Machine Learning in Financial Distress Prediction: A Review and Some Experimental Results

Chih-Fong Tsai
Assistant Professor
Department of Information Management
National Central University
cftsai@mgt.ncu.edu.tw

Presentation Outline

- Introduction
- Machine Learning
- Study 1: Reviews of Related Work
- Study 2: Feature Selection
- Study 3: Comparisons of different machine learning techniques
- Conclusion and Discussion

Introduction

- Background:

  第一年國科會研究計畫:支援向量機群於二分類財務金融決策支援之應用與研究 (NSC 94-2416-H-194-036-)

- Achievements: 2 SCI journals

Introduction

- 國科會3年期研究計畫: 應用機器學習技術建構最佳化財務預警模型 (NSC 96-2416-H-194-010-MY3) (2007/08/01~2010/07/31)
- Achievements: 2 SCI journals; 2 Book chapters; 4 SCI journal papers under review
- The scope of this presentation: the results first two years (2007/08/01~2009/07/31)

Introduction

- Enterprise risks:

  can come from uncertainty in financial markets, project failures, legal liabilities, credit risk, accidents, etc.
- Enterprise Risk management (ERM):

  includes the methods and processes to manage risks and minimize unfortunate events

Introduction

- The focus of this project: financial distress prediction (e.g. Asian and global financial crisis in 1997 and 2008)
- Incorrect decision-making in financial institutions is very likely to cause financial crises and distress.
- Two financial decision-making problems

  - bankruptcy prediction

  - credit scoring

Introduction

- Bankruptcy prediction and credit scoring are two-class classification/prediction problems.
- Financial institutions need an effective prediction model for:
- bankruptcy prediction – whether new customers (either individuals or companies) will go bankrupt (bankruptcy vs. non-bankruptcy)
- credit scoring – whether the new customers have good credit for issuing loans (good credit vs. bad credit)

Introduction

- This prediction model once developed can be regarded as a early warning system.
- To develop a bankruptcy prediction/credit scoring model, two types of different techniques have been used:

  - (traditional) statistical techniques, e.g. logistic regression, discriminant analysis, etc.

  - machine learning (data mining) techniques, e.g. artificial neural networks, decision trees, etc.

Introduction

- Related studies have shown that machine learning techniques can provide better performances than statistical ones.
- Particularly, the most widely used technique is artificial neural networks, trained by backpropagation learning algorithm, usually called multi-layer perceptron (MLP).

Machine Learning

- Pattern Classification:

  - Training/learning stage: the mapping between input-output examples

  - classification stage: classify an unknown instance into one of the learned class labels in the training set

- Single learning classifiers: e.g. artificial neural networks, decision trees, support vector machines, etc.

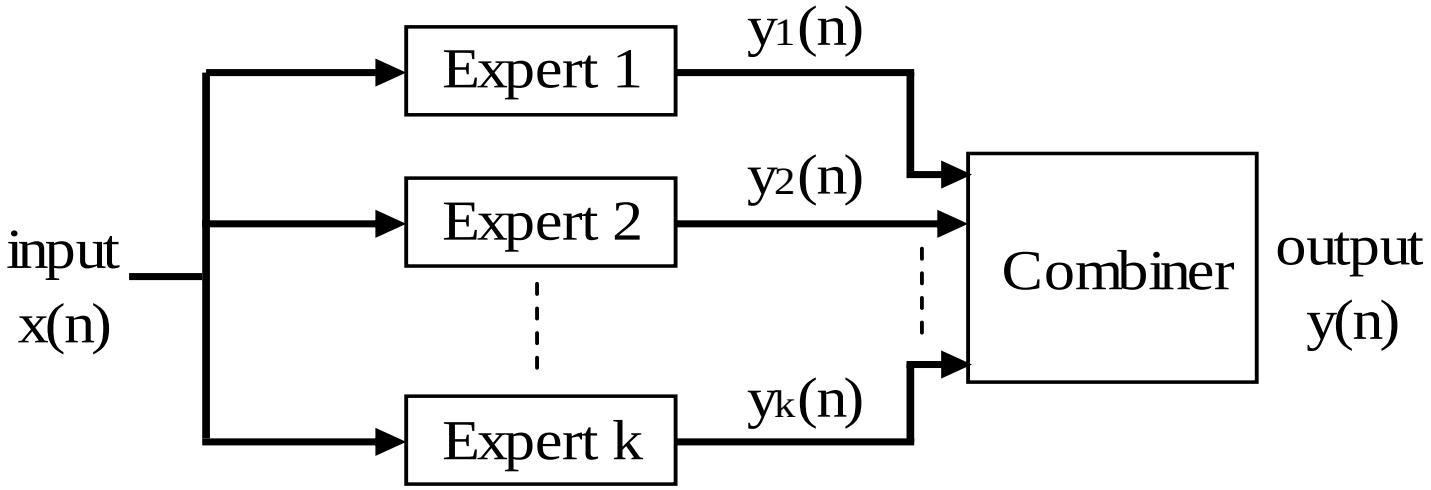
Machine Learning

- Advanced machine learning techniques:

  - classifier ensembles

  - hybrid classifiers: (a) cluster + classifier; (b) cascaded classifiers; (c) integrated classifiers

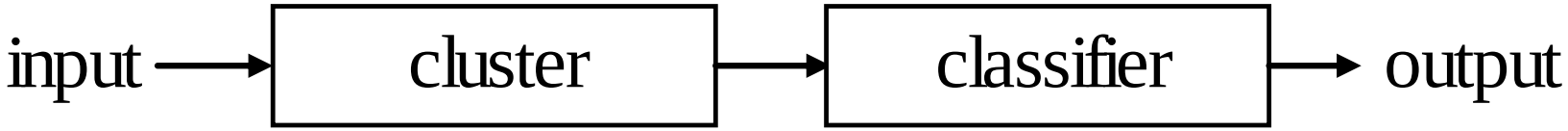
Classifier Ensembles



- The performance of classifier ensembles is usually better than the one of the best single classifier used in isolation
- The combination methods, e.g. (weighted) voting, boosting, bagging, etc.

Hybrid Classifiers

- Cluster + Classifier



- The clustering-based approach is used to pre-process the input samples in order to eliminate unrepresentative training examples from each class. Then, the clustering results are used as training examples for classifier design.

Hybrid Classifiers

- Cascading different classifiers, i.e. classifier + classifier
- The first level classifier is trained for a specific problem and the output of this classifier is the input for the second level classifier and so on.
- For example, the neuro-fuzzy approach

Hybrid Classifiers

- Integrating two different techniques
- For example, the first one aims at optimizing the learning performance (i.e. parameter tuning) of the second model for prediction, e.g GA-SVM.

