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

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.

Text & Multimedia Mining

Digital Data Accumulation

Medical images

Photos & videos

E-mail, memos, text documents, images

Document Categorization

Single-label classification and Multi-label classification
Classification = Learning a Model

Training Set (labeled) → Classification Model

New unlabeled data → Labeling = Classification

Framework

Training Data → Derive Classifier (Model) → Estimate Accuracy

Labeled Data → Testing Data → Unlabeled New Data

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

Derive Classifier (Model)

Class 1

Neural Networks

Extract Features → Convert to bit-streams 1100101010100011001...

Input Layer (n nodes) → Hidden Layer (h nodes on k layers) → Output Layer (x node)

Back-propagation algorithm Adjusts internal weights
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

Basic Concepts

A transaction is a set of items: \( T=\{i_1, i_2, \ldots, i_n\} \)

\( T \subseteq I \), 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 \( P \subseteq I \), \( Q \subseteq I \), and \( P \cap Q = \emptyset \)

\( P \Rightarrow Q \) holds in \( D \) with support \( s \) and
\( P \Rightarrow Q \) has a confidence \( c \) in the transaction set \( D \).

\[
\text{Support}(P \Rightarrow Q) = \text{Probability}(P \cup Q) \\
\text{Confidence}(P \Rightarrow Q) = \text{Probability}(Q/P) 
\]
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:
  \[ F_\alpha \land F_\beta \land F_\gamma \land \ldots \land F_\delta \Rightarrow \text{class} \]
- We used a constrained version of frequent itemset mining:
  - The class label has to be part of any frequent itemset
  - The class label is a consequent, and all other items are the antecedent of a rule

\[ \{A, B, C, \text{Class}\} \]

How do Associative Classifiers Work?

\[ \{\text{Tid}, \text{Item}_1, \text{Item}_2, \ldots, \text{Item}_n\} \Rightarrow \text{Frequent k-itemsets} \]

\[ \{\text{Tid}, \text{Item}_m, \ldots, \text{Item}_n\} \]

\[ \Rightarrow \text{Rules} \]

\[ \{\text{Itemset} \Rightarrow \text{Itemset}\} \]

Constrained Association Rules

Modeling documents

\( \{\text{bread, milk, beer,}\ldots\} \rightarrow (\text{Bread, milk}) \rightarrow \text{Bread} \Rightarrow \text{milk} \)

\( \{\text{term}1, \text{term}2,\ldots,\text{Ca}\} \rightarrow (\text{term}2, \text{Ca}) \rightarrow \text{term}2 \Rightarrow \text{Ca} \)

\( \{f1, f2,\ldots,\text{Ca}\} \rightarrow (f3, f5, \text{Ca}) \rightarrow f3^f5 \Rightarrow \text{Ca} \)

General Approach

Model input data into transactions

\[ \text{Rule Generation} \]
\[ \text{Rule Pruning} \]
\[ \text{Rule Selection} \]

Unlabeled new objects also modeled into transactions

Set of transactions \( \rightarrow \) Set of rules \( \rightarrow \) Set of rules

Transactions (Training Data) \( \rightarrow \) Association Rules \( \rightarrow \) Pruned Rules \( \rightarrow \) Applicable Rules \( \rightarrow \) Selected Rules

New object \( \rightarrow \) New object labeled
Association Rules - Classification for all Categories

- CBA (1998) [Apriori - confidence]
- CMAR (2001) [FP-Growth – Qui2]
- ARC-AC (2001) [Apriori – confidence vote]

New objects

Put objects in its predicted class

Association Rules for all Categories

Category 1

Category i

Category n

Associative Classifier

ARC-AC

Association Rules - Classification by Category

ARC-BC (2002)

Category 1

Category i

Category n

New objects

Put objects in its predicted class

Association Rules for Category 1

Association Rules for Category i

Association Rules for Category n

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
  - naive selection function
- AC models consist of a large number of rules
  - harder selection
  - redundant, uninteresting rules
  - longer classification time
  - difficult to manually revisit rules

Pruning Rules

Solution: Pruning Techniques

- Removing low ranked specialized rules:
  \[ R_1 : F_1 \Rightarrow C \text{  Confidence } 90\% \]
  \[ R_2 : F_1 \land F_2 \Rightarrow C \text{  Confidence } 80\% \]

- Eliminate conflicting rules (for single-class classification):
  \[ F_1 \Rightarrow C_1 \land F_1 \Rightarrow C_2 \]

- Database coverage;
Classification Stage

Let S be the classification system

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

<table>
<thead>
<tr>
<th>f1</th>
<th>C1 confidence 0.9</th>
</tr>
</thead>
<tbody>
<tr>
<td>f3 &amp; f4</td>
<td>C2 confidence 0.85</td>
</tr>
<tr>
<td>f4</td>
<td>C1 confidence 0.75</td>
</tr>
<tr>
<td>f7</td>
<td>C3 confidence 0.5</td>
</tr>
<tr>
<td>f9</td>
<td>C3 confidence 0.5</td>
</tr>
</tbody>
</table>

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.

