

# 利用粗糙集理论提高 SVM 预测系统的实时性

冯利军<sup>1,2</sup>, 李书全<sup>1</sup>, 宋连友<sup>3</sup>

(1. 天津财经大学, 天津 300222;  
2. 河北农业大学, 河北 保定 071001;  
3. 沧州师范专科学校, 河北 沧州 061001)

**摘要:**支持向量机是一种新的机器学习方法,它具有良好的推广性和分类精确性。但是在利用支持向量机的分类算法处理实际问题时,该算法的计算速度较慢、处理问题效率较低。文中介绍了一种新的学习算法,就是将粗糙集和支持向量机相结合,利用粗糙集对支持向量机的训练样本进行预处理,从而缩短样本的训练时间,提高基于 SVM 预测系统实时性。文中最后利用该方法进行了数据试验,试验结果表明了该方法可以大大缩短样本的训练时间,提高基于支持向量机处理预测系统的效率。从而也证明了该方法的有效性。

**关键词:**粗糙集;支持向量机;预测

**中图分类号:**TP18

**文献标识码:**A

**文章编号:**1673-629X(2006)09-0030-02

## Improving Real-Time Character of Prediction System Based on SVM Using RS Theory

FENG Li-jun<sup>1,2</sup>, LI Shu-quan<sup>1</sup>, SONG Lian-you<sup>3</sup>

(1. Tianjin University of Finance and Economics, Tianjin 300222, China;  
2. Agriculture University of Hebei, Baoding 071001, China;  
3. Cangzhou Teacher's College, Cangzhou 061001, China)

**Abstract:** Support vector machine is a kind of new machine learning method. This method has good generality capability and better classification accuracy. But when solve real problem using support vector machine, its computation rate is slow and its efficiency is low. Introduce a kind of method that improves the real-time character of prediction system based on SVM in this paper. That can shorten the training time of prediction system based on SVM by preprocessing the training sample of SVM using rough sets theory. At last, carried on data experiments using this method in this paper. The experiments result indicated that this method can shorten the training time greatly and improve the efficiency of prediction system based on support vector machine. Consequently the experiments result proved the validity of this method.

**Key words:** rough sets; support vector machine; prediction

### 0 引言

支持向量机(SVM)是在统计学习理论的基础上发展起来的一种新的机器学习方法,它基于结构风险最小化原则,能有效地解决过学习问题,具有良好的推广性和较好的分类精确性。目前,SVM在许多领域的分类和回归方面起着越来越重要的作用。比较成熟的应用有人脸识别、语音识别和医疗诊断等<sup>[1]</sup>。然而,SVM有一个不足之处就是它在对训练样本进行处理时,不能确定数据中哪些知识是冗余的,哪些是有用的,哪些作用大,哪些作用小。这

样,就可能造成建立在二次规划基础上的 SVM 算法在对样本数据进行训练时耗费较长的时间,降低 SVM 预测系统的实时性。为了解决这一问题,可以考虑采用一定的方法对需要训练的大量样本数据进行预处理,剔除掉冗余、无用的信息,将留下的数据作为 SVM 的训练样本,这样就可以大大缩短样本的训练时间,提高 SVM 预测系统的效率。粗糙集(Rough Sets,简称 RS)理论为这一问题的解决提供了可能。

### 1 RS 理论概述

RS 是波兰数学家 Z. Pawlak 为开发自动规则生成系统及研究软件计算问题于 1982 年提出的。它是一种处理不精确、不确定和不完全数据的新的数学方法。由于它在机器学习、知识发现、数据处理、决策支持与分析、专家系统、归纳推理和模式识别等方面的广泛应用,现已成为一

收稿日期:2005-12-09

基金项目:天津市教委十五综合投资项目(2004BA11)

