.7
micro-avg	**82.1**	81.8	81.1	72.0	79.9	79.4	82.3	73.3	85.4	86.3
macro-avg	76.74	**78.24**	76.32	65.21	79.14	77.78	82.05	67.07	84.58	86.01

Precision/Recall-breakeven point on ten most populated Reuters categories for ARC-BC and most known classifiers

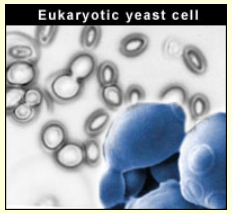
KDD Cup 2002

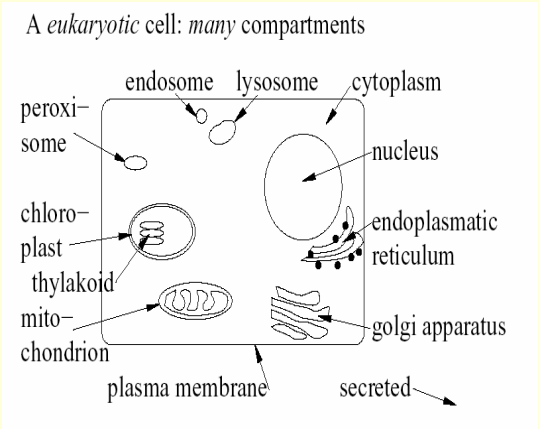
Yeast Gene Regulation Prediction

- A list of test set genes (many missing values) (Instances: 127 positive; 4380 negative)
- Protein / protein interaction
- MEDLIN text abstracts describing genes of interest.



Localization of Proteins





Presentation Outline

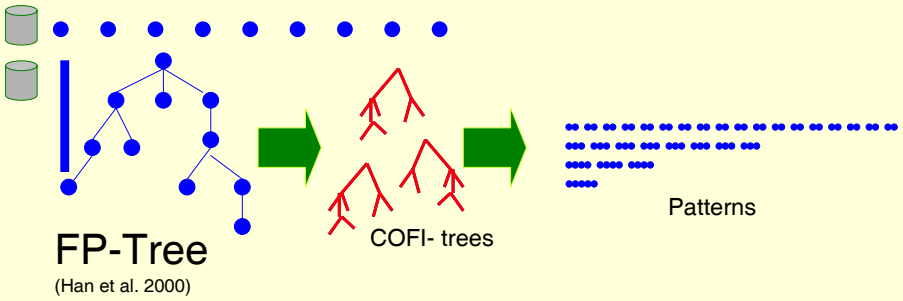
- Motivation
- What is Classification?
- What are the Challenges?
- The Associative Classifier
- Breast Cancer Detection
- Other Examples
- Dealing with efficiency & effectiveness

Improving Efficiency (Rule Generation)

COFI (El-Hajj & Zaïane 2003)

Leap (Zaïane & El-Hajj 2005)

- 2 I/O scans
- Reduced candidacy generation
- Low memory requirements (small footprint)
- One order of magnitude faster than Apriori and FP-Growth



But Efficiency is not a major issue since Learning is off-line.

Improving Effectiveness

- What about the strength of rules?
- What about absence of features?
- What about repetition of features?

Can we Exploit this to improve the classifier?

Association Rules with Recurrent Items

Negative Association Rules

More Pruning and selection Strategies

Negative Association Rules

PKDD'04 and DMKD at SIGMOD'04

- Generalized negative association rule is a rule containing a negation of an item

e.g.: $A \wedge \neg B \wedge \neg C \wedge D \rightarrow E \wedge \neg F$

- Confined negative association rules

$\neg X \rightarrow Y$ $X \rightarrow \neg Y$ $\neg X \rightarrow \neg Y$

Correlation Coefficient

Correlation coefficient:

$$\rho = \frac{Cov(X, Y)}{\sigma_X \times \sigma_Y}$$

Contingency table for binary variables:

	Y	¬Y	Σ_{row}
X	f_{11}	f_{10}	f_{1+}
¬X	f_{01}	f_{00}	f_{0+}
Σ_{col}	f_{+1}	f_{+0}	N

