NANYANG TECHNOLOGICAL UNIVERSITY

Identifying Simple Discriminatory Gene Vectors with An Information Theory Approach

presented by

Zheng Yun & Kwoh Chee Keong

Bioinformatics Research Center, School of Computer Engineering,
Nanyang Technological University
Nanyang Avenue, Singapore 639798

Tel: +65-67906613 Fax: +65-63162780

Email: pg04325488@ntu.edu.sg, asckkwoh@ntu.edu.sg

Supplements: http://www.ntu.edu.sg/home5/pg04325488/csb2005.htm

Outline

- Motivation of Feature Selection
- Current Feature Selection Methods
- Background Knowledge of Information Theory
- The Discrete Function Learning (DFL) Algorithm
- Experimental Results
- Discussions
- Conclusions

NANYANG TECHNOLOGICAL UNIVERSITY

Motivation of Feature Selection

- The curse of dimensionality
  - The run times of the inference algorithms
  - The number of samples
- There are irrelevant and redundant features in data sets.
  - The irrelevant and redundant features deteriorate the prediction performances.
- The principle of Occam's razor
  - According to this principle, a small subset of discriminatory features is more preferable to large number of features, if comparable prediction performances can be obtained.

NANYANG TECHNOLOGICAL UNIVERSITY

Categorization of Feature Selection Methods

- Single Feature Evaluation Methods

  A statistic is calculated, then a feature ranking list is provided in predefined order.

  Limitations: (i) Many *redundant* top features with similar gene expression patterns are selected to build the models, which introduces the risk of overfitting the training data sets.

  (ii) We do not know which gene is really relevant.

- Feature Subset Evaluation Methods

  A searching methods, like forward selection, is used to find optimal feature subset.

  Limitations: (i) Use *heuristic* scores, such as CFS (Correlation-based Feature Selection) and CSE (Consistency-based Subset Evaluation).

  (ii) Inefficient. The WSE (Wrapper Subset Evaluation) method chooses feature subset by performing cross validation for every feature subsets.

Related Work

- Fleuret (2004) and Vidal-Naquet and Ullman (2003), used the following criterion to choose feature subsets, X_i is good only if it carries information about Y, and if this information has not been caught by any of the $X_{(j)}$ already picked individually.

- Limitations: A new candidate feature is evaluated with respect to all selected features, one-by-one. However, it can not be known whether the existing features as a vector have captured the information carried by X_i or not. Redundant computation.

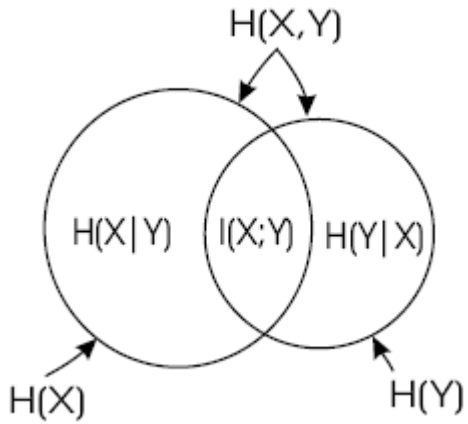
Background Knowledge of Information Theory

- Entropy, diversity of variable

$$H(X) = -\sum_{x} P(X = x) log P(X = x)$$

- Conditional Entropy, the remaining diversity

$$H(Y|\mathbf{X}) = -\sum_{\mathbf{x}} \sum_{y} p(\mathbf{x}, y) log p(y|\mathbf{x})$$



- Mutual Information, relation between vector or variables

$$I(\mathbf{X}; Y) = H(Y) - H(Y|\mathbf{X}) = H(\mathbf{X}) - H(\mathbf{X}|Y)$$

NANYANG TECHNOLOGICAL UNIVERSITY

Theoretical Foundation of Our Method

- The more variable, the information is provided about another variable, see Theorem 3.1 in the paper.

