华南理工大学
South China University of Technology

《网络信息检索》 课 程 论 文

(2018-2019 学年 第一学期)

聚类分析在支持向量机中的应用

学生姓名： 石望华

提交日期：2018 年 12 月 9 日 学生签名：

学 号	201630676843	座位编号	08
学 院	软件学院	专业班级	卓越班
课程名称	网络信息检索	任课教师	张芩
教师评语：			
本论文成绩评定： _____分			

聚类分析在支持向量机中的应用

石望华

摘要：聚类分析与支持向量机是两种经典的用于分类预测的机器学习方法，针对两种方法各自的特点，本文主要分析了在支持向量机中应用聚类分析的方法，通过结合基于K-means 聚类与偏最小二乘法的支持向量机预测模型、基于模糊聚类的完全二叉树支持向量机分类器两个具体应用场景进行讨论与研究，论述了将聚类分析应用于支持向量机的可行性与应用场景。

关键词：聚类分析；支持向量机；K-means 聚类；模糊聚类

0　引言

聚类分析指将物理或抽象对象的集合分组为由类似的对象组成的多个类的分析过程，是一组将研究对象分为相对同质的群组(clusters)的统计分析技术。支持向量机在解决小样本、非线性及高维模式识别中表现出许多特有的优势，并能够推广应用到函数拟合、模式识别、分类和回归分析等其他机器学习问题中。本文综合考虑两种方法各自的优缺点，比较详细地讨论了两个应用了两种方法的具体场景。

1　聚类分析简介

聚类分析指把数据集通过特定方法划分成多个组或簇的过程，簇内的对象具有很高的相似性，但不同簇之间的对象相似性很差，其本质是一种多元统计分析方法。

聚类分析的目标是在相似的基础上收集数据来分类。其基本思想是把全部需要分类的样本尽可能地分到指定的不同的组中，避免相似的样本分到不同的类中。

在机器学习中，分类称作监督学习，因为给定了类标号信息，即学习是受监督的，每个训练元素的类隶属关系是已知的。聚类则称为无监督学习，因为没有提供对象的初始所属类别[1]。

聚类分析主要有两类方法，一种是层次聚类法，一种是非层次聚类法。层次聚类主要包括合并法、分解法、画树状图法；非层次聚类法主要包括划分聚类法、谱系聚类法。传统的聚类方法有系统聚类法、分解法、加入法、动态聚类法、有序样品聚类法、有重叠聚类法和模糊聚类法等[2]。

2 支持向量机原理

支持向量机(Support Vector Machine，SVM)是 Corinna Cortes 和 Vapnik 等于 1995 年首先提出的，它在解决小样本、非线性及高维模式识别中表现出许多特有的优势，并能够推广应用到函数拟合、模式识别、分类和回归分析等其他机器学习问题中[3]。

在机器学习中，支持向量机（SVM，还称为支持矢量网络）是与相关的学习算法有关的监督学习模型，可以分析数据，识别模式，用于分类和回归分析。给定一组训练样本，每个标记为属于两类，形成一个训练集，运用于一个 SVM 训练算法建立的一个模型，然后分配新的实例为一类或其他类，使其成为非概率二元线性分类。举一个 SVM 模型的例子，在空间中的点的映射，使得不同的类别的例子是由一个尽可能宽的明显的差距表示。新的实施例则映射到相同的空间中，并被预测到基于它们落在所述间隙侧上的一个类别。

支持向量机将向量映射到一个更高维的空间里，在这个空间里建立有一个最大间隔超平面。在分开数据的超平面的两边建有两个互相平行的超平面。建立方向合适的分隔超平面使两个与之平行的超平面间的距离最大化。在这一假设下，平行超平面间的距离或差距越大，分类器的总误差越小。这就是支持向量机的原理简述。

支持向量机的关键在于核函数。低维空间向量集通常难于划分，解决的方法是将它们映射到高维空间。但这个办法带来的困难就是计算复杂度的增加，而核函数正好巧妙地解决了这个问题。也就是说，只要选用适当的核函数，就可以得到高维空间的分类函数。在 SVM 理论中，采用不同的核函数将导致不同的 SVM 算法。在确定了核函数之后，由于确定核函数的已知数据也存在一定的误差，考虑到推广性问题，因此引入了松弛系数以及惩罚系数两个参变量来加以校正。在确定了核函数基础上，再经过大量对比实验等将这两个系数取定，一项研究就基本完成，适合相关学科或业务内应用，且有一定能力的推广性。当然误差是绝对的，不同学科、不同专业的要求不一。

