Mini KDD Cup Report

黄一夫
软件学院
华东师范大学
上海, 中国
10092510437@ecnu.cn

***摘要*—Mini KDD Cup 是用所给的训练数据建立一定的分类器，然后再将该分类器运用到测试数据上来预测类标的过程。分类问题是一个普遍存在的问题，在数据挖掘，信息检索，机器学习等领域有着深厚的理论基础和广泛的实际应用。本报告中先对训练数据进行统计分析，然后采用特征选取，离散化等方式进行数据预处理，再通过测试从大量分类模型选取出性能较好的数种模型，最后采用投票原则对测试数据进行预测。通过实验我们发现，合适的数据预处理可以提升分类器的准确率，投票策略能得到鲁棒性更高的预测结果。**

关键词-Mini KDD Cup，分类，报告。

I. 介绍

分类任务就是确定对象属于哪个预定义的目标类。分类问题是一个普遍存在的问题，有许多不同的应用。例如：根据电子邮件的标题和内容检查出垃圾邮件，根据核磁共振扫描的结果区分肿瘤是恶性的还是良性的，根据星系的形状对它们进行分类[1]。

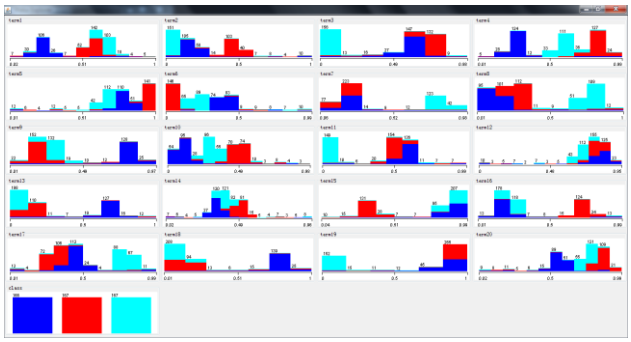
Mini KDD Cup 是用所给的训练数据建立一定的分类器，然后再将该分类器运用到测试数据上来预测类标的过程。实验中我们先对训练数据进行统计分析，总体上掌握数据的分布；然后采用特征选取，离散化等方式进行数据预处理，以此来提高分类模型的准确率；再通过测试大量的分类模型，从中选取出性能较好的数种，如 Naïve Bayes，KStar 等等；最后采用投票原则，结合之前选取的分类器，对测试数据进行综合预测。

我们对数据格式进行一定的转换，采用了 LIBSVM[2]，WEKA[3]等工具进行实验。通过实验我们发现，合适的数据预处理如特征选取，特征离散化等，可以提升分类器的准确率；投票原则的使用能得到鲁棒性更高的预测结果。

II. 方法和实验

A) 数据预处理

首先将训练数据 train.txt 转换成 arff 格式，导入 WEKA，得到如下的可视化。



从中我们可以得知 0 类 166 例，1 类 167 例，2 类 167 例，数据在各类上分布均匀。再对每个属性的值的范围进行统计如下。

1	0.017	0.996
2	0	0.999
3	0	0.977
4	0.012	0.992
5	0.01	1
6	0	0.99
7	0.058	0.975

8	0.008999999999999999	0.998
9	0.005	0.967
10	0	0.978
11	0	0.993
12	0.001	0.953
13	0	0.999
14	0.02	0.964
15	0.036	0.989
16	0.008	0.985
17	0.007	0.994
18	0.014	1
19	0	1
20	0.015	0.986

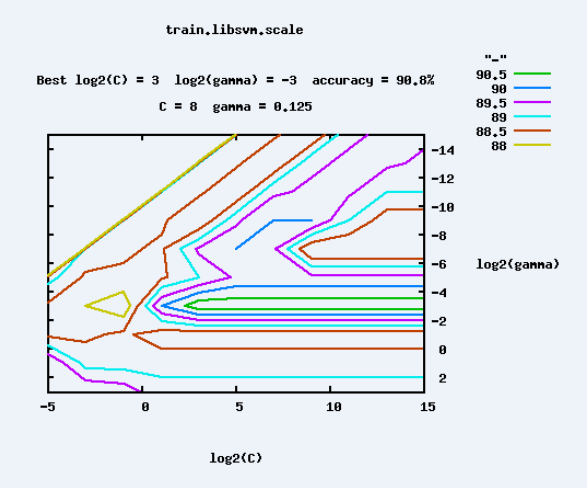
从上述数据中我们可以发现，每个维度的数据的值的范围几乎都在[0,1]，可免除归一化过程。

特征选择，离散化等数据预处理过程参见 C)部分。

B) LIBSVM

考虑到 SVM 具有坚实的统计学理论基础，并在许多实际应用中展示了大有可为的实践效用，我们首先采用了 LIBSVM 对数据进行分类。

我们选用 C_SVCSVM，RBF 核函数，其需要进一步确定的有参数 C 和 gamma。我们采用交叉验证的方式，在训练数据里，对最佳的 C 和 gamma 施行网格搜索。以下是搜索过程中绘制的轮廓图。



从图中可以知道在交叉验证精确度为 90.8%的情况下等到最佳参数 C=8，gamma=0.125.

