Computer Science 477

Rule Interestingness

Lecture 13

Rules

- Classification rules predict the value of a categorical attribute
  - Of particular importance
- More general problem:
  - Identify relationships between attribute values in a dataset.
- Identify rules that have a conjunction of ‘attribute = value’ terms on both their left- and right-hand sides
- More general than classification
  - Tests on the value of any attribute or combination of attributes

Example

- Financial Dataset
  - IF Has-Mortgage = yes AND Bank Account Status = In credit
  - THEN Job Status = Employed AND Age Group = Adult under 65
- Rules of this more general kind represent an *association* between the values of certain attributes
  - *Association Rules*
  - *Association Rule Mining* (ARM).
  - Also: *Generalized Rule Induction* (or GRI)

Confidence

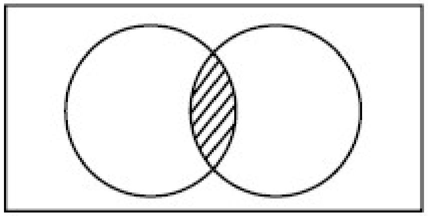
- Rules have a *confidence* value
  - Proportion of instances matched by its left- and right-hand sides combined
  - Divided by of the number of instances matched by the left-hand side on its own.
- Same measure as the predictive accuracy of a classification rule
  - 'Confidence' is more commonly used for association rules.
- Example:
  - IF Has-Mortgage = yes AND Bank Account Status = In credit THEN Job Status = Unemployed
  - Extractible, but very low confidence

Computation

- If there n attributes, each rule can have a conjunction of up to $n - 1$ ‘attribute = value’ terms on the left-hand side.
- Each of the attributes can appear with any of its possible values.
- Any attribute not used on the left-hand side can appear on the right-hand side
  - Also with any of its possible values.
- There are a very large number of possible rules of this kind.
- Generating all of these is very likely to involve a prohibitive amount of computation
  - Especially if there are a large number of instances in the dataset.

Measures of Rule Interestingness - Notation

- Rules always of the form
  - if LEFT then RIGHT
- Four measures
  - N_{LEFT} Number of instances matching LEFT
  - N_{RIGHT} Number of instances matching RIGHT
  - N_{BOTH} Number of instances matching both LEFT and RIGHT
  - N_{TOTAL} Total number of instances
- As a Venn Diagram



  - Instances matching LEFT, RIGHT and both LEFT and RIGHT

Measures of Rule Interestingness

- Confidence (Predictive Accuracy, Reliability)
  - N_{BOTH} / N_{LEFT}
  - The proportion of right-hand sides predicted by the rule that are correctly predicted
- Support
  - N_{BOTH} / N_{TOTAL}
  - The proportion of the training set correctly predicted by the rule
- Completeness
  - N_{BOTH} / N_{RIGHT}
  - The proportion of the matching right-hand sides that are correctly predicted by the rule

Interestingness - Illustration

- Assume values for a rule
- $N_{LEFT} = 65$
- $N_{RIGHT} = 54$
- $N_{BOTH} = 50$
- $N_{TOTAL} = 100$
- From these we can calculate the values of the three interestingness measures
- given in Figure 12.2.
- Confidence = $N_{BOTH}/N_{LEFT} = \frac{55}{65} = 0.77$
- Support = N_{BOTH} / N_{TOTAL} = 50/100 = 0.5
- Completeness = N_{BOTH} / N_{RIGHT} = 50/54 = 0.93

Interestingness - Illustration

- The confidence of the rule is 77%
- Correctly predicts for 93% of the instances in the dataset that match the right-hand side of the rule
- Correct predictions apply to as much as 50% of the dataset.
- A valuable rule.

Discriminability

- Another measure of interest
- Measures how well a rule discriminates between one class
- Defined:
- $1-(N_{LEFT}-N_{BOTH})/(N_{TOTAL}-N_{RIGHT})$
  - 1− (number of misclassifications produced by the rule) / (number of instances with other classifications)
- If the rule predicts perfectly
  - $N_{LEFT}=N_{BOTH}$
  - Value of discriminability is 1
- For the example given above, the value of discriminability is $1-(65-50)/(100-54)=0.67$.