ϕ correlation coefficient: (for binary variables)

$$\phi = \frac{f_{11} f_{00} - f_{10} f_{01}}{\sqrt{f_{+0} f_{+1} f_{1+} f_{0+}}}$$

Positive & Negative Rule Generation

- It generates all positive and negative association rules with strong correlation
- minsupp, minconf – user-defined
- correlation starts at ρ_{min}= 0.5
- The process of rules generation is apriori-like
  - $C_k = F_{k-1} \times F_1$
- For each pair X,Y, where X∪Y is itemset in C_k
  - correlation(X,Y) is computed

Positive & Negative Rule Generation

- if the correlation is **positive**:
  - $A \rightarrow B$ $\quad \neg A \rightarrow \neg B$
- if the correlation is **negative**:
  - $\neg A \rightarrow B$ $\quad A \rightarrow \neg B$
- if the rules have high confidence they are added to the discovered set of rules

Some ARC-PAN Results

Error Rate Comparison

Datasets	C4.5	CBA	ARC-PAN +R	ARC-PAN +R&-AR	ARC-PAN +R&-R
Breast	3.9	4.2	5.5	4.8	3.8
Diabetes	27.6	25.3	23.3	25.4	25.1
Heart	18.9	18.5	16.3	17.0	16.2
Iris	5.5	7.1	6.6	6.6	6.0
Led7	26.5	27.8	28.7	28.7	28.9
Pima	27.5	27.6	27.4	27.1	26.9

Support vs. Correlation

Strong Rules # rules	Strong Rules error	Correlated Rules # rules	Correlated Rules error
17,000	5.0	1,000	5.5
4,000	21.8	40	23.3
200,000	24.7	80	16.3
140	7.3	60	6.6
4,000	34.3	500	28.7
4,000	22.0	50	27.4

All positive and negative rules

Positive rules and rules of the form:¬F→C

Only positive rules

Repetition of Features

ICDE'00 and PAKDD'05

Model transactions of features not as binary

$\{Item_a, \ldots Item_k\}$

But as enumerations of repeated features

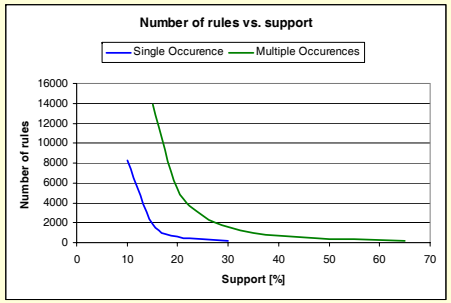
$\{\boldsymbol{\alpha} Item_a, \ldots \boldsymbol{\beta} Item_k\}$

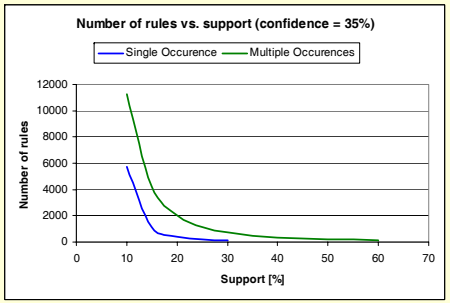
– Use reoccurring frequent itemset mining (ICDE 2000) to generate rules such as:

$\{\alpha\ Item_a, \ldots \beta\ Item_k\} \rightarrow \mathbf{Class_x}$

Nothing to do with quantitative association rules

Number of Rules



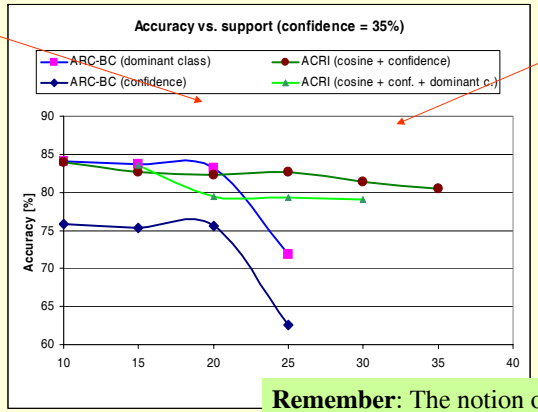


ACRI generates more rules

Accuracy

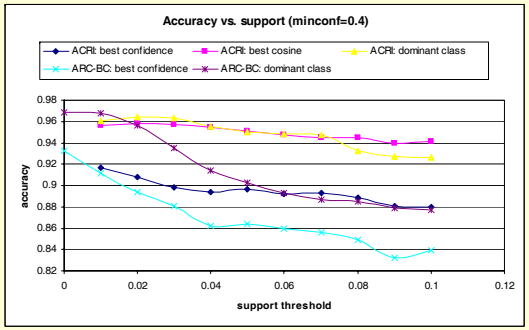
ARC-BC slightly better at 20% but dips later

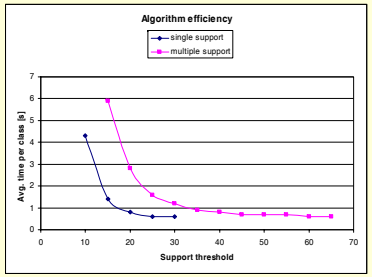
ACRI is stable



Remember: The notion of *support* for ACRI and ARC-BC is different.

Algorithm Efficiency



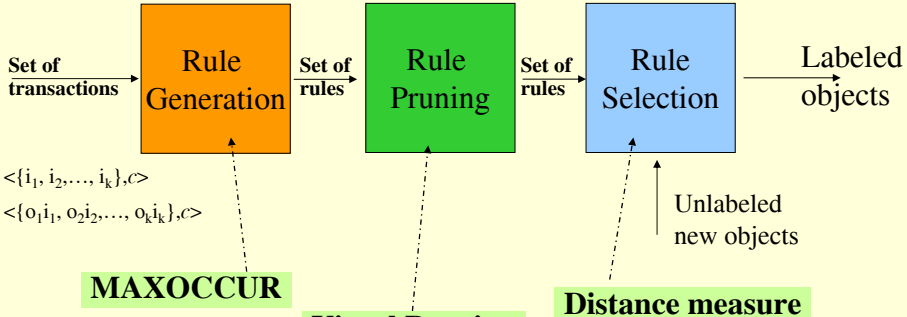


ACRI is slower because it deals with more rules.

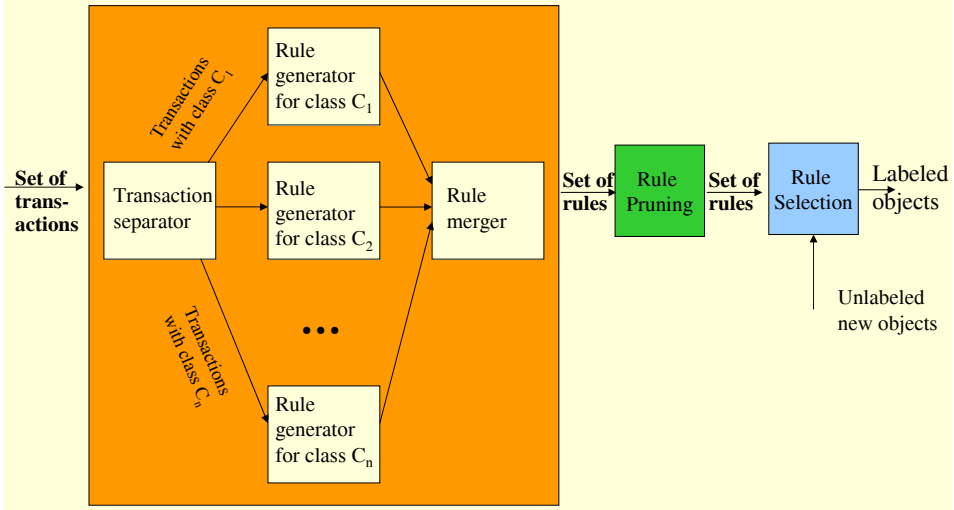
ACRI is consistent at low supports and outperforms ARC-BC.