对于线性问题[4]，SVM 算法的目的是得到最优超平面，将两类样本无错误地区分并且使两类的分类间隔最大。设分类面的方程为 $<w,x>+b=0$，为使样本集 $(x_i,y_i),i=1,2,...,n,x\in R^n,y\in\{+1,-1\}$ 满足：

$$y_i[(w\bullet x_i)+\mathrm{b}]-1\geq 0,\mathrm{i}=1,2,...,\mathrm{n} \qquad (1)$$

此时的分类间隔为 $\rho=\frac{2}{\|\mathrm{w}\|}$，分类间隔最大等价于 $\|\mathrm{w}\|^2$ 最小。其中使式(1)成立的样本叫做支持向量，满足条件(1)且满足 $\|\mathrm{w}\|^2$ 最小的分类面叫做最优分类面。使用 Lagrange 乘子方法解决这个问题，即在约束条件 $\sum_{i=1}^{n}a_iy_i=0$ 和 $a_i\geq 0$（a_i 为 Lagrange 乘子，i=1，2，…，n)下求解目标函数 Q(a)的最大值。

$$Q(a)=\sum_{i=1}^{n}a_i-\frac{1}{2}\sum_{i=1}^{n}\ \sum_{j=1}^{n}a_ia_jy_iy_j(\mathrm{x}_i\mathrm{x}_j) \qquad (2)$$

这是不等式约束下的二次函数寻优问题，存在唯一解。上述问题得到的最优决策函数是：

$$\mathrm{f}(x)=\mathrm{sgn}(\sum_{i=1}^{n}a_i^{*}y_i(x_i\mathrm{x})+b^{*}) \tag{3}$$

在线性不可分的情况下，可以在条件(1)中引入松弛变量$\xi_i \geq 0$成为：

$$\mathrm{y}_i[(wx_i+b]-1-\xi_i\geq 0,i=1,2,...,n \tag{4}$$

这样，将目标改为求：

$$(w,\xi)=\frac{1}{2}\|w\|^2+C\sum_{i=1}^{n}\xi_i \tag{5}$$

且$0\leq a_i\leq C$，其中 C 为惩罚因子，即综合考虑最少错分样本和最大分类间隔，这样就得到了广义最优分类超平面。

对于非线性问题，这里不再详细讨论。

3 应用 1：PM2.5 预测[5]

用基于 K-means 聚类与偏最小二乘法的支持向量机预测模型（K-PLS-LSSVM）可以进行 PM2.5 数据预测。根据 PM2.5 及其影响因子的内在属性，通过聚类建立各类预测模型，并利用偏最小二乘法筛选得到预测因子集，而后将其输入到支持向量机中加以训练，并进行预测将来的指标。

M2.5 浓度预测的 K-PLS-LSSVM 建模步骤如下：

第 1 步：收集包含历史 PM2.5 浓度和气象数据的建模数据。

第 2 步：根据式（6）将 PM2.5 浓度和气象参数分别归一化到[0,1]。

$$x_n=\frac{x-x_{\min}}{x_{\max}-x_{\min}} \tag{6}$$

式中：x、xn 分别为归一化前和归一化后的浓度序列值；$x_{\min}$、$x_{\max}$分别为原序列 x 的最小值和最大值。

第 3 步：利用式（7-11）计算出最佳聚类数 k，然后通过对气象属性聚类间接把 PM2.5 序列分成相似度较高的 k 类，并分别将各类样本数据作为 k 类模型的训练样本。

$$BWP(j,i)=\frac{b(j,i)-w(j,i)}{b(j,i)+w(j,i)} \tag{7}$$

$$\begin{cases} b(j,i)=\min\limits_{1\le k\le c,k\ne j}\left(\dfrac{1}{n_k}\sum\limits_{p=1}^{n_k}\| x_p^{(k)}-x_i^{(j)}\|^2\right) & (8)\\ w(j,i)=\dfrac{1}{n_j-1}\sum\limits_{q=1,q\ne i}^{n_j}\| x_q^{(j)}-x_i^{(j)}\|^2 & (9)\end{cases}$$