Rule Interestingness Measures: Lift and Leverage

- Number of rules with support and confidence greater than specified threshold still large.
- Need additional interestingness measures we can use to
  - Reduce the number to a manageable size
  - Rank rules in order of importance.
- Lift and Leverage
- The *lift* of rule $L \rightarrow R$ measures how many more times the items in L and R occur together in transactions than would be expected if the itemsets L and R were statistically independent
- *Leverage* example:
  - Suppose a population has an average response rate of 5%, but a certain model (or rule) has identified a segment with a response rate of 20%.
  - Then that segment would have a leverage of 4.0 (20%/5%).

Lift

- *Lift* of rule $L \rightarrow R$ measures how many more times the items in L and R occur together in transactions than would be expected if the itemsets L and R were statistically independent.
- The number of times the items in L and R occur together count($L \cup R$).
- The number of times the items in L occur is count(L).
- The proportion of transactions matched by R is support(R).
- If L and R are independent we would expect the number of times the items in L and R occurred together in transactions to be count(L) × support(R).
- Lift($L \rightarrow R$)= $\frac{\text{count}(L \cup R)}{\text{count}(L) \times \text{support}(R)}$

Other Formulations

- Lift($L \rightarrow R$)= $\frac{\text{count}(L \cup R)}{\text{count}(L) \times \text{support}(R)}$
- $= \frac{\text{support}(L \cup R)}{\text{support}(L) \times \text{support}(R)}$
- $= \frac{\text{confidence}(L \rightarrow R)}{\text{support}(R)}$
- $= \frac{n \times \text{confidence}(L \rightarrow R)}{\text{count}(R)}$
  - n is the number of transactions
- $= \frac{n \times \text{confidence}(R \rightarrow L)}{\text{support}(R)}$
- Lift($L \rightarrow R$)= Lift($R \rightarrow L$)

Lift Example

- Suppose a database of 2000 transactions and a rule $L \rightarrow R$ with the following counts

count(L)	count(R)	count($L \cup R$)
220	250	190

- Calculate:
- $\text{support}(L \rightarrow R) = \frac{\text{count}(L \cup R)}{2000} = 0.095$
- $\text{confidence}(L \rightarrow R) = \frac{\text{count}(L \cup R)}{\text{count}(L)} = 0.846$
- $\text{lift}(L \rightarrow R) = \text{confidence}(L \cup R) \times \frac{2000}{\text{count}(R)} = 6.91$

Lift Example

- The value of $\text{support}(R)$ measures the support for R in whole of the database.
- The itemset matches 250 transactions out of 2000, a proportion of 0.125.
- The value of $\text{confidence}(L \rightarrow R)$ measures the support for R if we only examine the transactions that match L.
- Here: $190/220 = 0.864$.
- So purchasing the items in L makes it $0.864/0.125 = 6.91$ times more likely that the items in R are purchased.
- Lift values greater than 1 are ‘interesting’.
- Indicate that transactions containing L tend to contain *R* more often than transactions that do not contain L.
- Although lift is a useful measure
  - Not always best
  - Sometimes a rule with higher support and lower lift can be more because it applies to more cases

Leverage

- Measures the difference between
  - The support for $L \cup R$ (i.e. the items in L and R occurring together in the database)
    - $\text{Support}(L \cup R)$.
  - The support that would be expected if L and R were independent
    - Frequencies (i.e. supports) of L and R are $\text{support}(L)$ and $\text{support}(R)$, respectively
- Formula
  - $\text{leverage}(L \rightarrow R) = \text{support}(L \cup R) - \text{support}(L) \times \text{support}(R)$.
- The value of the leverage of a rule is clearly always less than its support

Leverage Example

- The number of rules satisfying the support ≥ *minsup* and confidence ≥ *minconf* constraints reduced by setting a leverage constraint,
  - E.g. leverage ≥ 0.0001
  - Corresponds to an improvement in support of one occurrence per 10,000 transactions in the database.
- If a database has 100,000 transactions and we have a rule $L \rightarrow R$ with these support counts

$\text{count}(L)$	$\text{count}(R)$	$\text{count}(L \cup R)$
8000	9000	7000

- Values of support, confidence, lift and leverage can be calculated to be 0.070, 0.875, 9.722 and 0.063 respectively
  - (all to three decimal places)