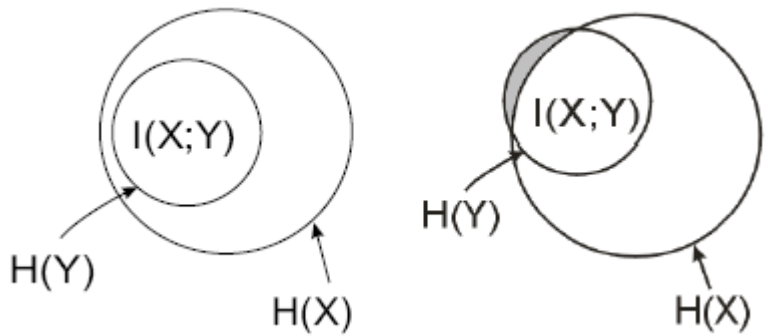
  Theorem 3.1 *$I(\{\mathbf{X}, Z\}; Y) \geq I(\mathbf{X}; Y)$, with equality if and only if $p(y|\mathbf{x}) = p(y|\mathbf{x}, z)$ for all $(\mathbf{x}, y, z)$ with $p(\mathbf{x}, y, z) > 0$.*

- To measure which subset of features is optimal.

  Theorem 3.2 *If the mutual information between $\mathbf{X}$ and Y is equal to the entropy of Y, i.e., $I(\mathbf{X}; Y) = H(Y)$, then Y is a function of $\mathbf{X}$.*

- The ϵ value method for noisy data sets.

  $$H(Y) - I(\mathbf{X}; Y) \leq \epsilon \times H(Y)$$



NANYANG TECHNOLOGICAL UNIVERSITY

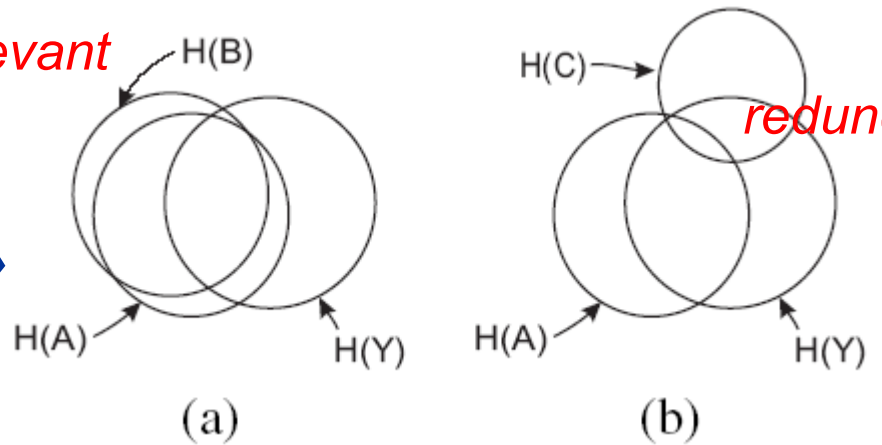
The Discrete Function Learning Algorithm

- Choosing feature subsets with the following criterion,

$$\begin{cases} X_{(1)} = & argmax_i I(X_i; Y), i = 1, \ldots, n \\ X_{(l)} = & argmax_Z I(\mathbf{X}, Z; Y), \end{cases}$$

Removing irrelevant features

Removing redundant features



(a) (b)

- Stopping the searching process based on Theorem 3.2, i.e., $I(\mathbf{X};Y) = H(Y)$ or $H(Y) - I(\mathbf{X};Y) <= \epsilon\, H(Y)$.

