数据挖掘Lab3实验报告

161220096 欧阳鸿荣

1.实验要求

在提供的数据集上，基于10折的交叉验证，使用J48(C4.8)，朴素贝叶斯，SVM，神经网络，kNN算法以及它们使用装袋(Bagging)的集成学习的版本对数据集训练。基于如accuracy，AUC等指标比较各种方法的表现。讨论各种方法的表现并且说明如何优化基于Bagging的kNN算法的性能。

2.实验配置

综合考虑数据集格式(arff)等多方面原因，本次实验采用Weka进行数据挖掘。使用的是Weka 3.8版本。

对于实验要求的算法，J4.8 (C4.5)、Naïve Bayes、神经网络和kNN在Weka中都有提供。其中，神经网络采用的是Weka中的Multilayer Perceptron多层感知机，kNN算法采用Weka中的IBk(K-nearest neighbours classifier)分类器。而对于Weka默认安装里缺少的SVM，使用的是weka库中的libSVM进行实验。

在实验中，在对比多个方法的任务中，统一采用原始数据，10-fold cross validation和weka中的默认配置进行分类器的训练。在对kNN的优化中，才考虑对数据预处理和修改默认配置等因素。同时，在实验过程中发现libSVM的表现很不稳定，因此对其进行数据预处理实验进行对比，从此发现了对数据标准化等预处理手段的重要性。

3.实验原理

3.1 J4.8 (C4.5)

3.2 Naïve Bayes

3.3 SVM

###

3.4 Neural Network

3.5 kNN

4.实验过程与实验结果

4.1 数据集基本情况

数据集	Attributes	Instances	数据规模
breast-w	10	699	6990
colic	23	368	8464
credit-a	16	690	11040
credit-g	21	1000	21000
diabetes	9	768	6912
hepatitis	20	155	3100
mozilla4	6	15545	93270
pc1	22	1109	24398
pc5	39	17186	670254
waveform-5000	41	5000	205000

4.2 五种方法的对比

该任务中，统一采用原始数据，10-fold cross validation和weka中的默认配置进行分类器的训练。其中J4.8 (C4.5)、Naïve Bayes采用Weka中默认版本，神经网络采用的是Weka中的Multilayer Perceptron多层感知机，kNN算法采用Weka中的IBk分类器。SVM使用的是weka库中的libSVM进行实验。

训练的模型和结果都放在`output`目录下，由于Markdown表格能力有限，结果备份在`实验三结果统计.xlsx`下。

1 breast-w 数据集

数据规模：6990	基础方法	基础方法	Bagging	Bagging	变化量	变化量
挖掘算法	Accuracy	ROC	Accuracy	ROC	Accuracy	ROC
J4.8 (C4.5)	94.56%	0.955	**96.28%**	0.985	1.72%	0.03
Naïve Bayes	**95.99%**	**0.986**	95.85%	**0.989**	-0.14%	0.003
SVM	95.71%	0.964	95.42%	0.973	-0.29%	0.009
Neural Network	95.28%	0.986	95.99%	**0.989**	0.72%	0.003
kNN	95.14%	0.973	95.85%	0.987	0.72%	0.014
最大值	95.99%	0.986	96.28%	0.989		
最优算法	Naïve Bayes	Naïve Bayes	J48	NB/NN		

2 colic 数据集

数据规模: 8464	基础方法	基础方法	Bagging	Bagging	变化量	变化量
挖掘算法	Accuracy	ROC	Accuracy	ROC	Accuracy	ROC
J4.8 (C4.5)	**85.33%**	0.813	**85.60%**	0.864	0.27%	0.051
Naïve Bayes	77.99%	0.842	77.99%	0.842	0.00%	0
SVM	72.55%	0.67	69.57%	0.692	-2.99%	0.022
Neural Network	80.43%	**0.857**	84.51%	**0.876**	4.08%	0.019
kNN	81.25%	0.802	81.25%	0.824	0.00%	0.022
最大值	85.33%	0.857	85.60%	0.876		
最优算法	J48	Neural Network	J48	Neural Network		

3 credit-a 数据集

数据规模：11040	基础方法	基础方法	Bagging	Bagging	变化量	变化量
挖掘算法	Accuracy	ROC	Accuracy	ROC	Accuracy	ROC
J4.8 (C4.5)	**86.09%**	0.887	**86.81%**	**0.928**	0.72%	0.041
Naïve Bayes	77.68%	**0.896**	77.83%	0.896	0.14%	0
SVM	55.51%	0.513	55.80%	0.535	0.29%	0.022
Neural Network	83.62%	0.895	85.07%	0.908	1.45%	0.013
kNN	81.16%	0.808	81.30%	0.886	0.14%	0.078
最大值	86.09%	0.896	86.81%	0.928		
最优算法	J48	Naïve Bayes	J48	J48		

4.credit-g 数据集

数据规模: 21000	**基础方法**	**基础方法**	**Bagging**	**Bagging**	**变化量**	**变化量**
挖掘算法	Accuracy	ROC	Accuracy	ROC	Accuracy	ROC
J4.8 (C4.5)	70.50%	0.639	73.30%	0.753	2.80%	0.114
Naïve Bayes	**75.40%**	**0.787**	74.80%	**0.787**	-0.60%	0
SVM	68.70%	0.491	68.60%	0.49	-0.10%	-0.001
Neural Network	71.50%	0.73	**76.10%**	0.776	4.60%	0.046
kNN	72%	0.66	72.10%	0.694	0.10%	0.034
最大值	75.40%	0.787	76.10%	0.787		
最优算法	Naïve Bayes	Naïve Bayes	Neural Network	Naïve Bayes		

