基于最大信息系数和凝聚层次聚类的特征选择方法在软件缺陷预测中的应用

软件工程国家重点实验室

徐洲

武漢大學

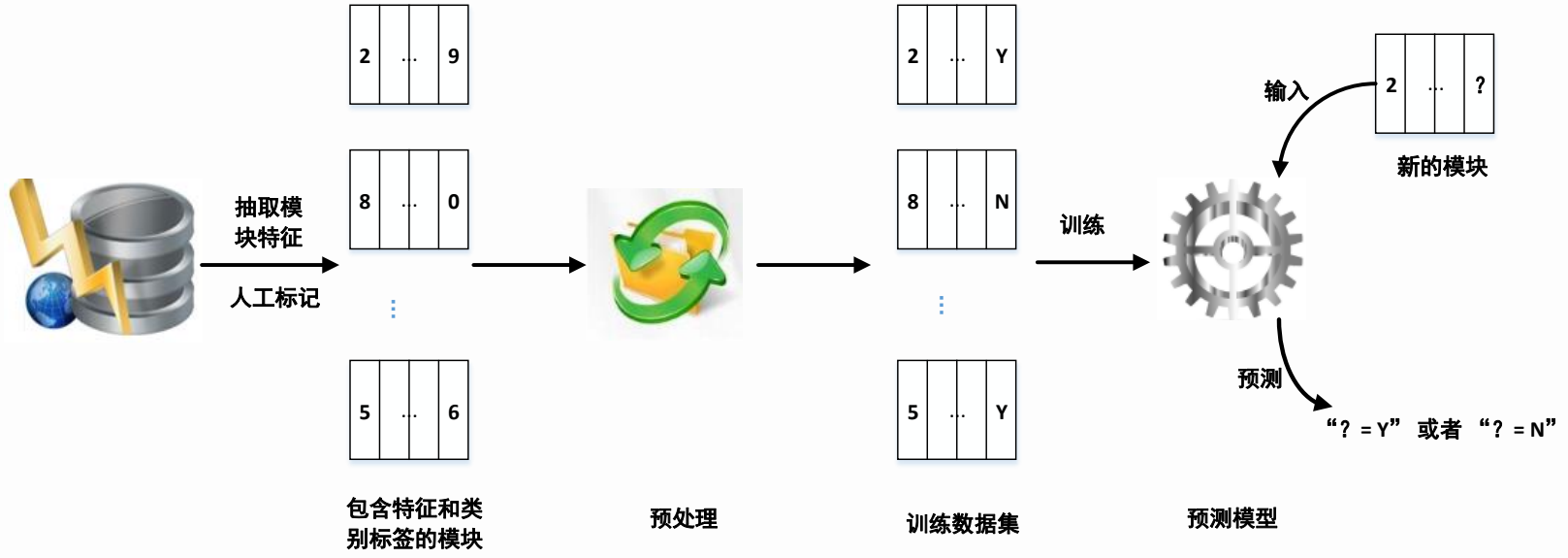
contents

- 研究背景
- 基础知识
- 方法框架
- 结果分析
- 总结展望

contents

背景介绍

- 版本控制系统、缺陷追踪系统等工具在软件开发过程中的普及，使我们可以获得大量的软件缺陷相关数据。如何有效分析这些丰富的数据，构建缺陷预测模型，来提高软件质量，合理分配测试资源，已成为软件质量保证的研究热点。



- 虽然已有的研究提出了很多缺陷预测的模型，但是软件缺陷数据集中存在的不相关和冗余特征会在一定程度的影响这些预测模型的性能。而且过多的特征会增加模型训练时间和复杂度。

- 特征选择方法通过评价特征对分类模型的贡献，可以过滤掉数据集中存在的不相关和冗余特征，得到一个精简的特征子集，能够有效解决以上问题。

现有方法的不足：

- 现有的软件缺陷预测中的特征选择方法大多数关注的是如何寻找与类标签相关性高的特征，很少考虑特征之间的冗余性。

- 软件缺陷特征与类标签之间往往存在很复杂的关系。

- Liu et al.[1] 和 Chen[2]分别提出了一种基于特征排序和特征聚类的特征选择方法来过滤掉不相关和冗余特征，但是他们在对特征进行聚类之前要事先人为指定聚类的个数。

[1] S. Liu, X. Chen, W. Liu, et al. FECAR: A Feature Selection Framework for Software Defect Prediction 2014 IEEE 38th Annual Computer Software and Applications Conference (COMPSAC). IEEE Computer Society, 426-435, 2014.