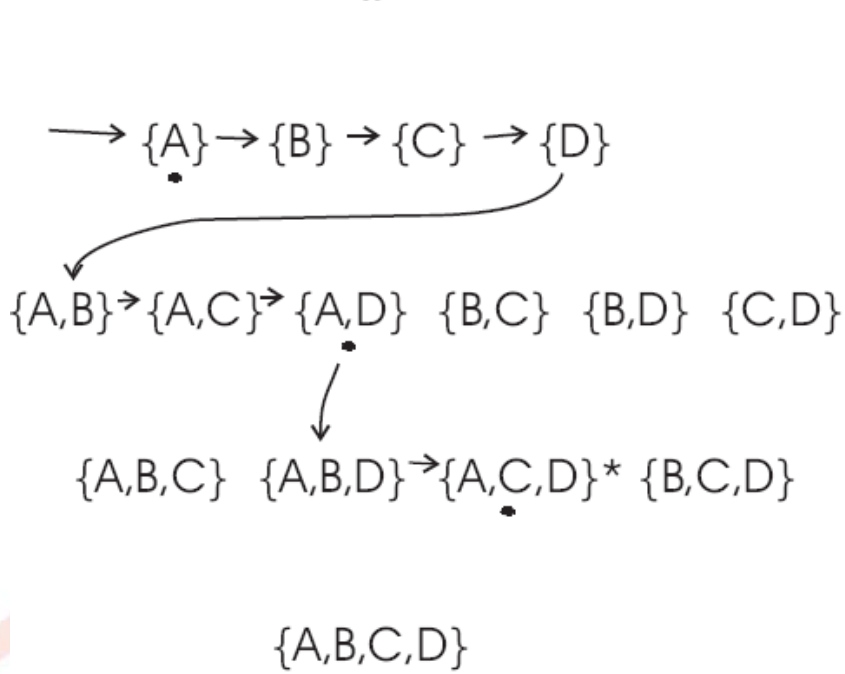
NANYANG TECHNOLOGICAL UNIVERSITY

The Discrete Function Learning Algorithm, Example

- Training data set, where $Y = (A \cdot C) + (A \cdot D)$

ABCD	*Y*	*ABCD*	*Y*	*ABCD*	*Y*	*ABCD*	*Y*
0000	0	0100	0	1000	0	1100	0
0001	0	0101	0	1001	1	1101	1
0010	0	0110	0	1010	1	1110	1
0011	0	0111	0	1011	1	1111	1

- The training process



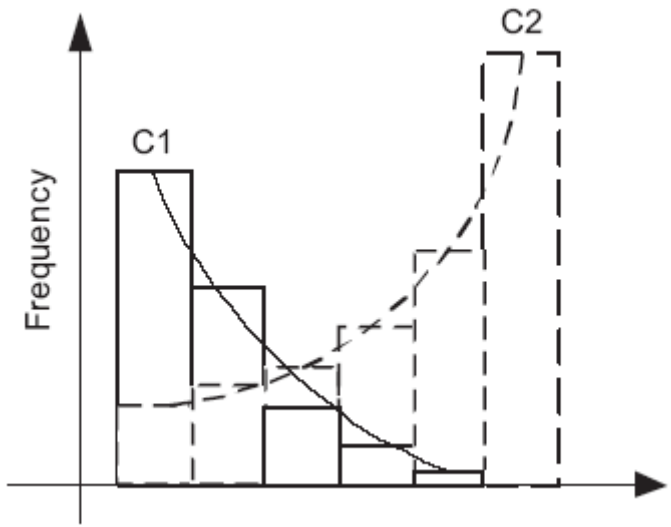
- The obtained classifier is the truth table of Y with counts of rules.

NANYANG TECHNOLOGICAL UNIVERSITY

Prediction Method

- Performing prediction with the weighted 1-Nearest-Neighbor algorithm based on Hamming distance.

  The count of rules in the training data sets is used to obtain statistically significant prediction.



NANYANG TECHNOLOGICAL UNIVERSITY

Data Sets

- Three gene expression profiles are chosen to validate our approach.

Data Set	Att.#	C.#	Trn.#	Tst.#	Lit.
ALL	7129	2	38	34	[11]
MLL	12582	3	57	15	[2]
DLBCL	7129	2	55	22	[30]

Discretization Results

- We use a widely used discretization method (Fayyad & Irani, 1993) based on entropy to discretize the selected data sets.
  - First, discretizing training data sets, then discretizing test data sets with the cutting points determined in the training data sets.
  - Some irrelevant features are assigned with one state value and can be removed from building classification models.

