Automating Exploratory Data Analysis for Efficient Data Mining

Jonathan D. Becher
becher@accrue.com

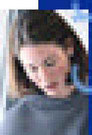
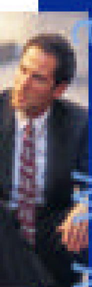
Pavel Berkhin
pavelb@accrue.com
Presenting Author

Edmund Freeman
eef@accrue.com

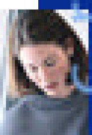
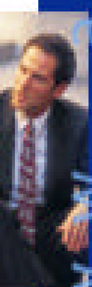
Accrue Software, Inc.
www.accrue.com
510-580-4500

Problem Space Overview

- Industrial issues
  - large data sets (Web data doubles every few months)
  - noisy data containing lots of missing/incorrect values
  - high cardinality categorical attributes
  - attributes irrelevant for particular mining purposes
- Goal
  - get model into production as quickly as possible
  - simplify, clean, and narrow the scope of data used
- Benefits
  - reduced CPU time for building a model
  - reduced CPU time for using a model
  - potentially increased model accuracy
  - increased explanatory power of the model

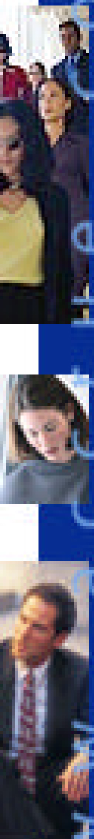




Exploratory Data Analysis (EDA)

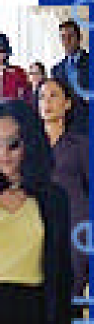
- Strategy
  - automate historically manual process
  - provide many tuning controls with intelligent defaults
  - focus on predictive problems
- Approach
  - identify inappropriate and suspicious attributes
  - select the most appropriate attribute encoding
  - create derived and transformed attributes
  - choose an optimal subset of attributes

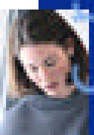
Inappropriate and Suspicious

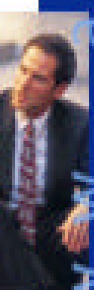


- Inappropriate: automatically excluded
  - *Constant*: only contains a single value
  - *Null*: has all Null (missing) values
  - *Near Null*: # Null values larger than threshold
  - *Many Values*: # unique values larger than threshold
- Suspicious: user determines if excluded
  - *Artifact*: association with target is greater than threshold
  - *Poor Predictor*: association is less than threshold
  - *Near Constant*: one value covers too many cases
  - *Few Values*: less distinct values than threshold
  - *Few Cases*: less distinct non-Null cases than threshold

Suspicious Attribute Example







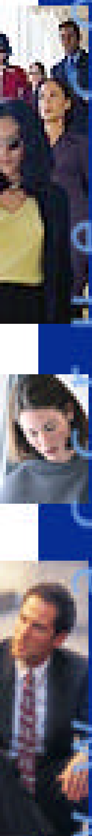
```
40. TopLevelDomain                          Categorical attribute
     Status: Near Constant

-Basic Stats - - - - - - - - - - - - - - - - - - - - - - -
  Number of distinct attribute values is        5
  Number of cases with non Null values is   22421 (100.000%)

-Association of original attribute with the target - - -
  InfoXT =  0.00237
  Chi2XT = 0.003041
  GoKrXT =        0

 Attribute values are:
 v1  =GOV                                   v2  =COM
 v3  =NET                                   v4  =ORG
 v5  =EDU

 Target values are:
 t1  =False  t2  =True
```

-Categorical attribute - Target distributions - - - - -

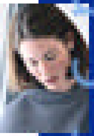
vVal	Cases	t1	t2	t1	t2	Total
v1	3	0.33	0.67	8e-5	1e-4	1e-4
v2	22225	0.50	0.50	1.00	0.99	0.99
v3	2	0.00	1.00	0.00	1e-4	8e-5
v4	6	0.50	0.50	2e-4	2e-4	2e-4
v5	185	0.20	0.80	3e-3	0.01	8e-3
Tot	22421	0.50	0.50			