[2] J. Chen, S. Liu, W. Liu, et al. A Two-Stage Data Preprocessing Approach for Software Defect prediction. Software Security and Reliability (SERE), 2014 Eighth International Conference on. 20 - 29. IEEE, 2014.

我们提出了一种基于特征排序和特征聚类相结合的特征选择方法。

不同的是：

(1)我们选用最大信息系数(MIC)作为软件缺陷特征与类标签的相关性度量。

(2)我们选用凝聚层次聚类(HAC)对特征进行聚类。该聚类方法可以通过聚类过程中的不一致系数来自动确定最终聚类个数。

contents

基础知识

- 最大信息系数MIC

MIC是Reshef et al.[1] 提出的一种用于探索大规模数据集中两个变量之间相关性的新型非参数统计方法，发表在2011年Science 。

MIC优点：

(1)MIC能够检测出两个变量之间复杂的关系[2]。

(2)MIC对噪声的鲁棒性。

[1] Reshef D N, Reshef Y A, Finucane H K, et al. Detecting novel associations in large data sets[J]. science, 2011, 334(6062): 1518-1524.

[2]Supporting Online Material for Detecting Novel Associations in Large Data Sets.

http://www.sciencemag.org/content/334/6062/1518/suppl/DC1

- MIC思想：如果两个变量X和Y之间存在关联，则可以用一组网格来划分这两个变量形成的散点图，使大多数的数据点落入网格的某些单元中。

对于一个变量集合大小为n的有限数据集D，X和Y分别表示两个变量，假设这两个变量的值分别被划分为x和y个箱，则称这样一次划分为x-by-y的网格。$D|_G$表示数据集D的数据点在网格G中单元上的分布。

对于一次特定的划分，最大互信息定义为:

$$I^*(D,x,y) = \max I(D|_G)$$

特征矩阵定义为：

$$M(D)_{x,y} = \frac{I^*(D,x,y)}{log\ (\min\{x,y\})}$$

MIC定义为：

$$MIC(D) = \max_{xy<B(n)}\{M(D)_{x,y}\}$$

凝聚层次聚类

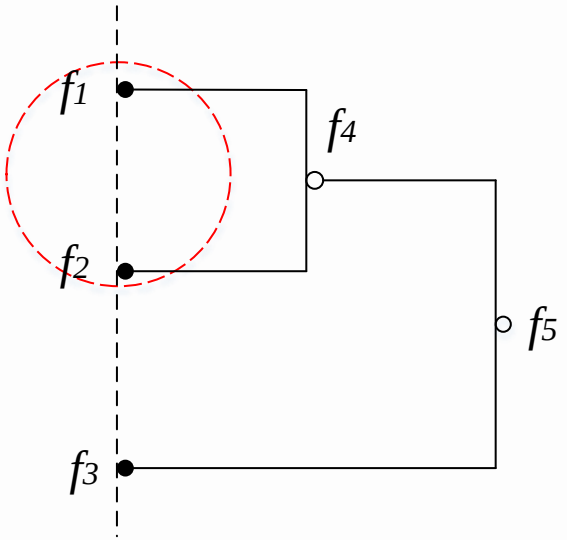
思想：

(1)每一个特征为一个簇，计算特征对之间的距离

(2)把距离最小的两个簇合并为一个新的簇

(3)重新计算簇与簇之间的距离

(4)重复(2)-(3)直到合并为一个簇



特征间的距离：相关系数

$$c_{ij}=\frac{\sum_{k=1}^{n}(x_{ik}-\bar{x}_i)(x_{jk}-\bar{x}_j)}{\sqrt{\sum_{k=1}^{n}(x_{ik}-\overline{x_i})^2\sum_{k=1}^{n}(x_{jk}-\overline{x_j})^2}}$$

