Concept Learning for Semantic Memory

Yongjia Wang
University of Michigan

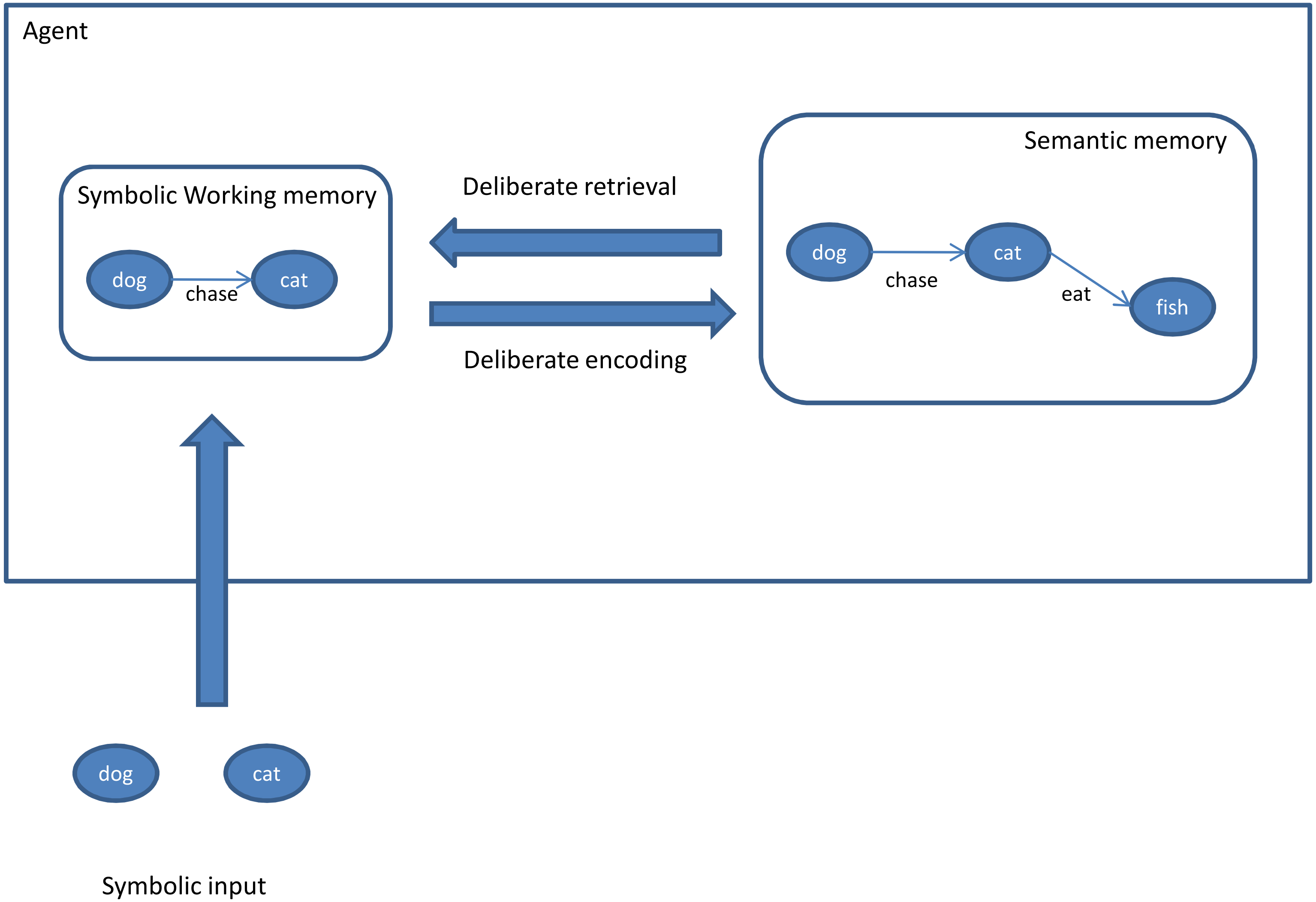
Outline

- Background & Motivation
- Algorithm
- Preliminary Evaluation
- Conclusion
- Nuggets and Coal

