论文标题：

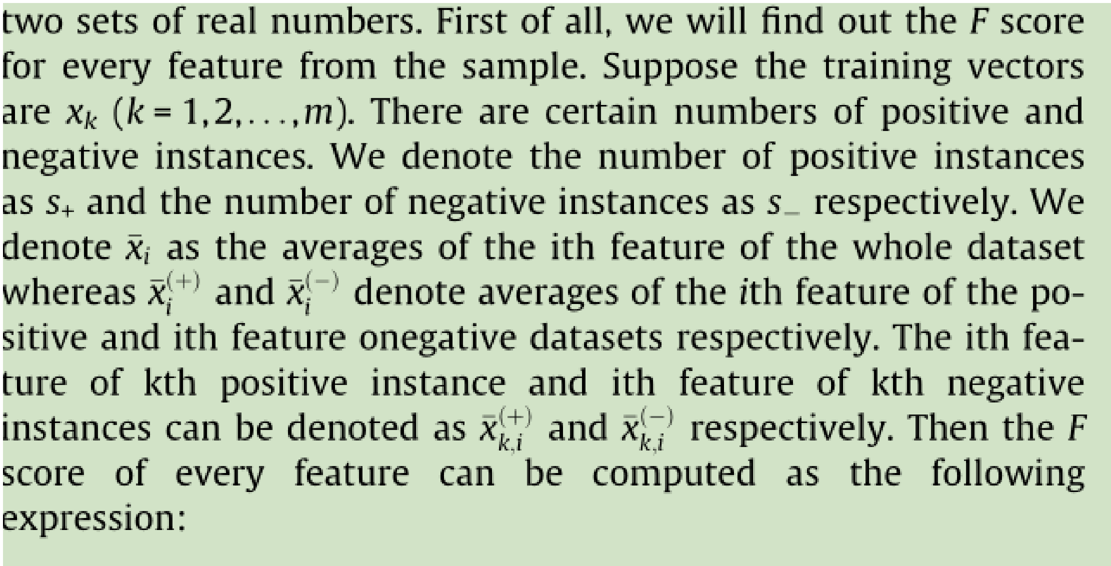
Computational time reduction for credit scoring An integrated approach based on support vector machine and stratified sampling method

论文主要内容：

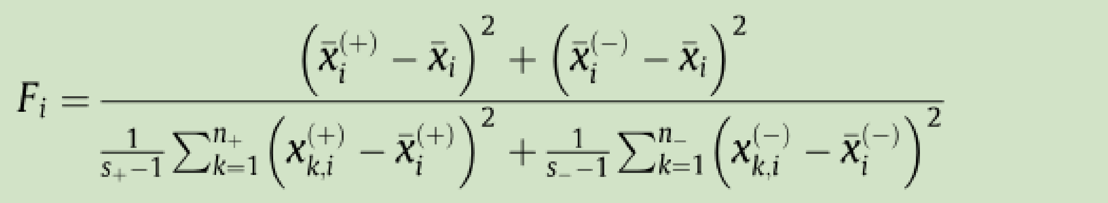
进行了SVM的简要介绍和推导，主要推导了hard-margin的svm，采用了较为常见的推导方式，将问题转化为一个二次规划(quadratic program)的问题，并介绍了svm模型在分类问题上的重要用途。

介绍了什么是分层抽样，并且介绍了分层抽样在现实中的应用及其科学性。

提出了一种进行特征选择的方式，引入了F值的概念，F值的定义如下;



two sets of real numbers. First of all, we will find out the F score for every feature from the sample. Suppose the training vectors are x_k $(k = 1, 2, \ldots, m)$. There are certain numbers of positive and negative instances. We denote the number of positive instances as s_+ and the number of negative instances as s_- respectively. We denote $\bar{x}_i$ as the averages of the ith feature of the whole dataset whereas $\bar{x}_i^{(+)}$ and $\bar{x}_i^{(-)}$ denote averages of the ith feature of the positive and ith feature onegative datasets respectively. The ith feature of kth positive instance and ith feature of kth negative instances can be denoted as $\bar{x}_{k,i}^{(+)}$ and $\bar{x}_{k,i}^{(-)}$ respectively. Then the F score of every feature can be computed as the following expression:



$$F_i = \frac{\left(\bar{x}_i^{(+)} - \bar{x}_i\right)^2 + \left(\bar{x}_i^{(-)} - \bar{x}_i\right)^2}{\frac{1}{s_+-1}\sum_{k=1}^{n_+}\left(x_{k,i}^{(+)} - \bar{x}_i^{(+)}\right)^2 + \frac{1}{s_--1}\sum_{k=1}^{n_-}\left(x_{k,i}^{(-)} - \bar{x}_i^{(-)}\right)^2}$$

在需要维度很大的数据时，为了避免“维数灾难”，常常需要对特征进行筛选，此方法是对于data进行分层抽样，在抽样得出的数据中将各个特征按照F值的进行排序，取出前K个特征并舍弃其余的特征，此方法的目的是保留相关性较小的特征，去掉相关度较大的特征。

论文优点：

能够将如此简单的内容写到8页之长，可见作者的写作功底之深厚。