Cluster-Grouping: From Subgroup Discovery to Clustering

Luc De Raedt, Albrecht Zimmermann
Machine Learning Lab
Albert-Ludwigs-University Freiburg

March 12, 2004

Outline

- Motivation
- Solution
- Statistical Metric Pruning
- Extension to Several Attributes
- Discussion
- Conclusion

Motivation

1. Unifying two KDD tasks:
   - Subgroup Discovery: Rule learning using WRAcc
   - Symbolic Clustering: Cobweb (using Category Utility)
   - Differences:
     - One fixed target **vs** Disjunction of all attributes
     - Rule learning **vs** Arbitrary instance assignment
     - Overlap **vs** Non-overlap
   - Similarities:
     - Finding groups of instances showing unexpected behavior with regard to target attribute(s)
     - Heuristic approach since number of possible partitions too large to exhaust
2. Getting away from the heuristic

Solution

1. View search process as trying to find the best rule(s) (having a disjunctive consequent) with regard to interestingness measure

    - For clustering fewer partitions that can be explored but
    - Rules are descriptions of found clusters, more easily understandable than Cobwebs representation
    - Allows mining for subsets of attributes

2. Use statistical metric pruning (later)

Solution - More Formal

- Given:
  - Set of instances $\mathcal{E}$
  - Set of attributes $\mathcal{A}$
  - Pattern language $\mathcal{L} = \{\bigwedge A_i = V_k^i \Rightarrow \bigvee A_j \mid A_i, A_j \in \mathcal{A}\}$
  - Interestingness measure $\sigma : \mathcal{L} \times \mathcal{E} \longrightarrow \mathbb{R}$
- Find: The antecedent(s) $\bigwedge A_i = V_k^i$ maximizing σ for fixed consequent $\bigvee A_i$

Solution - The Algorithm

- Separate-and-Conquer
- Find splitting criterion on $\mathcal{E}$, refine on subsets
- Best-first, dynamic threshold, length as tiebreaker
- Prune according to $promise$

Statistical Metric Pruning

- Applied to association rule mining by Morishita et al
- Motivation: rules fulfilling minimum support not necessarily meaningful, have to be judged after mining
- Also: too high minimum support might lead to loss of interesting rules, too low produces many useless rules

Statistical Metric Pruning (2)

a	b	c	d	e
1	1	1	1	1
1	1	0	1	1
1	0	1	0	1
1	0	0	0	1
0	1	1	0	1
0	1	0	0	1
0	0	1	1	1
0	0	0	1	1

$I \Rightarrow C$	Support	Confidence	Correlated?
$\{a\} \Rightarrow \{b\}$	25%	50%	no
$\{a, b\} \Rightarrow \{d\}$	25%	100%	yes
$\{a\} \Rightarrow \{e\}$	50%	100%	no

- Solution: make interestingness measure (entropy gain, χ^2 etc.) anti-monotonous by introducing upper bound on values specializations of rules can attain, prune accordingly

- Limited to one, fixed target, somewhat different from classical itemset mining

Statistical Metric Pruning - More In-Depth

- Assume association rule $I \Rightarrow C$

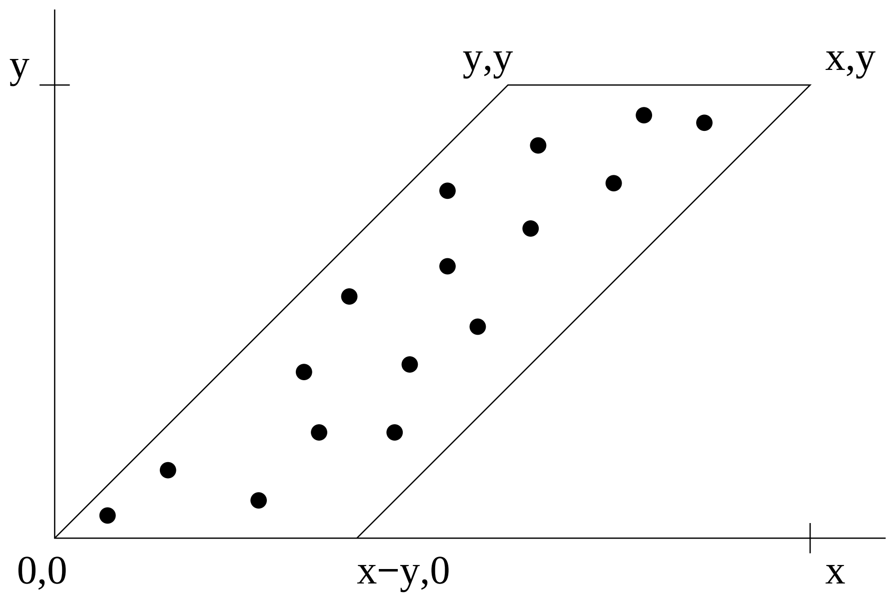
	C	$\neg C$	
I	$O_{IC} = y$	$O_{I\neg C}$	x
$\neg I$	$O_{\neg IC}$	$O_{\neg I \neg C}$	$n - x$
	m	$n - m$	n

