# 基于SVM的浮动车行驶模式判断模型

诸彤宇 郭胜敏 吕卫锋

(北京航空航天大学 软件开发环境国家重点实验室，北京 100191)

**摘 要**：浮动车在低速情况下存在两种行驶模式，如不能对上述模式进行准确区分，将严重影响浮动车实时路况计算的精度和效率. 研究和设计了一个基于支持向量机(SVM, Support Vector Machine)的浮动车行驶模式判断模型，并针对性地提出了一种简单的基于隶属度矩阵的特征评价和选择方法. 实验表明通过上述方法选择的特征子集所训练的分类器在测试样本集上具有92.6%的分类准确性；经过行驶模式分析后，浮动车系统的准确性有显著提升.

**关 键 词**：浮动车；采样区间；支持向量机；特征提取；隶属度矩阵

**中图分类号**：TP 391.4

**文献标识码**：A **文 章 编 号**：1001-5965(2008)08-0976-05

## SVM based float car driving mode classification model

Zhu Tongyu Guo Shengmin Lü Weifeng

(State Key Laboratory of Software Development Environment, Beijing University of Aeronautics and Astronautics, Beijing 100191, China)

**Abstract**: There are two kinds of driving modes of float car at low speed. The misjudgement of these modes will affect the accuracy and efficiency of the calculation of float car real-time traffic conditions seriously. A SVM(support vector machine) based float car driving mode classification model was studied and designed, and a novel membership matrix-based feature evaluation and selection method was proposed. The classifier whose features are selected through this method made a great classification accuracy of 92.6% in test samples. The float car driving mode analysis enhances the accuracy of exiting system evidently.

**Key words**: float car; sampling interval; support vector machine; feature selection; membership matrix

浮动车技术是近年来先进交通信息系统中所采用的获取实时道路交通信息的先进技术手段之一. 其基本原理是通过装备车载全球定位系统和无线通信系统的浮动车在其行驶过程中发送车辆位置、方向和速度等信息，在处理中心应用相关的计算模型和算法进行处理，最终得到浮动车所经道路的交通拥堵信息. 与传统方法相比，浮动车技术可以获得更精确、覆盖率更高的动态实时交通路况信息，故其日渐成为智能交通领域的热点研究课题.

在实际应用中，通常选择出租车或公交车作为浮动车，而这些车辆所具有的某些特殊行驶特征往往给路况计算[1]的准确性造成消极影响. 如出租车超过40%的时间处在停驶状态或在道旁及宾馆等地待客，系统则会根据此时车辆的行驶速度而做出该区域交通拥堵的误判. 鉴于系统采集的数据中充斥着大量的上述干扰数据，在进行进一步处理之前，对浮动车的行驶状态进行分析，甄别并剔除干扰数据，将极大提高路况信息计算的准确性并提升系统处理的整体效率.

鉴于目前尚无浮动车行驶模式对交通信息准确性及信息处理效率影响方面的研究报告，本文

**收稿日期**：2007-07-23

**基金项目**：国家863基金资助项目(2006AA12Z315)

**作者简介**：诸彤宇(1969—)，男，北京人，副教授，zhutongyu@nlsde.buaa.edu.cn.

在大量数据分析和实践的基础上,设计了一个基于支持向量机(SVM,Support Vector Machine)的浮动车行驶模式判断模型,并针对性地提出了一种基于隶属度矩阵的特征评价和选择方法,实验表明该方法所训练的分类器在测试样本集上具有良好的分类准确性,而且采用本文提出的模型对浮动车的行驶模式进行分析后,系统的准确性和处理效率均有显著提高.

## 1　支持向量机

SVM[2]是建立在统计学习的 VC 维(VC dimension)理论和结构风险最小化(structural risk minimization)原则基础上的. SVM 通过对分类精度(对特定样本的分类正确性)和分类能力(对任意样本进行无错误分类)进行折衷,以期使分类器获得最好的推广能力.

