Information theory methods for feature selection

Zuzana Reitermanová

Department of Computer Science
Faculty of Mathematics and Physics

Charles University in Prague, Czech Republic

Diplomový a doktorandský seminář I. – 11.11.2010

Outline

Introduction

Feature extraction

- An integral part of the data mining process.

Two steps

- Feature construction
- Feature selection

Introduction

Feature extraction

- An integral part of the data mining process.

Two steps

- Feature construction
  - Preprocessing techniques – standardization, normalization, discretization,...
  - Part of the model (ANN),...
  - Extraction of local features, signal enhancement,...
  - Space-embedding methods – PCA, MDS (Multidimensional scaling),...
  - Non-linear expansions
  - ...
- Feature selection

Feature selection

Why to employ feature selection techniques?

- ... to select relevant and informative features.
- ... to select features that are useful to build a good predictor

Moreover

- General data reduction – decrease storage requirements and increase algorithm speed
- Feature set reduction – save resources in the next round of data collection or during utilization
- Performance improvement – increase predictive accuracy
- Better data understanding
- ...

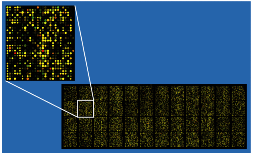
Advantage

- Selected features retain the original meanings.

Feature selection

Current challenges in Feature selection

- Unlabeled data
- Knowledge-oriented sparse learning
- Detection of feature dependencies / interaction
- Data-sets with a huge number of features (100 – 1000000) but relatively few instances ($\leq$ 1000)
  – microarrays, transaction logs, Web data,...



Feature selection

Current challenges in Feature selection

- Unlabeled data
- Knowledge-oriented sparse learning
- Detection of feature dependencies / interaction
- Data-sets with a huge number of features (100 – 1000000) but relatively few instances ($\leq$ 1000)
  – microarrays, transaction logs, Web data,...

NIPS 2003 challenge:

Dataset	Domain	Type	#Fe	%Pr	#Tr	#Val	#Te
ARCENE	Mass Spectrometry	Dense	10000	30	100	100	700
DEXTER	Text classification	Sparse	20000	50	300	300	2000
DOROTHEA	Drug discovery	Sparse binary	100000	50	800	350	800
GISETTE	Digit recognition	Dense	5000	30	6000	1000	6500
MADELON	Artificial	Dense	500	96	2000	600	1800

Feature selection

Basic approaches to Feature selection

- Filter models
  - Select features without optimizing the performance of a predictor
  - Feature ranking methods – provide a complete order of features using a relevance index
- Wrapper models
  - Use a predictor as a black box to score the feature subsets
- Embedded models
  - Feature selection is a part of the model training
- Hybrid approaches

Filter methods

Feature ranking methods

- Provide a complete order of features using a relevance index.
- Each feature is treated separately.

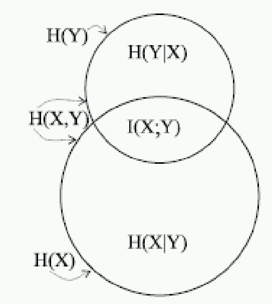
Many many various relevance indices

- Correlation coefficients – linear dependencies:
  Pearson: $R(i) = \frac{cov(X_i, Y)}{\sqrt{var(X_i)var(Y)}}$
  Estimate: $R(i) = \frac{\sum_k (x_k^i - \bar{x^i})(y_k - \bar{y})}{\sqrt{\sum_k (x_k^i - \bar{x^i})^2 \sum_k (y_k - \bar{y})^2}}$
  ...
- Classical test statistics – T-test, F-test, χ^2-test,...
- Single variable predictors (for example decision trees) – risk of overfitting
- Information theoretic ranking criteria – non-linear dependencies $\rightarrow$ ...

Relevance Measures Based on Information Theory

Mutual information

- (Shannon) Entropy:
  $H(X) = -\int_x p(x) log_2 p(x) dx$
- Conditional entropy: $H(Y|X) = \int_x p(x)(-\int_y p(y|x) log_2 p(y|x))dx$
- Mutual information:
  $MI(Y, X) = H(Y) - H(Y|X) = \int_x \int_y p(x, y) log_2 \frac{p(x,y)}{p(x)p(y)} dxdy$



Is MI for classification Bayes optimal?

