数据挖掘课程
作业三

刘扬 171850524

1.问题描述

在给定数据集上分别使用决策树、朴素贝叶斯、支持向量机、神经网络和 k 近邻的方法以及它们的 Bagging 版本进行分类，以正确率和 AUC 值作为指标比较不同方法的性能，并讨论如何提升 k 近邻 Bagging 版本的性能.

2.数据集描述

给定十个数据集， 其中包括 3 个不平衡的数据集和一个多分类数据集． 每个数据集包含的样本数量和属性个数均不同.

3.实验流程

使用 Weka3.8.4 对数据进行分类.

3.1 决策树

使用 weka 自带的 J48 分类器，参数使用默认值

单个模型

数据集	正确率	AUC
breast-w	94.6%	0.955
colic	85.3%	0.813
credit-a	86.1%	0.887
credit-g	70.5%	0.639
diabetes	73.8%	0.751
hepatitis	83.9%	0.708
mozilla4	94.8%	0.954
pc1	93.3%	0.668
pc5	97.5%	0.817
waveform-5000	75.8%	0.830

集成版本

数据集	正确率	AUC
breast-w	96.3%	0.985
colic	85.6%	0.864
credit-a	86.8%	0.928
credit-g	73.3%	0.753
diabetes	74.6%	0.798
hepatitis	83.9%	0.865
mozilla4	95.1%	0.976
pc1	93.6%	0.855
pc5	97.5%	0.959
waveform-5000	81.2%	0.949

3.2 朴素贝叶斯

使用 weka 自带的 NaiveBayes，参数使用默认值

单个模型

数据集	正确率	AUC
breast-w	96.0%	0.986
colic	78.0%	0.842
credit-a	77.7%	0.896
credit-g	75.4%	0.787
diabetes	76.3%	0.819
hepatitis	84.5%	0.860
mozilla4	68.6%	0.829
pc1	89.2%	0.650
pc5	96.4%	0.833
waveform-5000	80.0%	0.956

集成版本

数据集	正确率	AUC
breast-w	95.9%	0.989
colic	78.0%	0.842
credit-a	77.8%	0.896
credit-g	74.8%	0.787
diabetes	76.6%	0.817
hepatitis	85.8%	0.890
mozilla4	68.7%	0.830
pc1	88.9%	0.628
pc5	96.5%	0.845
waveform-5000	80.0%	0.956

3.3 支持向量机

使用 weka 自带的 SMO 分类器，参数使用默认值

单个模型

数据集	正确率	AUC
breast-w	97.0%	0.968
colic	82.6%	0.809
credit-a	84.9%	0.856
credit-g	75.1%	0.671
diabetes	77.3%	0.720
hepatitis	85.2%	0.756
mozilla4	83.2%	0.838
pc1	93.0%	0.500
pc5	97.2%	0.541
waveform-5000	86.7%	0.932

集成版本

数据集	正确率	AUC
breast-w	97.0%	0.975
colic	84.0%	0.868
credit-a	85.2%	0.888
credit-g	75.4%	0.754
diabetes	77.5%	0.747
hepatitis	85.8%	0.828
mozilla4	83.1%	0.849
pc1	93.1%	0.512
pc5	97.2%	0.572
waveform-5000	86.3%	0.954

3.4 神经网络

使用 weka 自带的 MultilayerPerceptron，参数使用默认值

单个模型

数据集	正确率	AUC
breast-w	95.3%	0.986
colic	80.4%	0.857
credit-a	83.6%	0.895
credit-g	72.1%	0.734
diabetes	75.4%	0.793
hepatitis	80.0%	0.823
mozilla4	91.2%	0.940
pc1	93.6%	0.723
pc5	97.1%	0.941
waveform-5000	83.6%	0.963

集成版本

数据集	正确率	AUC
breast-w	96.0%	0.989
colic	84.5%	0.876
credit-a	84.9%	0.908
credit-g	76.1%	0.776
diabetes	76.8%	0.822
hepatitis	84.5%	0.846
mozilla4	91.3%	0.945
pc1	93.3%	0.835
pc5	97.3%	0.954
waveform-5000	85.7%	0.969