作者简介:冯利军(1974-),男,河北涉县人,博士研究生,主要从事项目智能管理等研究;李书全,教授,主要从事项目管理、机器学习等方面的研究。

个热门的研究领域<sup>[2]</sup>。

RS理论以不可分辨关系划分所研究论域的知识,形成知识表达系统,利用上、下近似集逼近描述对象,通过知识约简,从而获得最简知识。知识约简是RS理论的核心内容之一。人类在对一个事物做出判断和决策时,并不是依据被判断事物的全部特性,而是根据事物的一个或几个最主要的特征做出判断。知识约简就是根据这一原理,剔除知识库中的冗余知识,简化判断规则。

假定所讨论的对象的论域为  $U$ ,  $U$  中的一种关系定义为  $R$ ,  $R$  可以是一种属性的描述,也可以是一个属性集合的描述;可以是定义一种变量,也可以是定义一种规则。当用  $R$  描述  $U$  中所有等价类簇时,可以表示为  $U/R$ 。若  $R$  是  $U$  上的划分,  $R = \{X_1, X_2, \dots, X_n\}$ ,  $(U, R)$  称为近似空间。用  $\text{des}_A\{X_i\}$  表示  $U$  上基于关系  $R$  的一个等价关系对  $X_i$  的基本集合的描述。例如,属性集  $A \subset R$ ,  $\text{des}_A\{X_i\} = \{(a, b) :: f(x, a) = b, x \in X_i, a \in A\}$ , 因此这里表示给定的集合  $X_i$  可用属性  $A$  和属性集  $V_A$  表示。

不可分辨关系是指事物有属性集  $P$  表示时,在论域  $U$  中的等价关系。例如,属性集  $P \subset R$ , 对象  $X, Y \in U$ , 对于每个  $Q \in P$ , 当且仅当  $f(X, a) = f(Y, b)$  时,  $X$  和  $Y$  是不可分辨的, 即有:  $\text{ind}(P) = \{(X, Y) \in U : a \in P, f(X, a) = f(Y, b)\}$ ,  $\text{ind}(P)$  的等价类称为知识  $P$  的基本概念或基本范畴。如果  $r \in P$ , 且  $\text{ind}(P) = \text{ind}(P - (r))$ , 就称属性  $r$  是属性  $P$  中可省略的, 否则称属性  $r$  是属性  $P$  中不可省略的。也就是说, 属性  $r$  对于描述对象的特征作用不大, 属于冗余属性, 剔除以后不会影响对象特征的描述。利用该方法, 就可以进行知识约简, 从众多的特征信息中, 提取有用的属性, 从而简化处理过程<sup>[3]</sup>。

## 2 利用 RS 理论对 SVM 的训练数据进行预处理

在基于 SVM 的预测系统中, 要对采集到的样本数据进行训练。如果 SVM 所处理的样本的维数较大, 就可能导致 SVM 的训练时间过长, 影响到预测系统的实时性。对于这个问题, 可以利用 RS 理论对样本数据进行预处理。因为 RS 在处理数据时有两个显著的优点: 一是 RS 不需要任何先验知识, 仅利用数据本身提供的信息即可; 二是 RS 能表达和处理不完备信息, 以不可分辨关系为基础, 侧重分类, 能在保留关键信息的前提下对数据进行约简并求得知识的最小表达, 能识别并评估数据之间的依赖关系, 揭示出概念简单的模式, 能从经验数据中获取易于证实的规则知识。这样, 经过 RS 处理后, 不但可以剔除数据中的冗余信息, 还可以降低样本的维数。然后将处理过的数据用于 SVM 预测系统, 则可以大大缩短样本的训练时间<sup>[4]</sup>。基于以上的理解, 可以设计出相应的系统简图, 如图 1 所示。

在图 1 中, 把 RS 对数据的预处理过程作为前置系统, 再根据 RS 预处理后的信息结构, 来构建 SVM 预测系统。具体的约简过程如下: 训练样本集首先从搜集的原始