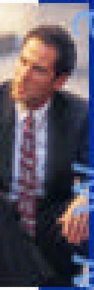
Encoding

- Determines most appropriate representation
  - continuous attributes are thresholded (discretize/quantize)
  - categorical attributes are grouped into smaller # of values
- Benefits
  - captures non-linear relationships (potentially more predictive)
  - increases explanatory power
- Issue
  - can cause fatal loss of some of detail in original attributes
- Solution: Target Dependency Analysis (TDA)
  - measures association b/ source and target before/after
  - optimizes reduction in association to find optimal encoding
    - thresholding: annealing, objective function is association measure
    - grouping: categorical clustering, minimize reduction in association



Encoded Attribute Example



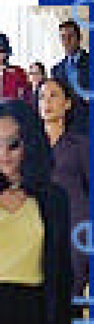
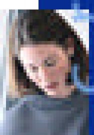


Age threshold 23.5 27.5 35.5 61.5 ; -- 4 cut points determined.

Education category { Doctorate Masters "Prof-school" } { Bachelors } { "Assoc-acdm" "Assoc-voc" "HS-grad" "Some-college" } { "10th" "11th" "12th" "1st-4th "5th-6th" "7th-8th" "9th" Preschool } ; -- 4 groups determined.

Sex category; -- No grouping (2 original values).

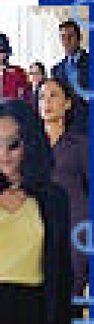
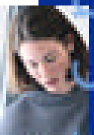
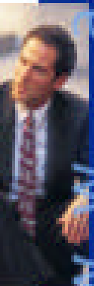
Figure 2. Optimized encodings determined by the EDA module


Target Dependency Analysis

- **Three available measures**
  - source attribute **X** with values **j=1:J**
  - target **Y** with values **q=1:Q**
  - joint distribution $\mathbf{P_{jq}}$, and marginal distributions $\mathbf{P_{j.}}$, $\mathbf{P_{.q}}$,

- **Mutual Information**
  - $\mathbf{I(X,Y) = \Sigma_{jq}\ P_{jq} \log(P_{jq} / P_{j.}P_{.q}) = H(Y) - H(Y|X)}$
  - $\mathbf{M(X,Y) = I(X,Y)/H(Y)}$ - normalized
- **Chi-squared (*Cramer's V*)**
  - $c^2\mathbf{(X,Y)} = \mathbf{N}\ S_{\mathbf{jq}}\ \mathbf{(P_{jq} - P_{j.}P_{.q})^2 / (P_{j.}P_{.q})}$
  - $\mathbf{V(X,Y)} = c^2\mathbf{(X,Y)/(N\ (min(Q,J)-1))}$ - normalized
- **Goodman-Kruskal**
  - Trivial classifier forecasts most frequent target value
  - Instead, choose most frequent forecast among all cases with the same **X** -value **j** (maximum likelihood forecast)
  - Goodman-Kruskal index is difference in error rate between trivial and X conditioned forecasters.

Encoded Attribute Example

3. Education Categorical attribute

-Basic Stats -

Number of distinct attribute values is 16

Number of cases with non Null values is 48842 (100.000%)

-Association of original attribute with the target - - -

InfoXT = 0.116

Chi2XT = 0.1339

GoKrXT = 0.07949

***Target values are*: t1 =<=50K t2 =>50K**

***Selected grouping* consists of:**

jVal	Cases	NumVal / Values covered				
j1	4085	3	Doctorate	Masters	Prof-school	
j2	8025	1	Bachelors			
j3	30324	4	Assoc-acdm	Assoc-voc	HS-grad	Some-college
j4	6408	8	10th	11th	12th	1st-4th
			5th-6th	7th-8th	9th	Preschool

-Association of discretized attribute with the target- -

InfoYT = 0.1097

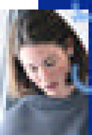
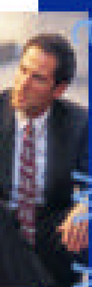
Chi2YT = 0.1269

