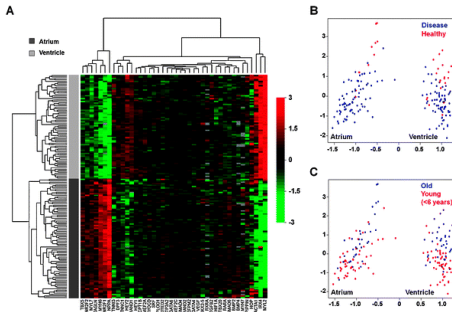


Feature Selection by Filters: A Unifying Perspective

Gavin Brown, University of Manchester

High-dimensional data : Gene expression

- high-throughput technology produces 1000s of measurements
- typical: $> 10,000$ features (columns), < 50 patients (rows)
- very very easy to overfit



High-dimensional data : Netflix

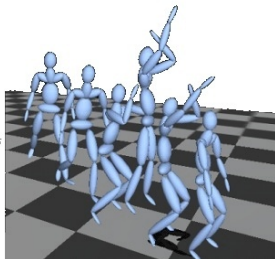
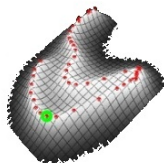
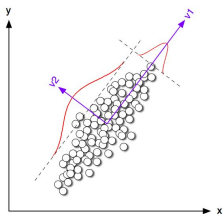
- 480,000 users rated 18,000 movies over several years
- $\approx 100,000,000$ ratings (i.e. sparse matrix)
- hugely expensive on time & memory



Plus: **Million dollar prize!**

Handling Hi-D data?

- Feature extraction - PCA, ICA, GP-LVM, etc
- $S = f(\Omega)$, usually $S = \mathbf{w}^T \Omega$.



Handling Hi-D data?

Feature **Extraction**

$$S = f(\Omega), \text{ usually } S = \mathbf{w}^T \Omega.$$

Feature **Selection**

$$S \subseteq \Omega.$$

With Feature Extraction, we lose the *meaning* of original features.

With Feature Selection, we identify meaningful *subsets* of originals.
Combinatorial optimization over space of possible feature subsets.

3 Ingredients of a FS algorithm

- 1 search algorithm (NP-hard problem)
- 2 objective (“score”) function
- 3 stopping criterion (e.g. pick k features)

3 Ingredients of a FS algorithm

- 1 search algorithm (NP-hard problem)
- 2 objective (“score”) function
- 3 stopping criterion (e.g. pick k features)

Where this talk is focused:

- Score functions based on mutual information are popular
- Numerous (heuristic) score functions published 1990-2009
- Main result of this work shows that almost all can be unified under a single perspective, derived from a single equation.

Outline

This talk:

- ① Basics of FS and Info Theoretic FS
- ② Key Component: Multivariate Mutual Information
- ③ A Natural Space of Feature Selection Algorithms
- ④ Conclusion (you already know it)

Feature Selection: Wrappers

Input: large feature set Ω

10 Identify candidate subset $S \subseteq \Omega$

20 While !stop_criterion()

 Evaluate **error of a classifier** using S .

 Adapt subset S .

30 Return S .

- Pros: excellent performance for the *chosen* classifier
- Cons: computationally and memory-intensive

Feature Selection: Filters

Input: large feature set Ω

10 Identify candidate subset $S \subseteq \Omega$

20 While !stop_criterion()

 Evaluate **utility function** J using S .

 Adapt subset S .

30 Return S .

- Pros: fast, provides generically useful feature set
- Cons: generally higher error than wrappers

Common Filter Criteria

Utility function J is some kind of statistical dependence measure.
Rank features by the value of J .

- Pearson's correlation coefficient.
- Chi-squared coefficient
- Bhattacharyya distance
- **Mutual Information**

Common Filter Criteria

Utility function J is some kind of statistical dependence measure.
Rank features by the value of J .

- Pearson's correlation coefficient.
- Chi-squared coefficient
- Bhattacharyya distance
- **Mutual Information**

Pearson can only detect linear relationships.

Chi-squared and Bhattacharyya assumes Gaussian variables.

Common Filter Criteria

Utility function J is some kind of statistical dependence measure.
Rank features by the value of J .

- Pearson's correlation coefficient.
- Chi-squared coefficient
- Bhattacharyya distance
- **Mutual Information**

Pearson can only detect linear relationships.

Chi-squared and Bhattacharyya assumes Gaussian variables.

Mutual information is non-parametric, and can detect arbitrary nonlinear relationships between X and Y :-)

Mutual Information

$I(X; Y)$ - measure of dependence between feature X and target Y .

Zero when X is independent of Y .

Increases as X and Y become dependent.

Mutual Information

$I(X; Y)$ - measure of dependence between feature X and target Y .

Zero when X is independent of Y .

Increases as X and Y become dependent.

Defined as KL divergence:

$$I(X; Y) = \sum_X \sum_Y p(xy) \log \frac{p(xy)}{p(x)p(y)}$$

Filters using Mutual Information

Rank features $X_i, \forall i$ by their values of $J = I(X_i; Y)$.