**Experimental Results**

<table>
<thead>
<tr>
<th>Method comparison</th>
<th>Minimum support: 25%; Minimum confidence: 50%;</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>VNC-AC('01)</td>
</tr>
<tr>
<td>Precision over 10 splits</td>
<td>0.80 ± 0.02</td>
</tr>
<tr>
<td>Recall over 10 splits</td>
<td>0.90 ± 0.03</td>
</tr>
</tbody>
</table>

**Feature Extraction**

- Mean
- Variance
- Skewness
- Kurtosis

Image Transaction (ImageID, F1NE,F2NE,F3NE,F4NE)

Image Transaction (ImageID, F1NW,F2NW,F3NW,...,F4NE)

**Experimental Results**

Precision over 10 splits

Recall over 10 splits

False positives and false negatives tend to zero
Experimental Results with Reuters

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

<table>
<thead>
<tr>
<th>Classifier</th>
<th>ARC-BC with δ=0.5</th>
<th>Bayes</th>
<th>Rocchio CL5</th>
<th>k-NN</th>
<th>Bigrams</th>
<th>SVM (poly)</th>
<th>SVM (rbf)</th>
</tr>
</thead>
<tbody>
<tr>
<td>acq</td>
<td>90.9</td>
<td>89.9</td>
<td>87.8</td>
<td>91.5</td>
<td>92.1</td>
<td>85.4</td>
<td>92.6</td>
</tr>
<tr>
<td>corn</td>
<td>69.6</td>
<td>82.3</td>
<td>83.9</td>
<td>71.5</td>
<td>82.2</td>
<td>87.7</td>
<td>77.9</td>
</tr>
<tr>
<td>crude</td>
<td>77.9</td>
<td>77.0</td>
<td>80.7</td>
<td>81.0</td>
<td>81.5</td>
<td>75.5</td>
<td>85.7</td>
</tr>
<tr>
<td>earn</td>
<td>92.8</td>
<td>89.2</td>
<td>86.6</td>
<td>95.9</td>
<td>96.1</td>
<td>96.1</td>
<td>97.3</td>
</tr>
<tr>
<td>grain</td>
<td>68.8</td>
<td>72.1</td>
<td>73.4</td>
<td>72.5</td>
<td>73.5</td>
<td>89.1</td>
<td>82.2</td>
</tr>
<tr>
<td>interest</td>
<td>70.5</td>
<td>70.1</td>
<td><strong>73.1</strong></td>
<td>58.0</td>
<td>72.5</td>
<td>49.1</td>
<td>74.0</td>
</tr>
<tr>
<td>money-fx</td>
<td>70.5</td>
<td>72.4</td>
<td>70.5</td>
<td>62.9</td>
<td>67.6</td>
<td>69.4</td>
<td>78.2</td>
</tr>
<tr>
<td>ship</td>
<td><strong>73.6</strong></td>
<td>73.2</td>
<td>63.0</td>
<td>78.7</td>
<td>84.1</td>
<td>80.9</td>
<td>79.2</td>
</tr>
<tr>
<td>trade</td>
<td>68.0</td>
<td>69.7</td>
<td><strong>69.8</strong></td>
<td>50.0</td>
<td>77.4</td>
<td>59.2</td>
<td>77.4</td>
</tr>
<tr>
<td>wheat</td>
<td>84.8</td>
<td><strong>86.5</strong></td>
<td>83.3</td>
<td>60.6</td>
<td>79.4</td>
<td>85.5</td>
<td>76.6</td>
</tr>
<tr>
<td>micro-avg</td>
<td><strong>82.1</strong></td>
<td>81.8</td>
<td>81.4</td>
<td>72.0</td>
<td>73.9</td>
<td>79.4</td>
<td>82.3</td>
</tr>
<tr>
<td>macro-avg</td>
<td>76.74</td>
<td><strong>78.24</strong></td>
<td>76.32</td>
<td>65.21</td>
<td>79.14</td>
<td>77.78</td>
<td>82.05</td>
</tr>
</tbody>
</table>

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

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

**FP-Tree**

- (Han et al. 2000)

**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.: \( A \land \neg B \land \neg C \land D \rightarrow E \land \neg F \)

- Confined negative association rules
  \( \neg X \rightarrow Y \)
  \( X \rightarrow \neg Y \)
  \( \neg X \rightarrow \neg Y \)

**Correlation Coefficient**

- Correlation coefficient:
  \[ \rho = \frac{\text{Cov}(X, Y)}{\sigma_X \times \sigma_Y} \]

- Contingency table for binary variables:

<table>
<thead>
<tr>
<th></th>
<th>Y</th>
<th>\neg Y</th>
<th>\Sigma_{ROW}</th>
</tr>
</thead>
<tbody>
<tr>
<td>X</td>
<td>( f_{11} )</td>
<td>( f_{10} )</td>
<td>( f_{1+} )</td>
</tr>
<tr>
<td>\neg X</td>
<td>( f_{01} )</td>
<td>( f_{00} )</td>
<td>( f_{0+} )</td>
</tr>
<tr>
<td>\Sigma_{col}</td>
<td>( f_{+1} )</td>
<td>( f_{+0} )</td>
<td>N</td>
</tr>
</tbody>
</table>

- \( \phi \) correlation coefficient:
  \( \phi = \frac{f_{11} f_{00} - f_{10} f_{01}}{\sqrt{f_{+0} f_{+1} f_{1+} f_{0+}}} \)
Positive & Negative Rule Generation

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

Positive & Negative Rule Generation

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

Some ARC-PAN Results

<table>
<thead>
<tr>
<th>Datasets</th>
<th>C4.5</th>
<th>CBA</th>
<th>ARC-PAN</th>
<th>ARC-PAN</th>
<th>ARC-PAN</th>
<th>Strong Rules</th>
<th>Correlated Rules</th>
<th>Support vs. Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>+R</td>
<td>+R</td>
<td>+R &amp; AR</td>
<td>+R &amp; RA</td>
<td>+R &amp; R</td>
<td># rules</td>
<td>error</td>
<td># rules</td>
</tr>
<tr>
<td>Breast</td>
<td>3.9</td>
<td>4.2</td>
<td>5.5</td>
<td>4.8</td>
<td>3.8</td>
<td>17,000</td>
<td>5.0</td>
<td>1,000</td>
</tr>
<tr>
<td>Diabetes</td>
<td>27.6</td>
<td>25.3</td>
<td>23.3</td>
<td>25.4</td>
<td>25.1</td>
<td>4,000</td>
<td>21.8</td>
<td>40</td>
</tr>
<tr>
<td>Heart</td>
<td>18.9</td>
<td>18.5</td>
<td>16.3</td>
<td>17.0</td>
<td>16.2</td>
<td>200,000</td>
<td>24.7</td>
<td>80</td>
</tr>
<tr>
<td>Iris</td>
<td>5.5</td>
<td>7.1</td>
<td>6.6</td>
<td>6.6</td>
<td>6.0</td>
<td>140</td>
<td>7.3</td>
<td>60</td>
</tr>
<tr>
<td>Led7</td>
<td>26.5</td>
<td>27.8</td>
<td>28.7</td>
<td>28.7</td>
<td>28.9</td>
<td>4,000</td>
<td>34.3</td>
<td>500</td>
</tr>
<tr>
<td>Pima</td>
<td>27.5</td>
<td>27.6</td>
<td>27.4</td>
<td>27.1</td>
<td>26.9</td>
<td>4,000</td>
<td>22.0</td>
<td>50</td>
</tr>
</tbody>
</table>

Repetition of Features

Model transactions of features not as binary
\( \{ \text{Item}_a, \ldots \text{Item}_k \} \)
But as enumerations of repeated features
\( \{ \alpha \text{Item}_a, \ldots \beta \text{Item}_k \} \)

- Use reoccurring frequent itemset mining (ICDE 2000) to generate rules such as:
  \( \{ \alpha \text{Item}_a, \ldots \beta \text{Item}_k \} \rightarrow \text{Class}_x \)

Nothing to do with quantitative association rules
### Number of Rules

- **ACRI generates more rules**

### Accuracy

- **ARC-BC slightly better at 20% but dips later**

### Algorithm Efficiency

- **ACRI is slower because it deals with more rules.**

### Summary

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

- **Model input data into transactions**

- **Set of transactions**
  - Rule Generation
  - Rule Pruning
  - Rule Selection

- **Set of rules**
  - Set of transactions
  - Labeled objects

- **Unlabeled new objects**

- **Distance measure**
  - Coverage
  - Cosine measure
  - Dominance
  - Support/confidence

- **MAXOCCUR**

- **Visual Pruning**
Summary (2)

Rule generator for class $C_1$

Rule generator for class $C_2$

Rule generator for class $C_n$

Transaction separator

Rule merger

Set of transactions

transactions

Set of rules

Unlabeled new objects

Labeled objects

Set of transactions

Rule selection

Rule pruning

Set of rules

Set of transactions

Summary (3)

Rule Performance

correct

not correct

Transactions with class $C_1$

Transactions with class $C_n$

Acc. #rules

Before 74% 4086

After 78% 180

Open Problems?

Learning

Rule generation

Rule pruning

Rule selection

New object

Set of transactions

Set of rules

Classification

New object labeled

Applicable rules

Selected rules

Modelling transactions to incorporate more information

Support threshold-free rule generation

Rule value measure

New heuristics and new pruning strategies

Pruned rules

New heuristics and new selection strategies

Questions

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.