- $\frac{H(Y|X)-1}{log_2 K} \leq e_{bayes}(X) \leq 0.5 * H(Y|X)$
- Kullback-Leibler divergence:
  $MI(X, Y) \simeq D_{KL}(p(x, y) \| p(y)p(x))$,
  where $D_{KL}(p_1 \| p_2) = \int_x p_1(x) log_2 \frac{p_1(x)}{p_2(x)} dx$

Relevance Measures Based on Information Theory

Mutual information

$MI(Y,X) = H(Y) - H(Y|X) = \int_x \int_y p(x,y) log_2 \frac{p(x,y)}{p(x)p(y)} dxdy$

Problem: $p(x), p(y), p(x,y)$ are unknown and hard to estimate from the data

Classification with nominal or discrete features

- The simplest case – we can estimate the probabilities from the frequency counts
- This introduces a negative bias
- Harder estimate with larger numbers of classes and feature values

Relevance Measures Based on Information Theory

Mutual information

$MI(Y,X) = H(Y) - H(Y|X) = \int_x \int_y p(x,y) log_2 \frac{p(x,y)}{p(x)p(y)} dxdy$

Problem: $p(x), p(y), p(x,y)$ are unknown and hard to estimate from the data

Classification with nominal or discrete features

- MI corresponds to the Information Gain (IG) for Decision trees
- Many modifications of IG (avoiding bias towards the multivalued features)
  - Information Gain Ratio $IGR(Y,X) = \frac{MI(Y,X)}{H(X)}$,
  - Gini-index, J-measure,....
- Relaxed entropy measures are more straightforward to estimate:
  - Renyi Entropy $H_\alpha(X) = \frac{1}{1-\alpha} log_2(\int_x p(x)^\alpha)dx$
  - Parzen window approach

Relevance Measures Based on Information Theory

Mutual information

$MI(Y,X) = H(Y) - H(Y|X) = \int_x \int_y p(x,y) log_2 \frac{p(x,y)}{p(x)p(y)} dxdy$

Problem: $p(x), p(y), p(x,y)$ are unknown and hard to estimate from the data

Regression with continous features

- The hardest case
- Possible solutions:
  - Histogram-based discretization:
    - MI is overestimated – depending on the quantization level
    - MI should be overestimated the same for all features
  - Approximation of the densities (Parzen window,...)
    - Normal distribution $\rightarrow$ correlation coefficient
    - Computational complexity
  - ...

Filter methods – Feature ranking methods

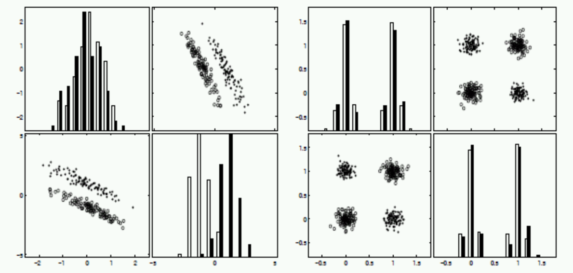
Advantages

- Simple and cheap methods, good empirical results.
- Fast and effective even in the case when the number of samples is smaller than the number of features.
- Can be used as preprocessing for more sophisticated methods.

Filter methods – Feature ranking methods

Limitations

- Which relevance index is the best?
- Select a redundand subset of features.
- A variable individually relevant may not be useful because of redundancies.
- A variable useless by itself can be useful together with others:



(a) (b)

Mutual information for multivariate feature selection

How to exclude both irrelevant and redundant features?

- Greedy selection of variables may not work well when there are dependencies among relevant variables.
- multivariate filter $MI(Y, \{X_1, ..., X_n\})$ is hard to approximate and compute
- $\rightarrow$ approximative MIFS algorithm and its variants:

MIFS algorithm

1. $X^* = argmax_{X \in A} MI(X, Y)$,
   $F \leftarrow \{X^*\}$, $A \leftarrow A \setminus X^*$
2. Repeat until $|F|$ is desired:
   $X^* = argmax_{X \in A}[MI(X, Y) - \beta \sum_{X' \in F} MI(X, X')]$,
   $F \leftarrow F \cup \{X^*\}$, $A \leftarrow A \setminus X$

Multivariate relevance criteria

Relief algorithms

- Based on the k-nearest neighbor algorithm.
- Relevance of features in the context of oders.
- Example of the ranking index (for multi-classification):
  $R(X) = \frac{\sum_i \sum_{k=1}^{K} |x_i - x_{M_k(i)}|}{\sum_i \sum_{k=1}^{K} |x_i - x_{H_k(i)}|}$, where
  $x_{M_k(i)}, k = 1, .., K$ K closest examples of the same class (nearest misses) in the original feature space
  $x_{H_k(i)}, k = 1, .., K$ K closest examples of a different class (nearest hits)
- Popular algorithm, low bias (NIPS 2003)