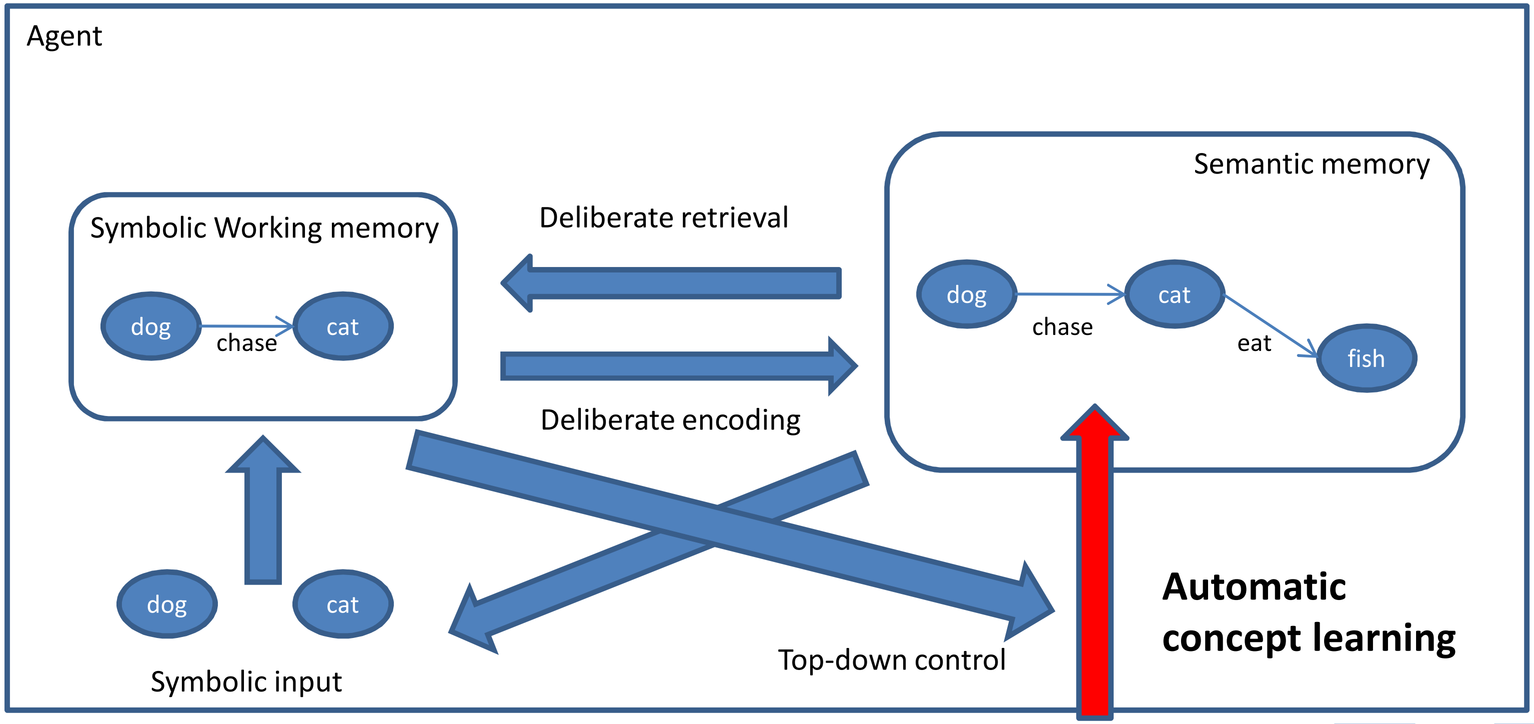
Semantic Memory

- Exploratory Research
- General characteristics of semantic memory
  - General facts
  - Abstract concepts
- Cognitive capabilities
  - Remembering and retrieving general facts
  - <u>Representing and learning abstract concepts</u>
  - Representing and learning world model

Semantic Learning



Semantic Learning



Subsymbolic inputs

Descriptive Boolean vector

fur	wing	bark	mew
1	0	1	0

Numeric vector

Relational graph

Image

…

Motivation

- Previous instance based approach
  - Sufficient for encoding and retrieving general facts interfaced with working memory
  - Cannot learning from sub-symbolic input
- Prototype based approach
  - Generate symbols from sub-symbolic input

Learning Paradigms

- Reinforcement learning
- Unsupervised learning
- Supervised learning
- Natural concept learning (Semi-supervised learning)
  - Unsupervised learning
    - Learn from input without class label
  - Supervised learning
    - Learn with class label
    - Externally supervised
    - Self supervised

Desired Algorithm Properties

- Semantic memory is the long term concept memory for a continuously learning agent
- Statistical learning
  - Robust against noisy environment
- Incremental
  - Continuously learning
- Scalable
  - Large amount of information
- Semi-supervised learning
  - Learn from both labeled and unlabeled input

Hierarchical Clustering Algorithm

- Adapted from COBWEB (D. Fisher)
- Major components
  - Clustering utility function
  - Local restructuring operators
  - Clustering space search
- Modification
  - Numeric attribute utility function
  - Local restructuring operators
  - Hash index based access (not evaluated)

UnitTransform(x) = $\frac{a^x}{1+a^x} \quad (0 \sim \infty) \rightarrow (0 \sim 1)$

a is a constant > 1, 1.1 seems to be good

PartitionUtility = $\sum_{i=1}^{n} P(C_i) \bullet CU(C_i) \Big/ PENALTY(n)$

= (.67 * 1.67 + .33 *2.08)/2

= 0.9

MAXIMIZE

Normalized by parent standard deviation

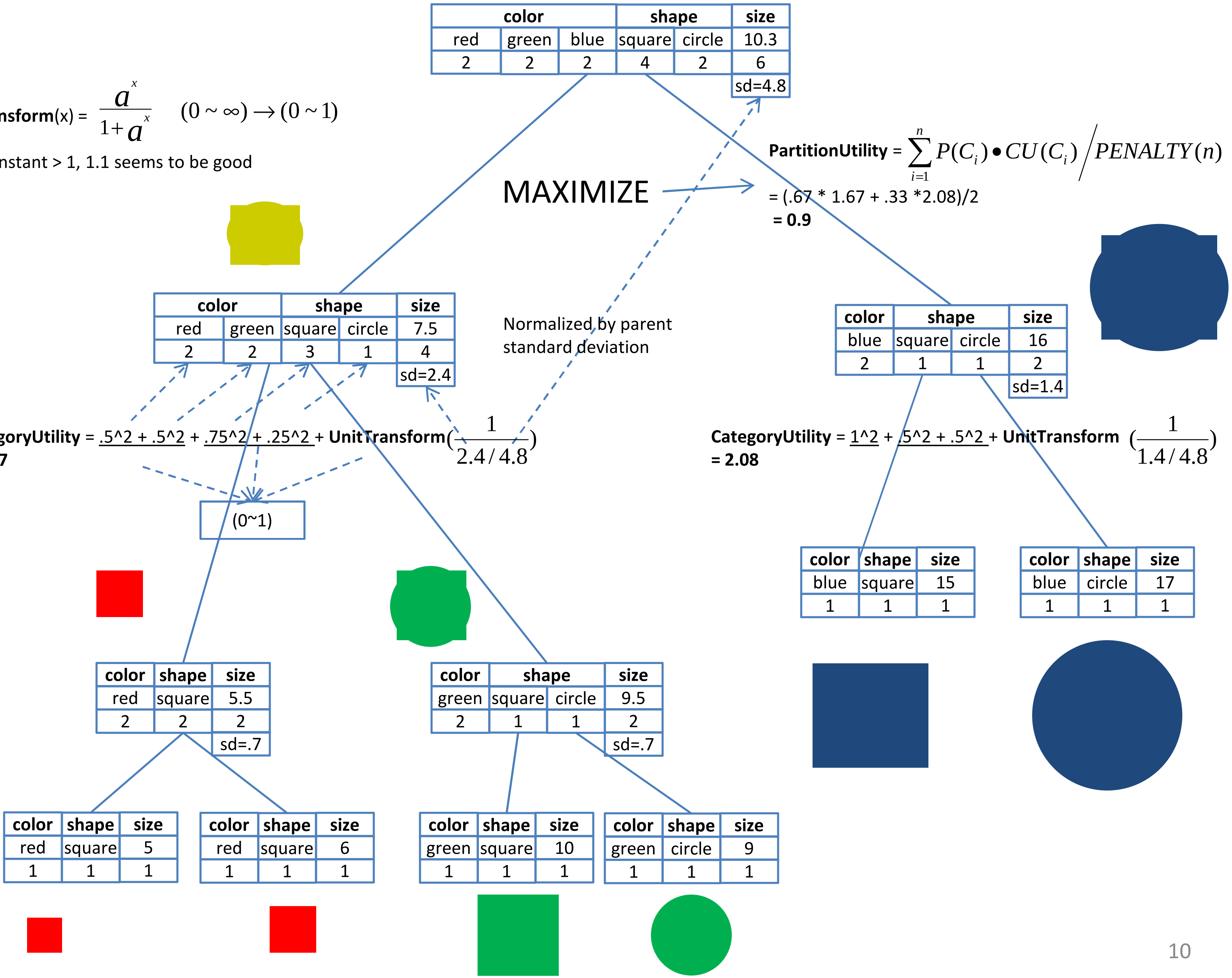
CategoryUtility = .5^2 + .5^2 + .75^2 + .25^2 + **UnitTransform**$\left(\frac{1}{2.4/4.8}\right)$

