Random Ordinality Ensembles
An Ensemble Method for Multi-Valued Categorical Data

Amir Ahmad
Gavin Brown

School of Computer Science,
University of Manchester, U.K.

Contents

- **What is "Random Ordinality"?**
- **Where/how it is used**
  - With a single decision tree
  - With an ensemble of decision trees
- **Results with various datasets**
  - RO significantly outperforms many popular algorithms
  - RO resists data fragmentation
- **Why does it work?**
- **Extensions / Conclusions**

Categorical data

	Outlook	Temperature	Humidity	Wind	Play Tennis?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

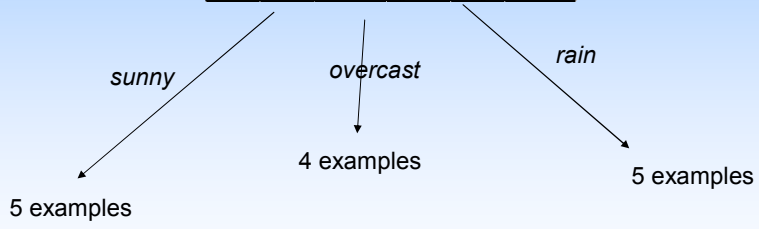
For some types of attributes, a natural "ordering" is present
(hot > mild > cool) … or (strong > weak)

What about (sunny > rain)? Or should it be (rain > sunny?)

How should we learn a decision tree in this case?

Decision trees: multi-way splits vs binary slits

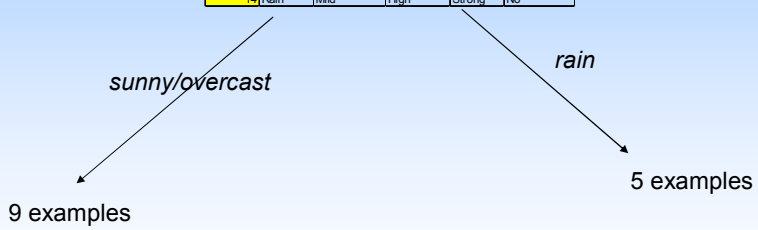
	Outlook	Temperature	Humidity	Wind	Play Tennis?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No



"Data Fragmentation" → unreliable lower branches

Decision trees: multi-way splits vs binary splits

	Outlook	Temperature	Humidity	Wind	Play Tennis?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No



Binary split → more reliable lower branches ☺

HUGE number of possible binary splits

Attribute with |A| possible values → $2^{(|A|-1)} - 1$ possible splits.

With |A| = 3 → possible splits 3

With |A| = 10 → possible splits 511

Computationally Expensive

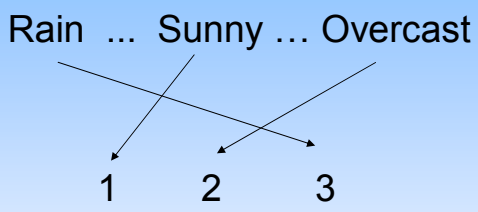
What is Random Ordinality?

- Some data does not have a natural “ordering”
- Multi-way splits cause data fragmentation

- We can **impose** an ordering, **randomly**
- **….**and do a binary split

	Outlook	Temperature	Humidity	Wind	Play Tennis?
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Randomly imposing an ordering - convert into ordinal data



Sunny < Overcast < Rain

Allows binary split : “Sunny OR Overcast” versus “Rain”

Data will be treated as continuous hence binary splits.

Results with a single tree

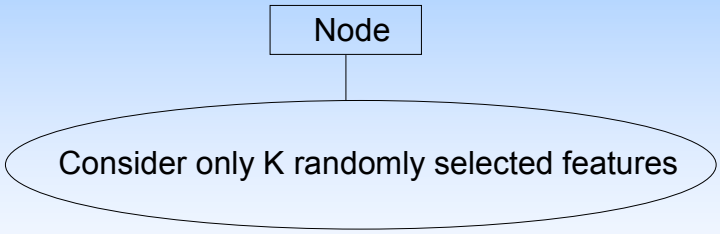
Dataset	Decision tree (J48) with original data, error in %	RO with 348 tree error in %
Promoter	28.5	**25.3**
Hayes-Roth	25.3	**21.7**
Breast Cancer	35.9	**33.4**
Monks-1	**15.9**	26.1
Monks-2	49.6	**32.3**
Monks-3	**0**	0.1
Balance	31.4	**26.6**
Soyalarge	**8.7**	10.5
Tic-tac-toe	18.4	**12.4**
Car	9.2	**6.5**
DNA	8.9	**8.5**
Mushroom	**0**	0.2
Nursery	3.6	**2.2**

Error of C4.5 decision trees with multi-way splits and average accuracy RO trees. **For 9 out of 13 datasets, on average RO trees perform better than multi-way split trees**.