Study 1

- Comparisons of Related Work:

  - types of classifier design

  - single classifiers

  - hybrid classifiers

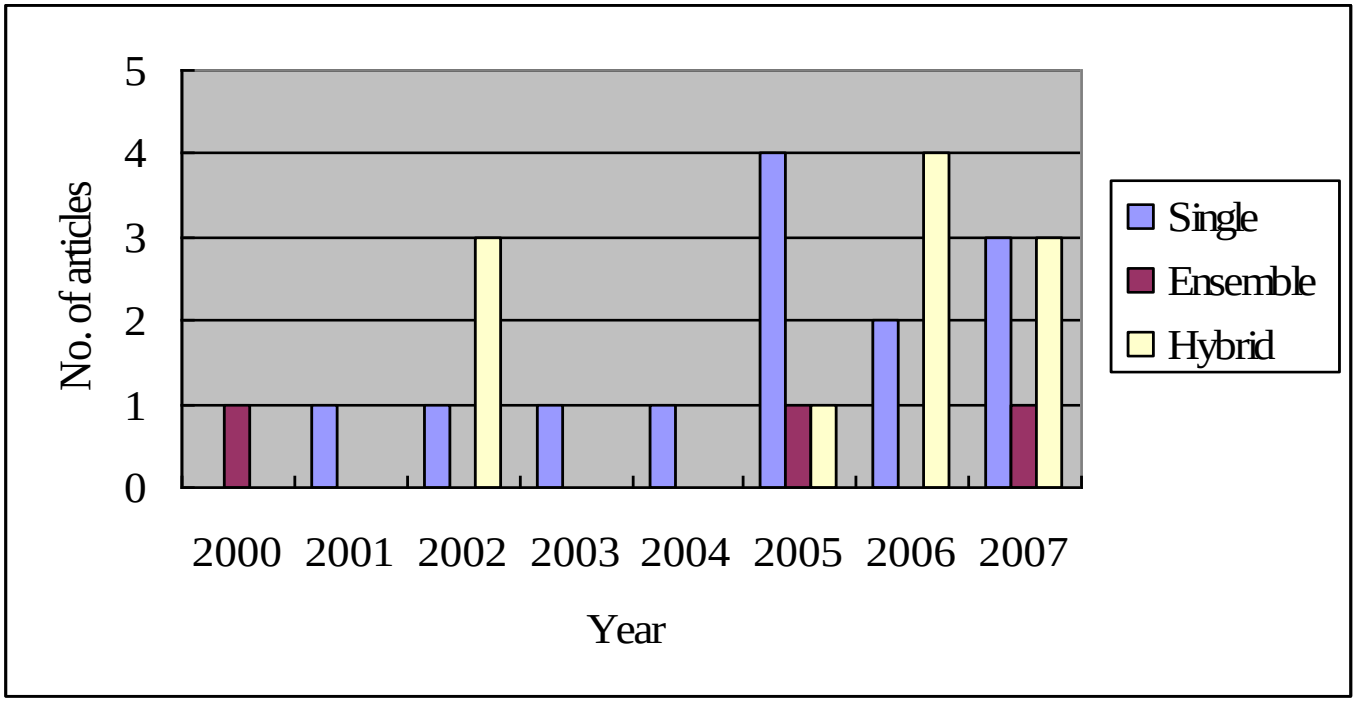
  - baseline classifiers

  - datasets, prediction accuracy, and feature selection, etc.

Comparisons of Related Work

- 27 Journal articles from 2000-2007 (Accomplishment of Year 1)

	Single	Ensemble	Hybrid
No. of articles	13	3	11



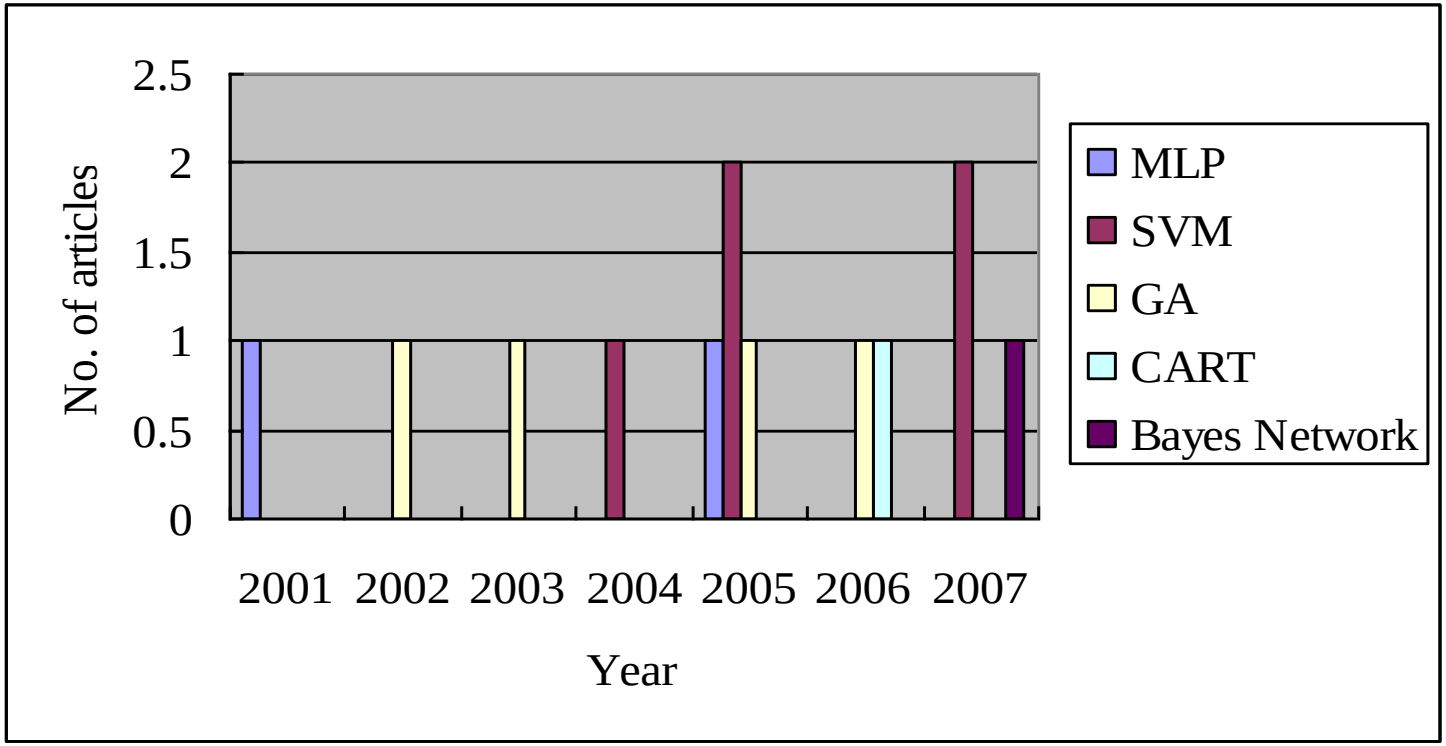
Comparisons of Related Work

- Few studies consider ensemble classifiers although it is likely outperform the single ‘best’ classifiers.
- On the other hand, hybrid classifiers are widely used to compare with single classifiers, especially in the year of 2006 and 2007.

