LibSVM（java版）的使用

0 准备工作

1）下载一个LibSVM，解压之，在libsvm的文件夹下可以看到有多种语言的实现，本篇基于java;

2）了解一下SVM的原理还是非常必要的，不然都不知道参数是啥意思。。我看过一篇SVM入门的博客不错，里面有入门十讲，分享之。

3）要准备好符合LibSVM输入格式的数据文件。

1 LibSVM的使用介绍

0）LibSVM输入格式：label index1:属性值 ;index2:属性值 index3:属性值 ..... indexn:属性值。

其中，label为你的类别号，随便你怎么设置，比如体育类label为0，军事类label为1；

index的目的其实就是标识一下你这个属性值是属于哪个特征的，你可以给它“安倍”或者“三胖”，随便你；

文本分类中属性值是某个特征词的权重值，不能随意改变，值得一提的是，如果你的属性值为0，那么就可以把这个省略，比如index2的属性值为0，那么我就可以写label index1:value index3:value....indexn:value。 你可能会想，这样省略不会造成前后关系改变从而LibSVM不能正确区分是哪一个特征词的属性值吗？index！我们还有index，index就有标识是属于哪个特征词的作用！这样的省略会大大提高LibSVM的速度！

每一篇文章都这样表示，最后弄到一个文件里面去，这就是LibSVM的输入文件了！

在上一篇经过TFIDF赋权值处理写入文件的时候，我们就可以按照这种格式生成文件了。

1）首先导入libsvm.jar包到你的工程，我的jar包在“E:\libsvm\libsvm-3.20\java\libsvm.jar”；

2）在“E:\libsvm\libsvm-3.20\java\”下可以看到有svm_train.java、svm_predict.java、svm_toy.java、svm_scale.java，我们只用到前两个，把它们粘贴到你的工程下。这四个java文件从名字也可以看出来，作用分别是训练、预测、画图、归一化（就是将文件弄成LibSVM需要的格式）；

3）写一段程序调用其svm_train.java、svm_predict.java完成我们需要的工作。在这个过程中我们还需要修改svm_train.java和svm_predict.java的代码以达到我们的需要。下面将详细说明，秉承有图can BB的真理，对于每一步尽量会有附图或代码。

3.1）调用svm_train.java、svm_predict.java，下面是我写的一段程序，我把它精简了一下。

```
public class TrainAndTestByLibSVM {

        //参数设置和满足LibSVM输入格式的训练文本

        public String[] str_trained = {"-g","2.0","-c","32","-t","2","-
m","500.0","-h","0","E:\\test\\train\\IF_IDF\\allTrainVSM.txt"};

        private String str_model =
"E:\\test\\train\\IF_IDF\\allTrainVSM.txt.model";     //训练后得到的模型文件

        private String testTxt = "E:\\test\\test\\IF_IDF\\allTestVSM.txt";

        private String[] str_result = {testTxt, str_model,
"E:\\test\\Res.txt"};

        private static TrainAndTestByLibSVM libSVM = null;

         * 私有化构造函数，并训练分类器，得到分类模型

        private TrainAndTestByLibSVM(){

        public static TrainAndTestByLibSVM getInstance(){

                        libSVM = new TrainAndTestByLibSVM();
```

```
        public void trainByLibSVM(){
                        //训练返回的是模型文件，其实是一个路径，可以看出要求改
svm_train.java
                        <span style="color:#cc0000;">str_model =
svm_train.main(str_trained);</span>
                } catch (IOException e) {
                        // TODO Auto-generated catch block
        public double tellByLibSVM(){
                        //测试返回的是准确率，可以看出要求改svm_predict.java
                        <span style="color:#cc0000;">accuracy =
svm_predict.main(str_result);</span>
                } catch (IOException e) {
                        // TODO Auto-generated catch block
        public static void main(String[] args){
                TrainAndTestByLibSVM tat =
TrainAndTestByLibSVM.getInstance();
                System.out.println("正在训练分类模型。。。。");
                System.out.println("正在应用分类模型进行分类。。。。");
```