$$avgBWP(k)=\frac{1}{n}\sum_{j=1}^{k}\sum_{i=1}^{n_j}BWP(j,i) \quad (10)$$

$$k_{opt}=\arg\max_{2\le k\le n}\{\text{avgBWP(k)}\} \quad (11)$$

式（7）中：b(j,i)为最小类间距离，表示类 j 中的第 i 个样本到其他每个类中样本平均距离的最小值；w(j, i)为类内距离，表示类 j 中的第 i 个样本到类 j 中的其他剩余样本的平均距离。b 和 w 的计算公式如式（8）、（9）所示。

式（8）、式（9）中，k 和 j 为类标；n_k、n_j 分别为类 k、j 的样本数；$x_p^{(k)}$ 为类 k 中的第 p 个样本；$x_i^{(j)}$ 为类 j 中的第 i 个样本。

对式（10）、（11）的解释：BWP 指标值越大，表示类间距离越大，单个样本的聚类效果就越好。显然，当数据集中所有样本的平均 BWP 指标值越大，该数据集的聚类效果也越好，此时对应的聚类数则为最佳聚类数。avgBWP(k)表示数据集聚成 k 类时的平均 BWP 指标值，k_{opt} 为最佳聚类数。

第 4 步：对第 3 步得到的 k 类 PM2.5 序列分别利用偏最小二乘回归法提取主成分，筛选得到各自的最优预测因子集，作为模型输入向量。

第 5 步：将第 4 步得到的最优预测因子集输入各个 LSSVM 模型进行训练，利用粒子群优化算法（PSO）来优化这 k 类 LSSVM 回归模型中的惩罚系数 γ 和核函数 σ^2。

第 6 步：利用已通过检验的模型进行预测，输出结果。

整个过程的模型流程图如图 1 所示。

通过以下指标可以对预测模型的精度进行评估：

平均绝对误差：$$\text{MAE}=\frac{1}{n}\sum_{i=1}^{n}|y_i-\hat{y}_i| \quad (12)$$

平均绝对百分误差：$$MAPE=\frac{100}{n}\sum_{i=1}^{n}\left|\frac{y_i-\hat{y}_i}{y_t}\right| \quad (13)$$

均方根误差：$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(\mathrm{y}_t - \hat{y}_t)^2}$ （14）