3.5 k 近邻

使用 weka 自带的 IBK，参数使用默认值

单个模型

数据集	正确率	AUC
breast-w	95.1%	0.973
colic	81.3%	0.802
credit-a	81.2%	0.808
credit-g	72.0%	0.660
diabetes	70.2%	0.650
hepatitis	80.6%	0.653
mozilla4	89.0%	0.877
pc1	92.1%	0.740
pc5	97.3%	0.932
waveform-5000	73.6%	0.802

集成版本

数据集	正确率	AUC
breast-w	95.9%	0.987
colic	81.3%	0.824
credit-a	81.3%	0.886
credit-g	72.1%	0.694
diabetes	71.1%	0.725
hepatitis	81.3%	0.782
mozilla4	88.9%	0.928
pc1	91.1%	0.793
pc5	97.4%	0.953
waveform-5000	74.5%	0.900

3.6 集成版本的 kNN 方法参数优化

bagging 的 kNN 在 diabetes 数据集上表现最差， 选择此数据集进行参数优化.

对于 kNN 的距离度量函数、k 值和 Bagging 抽样构成的数据集大小三个参数就行调整.

距离度量函数

方法	正确率	AUC
ChebyshevDistance	71.2%	0.730
EuclideanDistance	71.1%	0.725
FilteredDistance	69.0%	0.696
ManhattanDIstance	69.8%	0.717
MinkowskiDistance	71.1%	0.725

k值

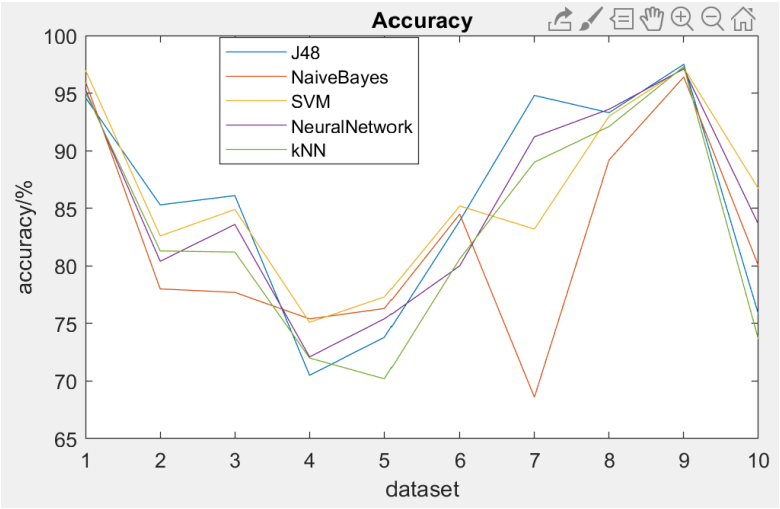
k值	正确率	AUC
1	71.2%	0.730
2	72.9%	0.755
3	73.4%	0.769
4	72.8%	0.770
5	73.0%	0.772
6	72.1%	0.777
7	72.7%	0.777
8	72.5%	0.779
9	73.4%	0.782
10	72.4%	0.785
11	73.2%	0.787
12	72.9%	0.791
13	73.7%	0.793
14	73.3%	0.793
15	73.2%	0.794
16	73.6%	0.796
17	73.6%	0.797
18	73.7%	0.798
19	73.4%	0.797
20	73.3%	0.798
21	73.2%	0.799
22	73.0%	0.798
23	72.7%	0.798
24	73.0%	0.798

数据集大小

大小	正确率	AUC
100%	71.2%	0.730
90%	70.4%	0.732
80%	70.7%	0.731
70%	70.7%	0.740
60%	71.4%	0.746
50%	71.2%	0.751
40%	71.6%	0.758
30%	72.0%	0.764
20%	71.9%	0.747
10%	70.6%	0.746