簇与簇的距离：平均距离法

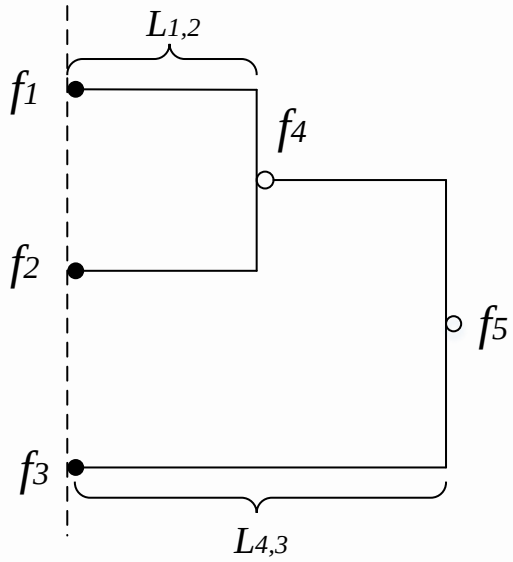
$$D_{ab}=\frac{1}{n_a n_b}\sum_{x_i\in M_a,x_j\in M_b}d_{ij}$$

不一致系数

不一致系数(Inconsistency Coefficient)是用来量化表示一次连接(合并)的相对一致性。

不一致系数的值可以通过比较当前连接的距离和该连接的邻居连接距离的平均值计算得到。

$$\mathrm{IC}(L_{curr}) = \frac{D_{L_{curr}} - \mathrm{avg}(D_{L_{neighbor}})}{\mathrm{Std}(D_{L_{neighbor}})}$$

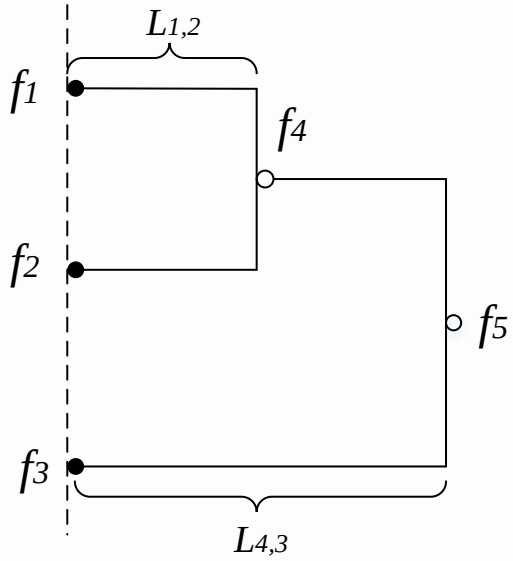


不一致系数

f_1和 f_2之间的距离 $D_{1,2} = d_{1,2} = 1 - c_{1,2}$

f_3和 f_4之间的距离 $D_{3,4} = (d_{1,3} + d_{2,3})/2$

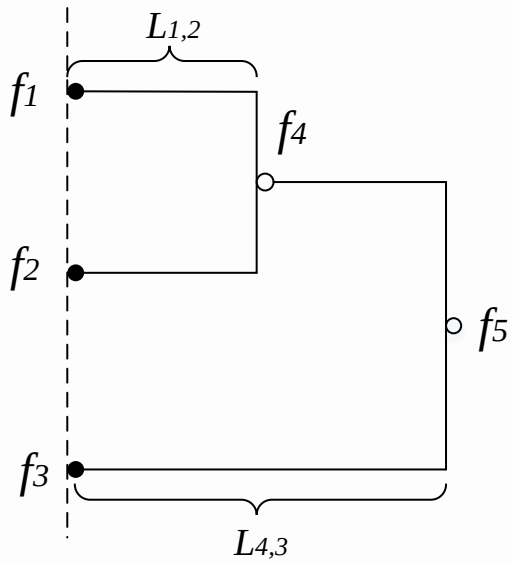
连接$L_{4,3}$的不一致系数 $IC(L_{4,3}) = \frac{D_{3,4} - \frac{D_{1,2}+D_{3,4}}{2}}{std}$