= 1.67

(0~1)

CategoryUtility = 1^2 + .5^2 + .5^2 + **UnitTransform** $\left(\frac{1}{1.4/4.8}\right)$

= 2.08



Algorithm Properties Revisited

- Statistical learning
- Incremental
- Scalable
  - Hierarchy (log n)
- Semi-supervised learning
  - Unsupervised learning: incremental clustering
  - Weak supervised learning: assign class label
  - Stronger supervised learning: class label can participate in clustering utility evaluation

Preliminary Evaluations

Purpose	Evaluation	Result
Unsupervised Clustering	Qualitative	?
Compare clustering with instance based learning	Prediction Accuracy	?
Compare different degrees of prior unsupervised learning	Prediction Accuracy	?

- Instance (exemplar) based learning
  - Naïve implementation
  - Linear complexity to find nearest neighbor (best partial match)
- Types of data
  - Symbolic to numeric features
  - Low dimension to high dimension vector input

Evaluation on Artificial data set

Symbolic values

	Component1			Component2			20 Random features with values (1~3)
class a	0	0	0	0	0	0	…
class b	0	0	0	1	1	1	…
class c	1	1	1	1	1	1	…
class d	1	1	1	2	2	2	…
class e	2	2	2	2	2	2	…
class f	2	2	2	0	0	0	…
	Signal						Noise

Training and Testing

Unsupervised

0	0	0	1	1	1	…

Learning structure of input without label

Supervised

0	0	0	1	1	1	…

\+ class b

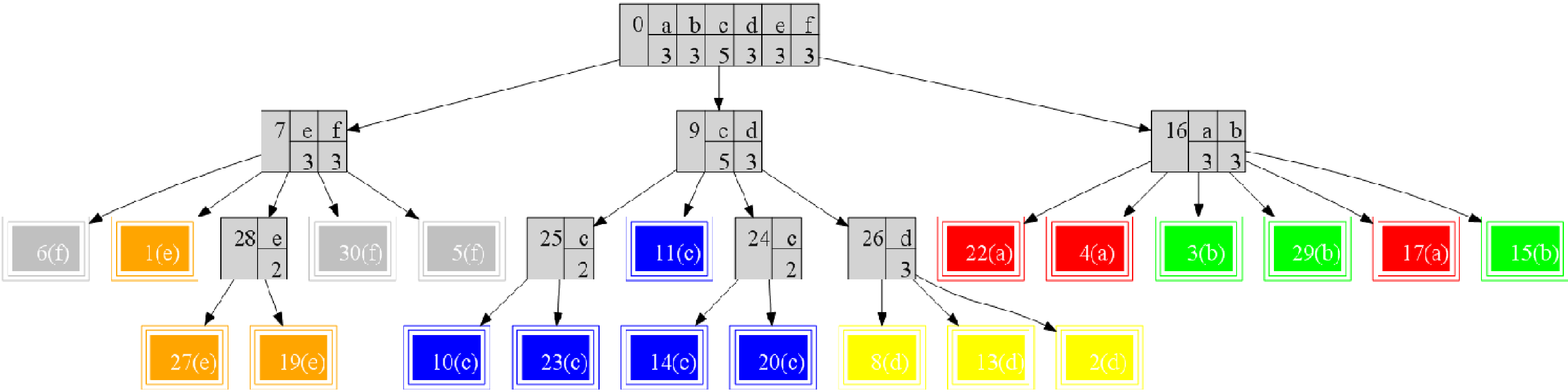
Learn the ‘meaning’ of the concepts

Prediction

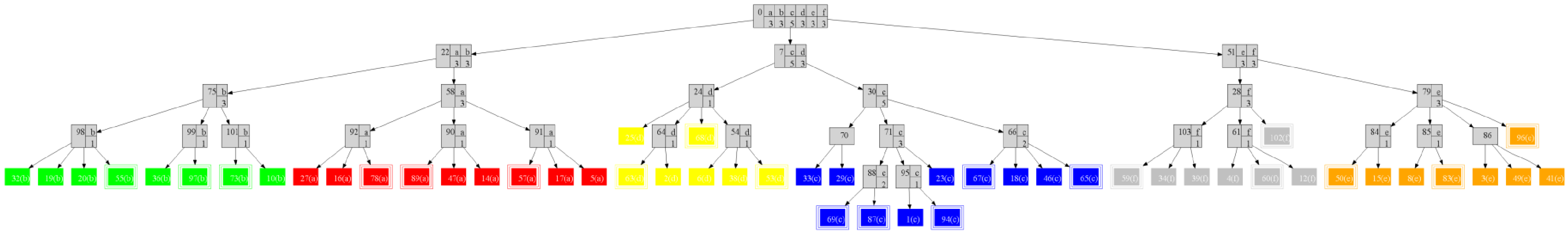
0	0	0	1	1	1	…

-> class ?