Wrapper methods

Multivariate feature selection

- Maximize the relevance of a subset of features $\bar{X}$: $R(Y, \bar{X})$
- Use a predictor to measure the relevance (i.e. accuracy).
  - A validation set must be used to achieve a useful estimate
  - K-fold cross-validation,...
  - A useful accuracy estimate on a separate testing set
- Employ a search strategy
  - Exhaustive search
  - Sequential search (growing/prunning),...
  - Stochastiic search (Simulated Annealing, GA,...)

Limitations

- Slower than the filter methods
- Tendency to overfitting – discrepancy between the evaluation score and the ultimate performance
- No valid good empirical results (NIPS 2003)
- High variance of the results

Embedded methods

- Feature selection depends on the predictive model (SVM, ANN, DT,...)
- Feature selection is a part of the model training
  - Forward selection methods
  - Backward elimination methods
  - Nested methods
  - Optimization of scaling factors over the compact interval $[0, 1]^n$ – regularization techniques

Advantages and limitations

- Slower than the filter methods
- Tendency to overfitting if not enough data is available
- Outperform filter methods if enough data is available
- High variance of the results

Ensemble learning

- Help the model-based (wrapper and embedded) methods
  - fast, greedy and unstable base learners (Decision trees, Neural networks,...)
- Robust variable selection
  - Improve feature set stability.
  - Improve stability generalization stability.

Parallel ensembles

- Variance reduction
- Bagging
  - Random forest,...

Serial ensembles

- Reduction of both bias and variance
- Boosting
  - Gradient tree boosting,...

Random forests for variable selection

Random forest (RF)

- Select a number $n \sim \sqrt{N}$, N is the number of variables.
- Each decision tree is trained on a bootstrap sample (about two-third of the training set).
- Each decision tree has maximal depth and it is not pruned.
- At each node, n variables are randomly chosen and the best split is considered on these variables.
- CART algorithm
- Grow trees until no more generalization improvement.

Random forests for variable selection

Variable importance measure for RF

- Compute an importance index for each variable and for each tree $M(X_i, T_j) = \sum_{t \in T_j} \triangle I_G(x_i, t)$,
  - $\triangle I_G(x_i, t)$ is the decrease of impurity due to an actual (or potential) split on variable x_i:
    $\triangle I_G(x_i, t) = I(t) - p_L I(t_L) - p_r I(t_R)$,
  - Impurity for regression: $I(t) = \frac{1}{N(t)} \sum_{s \in t} (y_s - \bar{y})^2$
  - Impurity for classification: $I(t) = Gini(t) = \sum_{y_i \neq y_j} {p_i}^t {p_j}^t$
- Compute the average importance of each variable over all trees: $M(x_i) = \frac{1}{N_T} \sum_{j=1}^{N_T} M(x_i, T_j)$
- Optimal number of features is selected by trying "cut-off points"

Random forests for variable selection

Advantages

- Avoid over-fitting in the case when there are more features than examples.
- More stable results.

NIPS 2003 Challenge results

- Top ranking challengers used a combination of filters and embedded methods.
- Very good results of methods using only filters, even simple correlation coefficients.
- Search strategies were generally unsophisticated.
- The winner was a combination of Bayesian neural networks and Dirichlet diffusion trees
- Ensemble methods (Random trees) were on the second and third position.

NIPS 2003 Challenge results

(a) **December 1st 2003 challenge results.**

Method (Team)	Score	BER	AUC	Fe	Pr
BayesNN-DFT (*Neal/Zhang*)	88.0	6.84 (1)	97.22 (1)	80.3	47.8
BayesNN-DFT (*Neal/Zhang*)	86.2	6.87 (2)	97.21 (2)	80.3	47.8
BayesNN-small (*Neal*)	68.7	8.20 (3)	96.12 (5)	4.7	2.9
BayesNN-large (*Neal*)	59.6	8.21 (4)	96.36 (3)	60.3	28.5
RF+RLSC (*Torkkola/Tuv*)	59.3	9.07 (7)	90.93 (29)	22.5	17.5
final2 (*Chen*)	52.0	9.31 (9)	90.69 (31)	24.9	12.0
SVMBased3 (*Zhili/Li*)	41.8	9.21 (8)	93.60 (16)	29.5	21.7
SVMBased4 (*Zhili/Li*)	41.1	9.40 (10)	93.41 (18)	29.5	21.7
final1 (*Chen*)	40.4	10.38 (23)	89.62 (34)	6.2	6.1
transSVM2 (*Zhili*)	36.0	9.60 (13)	93.21 (20)	29.5	21.7
BayesNN-E (*Neal*)	29.5	8.43 (5)	96.30 (4)	96.8	56.7
Collection2 (*Saffari*)	28.0	10.03 (20)	89.97 (32)	7.7	10.6
Collection1 (*Saffari*)	20.7	10.06 (21)	89.94 (33)	32.3	25.5