不一致系数增量

如果一次连接的不一致系数较前一次连接的不一致系数有增加，表明前一次连接的效果好，增加的幅度越大，表明前一次的聚类效果越好。所以可以参照不一致系数的变化来确定最终的聚类个数。

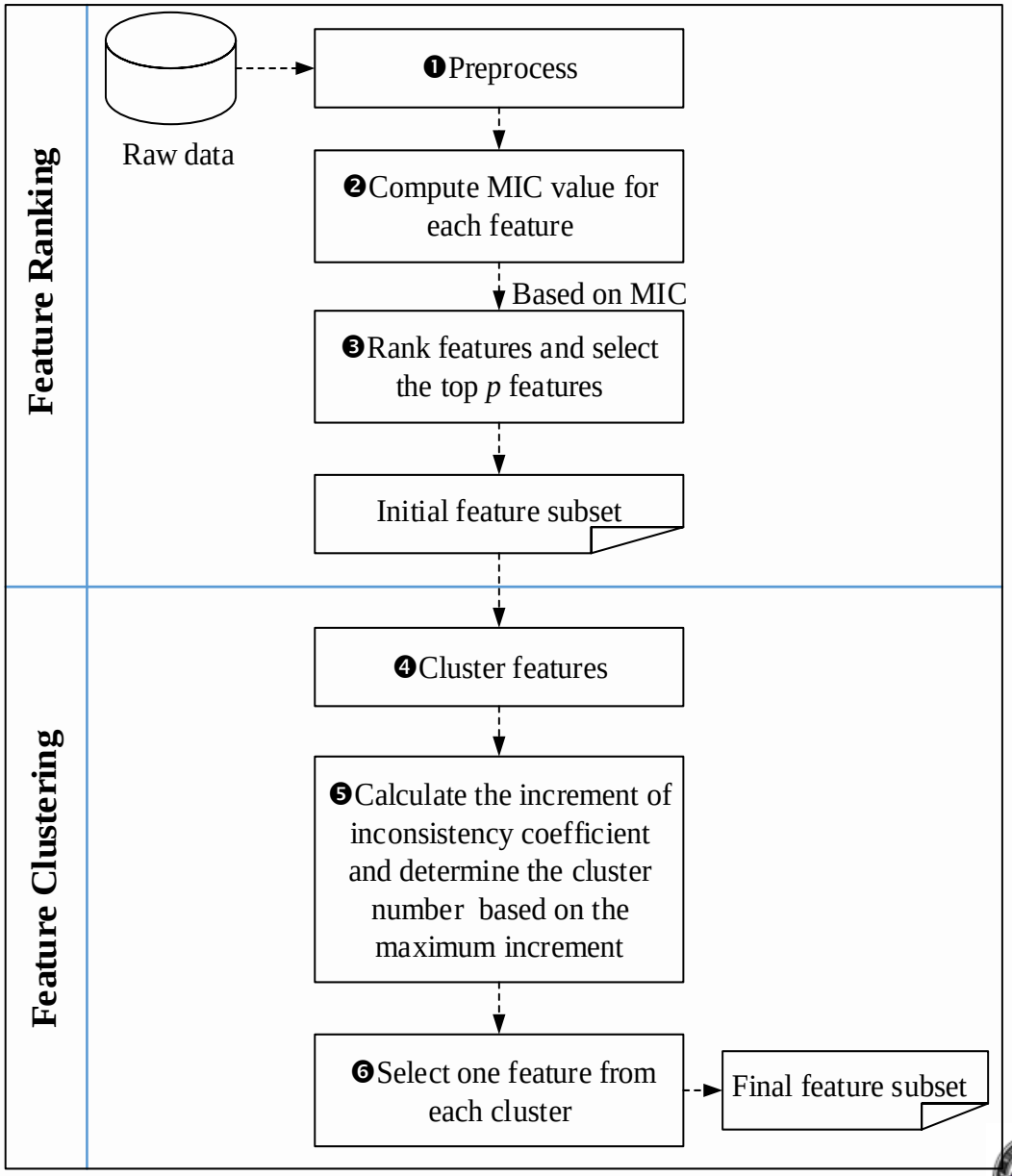
$$\Delta_{L_{curr},L_{prev}} = \mathrm{IC}(L_{curr}) - \mathrm{IC}(L_{prev})$$