Two variants of RO ensembles

1- Combination of RO trees

2- Combining RO with Random Subspaces at the nodes - to increase diversity



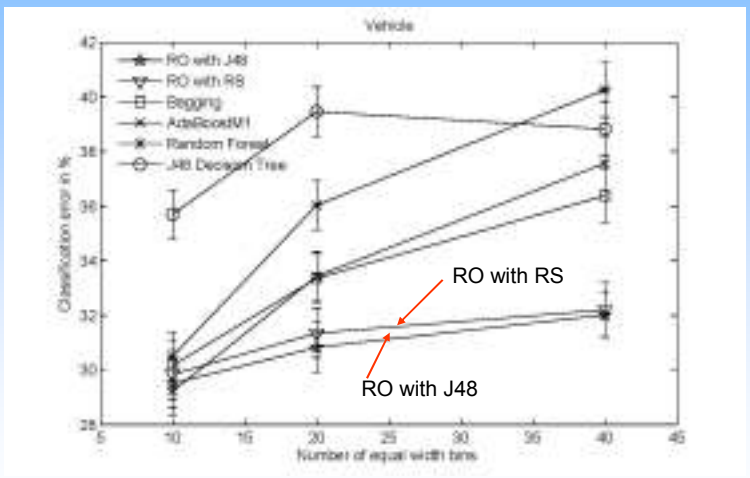
Examples – Random Forests.

Comparative study of various ensemble methods

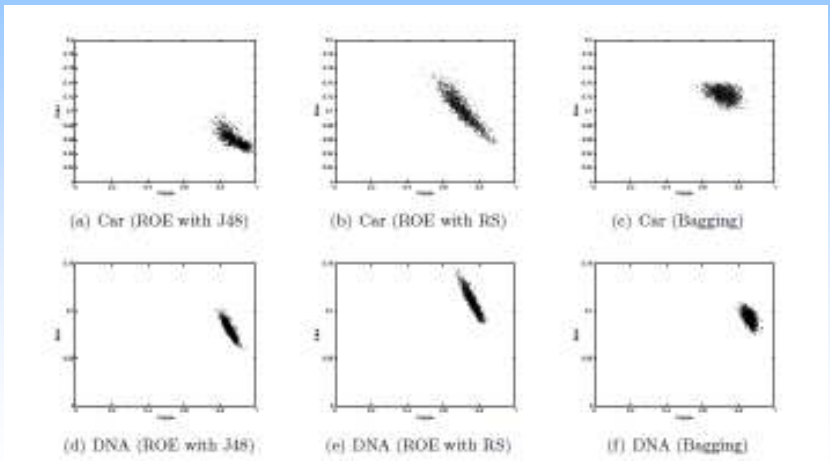
Dataset	RO with J48	RO with RS	Bagging	AdaBoostM1	Random Forests	Single Tree (J48)
Promoter	13.1(2)	**12.8(1)**	15.5(4)	19.6(5)	13.4(3)	28.5(6)
Hayes-Roth	16.9(2)	**15.9(1)**	22.8(4)	23.1(5)	22.2(3)	25.3(6)
Breast Cancer	30.3(3)	30.1(2)	**29.9(1)**	35.6(5)	32.4(4)	35.9(6)
Monks1	18.3(5)	**1.5(1)**	5.8(3)	5.9(4)	3.3(2)	15.9(6)
Monks2	33.9(2)	**30.9(1)**	46.9(3)	47.5(4)	50.4(6)	49.6(5)
Monks3	0(3.5)	0(3.5)	0(3.5)	0(3.5)	0(3.5)	0(3.5)
Balance	**19.4(1)**	20.0(2)	29.6(4)	30.3(5)	26.9(3)	31.4(6)
Soyalarge	8.8(5)	**7.3(1.5)**	8.2(4)	**7.3(1.5)**	7.9(3)	9.7(6)
Tic-tac-toe	6.6(3)	**3.4(1)**	10.0(5)	3.5(2)	8.6(4)	18.4(6)
Car	**4.1(1)**	4.2(2)	8.3(4.5)	5.9(3)	8.3(4.5)	9.2(6)
DNA	4.5(2)	**4.4(1)**	6.2(5)	5.1(3)	5.8(4)	8.9(6)
Mushroom	0.1(5.5)	0.1(5.5)	**0(2.5)**	**0(2.5)**	**0(2.5)**	**0(2.5)**
Nursery	1.0(2)	**0.9(1)**	2.8(5)	1.3(3)	2.6(4)	3.6(6)
Average Rank	2.8	1.8	3.8	3.6	3.3	5.5
Rank	**II**	**I**	**V**	**IV**	**III**	**VI**

Classification error in %, 5x2 cross fold testing, the size of the ensemble = 50.