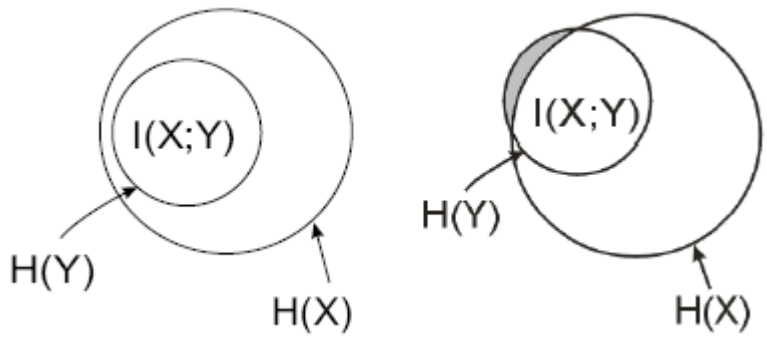
Data Set	Original #	# After Discret.	# Chosen by DFL (k^*)
ALL	7129	866	1
MLL	12582	4411	2
DLBCL	7129	761	1

The number of features with more than one values.

NANYANG TECHNOLOGICAL UNIVERSITY

Finding Optimal Models From Noisy Data Sets

- Gene expression data sets are noisy.
- For a given noisy data sets, the missing information of Y is determined.
- It is needed to find a good ϵ value, with which the DFL algorithm can find the optimal feature subsets.
  - We change the ϵ value from 0 to 0.6, with a step of 0.01. For each value, we train a model with the DFL algorithm, then validate its performance for the testing data sets.



NANYANG TECHNOLOGICAL UNIVERSITY

The Obtained Classifier

- The obtained classifiers are rules with their counts in the training data sets.
  - e.g., the classifier for the ALL data set is given below,

CST3	Class	Count
$(-\infty - 1419.5]$	ALL	27
$(1419.5 - \infty)$	AML	10
$(-\infty - 1419.5]$	AML	1

Prediction Details

CST3 (Cystatin C, M27891) is one of the 50 genes most highly correlated with the ALL-AML class distinction (Golub et al. 1999).