contents

- 研究背景
- 基础知识
- **方法框架**
- 结果分析
- 总结展望

方法框架 MICHAC



contents

- 研究背景
- 基础知识
- 方法框架
- **结果分析**
- 总结展望

结果分析

数据集：

Project	# features	# instances	# defective instances	% defective instances
CM1	38	344	42	12.2%
JM1	22	9593	1759	18.3%
KC1	22	2096	325	15.5%
KC2	16	522	107	20.5%
MC1	39	9277	68	0.7%
MC2	40	127	44	34.6%
MW1	38	253	27	10.7%
PC1	38	759	61	8.0%
PC3	38	1125	140	12.4%
PC4	38	1399	178	12.7%
PC5	39	1711	471	27.5%

评价指标

- Precision

  正确预测为有缺陷的模块数占所有被预测为有缺陷的模块数的比例

- Recall

  正确预测为有缺陷的模块数占真正有缺陷模块数的比例

- F-measure = (2*precision*recall) / (precision + recall)

  调和平均值

- AUC

  ROC曲线面积

■ F-measure 和 AUC 的值越大，模型性能越好

分类模型

- 朴素贝叶斯
- 随机森林
- RIPPER

对比方法

- MICHAC vs 特征排序方法(卡方CS、增益率GR、ReliefF RF)
- MICHAC vs 基于特征排序和特征聚类的方法(TC和FECAR)
- MICHAC vs MIC

实验结果1

Model	Metric	Full	MICHAC	CS	GR	RF	TC	FECAR
朴素贝叶斯	P	0.407	0.427	0.440	**0.454**	0.350	0.410	0.442
	R	**0.429**	0.397	0.332	0.366	0.424	0.342	0.376
	F	0.350	**0.373**	0.359	0.360	0.355	0.346	0.367
	W/D/L	7/1/3		8/1/2	7/0/4	6/0/5	6/1/4	6/0/5
	AUC	0.760	**0.771**	0.727	0.733	0.738	0.759	0.750
	W/D/L	8/1/2		10/0/1	10/0/1	10/0/1	7/0/4	10/0/1
随机森林	P	0.540	**0.560**	0.488	0.498	0.468	0.526	0.501
	R	0.288	0.311	0.302	0.311	0.258	0.287	**0.321**
	F	0.372	**0.392**	0.366	0.377	0.326	0.367	0.385
	W/D/L	4/1/7		8/1/2	8/0/3	9/0/2	7/0/4	7/0/4
	AUC	**0.815**	0.807	0.774	0.779	0.777	0.794	0.785
	W/D/L	4/0/7		10/0/1	11/0/0	9/0/2	8/0/3	7/0/4
RIPPER	P	0.488	**0.550**	0.535	0.536	0.423	0.498	0.530
	R	0.268	**0.280**	0.266	0.255	0.171	0.237	0.267
	F	0.333	**0.351**	0.345	0.334	0.230	0.309	0.344
	W/D/L	7/0/4		4/0/7	7/0/4	9/0/2	8/0/3	5/0/6
	AUC	0.605	**0.612**	0.608	0.601	0.568	0.595	0.608
	W/D/L	7/0/4		6/0/5	7/0/4	8/0/3	7/0/4	6/0/5