4.实验结果分析

4.1 不同模型在不同数据集上的正确率

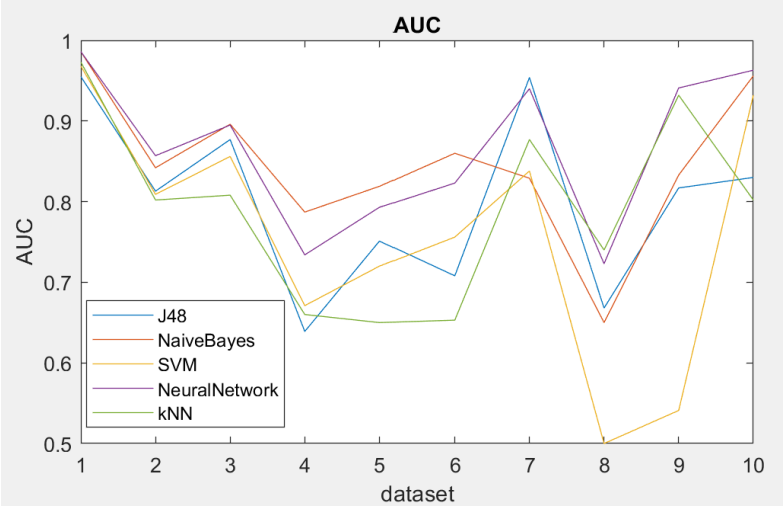


(1)不同模型在不同数据集上的正确率趋势大致相同.

(2)综合所有给定的数据集来看，表现最好的是 J4.8，其次是 SVM;
表现最差的是朴素贝叶斯方法，其次是神经网络.

(3)比较特殊的是朴素贝叶斯模型在第 7 个数据集上的正确率明显低于其他模型. 第 7 个数据集中的属性 start 与 end 之间并不相互独立，不满足朴素贝叶斯的假设，所以朴素贝叶斯效果明显比其他模型差.

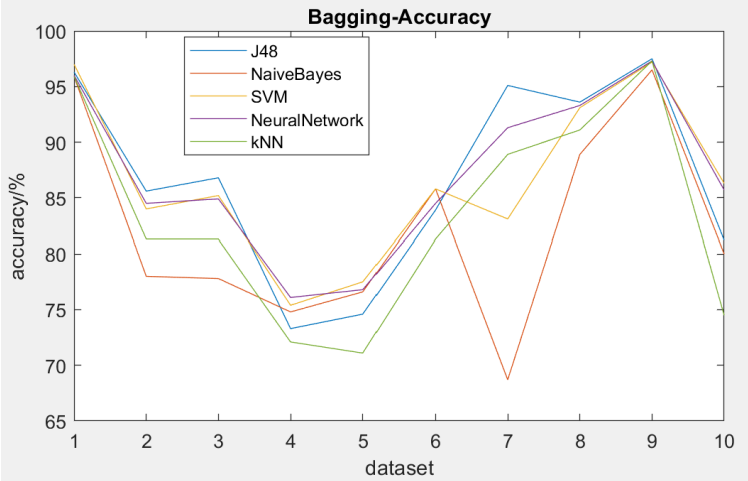
4.2 不同模型在不同数据集上的 AUC



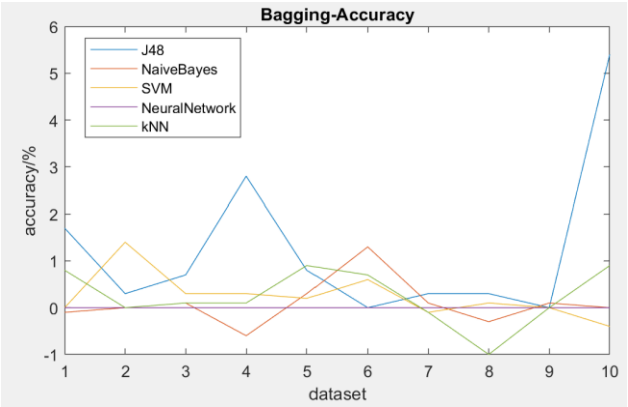
(1)在所有给定的数据集上，表现较好的是朴素贝叶斯和神经网络;
表现较差的是 SVM 与 kNN.

(2)观察到 SVM 在第 8 9 个数据集上 AUC 值很小，在第 8 个数据集上 AUC 值为 0.5，这意味着 SVM 对于这个数据集不起任何作用. 这两个数据集都是非常不平衡的数据集，SVM 只需要把所有样本都判定为样例比较多的类，就可以达到 90%以上的正确率，实际上 SVM 并没有真正达到分类的效果.