由此训练出的 SVM 模型中支持向量总数为 86。在各类中的具体分布如下。

	0	1	2
SV	29	28	29
rho	0.106379	-0.51736	0.446619

将该模型运用到 200 例测试数据上去，得到分类结果如下。

	0	1	2
Distribution	64	69	67

从结果中可以看出，预测结果在各类上分布相对均匀，推知预测结果较好。

C) WEKA

在上个步骤中，我们使用 LIBSVM 对训练数据进行了建模并对测试数据进行了预测。但是，考虑到单个分类器能力的局限性，我们在 WEKA 上进一步测试更多的分类器。

首先，我们对属性进行离散化，将各个属性的值从[0,1]分割为数段离散的区间。对于某些属性在离散化之后，其类别在不同区间的分布就相对明显。

然后，我们对属性进行选择，使用过滤器过滤掉多余的属性。

经上两步，得到如下数据。

@attribute term2 {'\'(-inf-0.0865]\'','\'(0.0865-0.374]\'','\'(0.374-0.5915]\'','\'(0.5915-inf)\''}

@attribute term3 {'\'(-inf-0.0905]\'','\'(0.0905-0.2325]\'','\'(0.2325-0.473]\'','\'(0.473-0.649]\'','\'(0.649-0.6595]\'','\'(0.6595-0.7065]\'','\'(0.7065-0.839]\'','\'(0.839-inf)\''}

@attribute term4 {'\'(-inf-0.361]\'','\'(0.361-0.7035]\'','\'(0.7035-0.92]\'','\'(0.92-inf)\''}

@attribute term5 {'\'(-inf-0.4625]\'','\'(0.4625-0.71]\'','\'(0.71-0.7405]\'','\'(0.7405-0.8945]\'','\'(0.8945-inf)\''}

@attribute term6 {'\'(-inf-0.1085]\'','\'(0.1085-0.305]\'','\'(0.305-0.442]\'','\'(0.442-inf)\''}

@attribute term8 {'\'(-inf-0.0155]\'','\'(0.0155-0.1895]\'','\'(0.1895-0.2155]\'','\'(0.2155-0.3875]\'','\'(0.3875-0.581]\'','\'(0.581-0.9485]\'','\'(0.9485-inf)\''}

@attribute term10 {'\'(-inf-0.1955]\'','\'(0.1955-0.392]\'','\'(0.392-0.5795]\'','\'(0.5795-inf)\''}

@attribute term16 {'\'(-inf-0.1975]\'','\'(0.1975-0.525]\'','\'(0.525-0.821]\'','\'(0.821-inf)\''}

@attribute term17 {'\'(-inf-0.109]\'','\'(0.109-0.376]\'','\'(0.376-0.6955]\'','\'(0.6955-0.895]\'','\'(0.895-inf)\''}

从中可以发现 9 个属性 2，3，4，5，6，8，10，16，17 被保留，并且每个属性按照分布特性进行了相应的离散化。

考虑到单一分类器构建模型的局限性，选择投票的方法；选择了精确度较高，不同类别的 10 个分类器；一同投票决定最后分类结果，这里选用 Average of probability 的方式，没有选择 Majority vote 的方式是因为在实验中其准确率较低，而且直观感觉用概率方式给出的结果更加精准。

在 10 个分类器投票方式下，以单一分类器准确率高低为原则，逐步减少参与投票的分类器个数；在多种组合实验下，选择了准确率等参数相对较优的以下分类器组合：

weka.classifiers.bayes.NaiveBayes

weka.classifiers.lazy.KStar -B 20 -M a

weka.classifiers.lazy.LBR

weka.classifiers.rules.DecisionTable -X 1 -S "weka.attributeSelection.BestFirst -D 1 -N 5"

weka.classifiers.rules.PART -M 2 -C 0.25 -Q 1

weka.classifiers.trees.NBTree

weka.classifiers.trees.SimpleCart -S 1 -M 2.0 -N 5 -C 1.0

建立在这 6 个分类器的基础上，我们进行投票分类。

	0	1	2
Distribution	69	66	65

从结果中可以看出，预测结果在各类上分布相对均匀，推知预测结果较好。

最终的预测结果如递交文档中所示。

III. 结论

本次 Mini KDD Cup 为多类扁平分类问题。实验中，我们首先对数据的分布进行了讨论。然后，使用了 LIBSVM，通过交叉验证，网格搜索的方式来寻求最佳参数，建立 SVM 模型；使用了 WEKA，通过特征选择，特征离散化来提高分类器的精确度，并采用了投票原则。

鉴于分类模型的选取很大程度在于分类器使用的经验和对数据分布的掌握。在未来的工作里，我们会进一步学习使用多种分类器来对实际数据进行分类，以提高对分类器的理解；对低维度的数据，对其进行 MDS 投影到二维空间进行可视化来交互分类，也是我们值得考虑的方法。

引用

[1] Pang-Ning Tan , Michael Steinbach , Vipin Kumar, Introduction to Data Mining, (First Edition), Addison-Wesley Longman Publishing Co., Inc., Boston, MA, 2005

[2] Chih-Chung Chang and Chih-Jen Lin, LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, 2:27:1--27:27, 2011. Software available at http://www.csie.ntu.edu.tw/~cjlin/libsvm

[3] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, Ian H. Witten (2009); The WEKA Data Mining Software: An Update; SIGKDD Explorations, Volume 11, Issue 1.