- The ALL data set

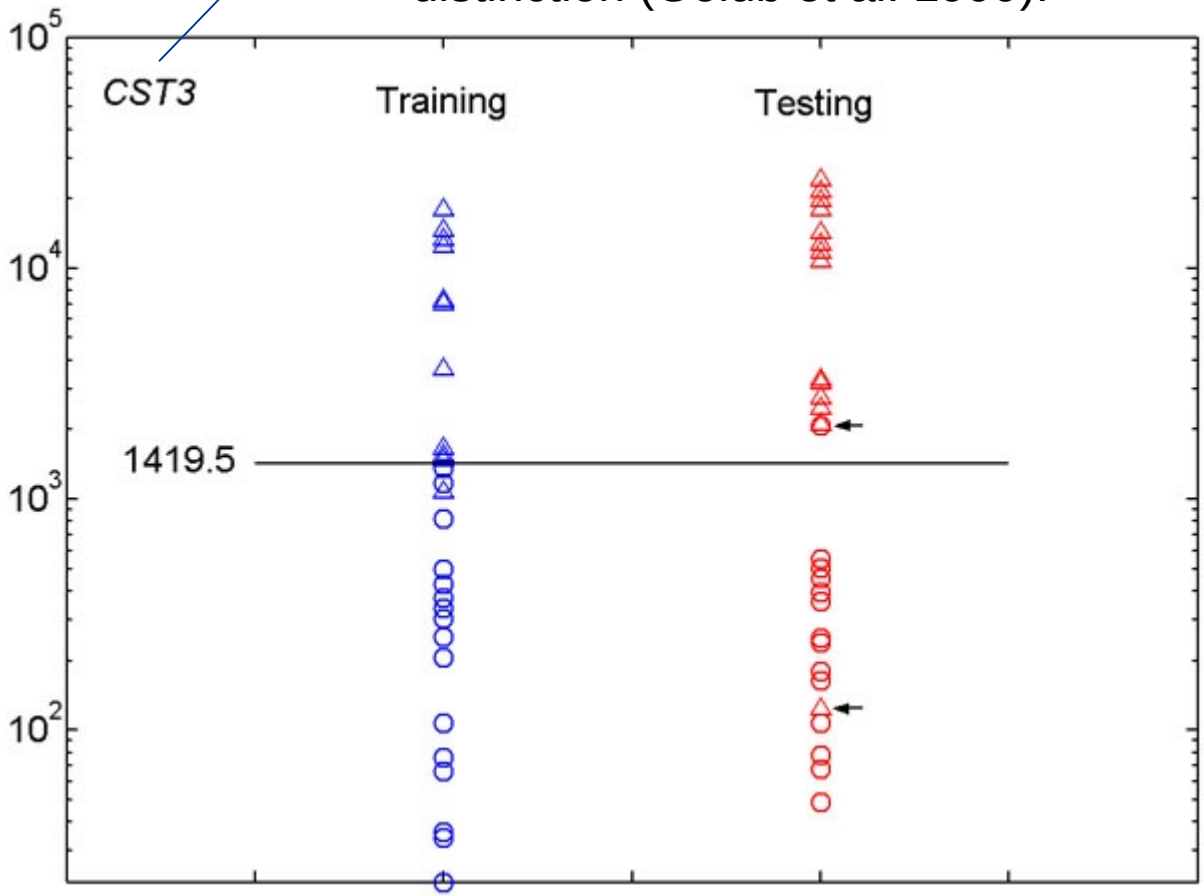
legend,

circles: ALL

triangles: AML

blue: training samples

red: testing samples

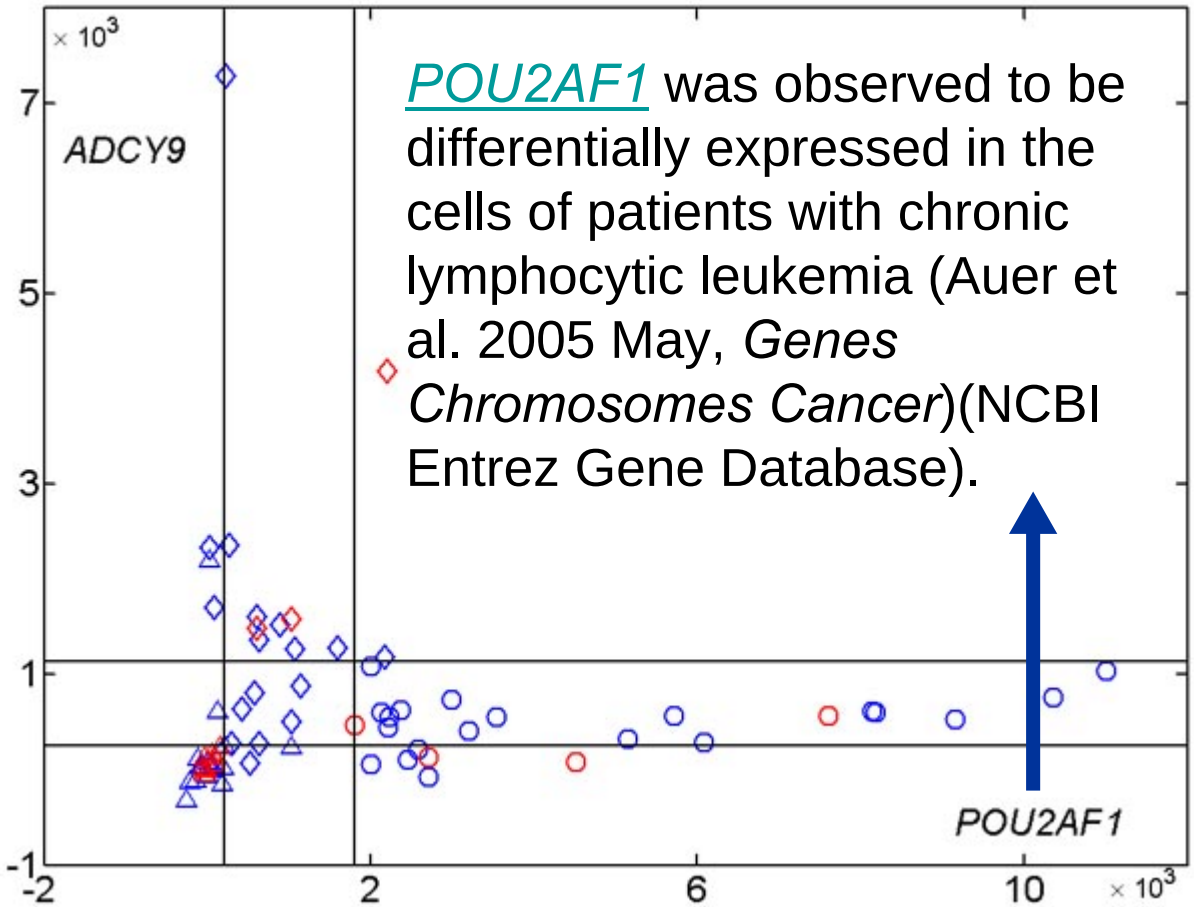


NANYANG TECHNOLOGICAL UNIVERSITY

Prediction Details (cont.)

- The MLL data set

legend,
circles: ALL
triangles: AML
diamonds: MLL
blue: training samples
red: testing samples



POU2AF1 was observed to be differentially expressed in the cells of patients with chronic lymphocytic leukemia (Auer et al. 2005 May, *Genes Chromosomes Cancer*)(NCBI Entrez Gene Database).

NANYANG TECHNOLOGICAL UNIVERSITY

Prediction Performances Compared With Other Methods

- The results of other compared algorithms are obtained with the *Weka* software.

Table 6. The comparison of prediction errors from the DFL algorithm and some well-known classification methods. The numbers shown are the incorrect predictions on discretized/continuous data sets.

Data Set	DFL[1]	C4.5	NB	1NN	kNN[2]	SVM
ALL	2	3/3	5/4	8/9	6/11	6/5
MLL	0	2/3	2/0	2/3	3/2	0/0
DLBCL	1	1/4	1/4	1/4	1/2	1/1
Average	1	2/3	3/3	4/5	3/5	2/2

NANYANG TECHNOLOGICAL UNIVERSITY

Prediction Performances Compared With Results in Literat.