Comparisons of Related Work

- Single classifiers:

	MLP	SVM	GA	Bayes Network	CART
No. of articles	2	5	4	1	1



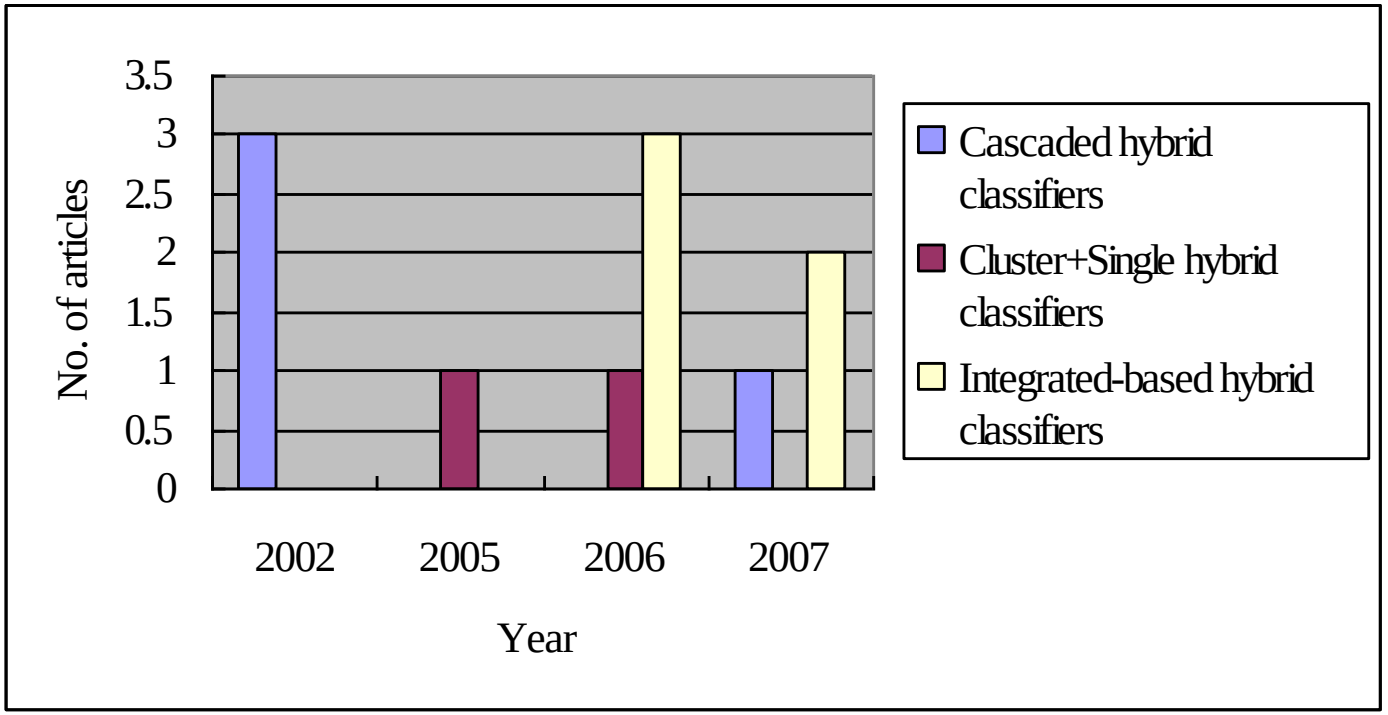
Comparisons of Related Work

- SVM is getting more considered for single classifier design.
- However, MLP is less used recently. This may be because MLP has been widely used in the area of bankruptcy prediction since the 1990s. This leads to MLP could be the potential candidate for the baseline classifier (see later Table).

Comparisons of Related Work

- Hybrid classifiers

	Cascaded hybrid classifiers	Cluster + Single hybrid classifiers	Integrated-based hybrid classifiers
No. of articles	13	3	11



Comparisons of Related Work

- Integrated-based hybrid classifiers are the most considered hybrid classifier design approach.
- On the other hand, cascaded hybrid classifiers are mostly used in 2002.
- Combining a clustering method with a single classifier has been recently studied.

Comparisons of Related Work

- Baselines

Baseline Classifiers	'07	'06	'05	'04	'03	'02	'01	'00	Total
MLP	3	2	5	1	1	1	0	1	14
SVM	2	2	0	0	0	0	0	0	4
LDA/MDA	1	2	1	0	0	2	0	1	10
LR	4	5	2	0	0	1	0	1	13
CART/C4.5	3	0	1	0	0	0	0	1	5
GA	0	0	0	0	0	1	0	0	1
Rough Sets	0	1	1	0	0	1	0	0	3
SOM	0	1	1	0	0	0	0	0	2
Ensemble	0	1	0	0	0	0	0	0	1

Comparisons of Related Work

- MLP is the most widely used baseline classifier.
- For statistical methods, LR and LDA/MDA are the most and second most used baseline classifiers respectively.
- On the other hand, SVM has also been considered recently for model comparisons.

Comparisons of Related Work

- For the 14 articles using ensemble classifiers or hybrid classifiers, their baseline classifiers are only based on some of the above single classifiers.

Comparisons of Related Work

- Datasets

	‘07	‘06	‘05	‘04	‘03	‘02	‘01	‘00	Total
Australia	3	0	3	0	0	0	0	1	7
Benelux	1	3	0	0	0	0	0	0	4
German	3	0	3	0	0	0	0	1	7
Greek	0	1	0	0	0	0	0	0	1
Korea	0	1	3	0	1	1	0	0	6
NASDAQ	1	0	0	0	0	0	0	0	1
Norwegian	0	1	0	0	0	0	0	0	1
Shang-hai Stock	1	0	0	0	0	0	0	0	1
Taiwan	1	2	0	1	0	1	0	0	5
US	0	0	0	1	0	2	1	0	4

Comparisons of Related Work

- Only 14 out of the 41 experiments are based on the public datasets (i.e. Australian and German Credit datasets). Most of the studies use their own collected datasets for experiments.

Comparisons of Related Work

- Datasets & the size of training and testing examples

Datasets	*Training : Testing*
Australian	17:3; 9:1; 2:1
Benelux	3:1; 2:1
German	17:3; 9:1; 2:1
Greek	02:01
Korea	3:1; 4:1; 7:3; 9:1
NASDAQ	10:01
Norwegian	03:01
Shang-hai Stock	N/A
Taiwan	3:1/2:1/9:1
US	5:3/1:1

Comparisons of Related Work

- Different studies use different sizes of training and testing data for classifier design and evaluation.
- By doing this, it is very difficult to conclude which developed classifiers are better for bankruptcy prediction.

Comparisons of Related Work

- Feature Selection

	‘07	‘06	‘05	‘04	‘03	‘02	‘01	‘00	Total
Yes	2	2	2	2	1	2	1	0	12
No	12	5	8	0	0	2	0	2	29

- Not all studies perform feature selection before classifier training.
- Many experiments using non-public datasets involve the feature selection procedure for bankruptcy prediction.
- On the other hand, studies using public datasets do not take feature selection into account (see next Table)

Comparisons of Related Work

- Datasets vs. Prediction Accuracy vs. Feature Selection

Datasets	‘07	‘06	‘05	‘04	‘03	‘02	‘01	‘00
Australian	85.9% (N)		99.2%(N)					86.7%(N)
	86.9% (N)		88.3%(N)					
	85.7%(N)		87.5%(N)					
Benelux	72.9%(N)	89.3%(N)						
	73.3%(N)	72.5%(N)						
	96.5%(N)	75.1%(N)						
German	73.4%(N)		99.3%(N)					75.7%(N)
	77.9%(N)		77.3%(N)					
	88.2%(N)		76.3%(N)					
Korea		80.3%(Y)	82.1%(N)		93.3%(Y)	80.8%(Y)		
			83.9%(Y)					
			76.7%(Y)					
NASDAQ	81.9%(Y)							
Shang-hai	94.6%(Y)							
Taiwan	98.3%(N)	79.2%(N)		79.7%(Y)		77%(N)		
	99.4%(N)							
US				80.4%(Y)		80.3%(N)	85.5%(Y)	
						65.9%(Y)		

Comparisons of Related Work

- For the public datasets (i.e. Australian and German Credit datasets), prediction accuracy does not show significant improvement from 2000 to 2007.
- For the non-public datasets, on the other hand, it is difficult to compare directly in terms of prediction accuracy.