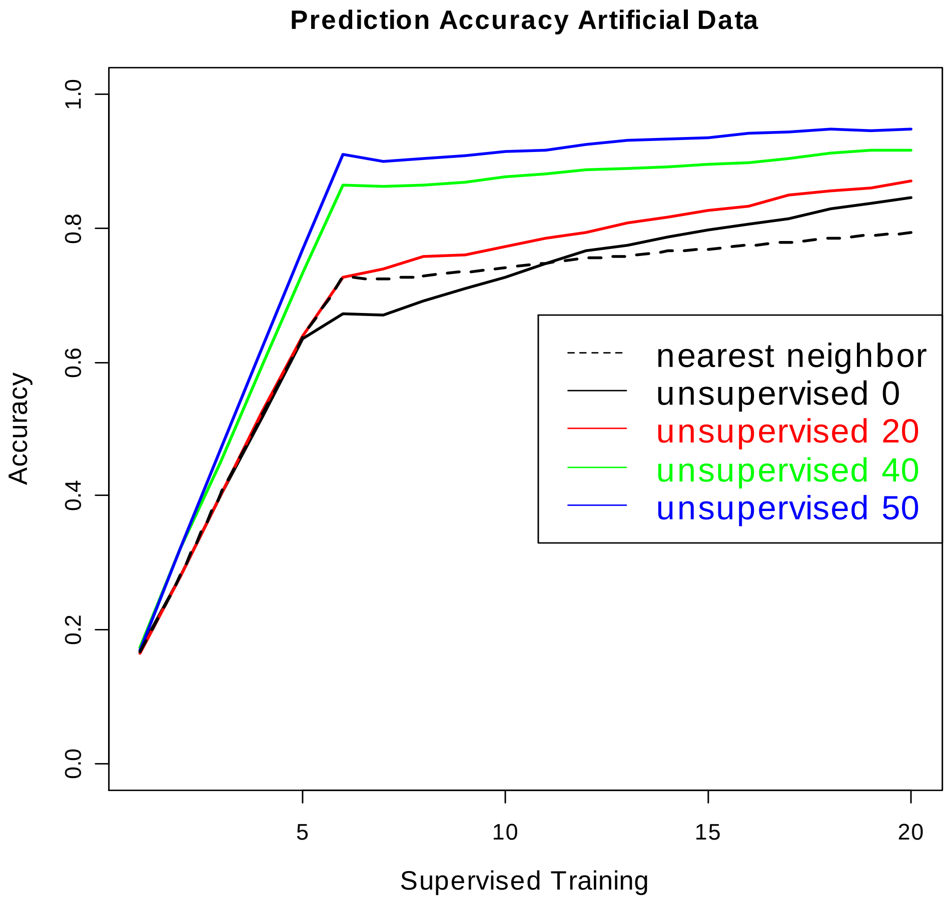
Supervised learning after 20 instances



Unsupervised learning of 30 and then supervised learning of 20



Result



Prediction Accuracy Artificial Data
Accuracy
0.0
0.2
0.4
0.6
0.8
1.0
5
10
15
20
Supervised Training
nearest neighbor
unsupervised 0
unsupervised 20
unsupervised 40
unsupervised 50

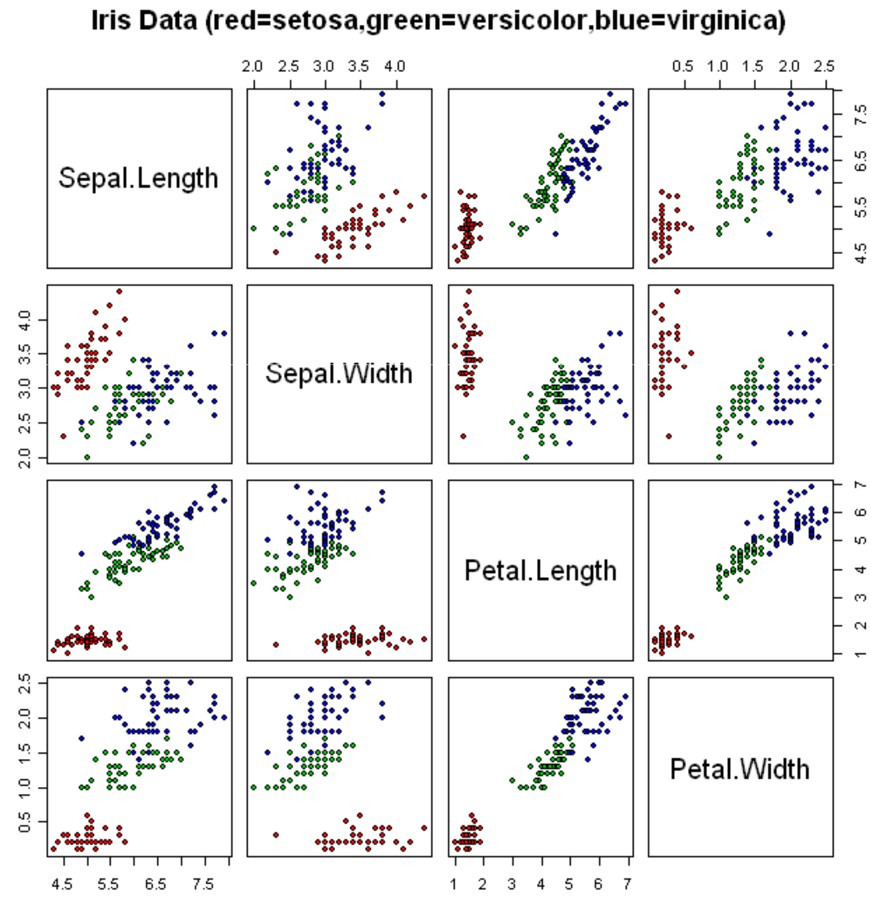
Artificial Data Evaluations

Purpose	Evaluation	Result
Unsupervised Clustering	Qualitative	+
Compare clustering with instance based learning	Prediction Accuracy	+
Compare different degrees of prior unsupervised learning	Prediction Accuracy	+