以二分类问题为例,对于训练样本集$(\boldsymbol{x}_i, y_i)$,其中 $i=1,2,\cdots,n$,$\boldsymbol{x}\in\boldsymbol{R}^d$,$y=\{+1,-1\}$. SVM 通过寻找最优分类面方程 $\boldsymbol{x}\cdot\boldsymbol{\omega}+b=0$,不仅确保两类能准确分开,还要使分类间隔最大. 其中线性可分样本集$(\boldsymbol{x}_i, y_i)$需满足

$$y_i(\boldsymbol{\omega}\cdot\boldsymbol{x}_i+b)-1\geqslant 0 \tag{1}$$

最优分类面问题可以转化为约束优化问题,即在式(1)成立的前提下求 $\phi(\boldsymbol{\omega})=\|\boldsymbol{\omega}\|^2/2$ 的最小值,定义 Lagrange 函数

$$L(\boldsymbol{\omega},b,\alpha)=$$

$$\frac{1}{2}\|\boldsymbol{\omega}\|^2-\sum_{i=1}^{n}\alpha_i[y_i\cdot(\boldsymbol{\omega}\cdot\boldsymbol{x}_i+b)-1] \tag{2}$$

其中 $\alpha_i$ 为每个样本对应的 Lagrange 乘子,根据 Kuhn-Tucker 条件

$$\alpha_i[y_i\cdot(\boldsymbol{\omega}\cdot\boldsymbol{x}_i+b)-1]=0 \tag{3}$$

对于大多数样本来讲 $\alpha_i$ 的值为 0,定义 $\alpha_i\neq 0$ 对应的向量 $\boldsymbol{x}_i^*$ 为支持向量,综上得最优分类面权系数向量 $\boldsymbol{\omega}^*$ 和阈值 $b^*$ 分别为

$$\boldsymbol{\omega}^*=\sum_{i=1}^{l}\alpha_i y_i \boldsymbol{x}_i^*$$

其中,$\boldsymbol{x}_i^*$ 为支持向量;$l$ 为支持向量的个数.

$$b^*=-1/2[\boldsymbol{\omega}^*\cdot(\boldsymbol{x}_r^*+\boldsymbol{x}_s^*)]$$

其中 $\boldsymbol{x}_r^*$ 和 $\boldsymbol{x}_s^*$ 为分属两类的任意 2 个支持向量.

最优分类函数为

$$f(x)=\mathrm{sgn}[(\boldsymbol{\omega}^*\cdot\boldsymbol{x})+b^*] \tag{4}$$

对于线性不可分的情况,则可在式(1)中增加松弛变量 $\xi_i$;对于非线性问题,可以通过非线性变换转化为某个高维空间中的线性问题,在高维空间中求解最优分类面.

## 2　基于隶属度矩阵的特征选取

特征选择是分类器设计的一个关键步骤,其基本任务是从已知特征集中选择特征子集,使评价标准最优[3],以上定义可表述如下:

给定一个学习算法 $L$ 和一个数据集 $S$,数据集 $S$ 是具有 $p$ 个特征 $X_1,X_2,\cdots,X_p$ 的具有类别标记 $Y$ 的样本空间,特征选择算法即寻找使得某评价准则 $J=J(L,S)$最优的特征子集$\{X\}_{\mathrm{OPT}}$.

### 2.1　相关工作

在特征空间搜索的过程中,对于新产生的一个特征子集,需要依照某一标准对其进行评价,由评价结果决定下一步搜索方向或终止搜索. 常用的特征评价方法有信息增益、距离、一致性和正确性等.

近年来有学者将模糊集理论引入特征选择,基于模糊熵[4]和隶属度函数[5]的特征选择方法均取得了很好的实验效果. 在上述工作的基础上,本文提出了一种简化的基于隶属度矩阵的特征评价和选择算法.

### 2.2　隶属度矩阵

Zadeh 提出的模糊集理论是对传统集合理论的一种推广:一个元素不再严格地属于或不属于一个集合,而是以一定的程度属于某个集合,同时以不同的程度属于其它集合.

