Wind Predict

姓名：崔诗颖 日期：2017/03/31

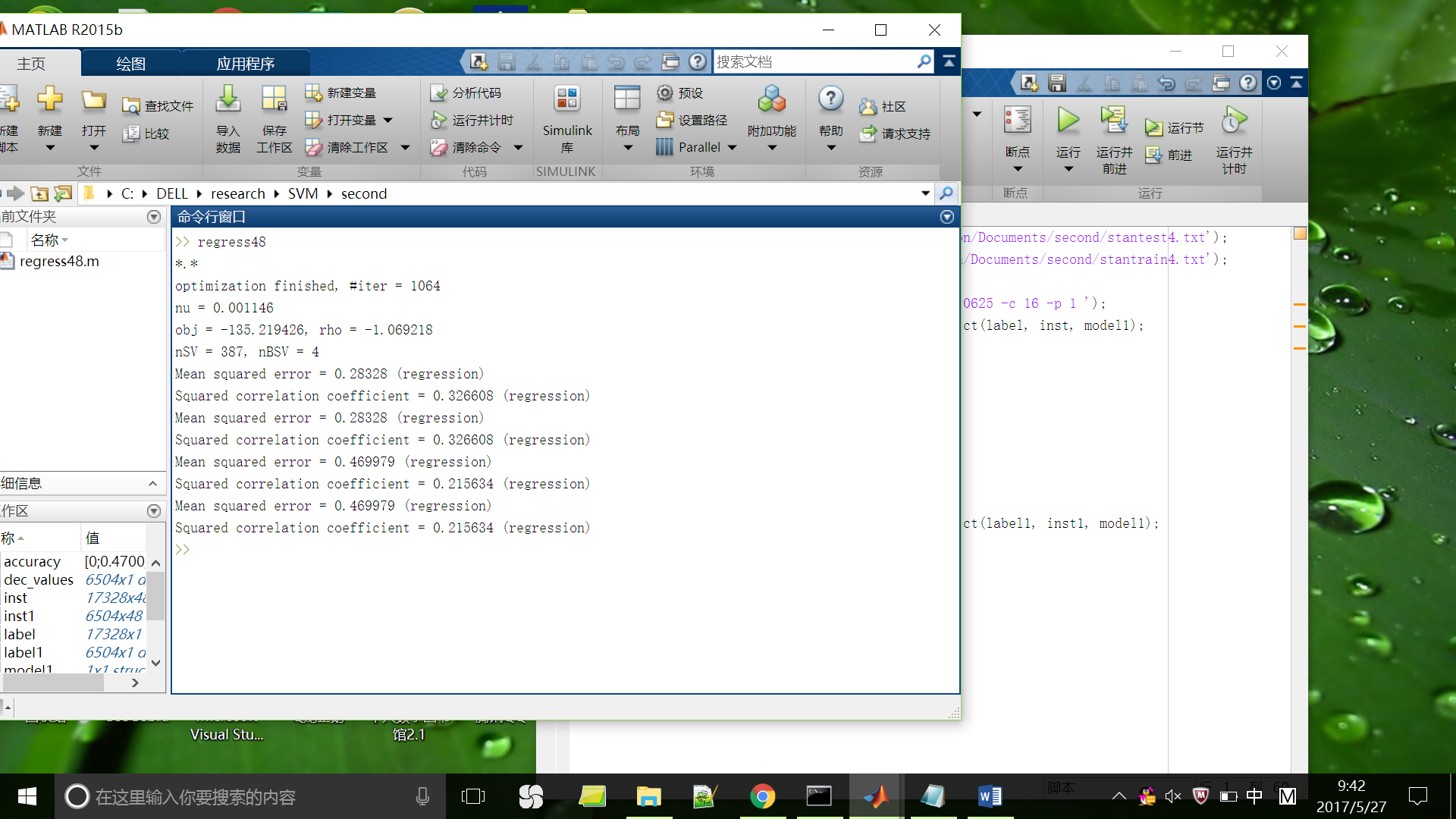
1 工作进展

本次工作分为下列3个实验。

* 1. 风速时间分区改为将一年分为4个风季

查阅了中国风检验标准后，将中国风速时间跨度变为其4个风季区，现对一月至三月进行分析，此时的训练集矩阵为17328\*48，测试集矩阵为6504\*48.