上面红色标注（不知道怎么回事，显示的时候变成了<span type color=..>,可以自己在代码里找一下）的是最重要的代码，可以以此作为查看LibSVM源代码的入口。还有一些参数设置，说一下这些参数的意义，

g是gamma值，属于高斯核里面的一个参数，如果是线性核就不必 设置该参数；

c是惩罚值，表征的 是对离群点的重视程度；

t是核函数的类型，2为高斯核，0位线性核；

m是表示你为LibSVM划多少MB的内存供其使用；

h不知道是啥意思，但是h=0时训练速度会变快，个人感觉内部迭代时的一种优化算法。

这些参数的设置怎么也得看看SVM的原理。。。这是一个痛苦的过程，此处应该响起《二泉映月》的音乐。。。。

但是，svm_train和svm_predict的的main函数是没有返回值的，为了达到我们的要求，必须修改svm_train和svm_predict的代码！！！

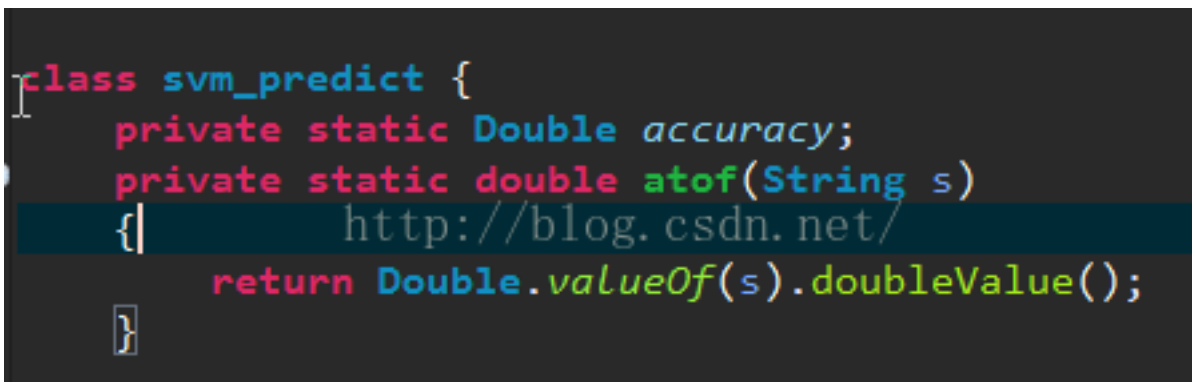
3.2）修改svm_train.java和svm_predict.java的代码

打开svm_train.java的代码找到main函数，把红色部分的代码添加或修改！

```
public static String main(String argv[]) throws IOException{
                svm_train t = new svm_train();
                <span style="color:#cc0000;">return model_file_name;
</span>
```

打开svm_predict.java的代码，这一部分需要改动的略多一点点，看图把不一样的代码添加或修改！

添加部分一：



添加部分二：

```
        else
        {
            System.out.print("Accuracy = "+(double)correct/total*100+
                  "% ("+correct+"/"+total+") (classification)\n");
            accuracy = (double)correct/total;
        }
    }

    private static void exit_with_help()
```


修改部分三：

main函数的返回类型要改成Double型；在main函数的末尾加上“return accracy；”。

经过上面的这些步骤之后，LibSVM就可以使用到我们的java工程下了！

3 利用LibSVM的Python工具找到最优参数

1）下载并安装Python，无需多言；

2）在下载的LibSVM文件夹下有一个tools文件夹，里面有一些Python的工具。查看Readme文档，可以找到工具的相应用法。我们使用grid.py文件来查找最优参数，其原理是对训练文件进行指定折数的交叉检验；

3）为了运行这个文件，还需要下载gnuplot的画图软件；

4）准备完成，在运行命令之前，在控制台cd进入tools文件下，最后运行下面的命令：

python grid.py -gnuplot F:\gnuplot463\gnuplot\bin\gnuplot.exe -v 10 -m 500 E:\test\train\IF_IDF\allTrainVSM.txt

-gnuplot后面是其安装地址；

-v后面表示的要做几折的交叉检验；

-m还是表示划分的内存；

最后是满足LibSVM输入格式的输入文件。

5）就等着吧。。。5折跑个半天，10折跑个一天。。。Python就是慢，Java是世界上最好的语言！笑cry。。。

用的愉快！

宿舍的呼喽声、磨牙声已经此起彼伏了。。上床睡觉！