Discussion & Conclusion

- Baseline classifiers.
- The chosen one single classifier (especially the statistical ones) for model validation may be no longer a good candidate as the baseline classifier. It would be valuable if different ensemble classifiers and hybrid classifiers are compared in terms of prediction accuracy.

Discussion & Conclusion

- The architecture of multiple classifiers.
- Designing more sophisticated classifiers via combining ensemble and hybrid classifiers can be examined. Since the idea of combining multiple classifiers is that individual classifiers should not compete each other but collaborate, it may be worth combining ensemble and hybrid classifiers for bankruptcy prediction.

Discussion & Conclusion

- Standard datasets.
- In addition to use private datasets for experiments, it is necessary to consider the above mentioned public datasets to fairly evaluate the developed models. By doing this, the future studies could regard previous experimental results for effective comparisons.

Discussion & Conclusion

- Feature selection.
- As there are numbers of feature selection approaches, the reviewed studies which consider feature selection only choose one specific method, it is not known which method perform the best especially under what classification techniques for bankruptcy prediction.

Study 2 Feature Selection

- Feature selection or dimensionality reduction
- Aim:

  - to filter out unrepresentative features (variables) for better mining results

The Research Question

- Which feature selection method is the best one for allowing the bankruptcy prediction models to provide the best performance?
- Five feature selection methods are compared:
  - t-test
  - correlation matrix
  - stepwise regression
  - principal component analysis
  - factor analysis

Experimental Design

- The datasets:

	Australian	German	Japanese	Bankruptcy dataset[1]	UC Competition[2]
No. of variables	14	20	15	33	39
No. of samples	690	1000	690	240	2528
Good/bad cases	307/383	700/300	307/383	128/112	2449/79

- Australian, German, and Japanese Credit datasets are available at: UCI Machine Learning Repository
- 1: http://www.pietruszkiewicz.com/
- 2: http://mill.ucsd.edu/

Experimental Design

- The prediction model is based on the multi-layer perceptron (MLP) neural network.
- This is because approximately 95% of business application studies utilize MLP (Smith and Gupta, 2000).
- The numbers of hidden nodes: 8, 16, 32, and 64
- The learning epochs: 50, 100, 200, and 400
- As a result, there are sixteen models to test each data set.
- Moreover, 5-fold cross validation is used.

Experimental Design

- Correlation matrix: 95% confident level to reach the significance between each variable
- FA and PCA: factor loadings equal to or greater than 0.5 as informative variables;
- Stepwise: 0.05 probability of *F*
- T-test: 95% level of significance

Experimental Design

- Evaluation method:

actual/predicted	Bad credit	Good credit
Bad credit	(a)	II (b)
Good credit	I (c)	(d)

- Accuracy = $\frac{a+d}{a+b+c+d}$
- Type I/II errors

Experimental Results

- The baseline MLP models

	Japanese	Australian	Bankruptcy dataset	German	UC Competition
Learning epoch	400	400	400	100	50
Hidden nodes	64	32	8	16	32
Average Accuracy	85.88%	81.93%	71.03%	74.28%	96.92%
Type I error	90.05%	21.89%	12.85%	55.39%	81.68%
Type II error	22.40%	13.89%	30.42%	9.63%	4.05%

Experimental Results

	t-test	Stepwise	Correlation matrix	FA	PCA	Baseline Models	*F value*
Japanese Credit							
Accuracy	63.53	82.64	60.16	74.22	74	85.88	*3.25**
Type I error	55.33	32.27	74.55	29.17	47.46	90.05	*3.521**
Type II error	17.29	6.77	3.49	23.75	10.37	22.4	*2.046*
Australian Credit							
Accuracy	89.27	84.74	89.31	86.08	89.93	81.93	*7.279***
Type I error	9.38	12.8	13.33	14.58	7.93	21.89	*7.949***
Type II error	11.72	16.71	8.33	13.6	11.53	13.89	*7.136***
Bankruptcy Dataset							
Accuracy	82.98	77	76.08	72.91	79.59	71.03	*3.219**
Type I error	7.69	37.27	22.76	22.5	16.55	12.85	*7.707***
Type II error	28.57	5.56	25.45	32.73	26	30.42	*8.132***
German Credit							
Accuracy	75.87	75.51	74.84	78.76	67.03	74.28	*33.002***
Type I error	61.28	51.34	54.36	48.69	84.92	55.39	*18.261***
Type II error	8.62	12.25	12.04	10.66	6.27	9.63	*2.348*
UC Competition							
Accuracy	97.25	96.33	96.7	97.3	96.47	96.92	*3.678**
Type I error	74.82	79.25	96.47	94	90	81.68	*7.811***
Type II error	0.16	0.35	0.04	0.08	0.13	4.05	*3.243**

Experimental Results

- All the three performance measures contain the high level of significant difference, except the Type II error of the Japanese Credit and German Credit datasets.

Experimental Results

- Ranking of the feature selection methods

	Accuracy	Type I error	Type II error
Japanese	*B>C*	*B>C*	-
Australian	*P>C>T>S>B*	*P>T>S>C>F>B*	*C>P>F>B>S*
Bankruptcy dataset	*T>B*	*T>P>B>S*	*S>C>P>T>B>F*
German	*F>T>S>C>B>P*	*F>S>C>B>T>P*	-
UC Competition	*F>T>S*	*T>S>P>F>C>B*	*C>B>S*

- T for t-test; S for Stepwise; C for correlation matrix; F for FA; P for PCA; B for Baseline model

Experimental Results

- We disregard the effect of the Japanese dataset and the accuracy of the Bankruptcy dataset because most of the five feature selection methods are not significantly different.
- By the other three datasets, the top 3 positions of the ANOVA results are considered and ranking scores are given: three points for the rank one, two points for rank two and one point for the last one.