1. 在修改季风跨度区间后，分别做分类与回归的预测。对回归进行探索确定标准化区间为[0,5],此时的方差最为明显。放入SVM中进行验证，标准化区间为[0,5]的均方根误差变大（由均方根误差计算公式，误差的无量纲化应除以标准化区间的比例），但可解释度效果好，表现为训练集与测试集的可解释度均为30%左右（优于训练集80%，测试集0.8%的情况—严重过拟合）。



观察nSV和nBSV的个数与迭代次数，能确定数据集中存在大量聚类现象，不适合用回归分析来预测。其与以一月为跨度的回归SVM共用同一种季风气候规律，但准确率相差的原因是因为，一个月的跨度风速相差不大，整体分布相似，而三个月的跨度中规律不同分布。据以上，可确定此问题应该是一个分类问题。

1. 采用多分类，考虑到之后要做时间序列上的预测，故不能跳过极端数据（会打断时间点的连续性）。此时得到的准确率较低，为56%左右，其原因为：极端数据所造成的极端不平衡。

用五折寻优找最优参数，C为4，g为0.0625. 标准化区间仍与回归一致，为0到5之间。

平衡数据时用简单的过抽样方法（即重复复制）时，发现不能提高准确率， 查阅文献，SVM算法对重复数据具有筛选性。

* 1. 数据平衡对多分类SVM

1. 对极端不平衡数据

小于3m/s（此分类为1）与大于24m/s（此分类为5）的数据及其少，占训练集总数的1.4%

对于这种数据，处理方法参见文献 Extreme Re-balancing for SVMs: a case study. 采用的是加权平衡的方式，给不同的类别加权。这篇文章针对的是二分类，用同样的加权原理对5分类进行加权处理，代入SVM中运行。

结果： 在训练集的准确率上为89.91%，与处理前比没有太大区别，但是在预测时，准确率提升到70.1488%，当用confusionmat时，能看到第1类与第5类的准确率从原来的0%提升到53% 【此处还要再想办法改进】，而第四类的准确率从原来的13%提升到64%。