Leverage Example

- $\text{Support} = 0.070, \text{confidence} = 0.875, \text{lift} = 9.722,$ $\text{leverage} = 0.063$
- Rule applies to 7% of the transactions in the database
- Rule is satisfied for 87.5% of the transactions that include the items in *L*.
- The latter value is 9.722 times more frequent than would be expected by chance.
- The improvement in support compared with chance is 0.063
  - Corresponding to 6.3 transactions per 100 in the database,
  - I.e. approximately 6300 in the database of 100,000 transactions

Piatetsky-Shapiro Criteria

- Criterion 1
  - The measure should be zero if $N_{BOTH} = (N_{LEFT} \times N_{RIGHT})/N_{TOTAL}$
  - Interestingness should be zero if the antecedent and the consequent are statistically independent
- Criterion 2
  - The measure should increase monotonically with N_{BOTH}
- Criterion 3
  - The measure should decrease monotonically with each of N_{LEFT} and N_{RIGHT}
- For criteria 2 and 3, it is assumed that all other parameters are fixed.

Piatetsky-Shapiro Criteria - Interpretation

- Criterion 2
  - If everything else is fixed the more right-hand sides that are correctly predicted by a rule the more interesting it is.
- Criterion 3
  - If everything else is fixed
    - (a) the more instances that match the left-hand side of a rule the less interesting it is.
    - (b) the more instances that match the right-hand side of a rule the less interesting it is.

Piatetsky-Shapiro Criteria - Interpretation

- The purpose of (a)
  - Give preference to rules that correctly predict a given number of right-hand sides from as few matching left-hand sides as possible
    - For a fixed value of N_{BOTH}, the smaller the value of N_{LEFT} the better).
- The purpose of (b)
  - Give preference to rules that predict right-hand sides that are relatively infrequent
    - Predicting common right-hand sides is easier to do).

Meaning of Criterion 1

- Antecedent and the consequent of a rule (i.e. its left- and right-hand sides) are independent.
  - Whether RHS predicted by chance.
- Total instances given by N_{TOTAL}
- Number of those instances that match the right-hand side of the rule is N_{RIGHT}
- So random prediction expects N_{RIGHT}/N_{TOTAL}
- If we predicted the same right-hand side N_{LEFT} times
  - (one for each instance that matches the left-hand side of the rule),
  - Expect that $N_{LEFT} \times N_{RIGHT}/N_{TOTAL}$

Meaning of Criterion 1

- If we predicted the same right-hand side N_{LEFT} times
  - (one for each instance that matches the left-hand side of the rule),
- Expect that $N_{LEFT} \times N_{RIGHT} / N_{TOTAL}$
- By definition the number of times that the prediction actually turns out to be correct is N_{BOTH}.
- If the number of correct predictions made by the rule is the same as the number that would be expected by chance the rule interestingness is zero.

Piatetsky-Shapiro Measure

- Interestingness measure: *RI*
  - Simplest measure that meets his three criteria.
- Defined by:
  - $RI = N_{BOTH} - N_{LEFT} \times N_{RIGHT} / N_{TOTAL}$
- *RI* measures the difference between the actual number of matches and the expected number if the left- and right-hand sides of the rule were independent.
- A value of zero would indicate that the rule is no better than chance.
- A negative value would imply that the rule is less successful than chance.
- The *RI* measure satisfies all three of Piatetsky-Shapiro's criteria.

Application to Classification - Chess

- Unpruned decision tree derived from the *chess* dataset (with attribute selection using entropy) comprises 20 rules.
- Example:
  - IF inline = 1 AND wr bears bk = 2 THEN Class = safe
  - $RI = N_{BOTH} - N_{LEFT} \times N_{LEFT}/N_{TOTAL}$
- For this rule
  - $N_{LEFT} = 162$
  - $N_{RIGHT} = 613$
  - $N_{BOTH} = 162$
  - $N_{TOTAL} = 647$

Application to Classification - Chess

- Confidence = $162/162 = 1$
- Completeness = $162/613 = 0.26$
- Support = $162/647 = 0.25$
- Discriminability = $1 - (162 - 162)/(647 - 613) = 1$
- $RI = 162 - (162 \times 613/647) = 8.513$
- Perfect values of confidence and discriminability are of little value here.
  - Always occur when (1) classification tree unpruned and (2) no clashes