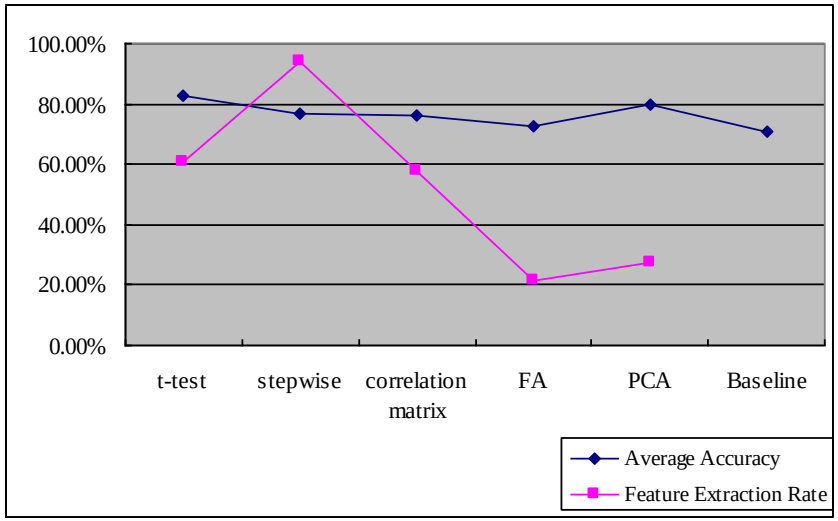
Experimental Results

- Ranking results

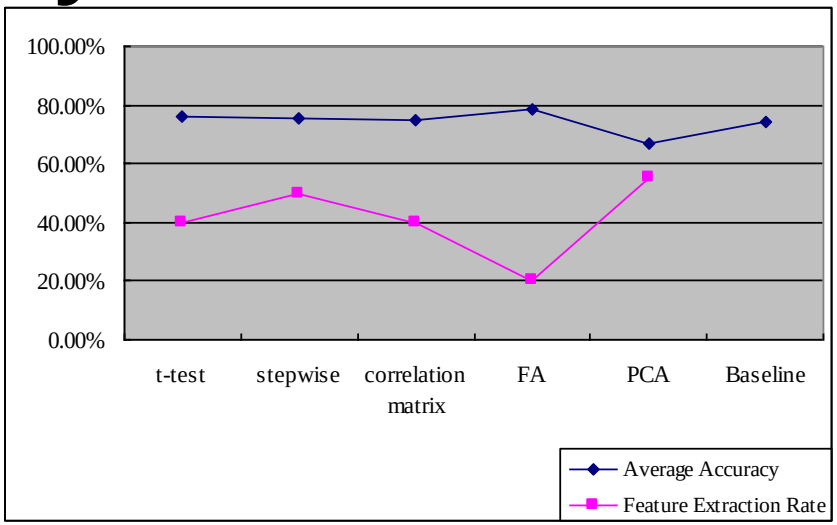
	Number 1	**Number 2**
Average accuracy	factor analysis	t-test
Type I error	t-test	stepwise
Type II error	correlation matrix	stepwise

- To sum up, t-test is the better feature selection method to provide higher prediction accuracy and reduce the Type I error.

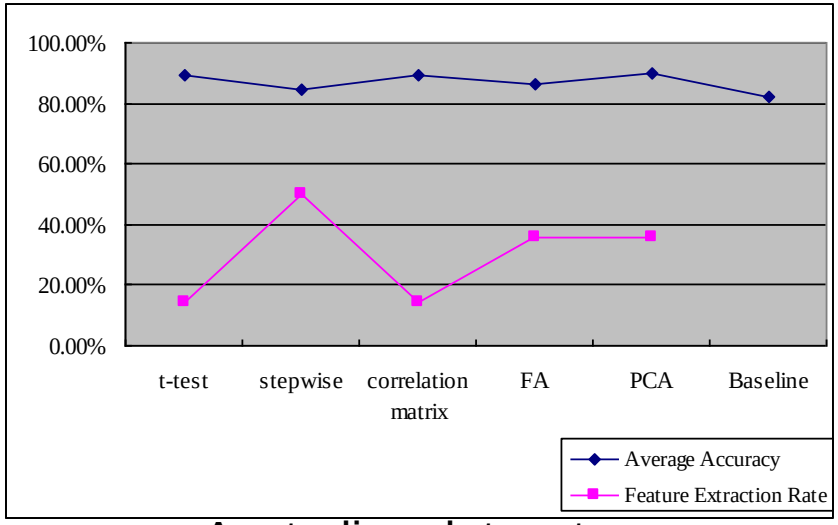
Feature Reduction Rate vs. Prediction Accuracy



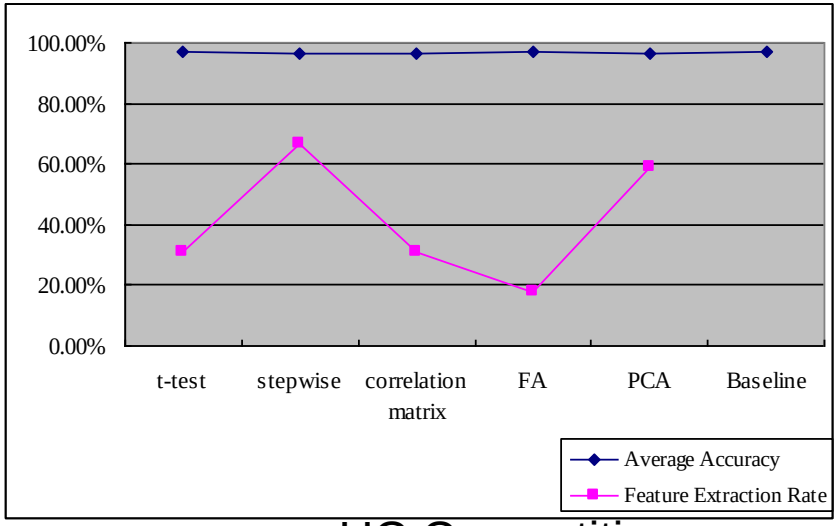
Bankruptcy dataset



German dataset



Australian dataset



UC Competition

Discussion

- Stepwise extracts the largest numbers of variables (i.e. the highest extraction rate), on average 65.5% variables are extracted.
- The other four feature selection methods are seldom distinct. The disparity of them is not larger than 45 %.
- FA reduces the lowest rate of irrelevant features. It provides the best result in the German data set, but has the worst result in the Bankruptcy dataset.
- For t-test, it lies on the middle position of the feature reduction rate and perform very good over the four datasets.

Future Work of Study 2

- Data reduction/outlier detection
- How many outliers (noisy data) removed can allow prediction models to perform the best?
- If the removed outliers are used as the validation data, which learning algorithm is the most stable?
- The priority of performing feature selection and data reduction?

Study-3 Comparisons of Different Machine Learning Techniques

- Motivation: As related work shows that classifier ensembles and hybrid classifiers outperform single classifiers, they have not been compared each other in bankruptcy prediction.

Experimental Setup

- The datasets: Australian, German, Japanese, Bankruptcy, UC Competition
- Single classifiers: MLP, CART, LR
- Classifier ensembles: homogeneous and heterogeneous classifier ensembles
- The combination methods: majority voting and weighted voting

Experimental Setup

- Hybrid classifiers: Self-Organizing Maps (SOM) and k-means are used as the clustering technique. SOM+MLP; SOM+CART; SOM+LR; k-means+MLP; k-means+CART; k-means+LR
- 2 x 2, 3 x 3, 4 x 4, and 5 x 5 SOMs are examined.
- In this paper, 5 x 5 SOM performs the best. That is, two clusters out of 25 contain the largest proportions of the bankruptcy and non-bankruptcy groups respectively.
- The k value of k-means is based on the result of SOM.