Retain the highest ranked features, discard the lowest ranked.

i	$J(X_i)$
35	0.846
42	0.811
10	0.810
654	0.611
22	0.443
59	0.388
...	...
212	0.09
39	0.05

Cut-off point decided by user, e.g. $|S| = 5$, so
 $S = \{35, 42, 10, 654, 22\}$.

Filters using Mutual Information

Problem: Highly correlated features are no use!

Suppose, from the previous slide, $S = \{35, 42, 10, 654, 22\}$, but 42 and 10 are almost identical!

We need **relevant** features, but not **redundant** features.

Filters using Mutual Information

Problem: Highly correlated features are no use!

Suppose, from the previous slide, $S = \{35, 42, 10, 654, 22\}$, but 42 and 10 are almost identical!

We need **relevant** features, but not **redundant** features.

Solution: Penalize correlations.

$$J(X_n) = I(X_n; Y) - \sum_{X_i \in S} I(X_n; X_i)$$

(Batitti, IEEE TNN 1994)

Filters using Mutual Information

Another Solution: 'Max Relevance Min Redundancy'
(Peng et al, IEEE PAMI 2005)

$$J(X_n) = I(X_n; Y) - \frac{1}{|S|} \sum_{X_i \in S} I(X_n; X_i)$$

Smooths out noise by averaging across S ?

Filters using Mutual Information

Another Another Solution: ‘Joint Mutual Information’
(Yang & Moody, NIPS 1999)

$$J(X_n) = \sum_{X_i \in S} I(X_n X_i; Y)$$

“How useful is X_n when paired with each of the existing features?”

The Confusing Literature

Why should we use any of these? How do they relate?

<i>Criterion</i>	<i>Full name</i>	<i>Authors</i>
MIFS	Mutual Information Feature Selection	Batitti 1994
KS	Koller-Sahami metric	Koller & Sahami 1996
JMI	Joint Mutual Information	Yang & Moody 1999
MIFS-U	MIFS-'Uniform'	Kwak & Choi 2002
IF	Informative Fragments	Vidal-Naquet 2003
FCBF	Fast Correlation Based Filter	Yu et al 2004
CMIM	Conditional Mutual Info Maximisation	Fleuret 2004
MRMR	Max-Relevance Min-Redundancy	Peng et al 2005
ICAP	Interaction Capping	Jakulin 2005
CIFE	Conditional Infomax Feature Extraction	Lin et al 2006
DISR	Double Input Symmetrical Relevance	Meyer 2006
MINRED	Minimum Redundancy	Duch 2006
IGFS	Interaction Gain Feature Selection	Akadi 2008

Let's start from scratch

Define relevance of a feature **set** $S = \{X_1, X_2, \dots, X_n\}$ as

$$I(X_{1:n}; Y).$$

Overall we know we want to maximize this.

Let's start from scratch

Define relevance of a feature **set** $S = \{X_1, X_2, \dots, X_n\}$ as

$$I(X_{1:n}; Y).$$

Overall we know we want to maximize this.

We know that we need to:

- have large values of $I(X_i; Y)$
- have small(ish?) values of $I(X_j; X_i)$
- measure information properties of N variables

Let's start from scratch

Define relevance of a feature **set** $S = \{X_1, X_2, \dots, X_n\}$ as

$$I(X_{1:n}; Y).$$

Overall we know we want to maximize this.

We know that we need to:

- have large values of $I(X_i; Y)$
- have small(ish?) values of $I(X_j; X_i)$
- measure information properties of N variables

Shannon's Mutual Info $I(X_1; X_2)$ is a function of two variables.

Not able to measure properties of multiple (N) variables.

Background: Multivariate Information Theory

Shannon's Mutual Info $I(X_1; X_2)$ is a function of two variables.

Generalized by **Interaction Information** (McGill 1954),

$$I(\{X_1, X_2, X_3\}) = I(\{X_1, X_2\}|X_3) - I(\{X_1, X_2\})$$

- Takes set argument S
- Reduces to Shannon's case with $|S| = 2$
- Conditional form obtained by marginalising over X_3
- General form for arbitrary size S defined recursively

$$I(\{S \cup X\}) = I(S|X) - I(S)$$

Multi-Variable Interactions



$$I(\{X_1, X_2, X_3\}) = I(\{X_1, X_2\}|X_3) - I(\{X_1, X_2\})$$

$$I(\{bart, lisa, homer\}) = I(\{bart, lisa\}|homer) - I(\{bart, lisa\})$$

A Novel Expansion Theorem

Theorem 1

Given a set of input features $S = \{X_1, \dots, X_n\}$, and a target Y , their Shannon mutual information can be expanded as

$$I(X_{1:n}; Y) = \sum_{T \subseteq S} I(\{T \cup Y\}), \quad |T| \geq 1.$$

The Shannon Mutual Information between $X_{1:n}$ and Y expands into a sum of Interaction Information terms. Note that $\sum_{T \subseteq S}$ should be read, “sum over all possible subsets T drawn from S ”.