式中，y_t 和 $\hat{y}_t$ 分别为真实值和预测值；n 为序列长度。

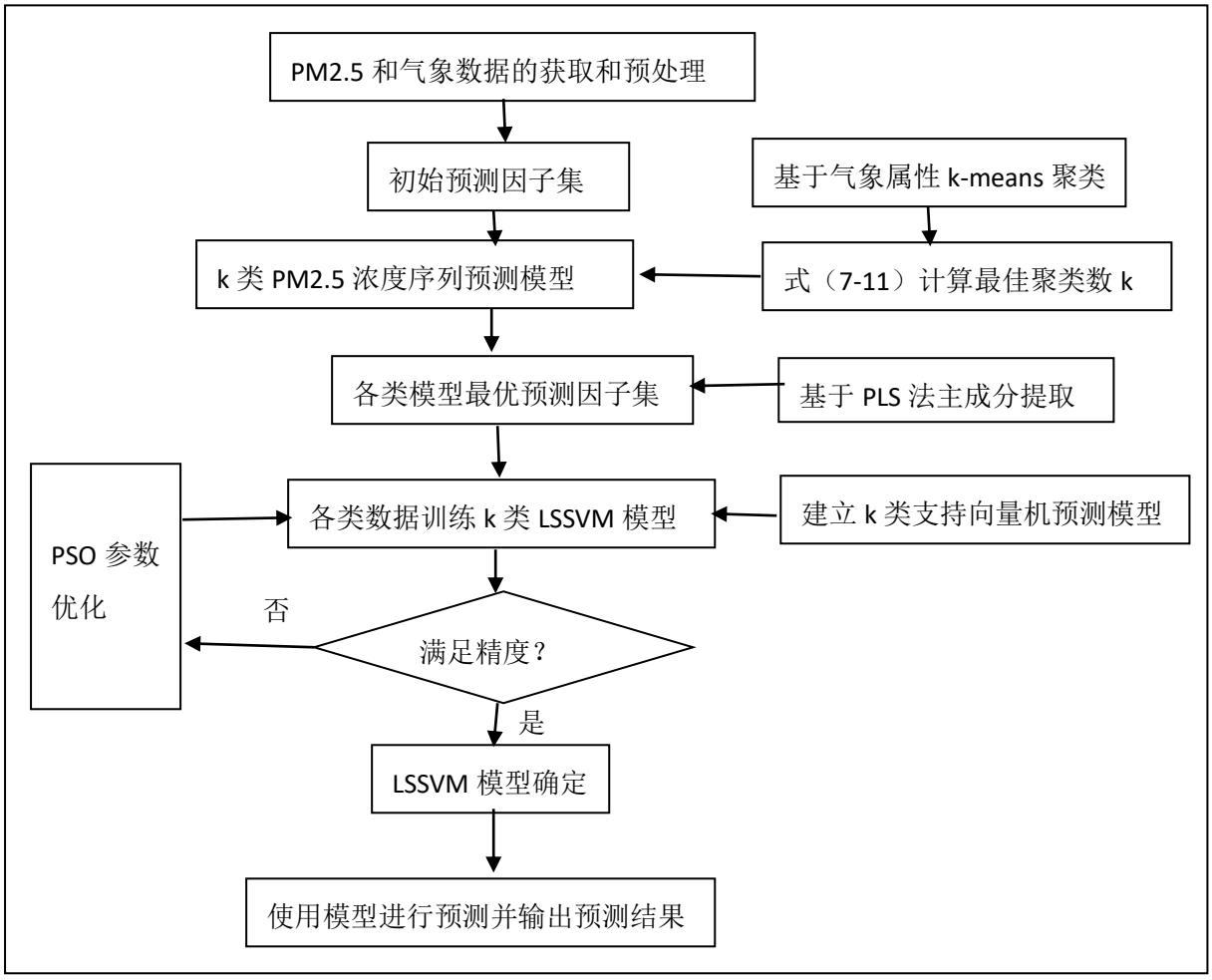

图 1 预测模型流程图

4 应用 2：完全二叉树支持向量机分类器[6]

首先介绍一下模糊聚类分析的概念：根据事物特性指标的模糊性，应用模糊数学的方法确定样本的亲疏程度来实现分类的方法称为模糊聚类分析。模糊 C 均值(Fuzzy C. Means，FCM)是由 Dunn 和 Bezdek 提出的一种聚类算法。模糊 C 均值聚类是从硬 C 均值(Hard C. Means，HCM)聚类算法上发展而来，它对数据采用柔性的模糊划分，对每个数据点属于某个聚类的程度用模糊隶属度描述。该算法首先随机选取若干聚类中心，所有数据点都被赋予对聚类中心一定的模糊隶属度，然后通过迭代方法不断修正聚类中心，迭代过程中以极小化所有数据点到各个聚类中心的距离与隶属度的加权和为优化目标。其基本思路为：将数据集 $X = \{x_1, x_2, ..., x_n\} \in R^{pn}$ 分为 c 类，任意样本 x_k 对 i 类的隶属度

为 μ_{ik}，分类结果用一个模糊隶属度矩阵 $\mathbf{U}=\{\mu_{ik}\}\in\mathbf{R}^{pn}$ 表示，*FCM* 通过最小化关于隶属度矩阵 $\mathbf{U}$ 和聚类中心 $\mathbf{V}$ 的目标函数 $J_m(\mathbf{U},\mathbf{V})$ 来实现：

$$J_m(\mathbf{U},\mathbf{V})=\sum_{i=1}^{c}\sum_{n=1}^{n}(\mu_{ik})^m {d_{ik}}^2(x_k,v_i) \qquad (15)$$

式中，$\mathbf{V}=\{v1,\mathrm{v}2,...,\mathrm{v}_c\}\in\mathbf{R}^{pc}$ 为 c 个聚类中心点集，$m\in[1,\infty)$ 为加权指数。

采用模糊 c 均值聚类和分裂层次法可以构建电力变压器故障诊断完全二叉树，模糊 C 均值聚类对数据采用柔性的模糊划分方法，表达了故障样本所属类的不确定性，反映了各故障性质类间的模糊相似性和关联性。由各类的聚类中心采用分裂层次法可以生成一棵近似完全二叉树，使相似的类聚集在一起，而在诊断时保证了上层节点分类器诊断的准确性，减少误差累积，从而提高诊断的准确率 。