模糊 C-均值聚类[6]算法是模糊集理论的一个典型应用. 算法定义了隶属度矩阵,用来描述每个样本对于不同类别的从属程度,同时在迭代过程中不断修改隶属度矩阵和各类别的聚类中心,以期达到最好的聚类效果. 当算法收敛时,隶属度矩阵体现出样本对不同类别隶属程度具有良好的区分性,即对于单个样本而言,只对某一类别具有较高的隶属度,而对其它类别的隶属度很低. 同时,各聚类中心的分布及聚类结果较其他聚类算法而言要准确.

一个具有良好分类特性的特征集,不仅能够使不同类别的样本能够准确得分开,还应当使各类别样本之间具有足够的区分度. 本文认为,隶属度矩阵兼顾了准确性与区分度的要求,因此将其作为特征选择算法的评价标准,在与分类算法低耦合的同时保持了一定的相关性,且能够提高整个特征选择算法的效率.

### 2.3　算法描述

#### 2.3.1　符号说明

为了更好地描述算法,特对下文涉及的符号作如下说明:

样本空间包含 $p$ 个不同的数据特征，这 $p$ 个数据特征有 $t=\sum_{i=1}^{p}C_p^i$ 种不同的组合方式，表示为 $\Theta=\{\theta_1,\theta_2,\cdots,\theta_t\}$，其中每一个 $\theta$ 称为一个特征子集.

$X=\{x_1,x_2,\cdots,x_n\}$ 是 $n$ 个训练样本组成的样本集，$\boldsymbol{V}(\theta_k)=\{\boldsymbol{v}_{k1},\boldsymbol{v}_{k2},\cdots,\boldsymbol{v}_{kn}\}$，$k=1,2,\cdots,t$，其中 $\boldsymbol{v}_{ki}$ 为选定特征集 $\theta_k$ 时样本 $i$ 对应的特征向量.

$c$ 为预定的类别数目，$\{\boldsymbol{m}_{k1},\boldsymbol{m}_{k2},\cdots,\boldsymbol{m}_{kc}\}$ 分别对应取特征集 $\theta_k$ 时每个类别的中心. $\mu_j(\boldsymbol{v}_i)$ 是第 $i$ 个特征向量对于第 $j$ 类的隶属度. 对单个样本而言，算法要求其对所有类别的隶属度之和为1，即

$$\sum_{j=1}^{c}\mu_j(\boldsymbol{v}_i)=1 \qquad i=1,2,\cdots,n \tag{5}$$

2.3.2 算法步骤

输入：训练样本集 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$，其中每个样本具有 $p$ 个不同的数据特征及人工标注类别；输出：$d$ 个数据特征，使分类效果最优或次优.

算法执行过程如下：

1) 对任意特征子集 $\theta_k$，$k=1,2,\cdots,t$，提取每个样本对应的特征向量 $\boldsymbol{v}_{ki}$，$i=1,2,\cdots,n$；

2) 将特征向量集 $\boldsymbol{V}(\theta_k)$ 按照人工标注类别进行分类，计算各类别中心 $\boldsymbol{m}_{kj}$，$j=1,2,\cdots,c$；

3) 根据下式计算隶属度矩阵：

$$\mu_j(\boldsymbol{v}_{ki})=\frac{(1/\|\boldsymbol{v}_{ki}-\boldsymbol{m}_{kj}\|^2)}{\sum_{l=1}^{c}(1/\|\boldsymbol{v}_{ki}-\boldsymbol{m}_{kl}\|^2)} \tag{6}$$

4) 根据隶属度矩阵分析特征选择的准确率和区分度.

准确率和区分度分别定义如下：

准确率 $\mathrm{cor}(\theta_k)$：单个样本对于每个类别的隶属度各不相同，如果其最大隶属度对应类别与人工标注类别相同，则分类正确，否则分类错误. 准确率为分类正确的样本占所有样本的比例.

区分度 $\mathrm{div}(\theta_k)$：就单个样本而言，如果对某个类别的隶属度与其他隶属度满足

$$\mu_j(\boldsymbol{v}_{ki})-\sum_{\substack{l=1\\l\neq j}}^{c}\mu_l(\boldsymbol{v}_{ki})>0.3 \qquad i=1,2,\cdots,n \tag{7}$$

则认为其具有良好的区分性，否则区分性不佳. 区分度为具有良好区分性的样本占所有样本的比例.

