KNN 近邻算法实验报告

王宗威

摘要：KNN 算法是一种基本分类与回归方法，是著名的模式识别统计方法，是最好的文本分类算法之一，在机器学习分类算法中占有相当大的地位。此次实验报告就 KNN 近邻算法的原理，算法以及代码实现结果进行汇报。

关键词：KNN 近邻算法；统计学习方法；分类；回归。

1. 原理：

给定一个训练数据集，对新的输入实例，在训练数据集中找到与该实例最临近的 k 个实例，这 k 个实力的多数属于某个类，就把该输入实力分为这个类。

2. 算法

输入：
训练数据集 T1={（x1，y1），（x2，y2），…，（xn，yn）}
测试数据集 T2={（**x1，y1**），（**x2，y2**），…，（**xn，yn**）}
令 K = 5
输出：测试数据集中所有实例 Xn 对应的类的 yn 的集合

（1） 根据给定的距离度量，在训练集 T 中找出与实例 x 最临近的 k 个点，涵盖这 k 个点的 x 的邻域记作 Nk（x）；
（2） 在 Nk（x）中根据多数表决规则决定 x 的类别 y
（3） 计算准确率，精确率，召回率和 F1

3. 实验结果

实验输出：
训练集数据分布为：
{'Iris-virginica': 34, 'Iris-setosa': 31, 'Iris-versicolor': 39}

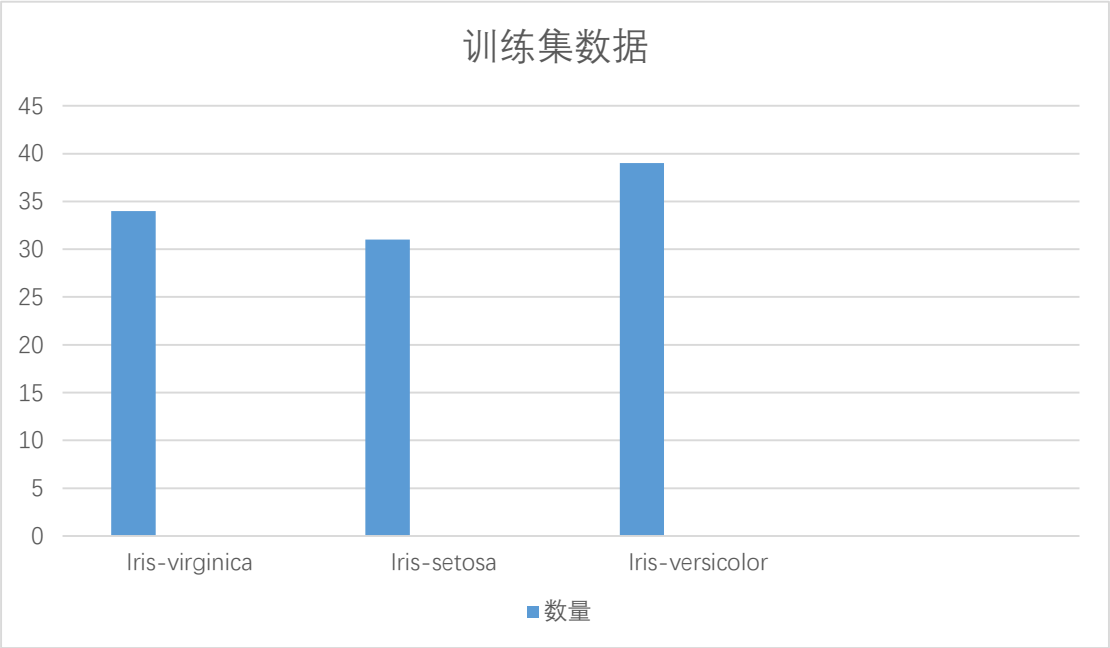

图 1

测试集数据分布为：

{'Iris-virginica': 17, 'Iris-setosa': 19, 'Iris-versicolor': 10}

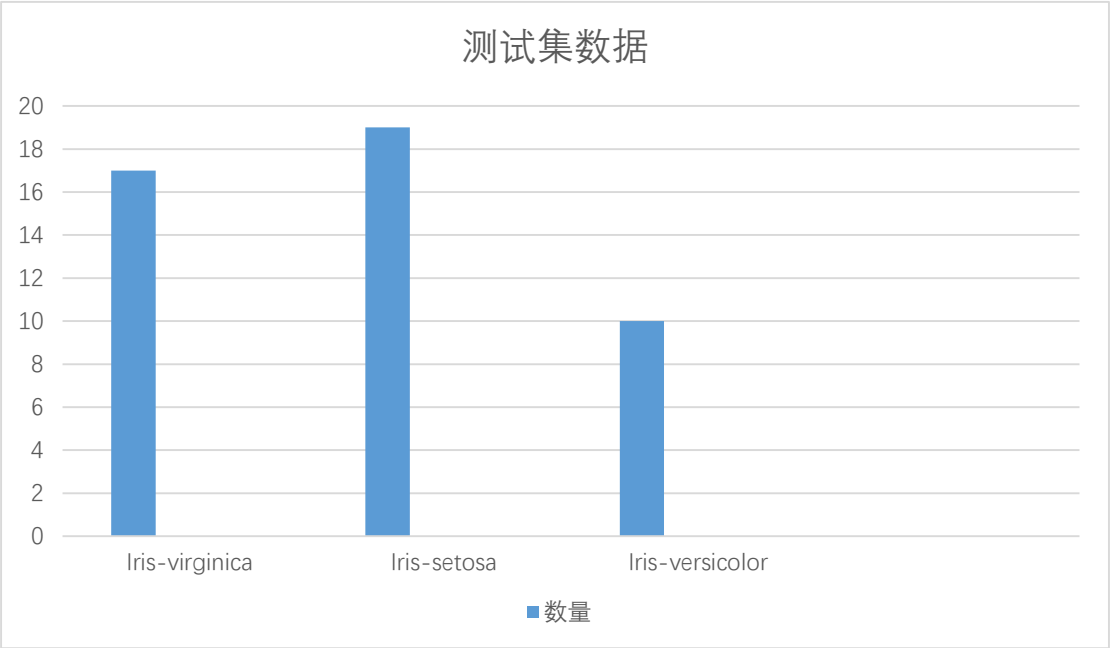

图 2

精确率，召回率，F1：
P=0.823，R=1.000，F1=0.933

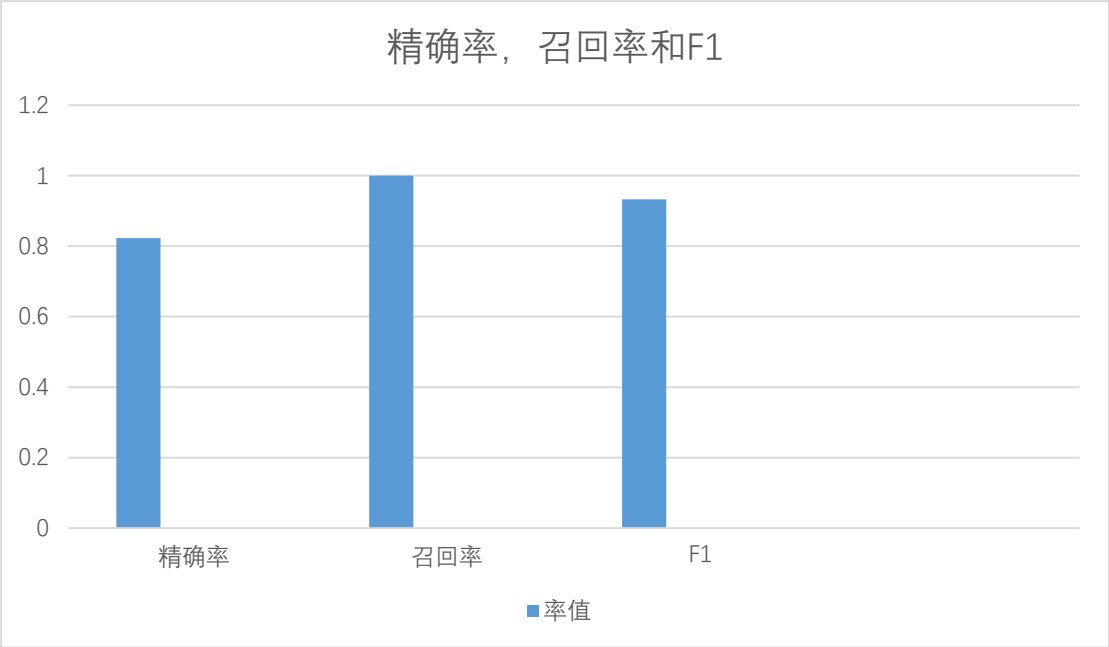

图 3

4. 结果分析

由于该数据集数据较少，在最后的精确率，召回率和 F1 值的表现上来看，KNN 近邻算法的表现还不错。