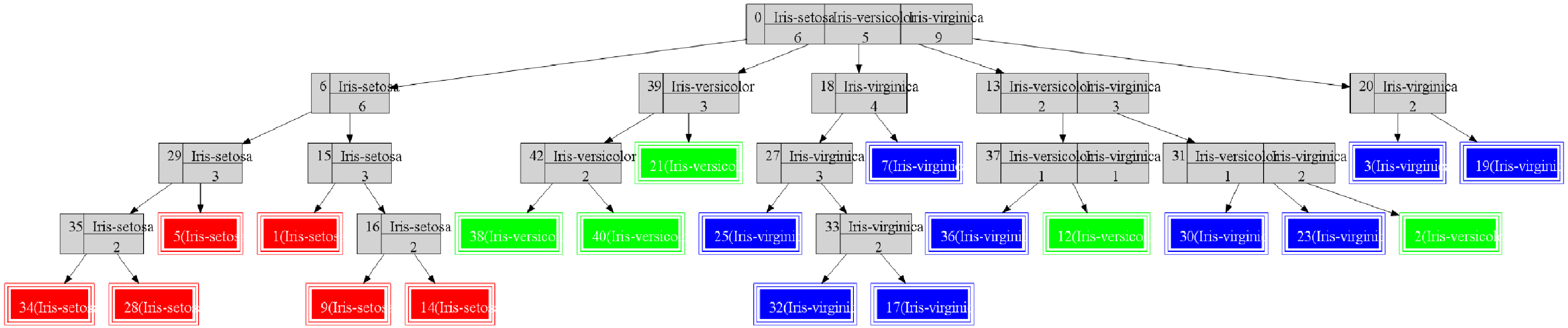
- High dimension symbolic vector

Iris Data Set

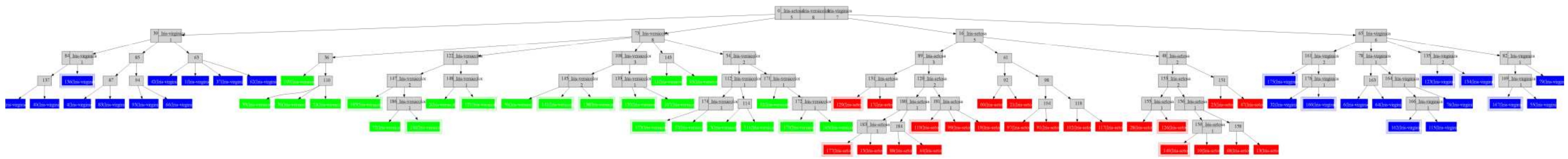
Fisher, R.A. (1936)



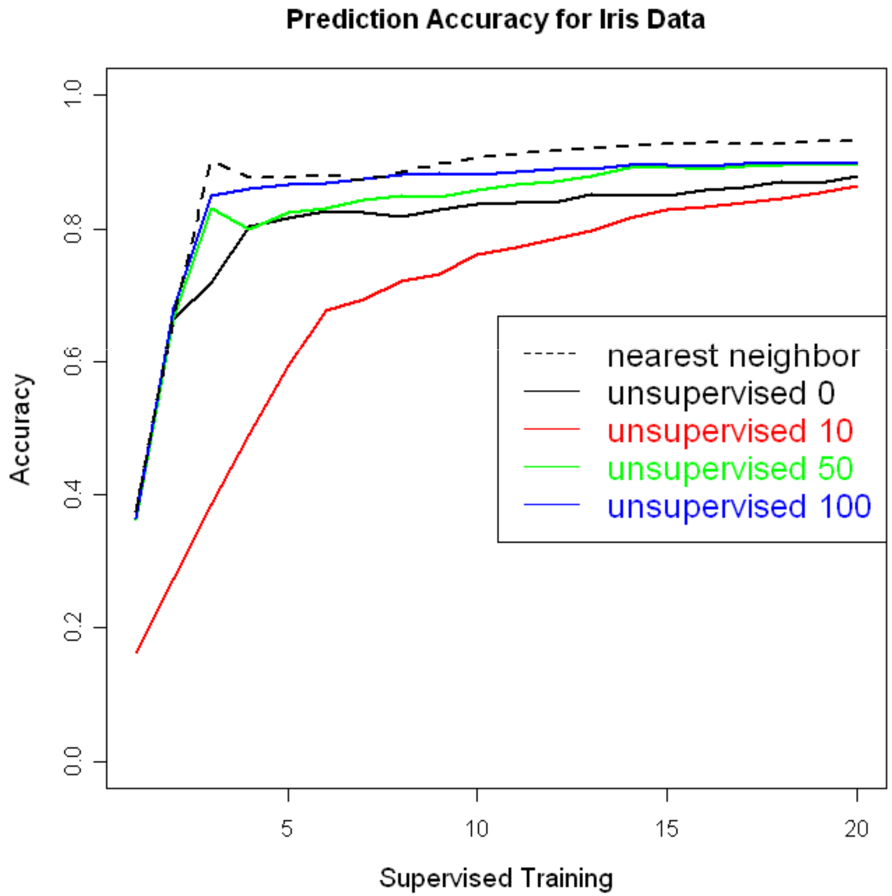
Supervised learning after 20 instances



Unsupervised learning of 50 and then supervised learning of 20



Result



Prediction Accuracy for Iris Data
Accuracy
0.0
0.2
0.4
0.6
0.8
1.0
5
10
15
20
Supervised Training
nearest neighbor
unsupervised 0
unsupervised 10
unsupervised 50
unsupervised 100