(b) **December 8th 2003 challenge results.**

Method (Team)	Score	BER	AUC	Fe	Pr
BayesNN-DFT (*Neal/Zhang*)	71.4	6.48 (1)	97.20 (1)	80.3	47.8
BayesNN-large (*Neal*)	66.3	7.27 (3)	96.98 (3)	60.3	28.5
BayesNN-small (*Neal*)	61.1	7.13 (2)	97.08 (2)	4.7	2.9
final_2-3 (*Chen*)	49.1	7.91 (8)	91.45 (25)	24.9	9.9
BayesNN-large (*Neal*)	49.1	7.83 (5)	96.78 (4)	60.3	28.5
final2-2 (*Chen*)	40.0	8.80 (17)	89.84 (29)	24.6	6.7
Ghostminer1 (*Ghostminer*)	37.1	7.89 (7)	92.11 (21)	80.6	36.1
RF+RLSC (*Torkkola/Tuv*)	35.4	8.04 (9)	91.96 (22)	22.4	17.5
Ghostminer2 (*Ghostminer*)	35.4	7.86 (6)	92.14 (20)	80.6	36.1
RF+RLSC (*Torkkola/Tuv*)	34.3	8.23 (12)	91.77 (23)	22.4	17.5
FS+SVM (*Lal*)	31.4	8.99 (19)	91.01 (27)	20.9	17.3
Ghostminer3 (*Ghostminer*)	26.3	8.24 (13)	91.76 (24)	80.6	36.1
CBAMethod3E (*CBA Group*)	21.1	8.14 (10)	96.62 (5)	12.8	0.1
CBAMethod3E (*CBA Group*)	21.1	8.14 (11)	96.62 (6)	12.8	0.1
Nameless (*Navot/Bachrach*)	12.0	7.78 (4)	96.43 (9)	32.3	16.2

NIPS 2003 Challenge results

Other (surprising) results

- Some of the top ranking challengers used almost all the probe features.
- Very good results for methods using only filters, even simple correlation coefficients.
- Non-linear classifiers outperformed the linear classifiers. They didn't overfit.
- The hyper-parameters are important. Several groups were using the same classifier (e.g. SVM) and reported significantly different results.

Conclusion

- Many different approaches to feature selection
- Best results obtained by hybrid methods

Advancing research

- Knowledge-based feature extraction
- Unsupervised feature extraction
- ...

References

- Guyon, I. M., Gunn, S. R., Nikravesh, M. and Zadeh, L., eds., **Feature Extraction, Foundations and Applications**. Springer, 2006.
- Huan Liu, Hiroshi Motoda, Rudy Setiono, Zheng Zhao, **Feature Selection: An Ever Evolving Frontier in Data Mining**, in JMLR: Workshop and Conference Proceedings, volume 10, pages 4–13, 2010
- Isabelle Guyon, André Elisseeff, **An Introduction to Variable and Feature Selection**, in JMLR: Workshop and Conference Proceedings, volume 3, pages 1157–1182, 2003
- Journal of Machine Learning Research, **http://jmlr.csail.mit.edu/**

References

- R. Battiti, **Using mutual information for selecting features in supervised neural net learning**, in: Neural Networks, volume 5(4), pages 537–550, July 1994.
- Kari Torkkola, **Feature extraction by non parametric mutual information maximization**, in: The Journal of Machine Learning Research, volume 3, pages 1415 – 1438, 2003.
- Francois Fleuret, **Fast Binary Feature Selection with Conditional Mutual Informationin**, in: The Journal of Machine Learning Research, volume 4, pages 1531 – 1555, 2004.
- Kraskov, Alexander; Stögbauer, Harald; Grassberger, Peter, **Estimating mutual information**, Physical Review E, volume 69, Issue 6, 16 pages, 2004