Resistance to data fragmentation
(*Vehicle data)*



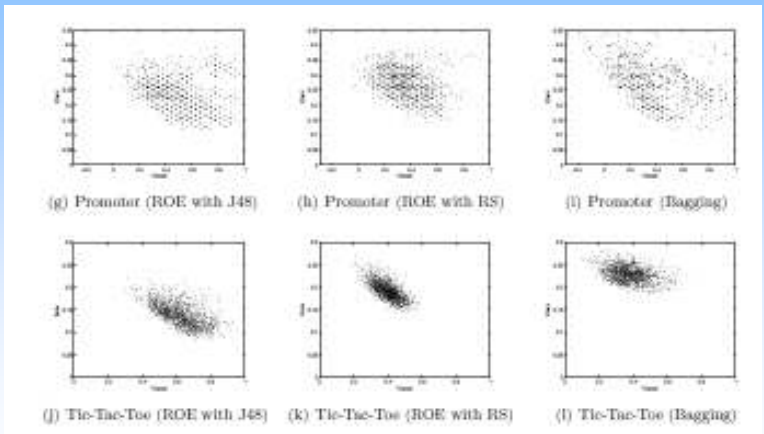
Kappa- Error Diversity Plots



(a) Car (ROE with J48) (b) Car (ROE with RS) (c) Car (Bagging)

(d) DNA (ROE with J48) (e) DNA (ROE with RS) (f) DNA (Bagging)

Kappa- Error Diversity Plots



(g) Promoter (ROE with J48) (h) Promoter (ROE with RS) (i) Promoter (Bagging)

(j) Tic-Tac-Toe (ROE with J48) (k) Tic-Tac-Toe (ROE with RS) (l) Tic-Tac-Toe (Bagging)

Theoretical study – information gain ratio

Let D is a 2 class (positive and negative) dataset such that it has same number of positive and negative examples.

Let A is a multi-valued attribute. It has |A| number of different values. These values have equal representation.

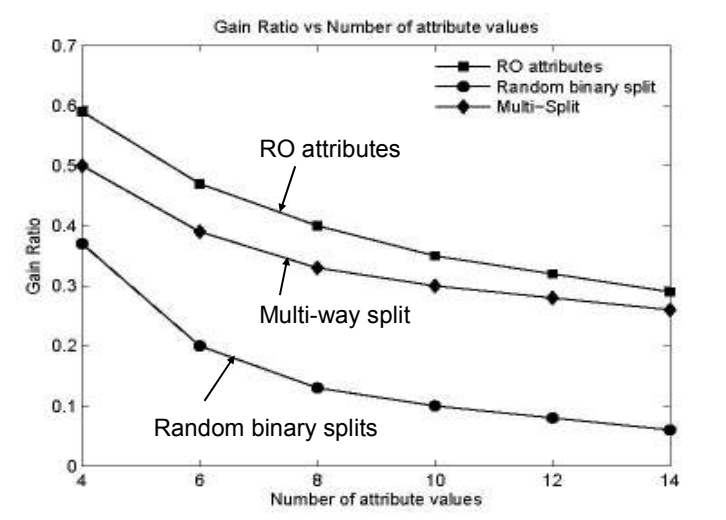
Half of these values correctly identify positive class, whereas rest of the values correctly identify negative class.

For example, if attribute values are (A, B, C, D, E, F),

p(class = positive|attribute value = A) = 1
p(class = positive|attribute value = B) = 1
p(class = positive|attribute value = C) = 1

p(class = negative|attribute value = D) = 1
p(class = negative|attribute value = E) = 1
p(class = negative|attribute value = F) = 1

Why does it work? **Better information gain ratio.**



We summarize RO ensembles as follows

- RO trees are generally more accurate as compared to normal decision trees.
- RO ensembles avoid the data fragmentation problem, and provide performance improvements over several standard ensemble methods
- Error-Diversity analysis suggests that RO is able to create accurate classifiers with reasonable diversity.
- RO is easy to implement. Parallel implementation of RO ensembles is also possible.

The "take-home" message of this work is that, *as categorical attribute values have no intrinsic order*, this property can be exploited to build an ensemble of diverse binary decision trees.

Thanks

Questions ?

Resistance to curse of dimensionality (*Segment data: 5 fold CV)*