Experimental Setup

- Combining cluster with classifier ensembles
- SOM + homogeneous and heterogeneous classifier ensembles respectively
- K-means + homogeneous and heterogeneous classifier ensembles respectively

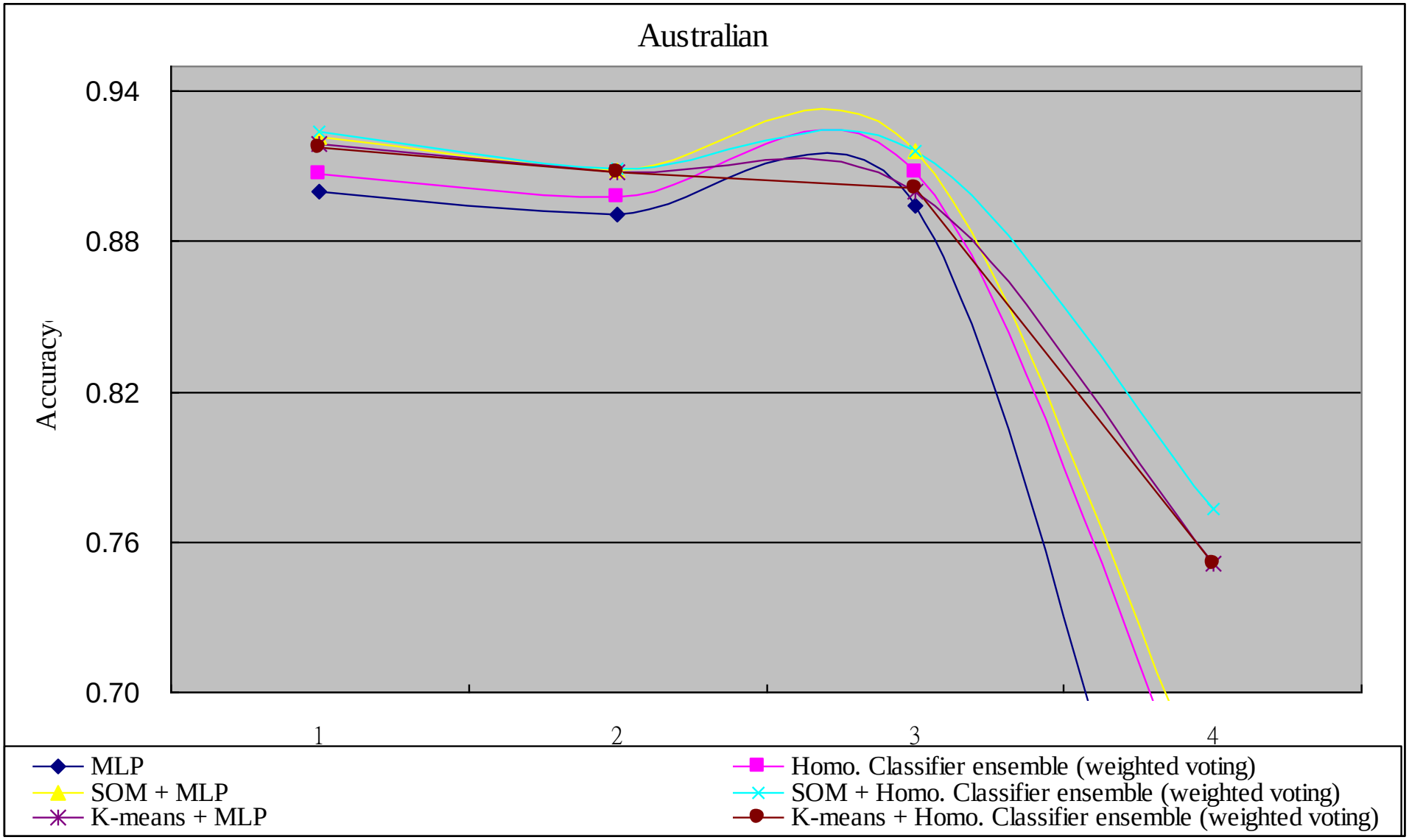
Experimental Setup

- In total, 21 different types of prediction models are developed, which fall into the following six categories
- Single classifiers
- Classifier ensembles
- SOM + single classifiers
- SOM + classifier ensembles
- k-means + single classifiers
- k-means + classifier ensembles

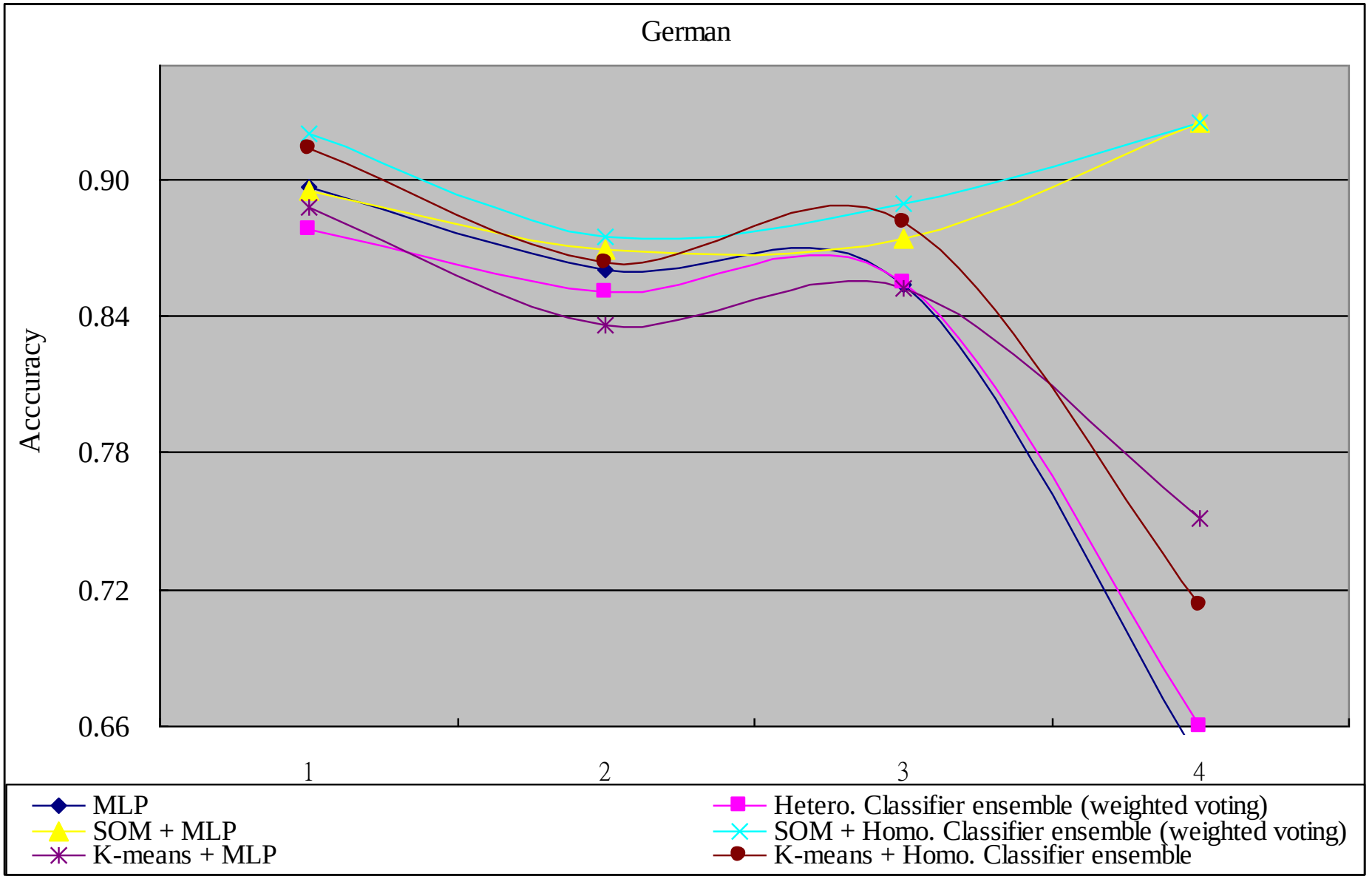
Experimental Setup

- Testing data are based on the training, testing, total data, and fuzzy testing data respectively (represented by 1, 2, 3, and 4 respectively).
- The fuzzy testing data are the outliers identified by SOM/k-means. That is, for the 5 x 5 SOM, 23 clusters are regarded as the fuzzy testing data.