如果存在未处理的特征子集，转1)，否则转5).

5) 对准确率和区分度进行线性加权来表示特征子集的分类效果，定义

$$\mathrm{auth}(\theta_k)=0.7\cdot\mathrm{cor}(\theta_k)+0.3\cdot\mathrm{div}(\theta_k) \tag{8}$$

若 $\mathrm{auth}(\theta_k)$ 大于阈值 threshold，则该特征子集按照 $\mathrm{auth}(\theta_k)$ 的权重对其包含的特征进行投票，最终得票数最多的 $d$ 个特征即为当选的特征.

## 3 浮动车行驶模式判断模型

### 3.1 问题描述

经过对浮动车原始数据的长期观察和分析，本文认为浮动车存在3种行驶模式：正常行驶模式，停驶及待客模式，拥堵及缓慢模式. 根据浮动车记录包含的即时速度信息可以较准确地辨别正常模式，但由于浮动车数据具有一定的误差，在低速情况下区分后两种模式具有一定的困难，需要结合浮动车的经纬度坐标变化，事件触发等其他数据特征，设计适当的分类器来判断.

本文设计了一个浮动车行驶模式判断模型，以期将后两种模式正确区分开来，为干扰数据的剔除提供决策依据. 分类器设计过程为：首先提取数据特征，将原始数据映射到特征空间；然后通过特征选择，对特征空间进行降维，选择对分类最有利的特征组合；最后在低维空间中进行 SVM 分类器的设计. 模型的体系架构如图1所示.

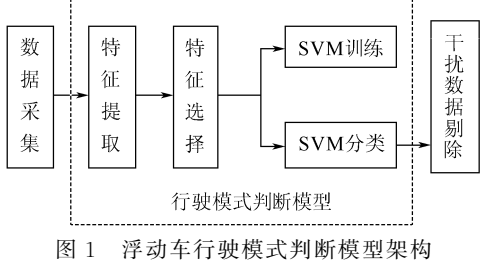



图1 浮动车行驶模式判断模型架构

### 3.2 数据特征提取

为了更好地描述数据，本文提出了采样区间的概念. 采样区间是满足一定条件的若干条浮动车记录的集合，文中其将作为一个整体来提取数据特征，同时采用基于熵的离散化[7]方法进行特征规约.

本文认为离散化的属性值对于分类器设计来讲仍过于粗糙，有必要从中进一步提取有利于分类的二次数据特征. 根据贝叶斯公式，在先验概率和类条件概率密度已知的情况下可以获知特征的后验概率. 对于离散化的每一维数据特征，可以获知其处在既定特征值时采样区间内均为干扰记录的概率. 例如：若在采样区间中发生触发事件，则该采样区间中均为干扰记录的概率为

$$P(\text{干扰}\mid\text{event})=\frac{p(\text{event}\mid\text{干扰})\cdot P(\text{干扰})}{\sum_{\text{state}=\text{干扰/正常}}p(\text{event}\mid\text{state})\cdot P(\text{state})} \tag{9}$$

综上,每一维数据特征对应一个属于[0,1]的后验概率值,每个样本对应一个 $p$ 维的特征向量.特征选择例程从这 $p$ 维特征向量中选取最有利于分类器设计的特征组合.

### 3.3　特征选择

特征选择是分类器设计的一个关键步骤,况且浮动车系统作为一个实时系统,车辆行驶模式判断模型需要以尽可能少的特征实现尽可能好的分类效果.本文在第二章提出了一种模糊策略的特征选取方法,对特征空间进行降维,保障分类器的设计质量.

### 3.4　SVM 分类

经过特征提取和选择后,浮动车模式判断模型基于 SVM 分类技术,在低维特征空间中将浮动车数据分为干扰和正常两类,后续处理根据分类结果将干扰数据剔除,从而提高整个系统的准确性和处理效率.

## 4　实验验证

本文选用 3 935 例训练样本及 2 850 例测试样本.实验分为两部分进行:第一部分考察本文提出的基于隶属度矩阵的特征选择方法,第二部分说明浮动车行驶模式判断模型对于系统准确性的贡献.