Iris Data Evaluations

Purpose	Evaluation	Result
Unsupervised Clustering	Qualitative	+
Compare clustering with instance based learning	Prediction Accuracy	-
Compare different degrees of prior unsupervised learning	Prediction Accuracy	+

- Low dimension numeric vector

Letter Recognition Data

David J. Slate (1991)

Letter Recognition Data

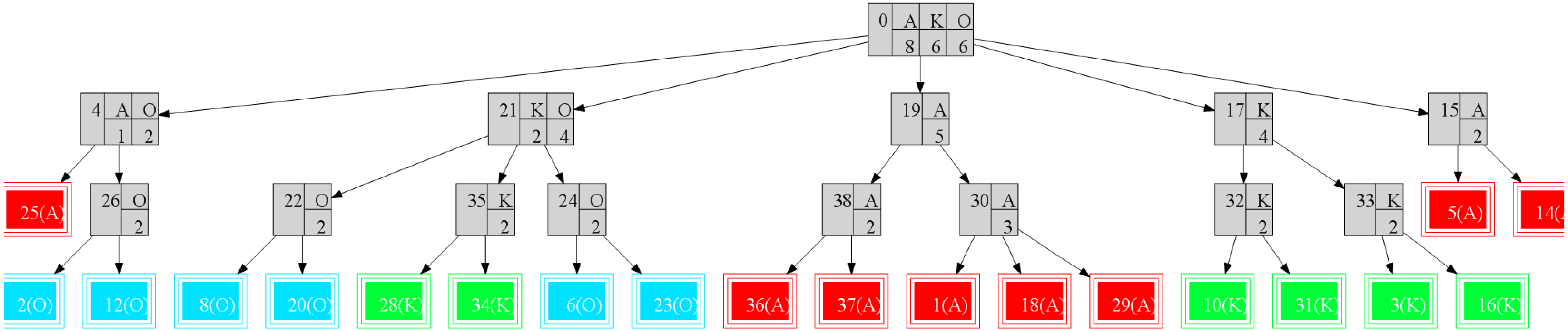
David J. Slate (1991)

```
Number of Instances: 20000

Number of Attributes: 17 (Letter category and 16 numeric features)

Attribute Information:
      1.     lettr   capital letter  (26 values from A to Z)
      2.     x-box   horizontal position of box      (integer)
      3.     y-box   vertical position of box        (integer)
      4.     width   width of box                    (integer)
      5.     high    height of box                   (integer)
      6.     onpix   total # on pixels               (integer)
      7.     x-bar   mean x of on pixels in box      (integer)
      8.     y-bar   mean y of on pixels in box      (integer)
      9.     x2bar   mean x variance                 (integer)
     10.     y2bar   mean y variance                 (integer)
     11.     xybar   mean x y correlation            (integer)
     12.     x2ybr   mean of x * x * y               (integer)
     13.     xy2br   mean of x * y * y               (integer)
     14.     x-ege   mean edge count left to right   (integer)
     15.     xegvy   correlation of x-ege with y     (integer)
     16.     y-ege   mean edge count bottom to top   (integer)
     17.     yegvx   correlation of y-ege with x     (integer)
```

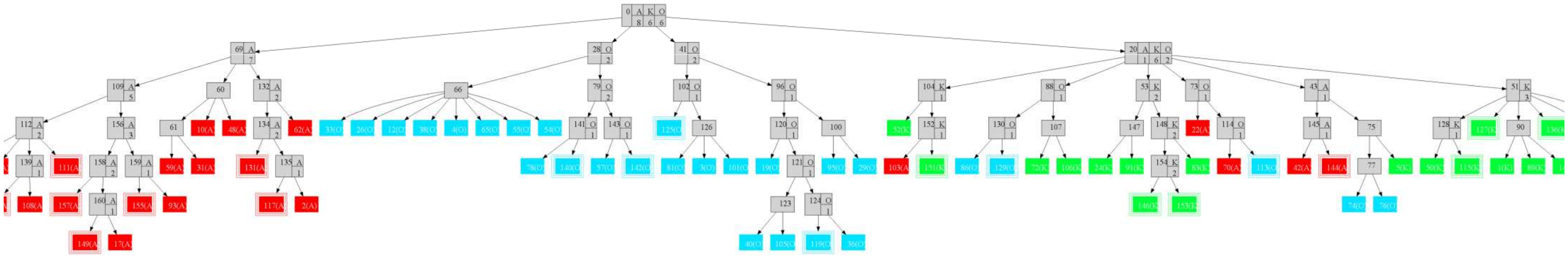
Easy Set and Difficult Set

- Difficult to test on entire data set
  - 26 classes
  - Diverse situations
  - Current implementation is not fast enough
- Tested on subpart of the data
  - Easy Set
    - A K O
  - Difficult Set
    - K R X