建立完全二叉树支持向量机分类器的思想是：先利用模糊 C 均值聚类求取每类样本聚类中心；再对各聚类中心基于分裂层次法构造一棵二叉树；然后在二叉树的每个结点处，根据聚类中心重新构造学习样本集；最后训练每个节点处的支持向量机子分类器，得到基于完全二叉树的支持向量机分类模型。

建模具体步骤如下：

第一步：对于具有 n 个样本 c 个类别的样本集 $S=\{x_i,y_i\},i=1,2,...,n$，依据物理联系，将其划分为 c 个子集 $\mathrm{S}=\{S_1,S_2,...,S_c\}$，利用模糊 C 均值聚类算法求取每个子集的模糊聚类中心 $C=\{\mathrm{c}_1,c_2,...,c_c\}$。

第二步：利用模糊 C 均值聚类将 $C=\{\mathrm{c}_1,c_2,...,c_c\}$ 聚类成两类 C_p 和 C_N，满足如下条件：$C_p,C_N\subset\mathrm{C}$，$C_p\cap C_N=\phi$，$C_p\cup C_N=C$。

第三步：正类样本集 P_1 由属于 C_p 的各聚类中心对应的训练样本构成，负类样本集 N_1 由属于 C_N 的各聚类中心对应的训练样本构成，二者满足 $P_1\cup N_1=S$，并训练生成支持向量机 SVM_1，形成二叉树顶层根节点。

第四步：将 C_p 聚类成两类，正类样本 P_2 和负类样本 N_2 由各聚类中心所对应的训练样本构成 ，此时 $P_2,N_2\subset P_1,P_2,N_2=\phi,P_2\cup N_2=P_1$；同理，将 C_N 聚类成两类，正类样本 P_3 和负类样本 N_3 由各聚类中心所对应的训练样本构成 ，此时 $P_3,N_3\subset N_1,P_3\cap N_3=\phi,P_3\cup N_3=N_1$。

第五步：由正类 P_2 和负类 N_2 训练生成支持向量机 SVM_2，由正类 P_3 和负类 N_3 训练生成支持向量机 SVM_3。

第六步：重复步骤四和五，直至构造第 c-1 个支持向量机 SVM_{c-1}。

第七步：由步骤 1-6 得到 c-1 个支持向量机 $SVM_1,...,SVM_{c-1}$，由此形成二叉树的中间结点。

通过以上过程建立的完全二叉树支持向量机分类器可以用来做电力变压器故障诊断模型，能很好地解决小样本学习问题，很好解决故障划分模糊、故障信息不明确导致故障识别困难的问题，同时保证了诊断准确度，可用于电力变压器诊断。其中用到的模糊聚类可以反映出各故障性质类间的模糊相似性和关联性。

5 总结

聚类方法有很多种，各有各的分类优缺点，如果能合适的把聚类分析方法应用于支持向量机，科学地构建预测、分类模型，结果往往能得到更高的准确率与更好的效果。

6 参考文献

[1]范佳健.微博评论信息的聚类分析[D].合肥：安徽大学，2017.

[2]w_ou.聚类分析[EB/OL].(2018-7-15)[2018-12-8]. https://baike.baidu.com/item/%E8%81%9A%E7%B1%BB%E5%88%86%E6%9E%90/3450227?fr=aladdin

[3]牵丝戏长.支持向量机[EB/OL].(2018-10-4)[2018-11-23].https://baike.baidu.com/item/%E6%94%AF%E6%8C%81%E5%90%91%E9%87%8F%E6%9C%BA/9683835?fr=aladdin

[4]刘夫成，高尚.基于聚类和支持向量机的个人信誉评估方法[J].信息技术，2013(2):42-47.

[5]喻其炳，李勇，白云，等.基于聚类分析与偏最小二乘法的支持向量机 PM2.5 预测[J].环境科学与技术，2017，40(6):157-164.

[6]李赢,舒乃秋.基于模糊聚类和完全二叉树支持向量机的变压器故障诊断[J].电工技术学报，2016，31(4):64-70.