- The column names E., Al., M. and k* stand for the number of incorrect predictions, the algorithm used, the relation measures used to do feature selection and the number of genes in the classifiers respectively.

	DFL	Methods in Literature				
Data Set	E.	E.	Al.	M.	k^*	Literature
ALL	2	5	WV	S2N	50	[11]
		2-4	SVM	S2N	1000	[10]
		3	EP	E	1	[21]
		0	kNN	MB	42	[34]
MLL	0	1	kNN[1]	S2N	40	[2]
		3	C45	χ^2	20	[19]
		1	SVM	χ^2	20	[19]
		1	kNN	χ^2	20	[19]
		0	PCL	χ^2	20	[19]
		0	NB	χ^2	20	[19]
DLBCL	1	NA				

NANYANG TECHNOLOGICAL UNIVERSITY

Comparison of Model Complexity

- The model from the C4.5 algorithm is comparable to our models (details available at the supplementary Table S4), but the performances of the C4.5 algorithm are not better than our method.
- The NB, 1NN, kNN and SVM algorithms build very complex models, using all genes of the data sets.
- The models in the literature are also more complex than the classifiers obtained by the DFL algorithm.

NANYANG TECHNOLOGICAL UNIVERSITY

Comparison of Efficiency

- Since all compared algorithms are implemented with the same Java language, and all experiments are performed on the same computer, the comparisons of their efficiency are meaningful.