4.3 不同模型的 Bagging 在不同数据集上的正确率



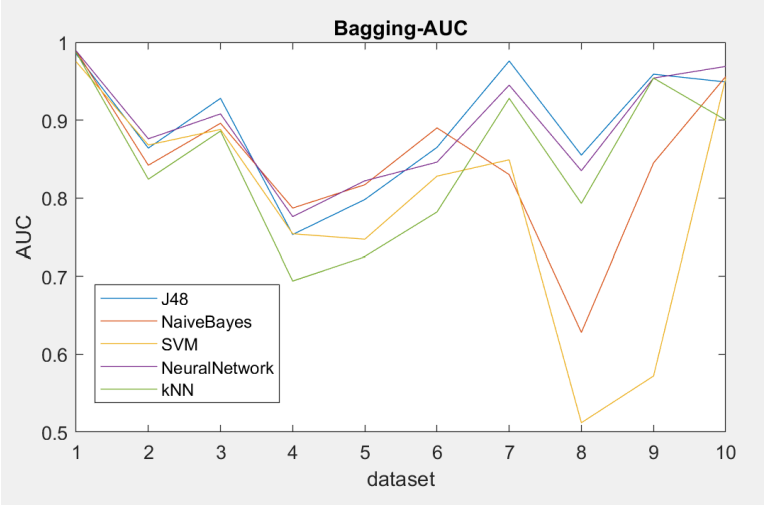
Bagging 与原模型正确率的差值：



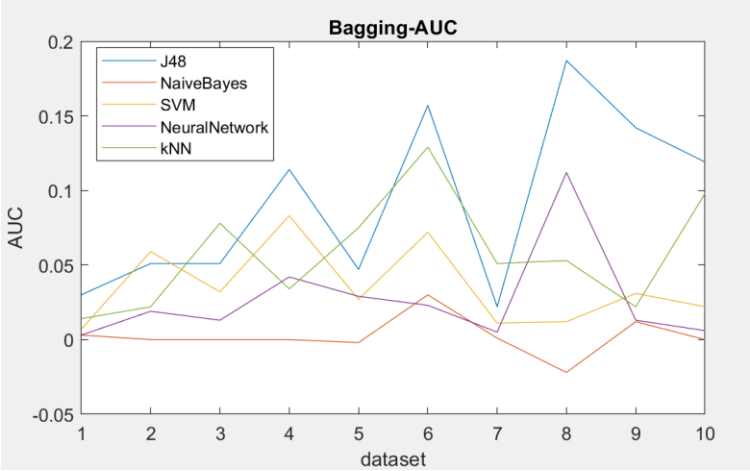
(1)对比不使用 Bagging 的模型， Bagging 对于正确率提升并不明显， 尤其是在不平衡的数据集上更多的会起到相反的效果.

(2)Bagging 对 J4.8 的影响最大，在给定的数据集上看， Bagging 不会降低 J4.8 的正确率，且整体上正确率的提升也大于其他几个算法

4.4 不同模型的 Bagging 在不同数据集上的 AUC



Bagging 与原模型 AUC 的差值：



(1)除了朴素贝叶斯方法外，Bagging 可以提高 AUC 值，在不平衡数据集上提升效果较明显.

(2)Bagging 对 J4.8 的 AUC 值提升最为明显.

5.总结

5.1 对不同方法的比较

(1)朴素贝叶斯方法处理属性相关的数据时，正确率较低;

(2)SVM 处理非常不平衡的数据时，AUC 值很低;

(3)Bagging 对于 J4.8 性能提升最明显;

(4)整体上看，集成版本的性能优于非集成版本;

(5)正确率与 AUC 值之间并无显著的正相关性；不平衡数据集可能会产生正确率较高的假象，此时以 AUC 为评价指标更好.

5.2 调整参数提升 kNN 集成版本的性能

(1)不同数据集最适合的距离度量函数不同，要结合数据分布与实验验证选择最合适的距离度量函数;

(2)适当降低抽样组成的新的数据集的大小可以提升 Bagging 的性能;

(3)随着 k 值增加，kNN 分类器的性能先上升后降低;

6.参考资料

https://blog.csdn.net/smilehehe110/article/details/53463650
https://zhuanlan.zhihu.com/p/41255464
https://www.openml.org/f/5300