### 4.1　数据特征说明

针对采样区间,本文提取了与行驶模式相关的 8 个数据特征(注:由于每条浮动车记录都包含 GPS 经纬度坐标信息,文中也将其称为采样点).

数据特征:①采样点个数,②位移矢量角度平均变化(°),③采样区间开始时间,④最大 GPS 速度(km/h),⑤相邻采样点平均时间间隔(s),⑥最小内切圆半径(m),⑦单位面积内采样点个数,⑧事件触发发生与否.

其中,相邻采样点构成一个位移矢量,相邻位移矢量构成一个夹角,故可定义位移矢量角度的平均变化来反映采样点的飘移程度.最小内切圆半径指各采样点与中心点(所有采样点经纬度坐标平均值)的最远距离.单位面积内分布的采样点个数指采样点个数与最小内切圆面积的商.

经过特征规约和基于贝叶斯公式的后验概率提取后,每个样本对应一个 8 维的特征向量,由特征选择例程进一步处理.

### 4.2　不同阈值对特征选取的影响

本文 2.3.2 节算法描述第 5)步中存在一个未知参数 threshold,在不同的阈值条件下,参与投票的特征子集数目不同,投票结果各异,实验统计结果如图 2 所示.当取阈值取 0.85 时,参与投票的特征子集偏少,可能会出现个别特征选取错误;当取阈值取 0.75 时,参与投票的特征子集很多,造成各特征得票数的均衡分布,给特征选取也造成一定的不确定性;取阈值为 0.80 时,各特征得票的差异很明显,特征选择结果也趋于合理.实验结果表明:当阈值设定使参与投票的特征子集占所有特征子集的 5% ～10% 时,特征选择将取得比较理想的效果.

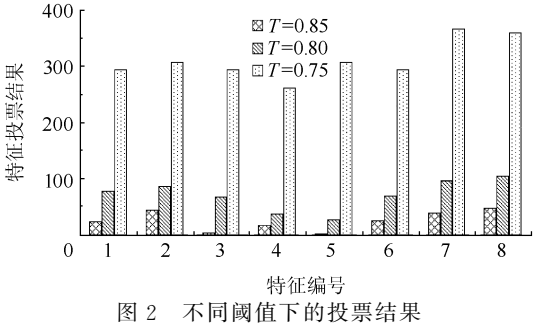



图 2　不同阈值下的投票结果

### 4.3　特征选择的有效性

特征选择的结果最终由分类算法的性能来评估,本节根据不同特征子集训练的分类器在测试样本集上的分类准确性来判断特征选择的效果.对于不同特征数目的特征子集,挑选分类效果最佳的 2 个特征组合,统计结果如表 1 所示.

**表 1　特征选择取向统计结果**

| 数征数目 | 分类准确性 | 数据特征选择 ① | ② | ③ | ④ | ⑤ | ⑥ | ⑦ | ⑧ |
|---|---|---|---|---|---|---|---|---|---|
| 3 | 0.910 | | √ | | | | | √ | √ |
| | 0.908 | | √ | | | | √ | √ | |
| 4 | 0.912 | | √ | √ | | | | √ | √ |
| | 0.912 | | √ | √ | | | √ | √ | |
| 5 | 0.926 | √ | √ | | | | √ | √ | √ |
| | 0.916 | √ | √ | √ | | | √ | √ | |
| 6 | 0.903 | √ | √ | √ | | | √ | √ | √ |
| | 0.836 | √ | √ | √ | | √ | √ | √ | |

分析发现,表 1 中各特征子集具有相同的特征选择取向.当特征空间中包含特征①②③⑥⑦⑧时,分类器在测试样本集上的准确性就高,上述分析与 4.2 节的结果一致.

### 4.4　不同特征数目下 SVM 分类表现

从表 2 可以看出,特征数目的变化会对分类器的设计造成一定的影响.综合分析几个最优的特征组合,我们可以看到当特征数目为 5 时分类器的准确性最高,依次是特征数目为 4 和 3 的分

类器，当特征区间有6个特征时，分类器的性能大幅降低.