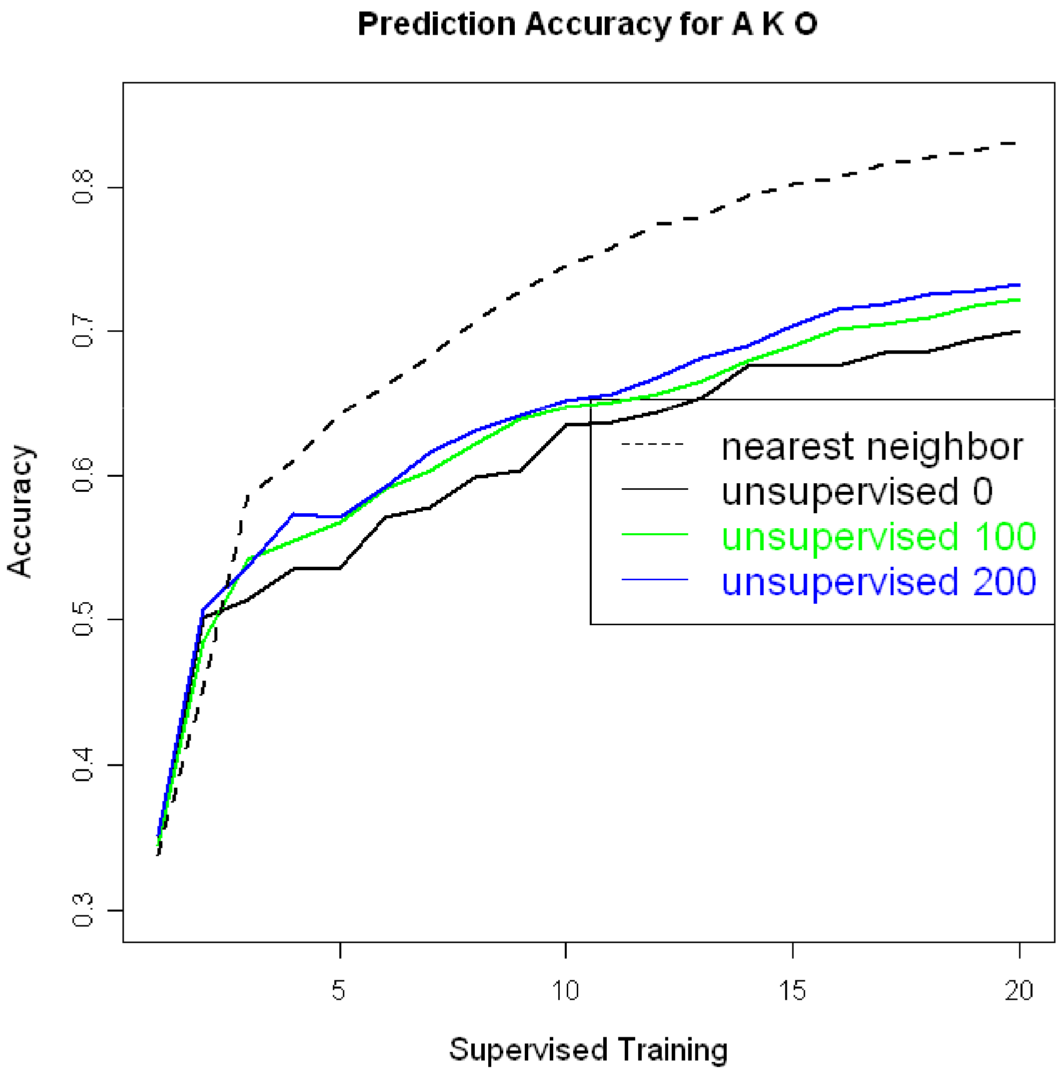
Supervised learning after 20 instances
Easy Set – A K O



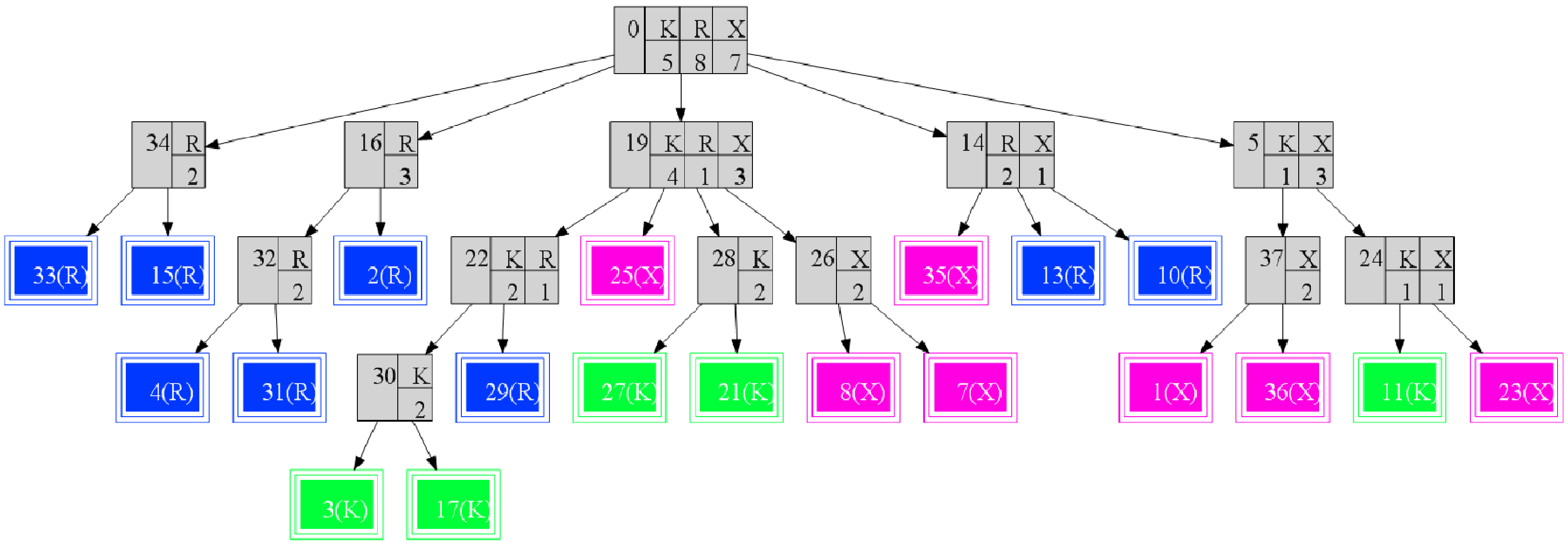
Unsupervised learning of 50 and then supervised learning of 20
Easy Set – A K O