Proof

Brown, AI-STATS 2009

A Novel Expansion Theorem

Intuition

$$\begin{aligned} I(X_{1:n}; Y) = & \text{0th order interactions} \\ & + \text{1st order} \\ & + \text{2nd order} \\ & + \dots \\ & + \text{nth order} \end{aligned}$$

Example

$$\begin{aligned} I(X_{1:3}; Y) = & I(\{X_1, Y\}) + I(\{X_2, Y\}) + I(\{X_3, Y\}) \\ & + I(\{X_1, X_2, Y\}) + I(\{X_1, X_3, Y\}) + I(\{X_2, X_3, Y\}) \\ & + I(\{X_1, X_2, X_3, Y\}). \end{aligned}$$

Order-0 means X_i does not interact with other features.

Order-1 means X_i interacts with one other feature.

Order-2 means X_i interacts simultaneously with two other features.

A Novel Perspective

A more precise statement of the problem, we want to maximise:

$$J = I(X_{1:n}; Y) - I(X_{1:n-1}; Y)$$

That is the difference in information, **with** and **without** X_n .
Involves high-dimensional probabilities, $p(x_1, x_2, \dots, x_n, y)$.

A Novel Perspective

A more precise statement of the problem, we want to maximise:

$$J = I(X_{1:n}; Y) - I(X_{1:n-1}; Y)$$

That is the difference in information, **with** and **without** X_n .
Involves high-dimensional probabilities, $p(x_1, x_2, \dots, x_n, y)$.

Objective:

$$\begin{aligned} J &= I(X_{1:n}; Y) - I(X_{1:n-1}; Y) \\ &= \text{approx} - \text{approx} \\ &= I(X_n; Y) - \sum_{k=1}^{n-1} I(X_n; X_k) + \sum_{k=1}^{n-1} I(X_n; X_k | Y). \end{aligned}$$

Existing Filters can be Re-written

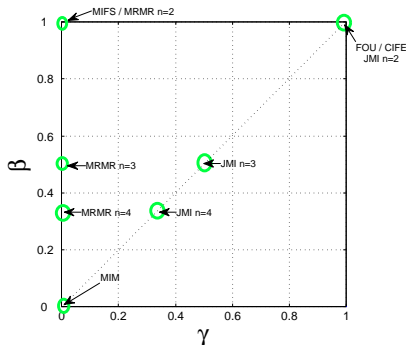
Joint Mutual Info, (Yang & Moody 1999)

$$J_{jmi} = \sum_{X_i \in S} I(X_n X_i; Y)$$

With some funky information calculus...

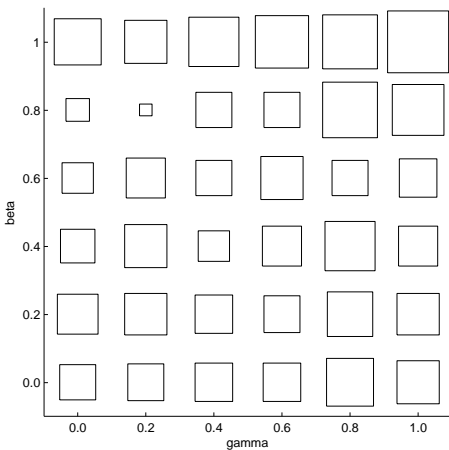
$$J_{jmi} = I(X_n; Y) - \frac{1}{n-1} \left\{ \sum_{k=1}^{n-1} I(X_n; X_k) - \sum_{k=1}^{n-1} I(X_n; X_k | Y) \right\}$$

A Space of Feature Filters



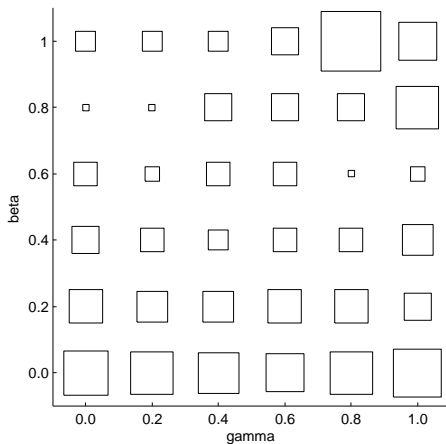
$$J = I(X_n; Y) - \beta \sum_{k=1}^{n-1} I(X_n; X_k) + \gamma \sum_{k=1}^{n-1} I(X_n; X_k | Y).$$

Exploring the Space



Lung Cancer data

Exploring the Space



SRBCT tumor data

Conclusion

Done:

- information theoretic clouds cleared, multivariate sunshine emerging.
- 12 separate methods united
- new space of novel criteria defined and explored

Ongoing work:

- Determining β/γ automatically from data
- Markov blanket methods
- links to ICA and information bottleneck
- Renyi entropy

References

A New Perspective for Information Theoretic Feature Selection, Gavin Brown. Intl Conf on Artificial Intelligence and Statistics, Florida April 2009.

An Information Theoretic Perspective on Multiple Classifier Systems, Gavin Brown. Intl Workshop on Multiple Classifier Systems, Iceland June 2009.

Multivariate Information Transmission, W. McGill, IEEE Trans. Information Theory, September 1954.

Feature Selection Literature