实验结果2

	朴素贝叶斯		随机森林		RIPPER	
Metric	MICHAC	MIC	MICHAC	MIC	MICHAC	MIC
P	**0.427**	0.405	**0.560**	0.539	**0.550**	0.544
R	**0.397**	0.347	**0.311**	0.310	**0.280**	0.270
F	**0.373**	0.309	**0.392**	0.388	**0.351**	0.347
W/D/L	11/0/0		7/0/4		6/0/5	
AUC	**0.771**	0.715	**0.807**	0.780	**0.612**	0.608
W/D/L	11/0/0		9/0/2		5/1/5	

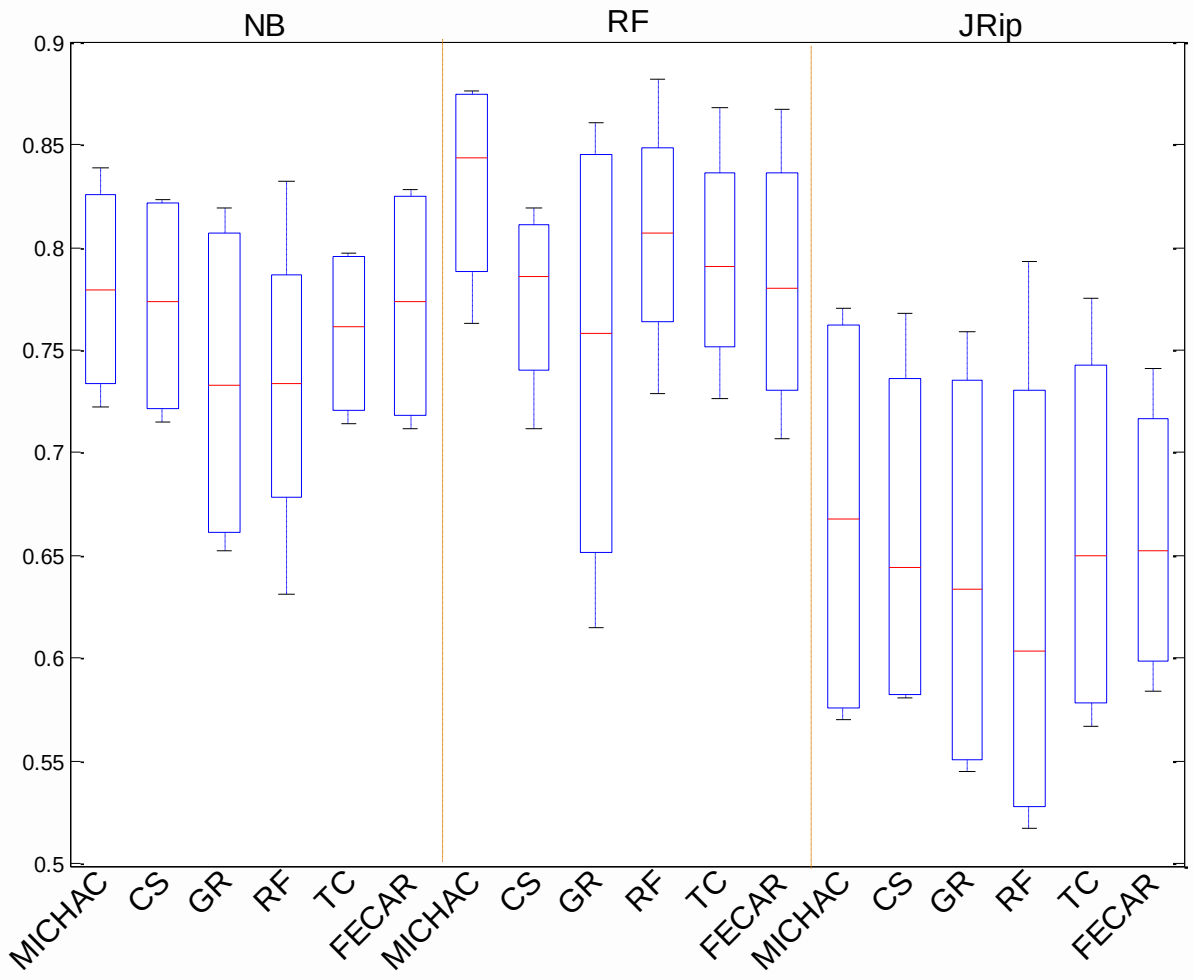
AEEEM数据集

Project	# features	# instances	# defective instances	% defective instances
Eclipse-JDT	77	997	206	20.7%
Equinox	77	324	129	39.8%
Mylyn	77	1862	245	13.2%
Eclipse-PDE	77	1497	209	14.0%

实验结果4

Model	Metric	Full	MICHAC	CS	GR	RF	TC	FECAR
朴素贝叶斯	P	0.538	0.549	**0.594**	0.567	0.491	0.547	0.582
	R	0.418	0.412	**0.419**	0.343	0.316	0.343	0.414
	F	0.468	0.469	**0.491**	0.426	0.376	0.420	0.481
	W/D/L	2/0/2		2/0/2	3/0/1	3/0/1	3/0/1	1/0/3
	AUC	0.771	**0.780**	0.772	0.734	0.733	0.759	0.772
	W/D/L	3/0/1		2/0/2	3/0/1	4/0/0	3/0/1	2/0/2
随机森林	P	**0.673**	0.667	0.610	0.528	0.535	0.601	0.608
	R	0.421	**0.425**	0.399	0.383	0.376	0.417	0.406
	F	0.490	**0.500**	0.471	0.427	0.426	0.475	0.465