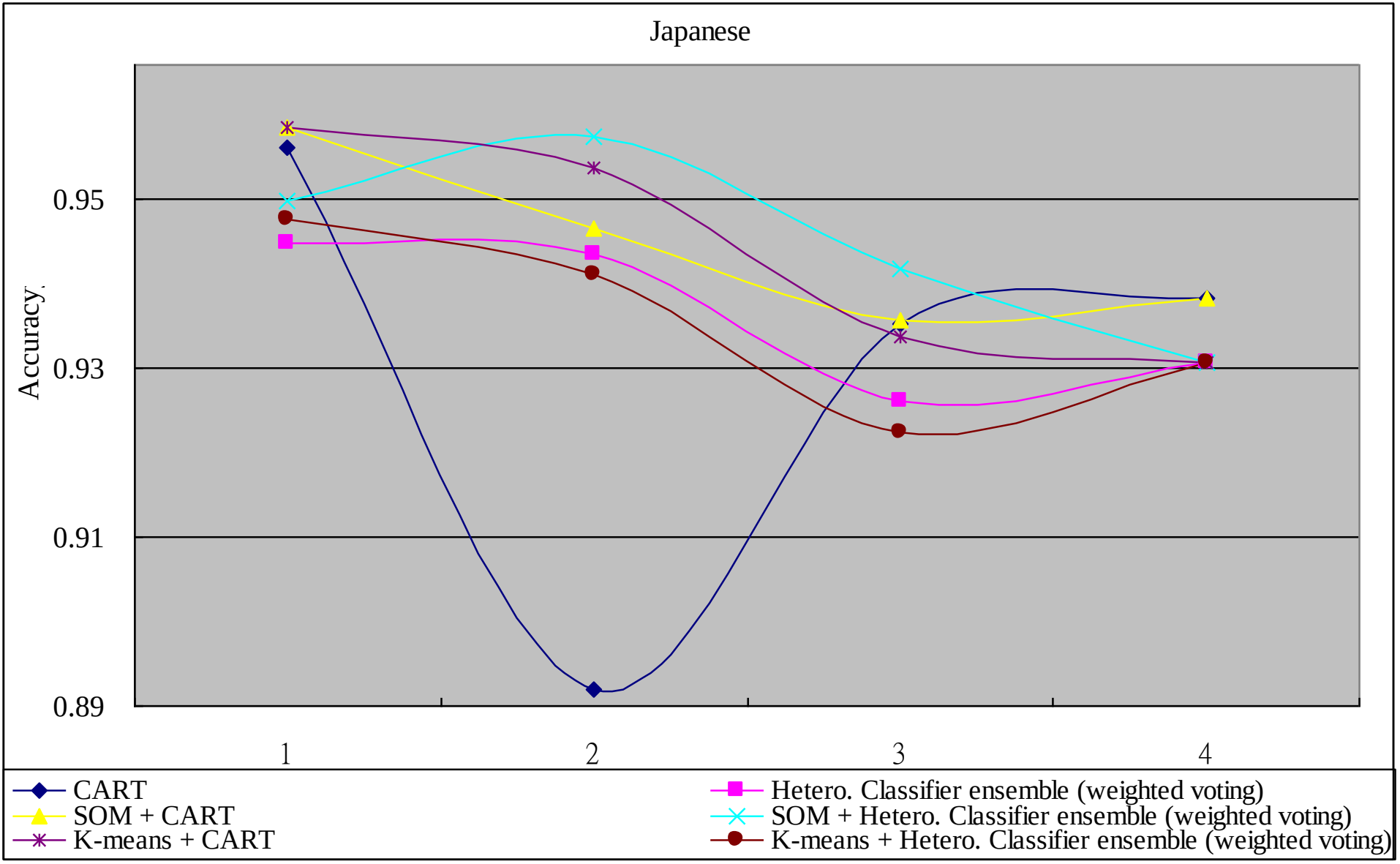
Results


Australian
0.94
0.88
0.82
0.76
0.70
Accuracy
1
2
3
4
MLP
SOM + MLP
K-means + MLP
Homo. Classifier ensemble (weighted voting)
SOM + Homo. Classifier ensemble (weighted voting)
K-means + Homo. Classifier ensemble (weighted voting)