### 4.5 浮动车系统性能评估

本文统计了北京市某浮动车实时交通路况信息处理系统(该系统拥有8 000辆浮动车)在应用浮动车行驶模式判断模型后，系统准确性的变化趋势. 2007年1～6月份的统计数据表明，在进行浮动车行驶模式分析及干扰数据剔除后，系统的准确率有5%～7%的提升，对比效果见图3.

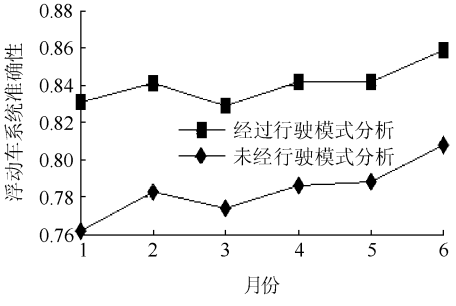



图3 系统准确性提升结果

## 5 结束语

本文分析了浮动车行驶模式对于交通信息处理准确性及处理效率的影响，设计并实现了一个基于SVM分类器的浮动车行驶模式判断模型，同时提出了一种基于模糊策略的特征评价和选择方法，取得了良好的实验效果. 浮动车系统运行情况表明：采用本文提出的模型对浮动车行驶模式进行分析并剔除干扰数据后，整个系统的准确性及处理效率有了显著提升.

## 参考文献 (References)

[1] Wu D D, Zhu T Y, Lü W F, et al. A heuristic map-matching algorithm by using vector-based recognition[C]// Proc of the International Multi-Conference on Computing in the Global Information Technology (ICCGI'07). Piscataway: IEEE, 2007:18－24

[2] Vapnik V N. The nature of statistical learning theory[M]. New York: Springer-Verlag, 1995:138－141

[3] Langley P. Selection of relevant features in machine learning [C]// Proc AAAI Fall Symposium on Relevance. New Orleans: AAAI Press, 1994:140－144

[4] Lee H M, Chen C M, Chen J M, et al. An efficient fuzzy classifier with feature selection based on fuzzy entropy[J]. IEEE Trans on Systems Man Cybernet, Part B, 2001, 31(3):426－432

[5] Uncu O, Turksen I B. Two step feature selection: approximate functional dependency approach using membership values[C]// Proceedings 2004 IEEE International Conference on Fuzzy System. Piscataway: IEEE, 2004:1643－1648

[6] Connon R L, Dave J V, Bezdek J C. Efficient implementation of the fuzzy c-means clustering algorithms[J]. IEEE Trans on Patten Analysis and Machine Intelligence, 1986, 8(2):248－255

[7] Han J W, Kamber M. Data mining concepts and techniques [M]. Los Altos: Morgan Kaufmann, 2001:132－136

(责任编辑：吴 琼)

(上接第955页)

[4] Michael R W. Infrared signature generation of airborne targets[C]//Proceedings of SPIE, Characterization, Propagation and Backgrounds III. Orlando, USA: SPIE, 1967:114－122

[5] 张建奇，方小平. 红外物理[M]. 西安：西安电子科技大学出版社，2004:53－54
Zhang Jianqi, Fang Xiaoping. Infrared physics[M]. Xi'an: Xidian University Press, 2004:53－54(in Chinese)

[6] 吴晗平. 红外辐射大气透过率的工程理论计算方法研究[J]. 光学 精密工程，1998，6(4):35－43
Wu Hanping. Research into theoretical calculation method on engineering of transmittance of infrared radiation through atmosphere[J]. Optics and Precision Engineering, 1998, 6(4):35－43(in Chinese)

[7] 王迪，李承芳，熊飞. 大气红外辐射传输的简便算法与MODTRAN的比较[J]. 光学技术，2006，32(增刊):293－298
Wang Di, Li Chengfang, Xiong Fei. The method of calculate the transmittance of infrared radiation through atmosphere and compared with MODTRAN[J]. Optical Technology, 2006, 32(Supp lement):293－298(in Chinese)

[8] Richard D Hudson. Infrared system engineering[M]. New York: John Wiley & Sons, INC, 1969

(责任编辑：娄 嘉)