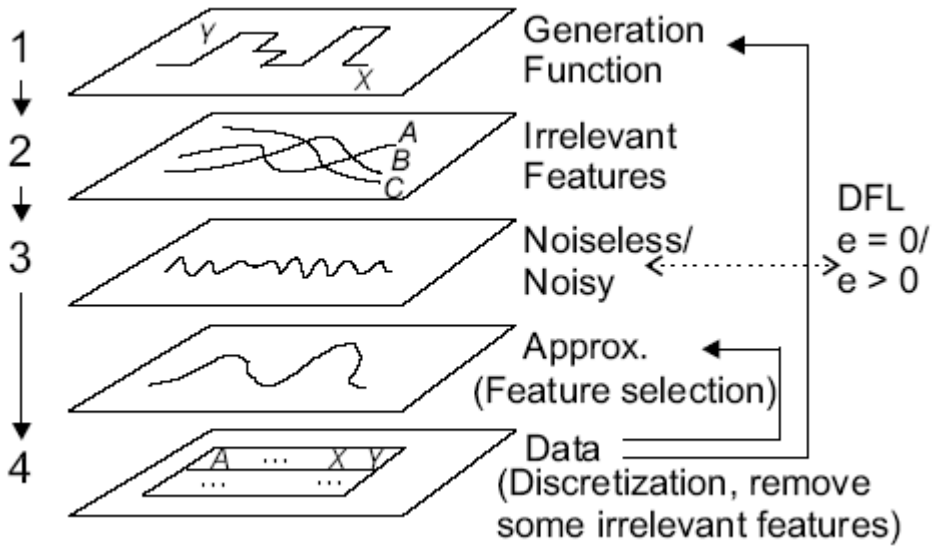
Table 8. The training times for discretized data sets of different classification methods. The unit is second.

Data Set	DFL	C4.5	NB	1NN	kNN	SVM
ALL	0.02	0.10	0.03	0.12	0.12	0.21
MLL	0.48	0.34	0.12	0.73	0.75	1.11
DLBCL	0.01	0.13	0.03	0.14	0.14	0.23

NANYANG TECHNOLOGICAL UNIVERSITY

Discussion: Comparison with Other Classification Methods

- The fundamental difference between the DFL algorithm and other classification methods lies in the underlying philosophy of the algorithms.



Other classification methods are trying to approximate the classification functions with complex models, like what have been done by the Multi-Layer Perceptrons and the SVMs with different kernels.

NANYANG TECHNOLOGICAL UNIVERSITY

Discussion: Comparison with Other Feature Selection Methods

- The stop criteria are different.
  - $I(\mathbf{X};Y) = H(Y)$ is used in the DFL algorithm
  - A predefined *k* is used in other methods
- Feature subset evaluation methods are different.
  - In DFL, new features are chosen with respect to the selected features as a vector.
  - In other methods, new features are chosen with respect to the selected features, one-by-one.
  - The existing methods are inefficient and less theoretically sound.
- The searching methods are different.
- Current methods based on information theory only deal with binary features.

NANYANG TECHNOLOGICAL UNIVERSITY

Conclusion

- The DFL algorithm can automatically find discriminatory feature subsets by using the criterion,

$$\begin{cases} X_{(1)} = & argmax_i I(X_i;Y), i = 1,\ldots,n \\ X_{(l)} = & argmax_Z I(\mathbf{X},Z;Y), \end{cases}$$

  The irrelevant features can be removed by maximizing $I(X_i;Y)$, while the redundant features can be removed by maximizing $I(\{\mathbf{X},Z\};Y)$.
- Optimal feature subsets $\boldsymbol{X}$ can be found by checking $I(\mathbf{X};Y) = H(Y)$.
- The DFL algorithm obtains comparable or more competitive prediction performances than those from literature and other well-known classification methods, however, with lower-complexity models.

NANYANG TECHNOLOGICAL UNIVERSITY

Acknowledgements

We thank Li Jinyan of Institute of Infocomm Research, Singapore, for his review on an early version of this paper.

We also thank the NTU-Compaq R&D research fund, for its partial support of this presentation.

NANYANG
TECHNOLOGICAL
UNIVERSITY

Questions and Suggestions

Thanks for your interests!

NANYANG TECHNOLOGICAL UNIVERSITY