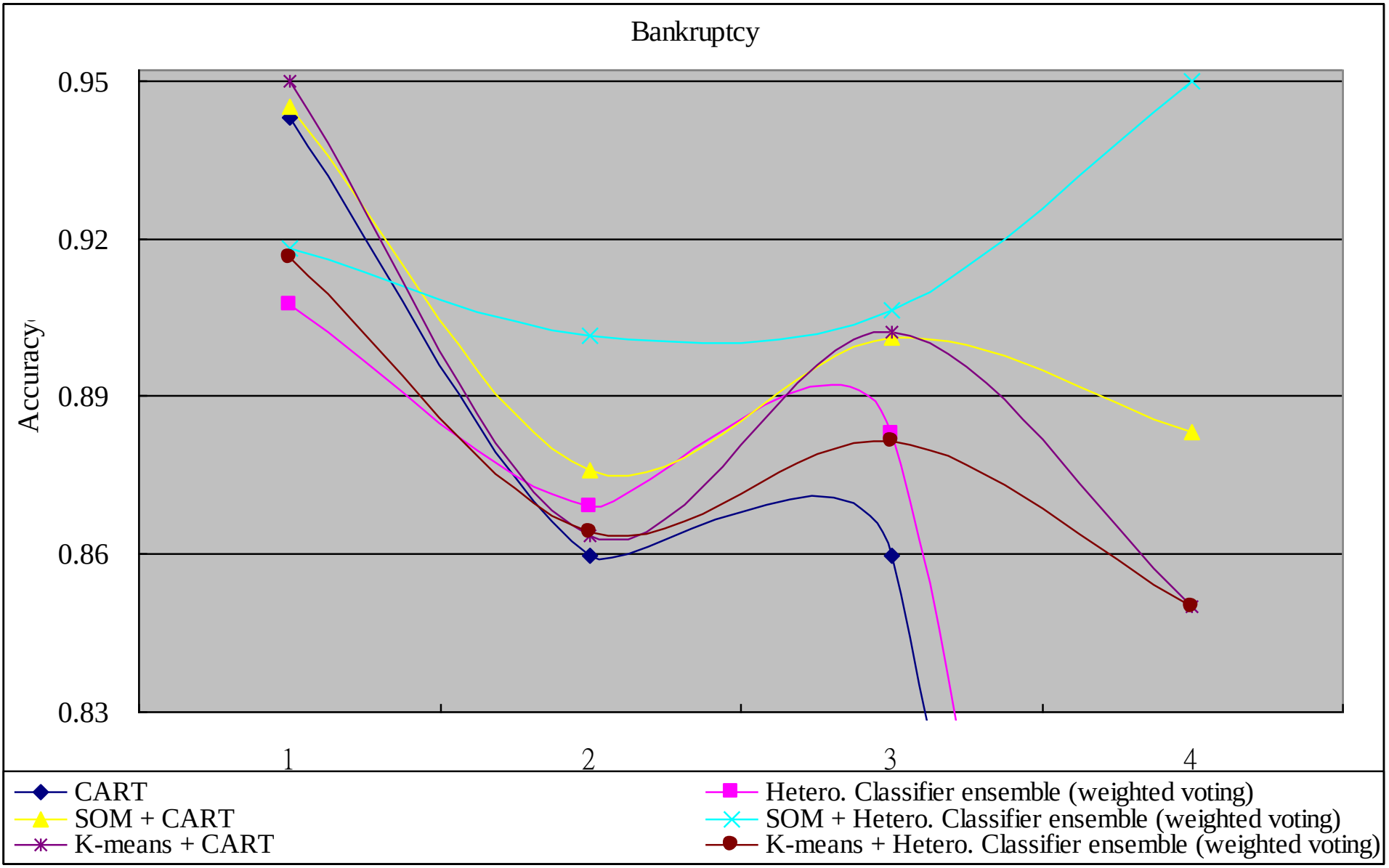
Results


German
Accuracy
0.90
0.84
0.78
0.72
0.66
1
2
3
4
MLP
SOM + MLP
K-means + MLP
Hetero. Classifier ensemble (weighted voting)
SOM + Homo. Classifier ensemble (weighted voting)
K-means + Homo. Classifier ensemble

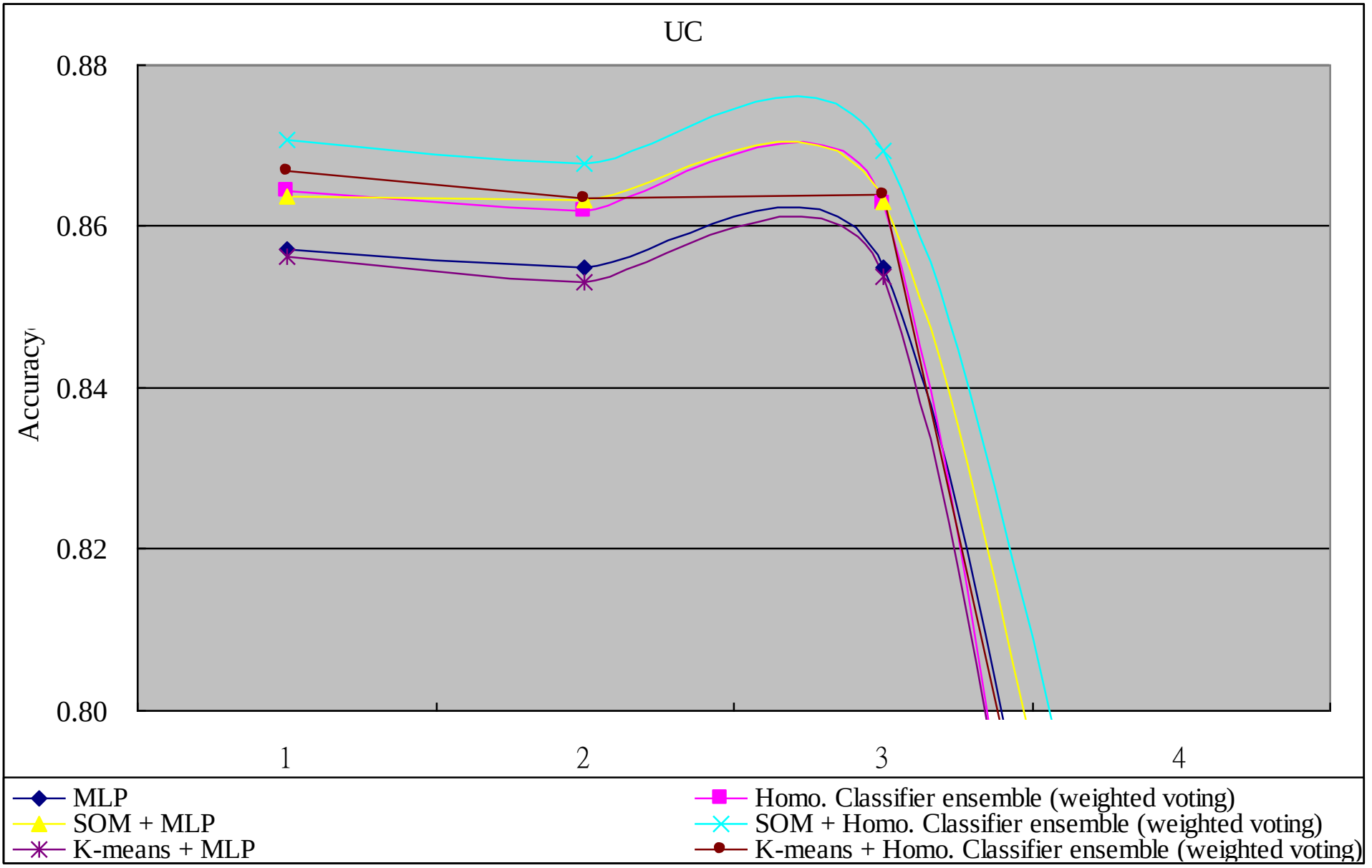
Results


Japanese
Accuracy
0.95
0.93
0.91
0.89
1
2
3
4
CART
SOM + CART
K-means + CART
Hetero. Classifier ensemble (weighted voting)
SOM + Hetero. Classifier ensemble (weighted voting)
K-means + Hetero. Classifier ensemble (weighted voting)

Results


Bankruptcy
0.95
0.92
0.89
0.86
0.83
Accuracy
1
2
3
4
CART
SOM + CART
K-means + CART
Hetero. Classifier ensemble (weighted voting)
SOM + Hetero. Classifier ensemble (weighted voting)
K-means + Hetero. Classifier ensemble (weighted voting)

Results


UC
0.88
0.86
0.84
0.82
0.80
Accuracy
1
2
3
4
MLP
SOM + MLP
K-means + MLP
Homo. Classifier ensemble (weighted voting)
SOM + Homo. Classifier ensemble (weighted voting)
K-means + Homo. Classifier ensemble (weighted voting)

Discussions

- For the single classifiers, MLP outperforms the other two classifiers (i.e. CART and LR) over three datasets (i.e. Australian, German, UC Competition).
- For classifier ensembles (majority voting and weighted voting) perform better than the single classifiers. Particularly, classifier ensembles based on weighted voting provide the best performance.

Discussions

- For hybrid classifiers, using SOM to construct a hybrid classifier can provide better performances than using *k*-means.
- The SOM based hybrid classifiers outperform the single classifiers.
- In addition, hybrid classifiers perform slightly better than classifier ensembles.

Discussions

- On average, combing SOM with (homogeneous/ heterogeneous) classifier ensembles can not only provide the best performances but also perform more stably when the fuzzy testing data are used.
- So, MLP may not be a competitive baseline model, but what should be the ‘best’ baseline model for future research?

Thank you for listening

Q & A