5.diabetes 数据集

数据规模: 6912	**基础方法**	**基础方法**	**Bagging**	**Bagging**	**变化量**	**变化量**
挖掘算法	Accuracy	ROC	Accuracy	ROC	Accuracy	ROC
J4.8 (C4.5)	73.83%	0.751	74.61%	0.798	0.78%	0.047
Naïve Bayes	**76.30%**	**0.819**	76.56%	0.817	0.26%	-0.002
SVM	65.10%	0.5	65.10%	0.5	0.00%	0
Neural Network	75.39%	0.793	**76.82%**	**0.822**	1.43%	0.029
kNN	70.18%	0.65	71.09%	0.725	0.91%	0.075
最大值	76.30%	0.819	76.82%	0.822		
最优算法	Naïve Bayes	Naïve Bayes	Neural Network	Neural Network		

6.hepatitis 数据集

数据规模：3100	基础方法	基础方法	Bagging	Bagging	变化量	变化量
挖掘算法	Accuracy	ROC	Accuracy	ROC	Accuracy	ROC
J4.8 (C4.5)	83.87%	0.708	83.87%	0.865	0.00%	0.157
Naïve Bayes	**84.52%**	**0.86**	**85.81%**	**0.89**	1.29%	0.03
SVM	79.35%	0.5	79.35%	0.492	0.00%	-0.008
Neural Network	80%	0.823	84.52%	0.846	4.52%	0.023
kNN	80.65%	0.653	81.29%	0.782	0.65%	0.129
最大值	84.52%	0.86	85.81%	0.89		
最优算法	Naïve Bayes	Naïve Bayes	Naïve Bayes	Naïve Bayes		

7.mozilla4 数据集

数据规模：93270	基础方法	基础方法	Bagging	Bagging	变化量	变化量
挖掘算法	Accuracy	ROC	Accuracy	ROC	Accuracy	ROC
J4.8 (C4.5)	**94.80%**	**0.954**	**95.11%**	**0.976**	0.32%	0.022
Naïve Bayes	68.64%	0.829	68.74%	0.83	0.10%	0.001
SVM	69.54%	0.537	69.82%	0.549	0.28%	0.012
Neural Network	91.19%	0.94	91.28%	0.945	0.10%	0.005
kNN	88.99%	0.877	88.86%	0.928	-0.13%	0.051
最大值	94.80%	0.954	95.11%	0.976		
最优算法	J48	J48	J48	J48		

8.pc1 数据集

数据规模：24398	基础方法	基础方法	Bagging	Bagging	变化量	变化量
挖掘算法	Accuracy	ROC	Accuracy	ROC	Accuracy	ROC
J4.8 (C4.5)	93.33%	0.668	93.60%	**0.855**	0.27%	0.187
Naïve Bayes	89.18%	0.65	88.91%	0.628	-0.27%	-0.022
SVM	93.51%	0.563	**93.87%**	0.574	0.36%	0.011
Neural Network	**93.60%**	0.723	93.33%	0.835	-0.27%	0.112
kNN	92.06%	**0.74**	91.07%	0.793	-0.99%	0.053
最大值	93.60%	0.74	93.87%	0.855		
最优算法	Neural Network	kNN	SVM	J48		

9.pc5 数据集

数据规模：670254	基础方法	基础方法	Bagging	Bagging	变化量	变化量
挖掘算法	Accuracy	ROC	Accuracy	ROC	Accuracy	ROC
J4.8 (C4.5)	**97.46%**	0.817	**97.53%**	**0.959**	0.06%	0.142
Naïve Bayes	96.42%	0.833	96.48%	0.845	0.06%	0.012
SVM	97.25%	0.548	97.22%	0.552	-0.03%	0.004
Neural Network	97.10%	**0.941**	97.31%	0.954	0.21%	0.013
kNN	97.29%	0.932	97.37%	0.953	0.08%	0.021
最大值	97.46%	0.941	97.53%	0.959		
最优算法	J48	Neural Network	J48	J48		

10.waveform-5000 数据集

数据规模: 205000	基础方法	基础方法	Bagging	Bagging	变化量	变化量
挖掘算法	Accuracy	ROC	Accuracy	ROC	Accuracy	ROC
J4.8 (C4.5)	75.08%	0.83	81.20%	0.949	6.12%	0.119
Naïve Bayes	80%	0.956	79.98%	0.956	-0.02%	0
SVM	**86.42%**	0.898	**86.02%**	0.939	-0.40%	0.041
Neural Network	83.56%	**0.963**	85.68%	**0.969**	2.12%	0.006
kNN	73.62%	0.802	74.46%	0.9	0.84%	0.098
最大值	86.42%	0.963	86.02%	0.969		
最优算法	SVM	Neural Network	SVM	Neural Network		

11.结果综合分析

根据上述结果，可以看出

4.3 归一化SVM的对比

4.4 kNN算法的调优

5.实验结果

6.实验感悟

7.参考链接