Summary

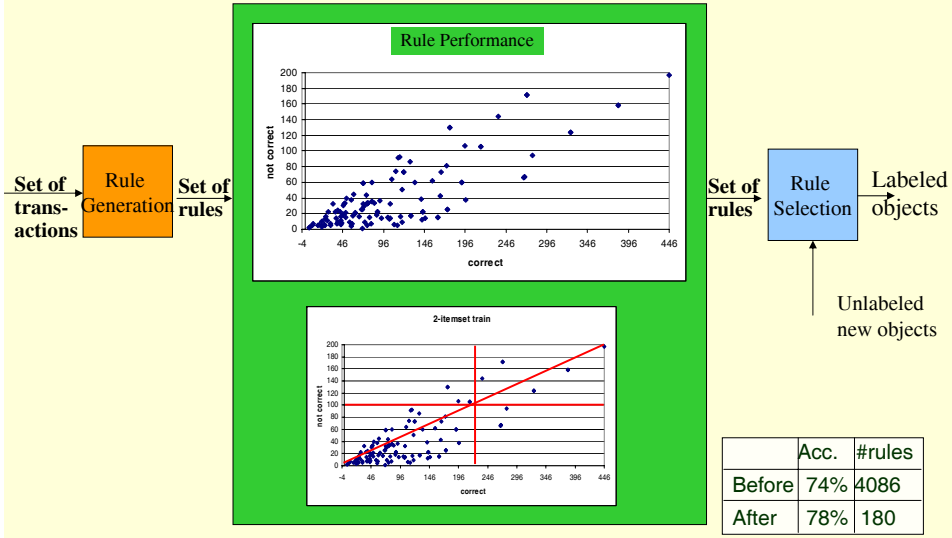
Model input data into transactions



Summary (2)

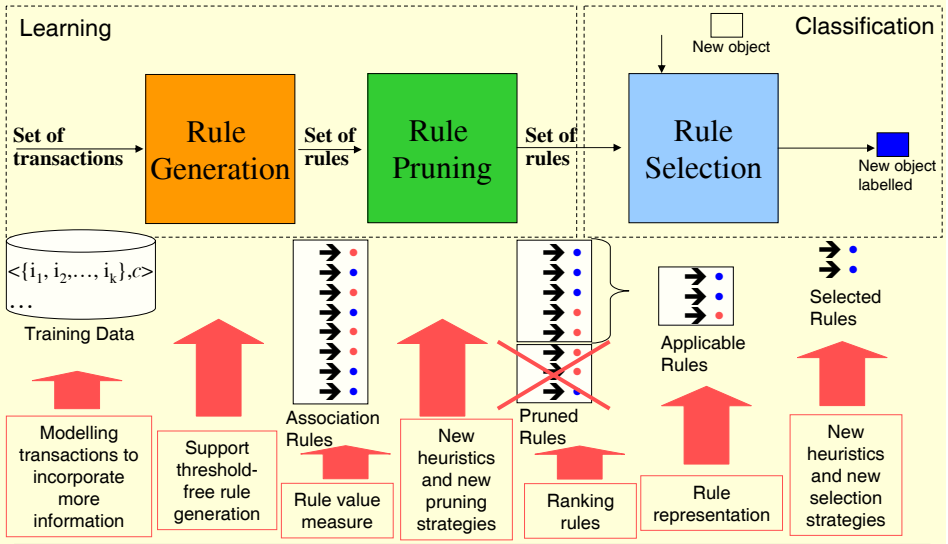


Summary (3)



	Acc.	#rules
Before	74%	4086
After	78%	180

Open Problems?



Thanks

This work was in collaboration with many graduate students (who did most of the work). Many thanks in particular to: Maria-Luiza Antonie, Alexandru Coman, Mohammad El-Hajj, Rafal Rak, and Wojciech Stach

But also to:

Andrew Foss, Chi-Hoon Lee, Stanley Oliveira, Jia Li, Yi Li, Yaling Pei, William Cheung, Weinan Wang, and many others.

Questions ?