- Interestingness measures such as entropy gain, χ^2 can be written as functions $f(x, y)$ if cover consequent, cardinality data set known

- Considered thus they are convex functions $\Rightarrow$ when defined on convex set take extreme values on the convex hull

Statistical Metric Pruning - More In-Depth (2)

- Current $\langle x, y\rangle$ puts constraints on future tuples $\Rightarrow$ maximum values attainable in the future can be computed (at $\langle y, y\rangle$, $\langle x - y, 0\rangle$)



Extension to Several Attributes

- Consider $I \Rightarrow C_1 \vee C_2 \vee \ldots \vee C_n$
- Interestingness measure becomes higher dimensional: $f(x, y_1, y_2, \ldots, y_n)$
- Same technique but dimension is increased
- Constraints on convex hull points more complex than in 2-dimensional case, simply projecting on 2-dimensional leads to losing points
- Number of hull points is in $O(2^n)$

Extension to Several Attributes - An Example

	C_1	$\neg C_1$	C_2	$\neg C_2$	
I	$O_{IC_1} = y_1$	$O_{I\neg C_1}$	$O_{IC_2} = y_2$	$O_{I\neg C_2}$	x
$\neg I$	$O_{\neg IC_1}$	$O_{\neg I\neg C_1}$	$O_{\neg IC_2}$	$O_{\neg I\neg C_2}$	$n - x$
	m_1	$n - m_1$	m_2	$n - m_2$	n

- Let $\langle x, y_1, y_2 \rangle = \langle 10, 2, 3 \rangle$
- Convex hull points (naïvely):

$$\begin{array}{cc} \langle 10, 2, 3 \rangle & \langle 3, 2, 3 \rangle \\ \langle 7, 2, 0 \rangle & \langle 2, 2, 0 \rangle \\ \langle 8, 0, 3 \rangle & \langle 3, 0, 3 \rangle \\ \langle 7, 0, 0 \rangle & \langle 0, 0, 0 \rangle \end{array}$$

Extension to Several Attributes - An Example (2)

- What about $\langle 8, 0, 1 \rangle$?
- Not covered by naïve convex hull

$$\begin{array}{cc} \langle 10, 2, 3 \rangle & \langle 8, 2, 3 \rangle \\ \langle 7, 2, 0 \rangle & \langle 2, 2, 0 \rangle \\ \langle 8, 0, 3 \rangle & \langle 3, 0, 3 \rangle \\ \langle 8, 0, 1 \rangle & \langle 0, 0, 0 \rangle \\ \langle 7, 0, 0 \rangle & \end{array}$$

Pros and Cons

- Pro:
  - Get rid of the black art:
    * No arbitrary beam size in rule learning, simply extend most promising
    * No re-runs in clustering, since global optimum found
  - In rule learning, don't consider all specializations, only promising ones
- Con:
  - Possibly many points to evaluate, probably only sensible if instance counting expensive
  - So far restricted to binary target attributes
  - Upper bound overestimates potential

Experimental Evaluation

- Compared to Cobweb on variety of UCI data sets
- Sometimes better, sometimes worse, always competitive
- On several sets fewer rules evaluated than number of instances
- Cobweb's better solutions not easily describable
- Cobweb sometimes stuck in local extrema

Future Work

- So far either multiple variables or multiple values, not both
  - For single variable, multi-value case approach is similar, hull more constrained
  - For multi-variable, multi-value possible to constrain sum of y_i but then ...
- Number of hull points quickly becomes too large
  - Identical $\langle x, y_1, y_2, \ldots \rangle$ have same upper bound: Look-up Table?
  - Given identical size of cover for different C_i, certain tuples can be considered symmetric
  - Since disjunction of rules, possible to sum maximum values of single rules $\longrightarrow$ impossible combinations considered $\Rightarrow$ upper bound too generous

Conclusions

- Framework unifying kdd tasks such as subgroup discovery, clustering
- Based on rule learning and statistical metric pruning
- Statistical metric pruning meaningful and effective
- Good results when compared to existing approach
- Open questions:
  - Multi-variable, multi-value tasks
  - Computational efficiency

Questions?

References

References

[Fisher 1987] Fisher, D.H. Knowledge acquisition via incremental conceptual clustering. *Machine Learning 2*, 129-172

[Gluck 1985] Gluck, M.A. and Corter, J.E. Information, uncertainty, and the utility of categories. *Proceedings of the Seventh Annual Conference of the Cognitive Science Society.* Irvine, CA. 1985

[Lavrac 1999] Lavrac, N., Flach, P., Zupan, B. Rule evaluation measures: A unifying view. *Proceedings of the 9th International Workshop on Inductive Logic Programming.* 1999

[Lavrac 2002] Lavrac, N., Flach, P., Zelezný RSD: Relational Subgroup Discovery through First-Order Feature Construction *Proceedings of the 12th International Conference on Inductive Logic Programming.* July 2002

[Morishita 2000] Morishita, S., Sese, J. Traversing Itemset Lattice with Statistical Metric Pruning. *Proceedings of the 19th ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems.* May 2000