	W/D/L	4/0/0		2/0/2	4/0/0	3/0/1	2/0/2	3/0/1
	AUC	**0.839**	0.832	0.776	0.748	0.806	0.794	0.784
	W/D/L	1/0/3		4/0/0	4/0/0	3/0/1	4/0/0	4/0/0
RIPPER	P	0.569	0.574	0.579	0.537	0.532	**0.590**	0.576
	R	0.393	0.398	**0.400**	0.346	0.302	0.390	0.381
	F	0.446	0.449	**0.458**	0.392	0.350	0.448	0.447
	W/D/L	3/0/1		1/0/3	3/0/1	3/0/1	2/0/2	2/0/2
	AUC	0.662	**0.669**	0.659	0.643	0.629	0.660	0.658
	W/D/L	3/0/1		2/0/2	4/0/0	3/0/1	2/0/2	2/0/2

实验结果5


NB
RF
JRip
0.9
0.85
0.8
0.75
0.7
0.65
0.6
0.55
0.5
MICHAC
CS
GR
RF
TC
FECAR
MICHAC
CS
GR
RF
TC
FECAR
MICHAC
CS
GR
RF
TC
FECAR

	朴素贝叶斯		随机森林		RIPPER	
Metric	MICHAC	MIC	MICHAC	MIC	MICHAC	MIC
P	0.549	**0.560**	**0.667**	0.582	**0.574**	0.539
R	**0.412**	0.406	**0.425**	0.393	**0.398**	0.375
F	**0.469**	0.467	**0.500**	0.450	**0.449**	0.430
W/D/L	2/0/2		4/0/0		3/0/1	
AUC	**0.832**	0.774	**0.832**	0.806	**0.669**	0.661
W/D/L	2/0/2		4/0/0		3/0/1	

contents

总结展望

- 首次将MIC作为相关性度量标准引入软件缺陷预测；
- 提出了一种基于特征排序和特征聚类的特征选择方法，在特征聚类阶段自动确定聚类个数；
- 实验表明该方法的有效性。

展望：

- 探索特征选择算法对不同分类模型性能的影响
- 采用含有更多特征的软件缺陷数据集

谢谢！