GoKrYT = 0.07949

Transformations

- Univariate
  - traditionally only benefit to continuous targets (regression)
  - correlation extended to continuous/categorical pairs
    - $y(x) = v*x^2 + (1-v)*x$,
    - $y(x) = 1/x$,
    - $y(x) = \exp(v*x)$,
    - $y(x) = \log(x)$,
    - $y(x) = x^v$,
    - $y(x) = (|x|+x)/2$
    - $y(x) = |x|$

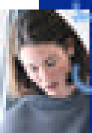
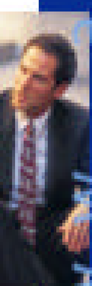
- Multivariate
  - useful for both classification and regression
  - functions of several continuous attributes, including linear combinations with undefined coefficients, ratios and products.

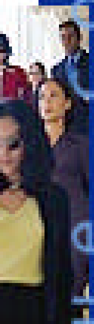
Selection: Markov Blanket (MB)

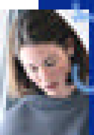
- Theoretical basis
  - expectation of Kullback Leibler (**KL**) distance between target distribution $\mathbf{P(Y=q|X_1=j_1,..X_k=j_k)}$, conditioned by joint distribution of all $\mathbf{k}$ source attributes, and target distribution conditioned by $\mathbf{s}$ selected attributes $\mathbf{X_1,..X_s}$, $\mathbf{P(Y=q|X_1=j_1,..X_s=j_s), s<k,}$

    $$d(\mathbf{X_{1:k},X_{1:s}}) = \Sigma_{\mathbf{j1,..,jk}} \mathbf{P_{j1,..,jk}\, KL(P_{q|j1,..,jk}} \,||\, \mathbf{P_{q|j1,..,js})},$$
    $$\mathbf{KL(P_q||R_q)} = \Sigma_{\mathbf{q}} \mathbf{P_q \log (P_q/R_q)}.$$
- Practice
  - computational feasibility: low dimensional blankets
  - attribute $\mathbf{X_0}$ associated w/ other attributes or blanket $\mathbf{X_{1:b}}$
  - if $d(\mathbf{X_{0:b},X_{1:b}})$ is small, $\mathbf{X_0}$ is well covered by its blanket and is a good candidate for exclusion.
- Implementation details
  - choice of the original blankets
  - exclusion criterion/schedule
  - recomputation of Markov blankets

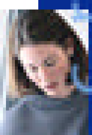




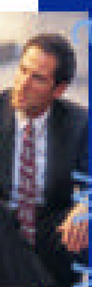
Selection: Inconsistency Rate (IR)





- Theoretical basis
  - generalization of Goodman-Kruskal previously described
- Practice
  - error rate of a trivial classifier which predicts the majority target outcome on each subset $X_1=j_1,..X_k=j_k$.
  - If omission of a certain attribute does not affect IR, error rate of this classifier remains intact without this attribute, and the attribute is a good candidate for exclusion.
- Implementation details
  - Backward selection process
  - Forward steps

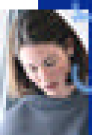
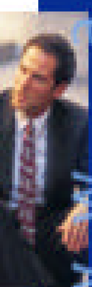




Attribute Selection Results

Table 1. Unoptimized Encoding vs EDA Encoding

		Unoptimized Encoding	EDA Encoding
No Selection	Attributes Used	593	593
	Training Time	151	130
	ROC	0.712	0.724
	Top 5% Lift	2.76	2.71
MB	Attributes Used	254	254
	Training Time	74	79
	ROC	0.729	0.742
	Top 5% Lift	3.28	3.39
IR	Attributes Used	16	16
	Training Time	44	40
	ROC	0.712	0.735
	Top 5% Lift	3.03	3.35

Conclusions

- EDA
  - historically manual preprocessing can be automated
  - native attribute representation can be improved
  - majority of attributes are unneeded by model

- Benefits to Data Mining
  - reduced CPU time for building a model
  - reduced CPU time for using a model
  - potentially increased model accuracy
  - increased explanatory power of the model

Note: all algorithms described are available commercially in Accrue Decision Series