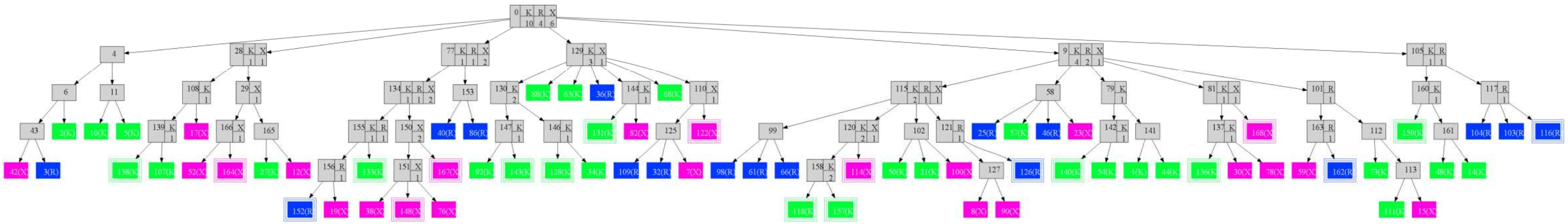
Result for easy set – A K O



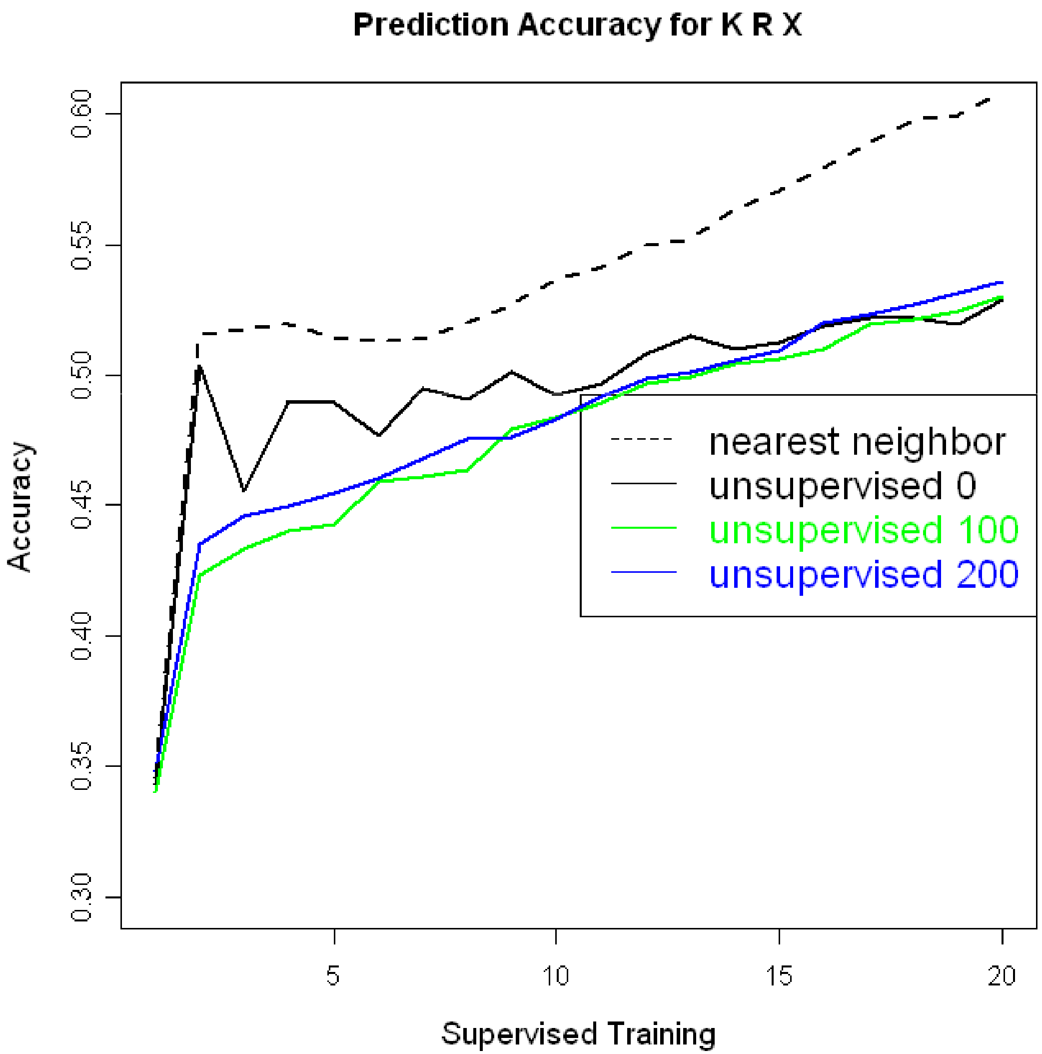
Supervised learning after 20 instances
Difficult Set – K R X



Unsupervised learning of 50 and then supervised learning of 20 Difficult Set – K R X



Result for difficult set – K R X



Letter Recognition Data Evaluations

Purpose	Evaluation	Result
Unsupervised Clustering	Qualitative	+
Compare clustering with instance based learning	Prediction Accuracy	?
Compare different degrees of prior unsupervised learning	Prediction Accuracy	?

- High dimension numeric vector

Conclusions

- Clustering is useful for filtering out ‘noisy’ features
  - Positive: Artificial data set
  - Negative: Iris data set.
- Quality of passive clustering directly depends on input features (slave of features)
  - Positive: All except K R X
  - Negative: K R X

Future Directions

- Adaptive feature selection
  - Generate and selection features
  - Clustering as guidance of feature selection
- Richer representation
  - Vector
  - Relational graph
  - Image
- Integration with Soar-RL
  - Provide abstract representation for symbolic TD learning

Nuggets and Coal

- Nuggets
  - Concept learning from subsymbolic input
  - Combine unsupervised and supervised learning
- Coal
  - Need feature selection
  - Need more realistic evaluation domain