此时的处理方法是整体粗略的，会在这个处理方法的基础上进一步平衡数据，提高准确率。

1. 对普通不平衡数据

区间类别为4的数据在所有数据中占比为11%，与类别为2与3的数据比例分别约为1：3与1：5，文献证明，SVM的分类比例倾斜为1：5时，已能显著影响分类准确率。

此时保持（1）中所加的权重不变，有两种方法探讨：

1. SVM Learning from Imbalanced Data by GA Sampling for Protein Domain Prediction

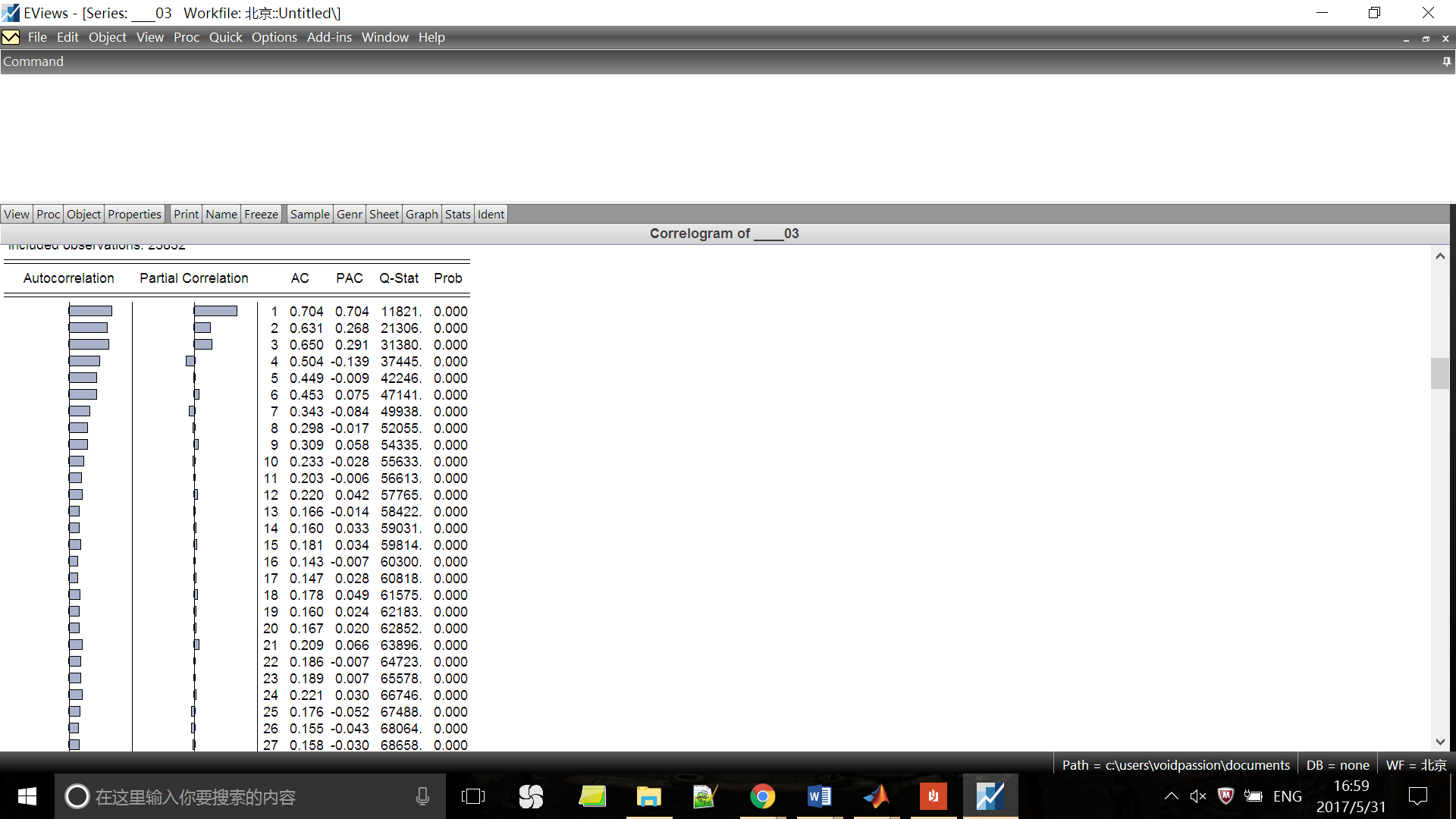
用遗传算法AUC拟合，反复迭代SVM【还未训练，需要放到超算里】

1. Multi-category Support Vector Machines: Theory and Application to the Classification of Microarray Data and Satellite Radiance Data

从理论上计算hinge loss function，在网上没有找到相关的算法【还未做】

* 1. 时间序列初步

用Eviews软件研究各个变量间的关系：



由时间序列分析知，没有截尾特征，自身不平滑，也不属于K阶单整。

用Granger因果关系分析得：风速与其他八个地方的风速、气压有因果解释关系，但是与风向、温度无关。

下面分情况为：

只用当地的预测风速来预测1小时后的风速：

得到结果为： 训练准确率为72%，测试准确率为68%；

而用当地的预测风速及其他8个地方的实时风速与气压来预测一小时后的风速平均值：

得到结果为： 训练准确率为77%，测试准确率为72%；

用当地的预测风速及其他8个地方的实时风速，气压的微分来预测一小时后的风速平均值：

得到结果为： 训练准确率为86%，测试准确率为81%

Conclusion：确定了用Granger因果关系检验中的解释变量作为特征输入。

* 1. 下周计划 [周五考完期末后进行]

1. 将近段时间爬虫的风速数据进行线性插值处理，做好所有输入数据的预处理；
2. 将增量SVM的代码调试好，并用近段时间的各地风速导入增量中看结果；
3. 将数据平衡的两种上述提到的方法试验看结果；
4. 将分类的结果完成后，用到时间序列分析中查看结果，并在时间序列中实现一小时、一天的结果【待定】
5. 将分类算法应用到其他的月份、其他的voronoi元胞区查看结果；