Interestingness for all Rules

Rule	N_{LEFT}	N_{RIGHT}	N_{BOTH}	Conf	Compl	Supp	Discr	RI
1	2	613	2	1.0	0.003	0.003	1.0	0.105
2	3	34	3	1.0	0.088	0.005	1.0	2.842
3	3	34	3	1.0	0.088	0.005	1.0	2.842
4	9	613	9	1.0	0.015	0.014	1.0	0.473
5	9	613	9	1.0	0.015	0.014	1.0	0.473
6	1	34	1	1.0	0.029	0.002	1.0	0.947
7	1	613	1	1.0	0.002	0.002	1.0	0.053
8	1	613	1	1.0	0.002	0.002	1.0	0.053
9	3	34	3	1.0	0.088	0.005	1.0	2.842
10	3	34	3	1.0	0.088	0.005	1.0	2.842
11	9	613	9	1.0	0.015	0.014	1.0	0.473
12	9	613	9	1.0	0.015	0.014	1.0	0.473
13	3	34	3	1.0	0.088	0.005	1.0	2.842
14	3	613	3	1.0	0.005	0.005	1.0	0.158
15	3	613	3	1.0	0.005	0.005	1.0	0.158
16	9	34	9	1.0	0.265	0.014	1.0	8.527
17	9	34	9	1.0	0.265	0.014	1.0	8.527
18	81	613	81	1.0	0.132	0.125	1.0	4.257
19	162	613	162	1.0	0.264	0.25	1.0	8.513
20	324	613	324	1.0	0.529	0.501	1.0	17.026

Chess Interestingness Results

Rule	N_{LEFT}	N_{RIGHT}	N_{BOTH}	Conf	Compl	Supp	Discr	*RI*
1	2	613	2	1.0	0.003	0.003	1.0	0.105
2	3	34	3	1.0	0.088	0.005	1.0	2.842
3	3	34	3	1.0	0.088	0.005	1.0	2.842
4	9	613	9	1.0	0.015	0.014	1.0	0.473
5	9	613	9	1.0	0.015	0.014	1.0	0.473
6	1	34	1	1.0	0.029	0.002	1.0	0.947
7	1	613	1	1.0	0.002	0.002	1.0	0.053
8	1	613	1	1.0	0.002	0.002	1.0	0.053
9	3	34	3	1.0	0.088	0.005	1.0	2.842
10	3	34	3	1.0	0.088	0.005	1.0	2.842
11	9	613	9	1.0	0.015	0.014	1.0	0.473
12	9	613	9	1.0	0.015	0.014	1.0	0.473
13	3	34	3	1.0	0.088	0.005	1.0	2.842
14	3	613	3	1.0	0.005	0.005	1.0	0.158
15	3	613	3	1.0	0.005	0.005	1.0	0.158
16	9	34	9	1.0	0.265	0.014	1.0	8.527
17	9	34	9	1.0	0.265	0.014	1.0	8.527
18	81	613	81	1.0	0.132	0.125	1.0	4.257
19	162	613	162	1.0	0.264	0.25	1.0	8.513
20	324	613	324	1.0	0.529	0.501	1.0	17.026

- Judging by the *RI* values, only the last five rules are of interest.
- They are the only rules (out of 20) that correctly predict the classification for at least four instances more than would be expected by chance.
- Rule 20 predicts the correct classification 324 out of 324 times.
  - Support value is 0.501
  - i.e. it applies to over half the dataset, and its completeness value is 0.529.
- By contrast, Rules 7 and 8 have *RI* values as low as 0.053,
  - i.e. they predict only slightly better than chance.

What we do with Measures

- Might prefer only to use rules 16 to 20.
- Unwise
- Result: a tree with only five branches
  - Unable to classify 62 out of the 647 instances in the dataset

Conflict Resolution

- When several rules predict different values for one or more attributes of interest for an unseen test instance.
- Rule interestingness measures give one approach to handling this.
- Might decide to use only the rule with the highest interestingness value,
- Or the most interesting three rules
- Or more ambitiously we might decide on a ‘weighted voting’ system that adjusts for the interestingness value
- Or values of each rule that fires.

Summary

- Problem: of finding any rules of interest that can be derived from a given dataset
  - Not just classification rules as before.
- Known as Association Rule Mining or Generalized Rule Induction.
- Requires measures of rule interestingness and criteria for choosing between measures.
- Monday: An algorithm for finding the best N rules that can be generated from a dataset using a new measure:
  - *J*-measure of the information content of a rule
- Also a 'beam search' strategy.