数据中产生, 然后将条件属性值进行量化。量化后的属性值构成一张二维表格, 每一行描述一个对象, 每一列描述对象的一个属性, 属性分条件属性和决策属性。决策表约简包括条件属性约简和决策规则约简。条件属性约简就是去某一属性后, 考察决策表的相容性, 如果去掉该属性后决策表是相容的, 就去掉该属性, 直到决策表最简为止。决策规则约简就是在条件属性简化后的决策表中, 去掉样本集中的重复信息, 考察剩下的训练集, 每一条规则中哪些属性值是冗余的, 去掉冗余信息和重复信息后, 就得到了最小决策算法。也可以先约简每一决策规则, 再简化条件属性, 从而得到最小条件属性集。采用约简得到的最小条件属性集及相应的原始数据重新形成新的训练样本集, 该样本集除去了所有不必要的条件属性, 仅保留了影响预测精度的重要属性。用约简后形成的训练样本对 SVM 训练。最后输入按照最小条件属性集及相应的原始数据形成的新的测试样本集, 对系统进行测试, 输出预测结果。

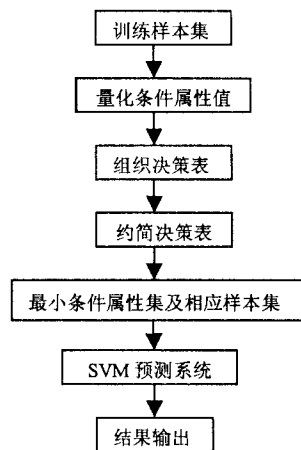


图 1 利用 RS 进行数据预处理的 SVM 预测系统简图

## 3 数据实验

文中利用心脏病诊断的例子来进行数据实验。采用的数据是美国 Cleveland Heart Disease Database 提供的。在该数据里, 共对 299 个病人进行了彻底的临床检测, 确诊了他们是否有心脏病。同时, 记录了他们的年龄、胆固醇等 125 项指标<sup>[5]</sup>。采用其中的 289 个病人的记录作为 SVM 的训练集, 剩余的 10 个病人的记录作为检测数据。在 MATLAB7 中运行程序发现, 在对训练集进行预处理前, 要完成整个训练及检测任务需要耗时 302s, 且预测结果准确率偏低。当用 RS 对 SVM 的训练集进行预处理后, 推断病人是否患有心脏病的指标由原来的 125 个减少为 4 个, 整个训练及检测任务耗时 233s, 准确率 100%。结果如表 1 所示。

在表 1 中, +1 表示有心脏病, -1 表示没有心脏病。从整个实验结果可以看出, 利用 RS 理论对 SVM 的训练数据进行预处理后, SVM 预测系统的实时性及预测精度都大大提高了。可见, 利用一些辅助工具对 SVM 进行适

(下转第 34 页)

```

virtual CJiuG();
bool MoveLeft(JGState * src,JGState * result); //左移
bool MoveRight(JGState * src,JGState * result); //右移
bool MoveUp(JGState * src,JGState * result); //上移
bool MoveDown(JGState * src,JGState * result); //下移
bool Compare(JGState * src1,JGState * src2); //比较两个状态是否相等
int ComputeFn(JGState * cur,JGState * dest); //估价函数的计算,我们采用了 Pn
bool Search(); //用 A* 算法搜索最优解
};

```

本程序的关键是用 A\* 算法来搜索最优解,所以程序中的核心部分是 Search()函数的实现。设计思路如下:

首先比较初始状态和目标状态是否相同,如果相同则搜索成功并且退出,不相同则将起始结点加入到 Open 表中去,然后搜索 Open 表中估计值最小的结点。在这里采用的启发函数是  $h(n) = p(n)$ ,即每一个将牌与其目标位置之间距离的总和,在程序中是 ComputeFn()函数:

```

int CJiuG::ComputeFn(JGState * cur,JGState * dest)
{
    int xcur[9],ycur[9],xdest[9],ydest[9]; //保存 9 个坐标
    int i,j;
    int result=0;
    for(i=0;i<3;i++)
    {
        for(j=0;j<3;j++)
        {
            xcur[cur->state[i][j]]=i;
            ycur[cur->state[i][j]]=j;
            xdest[dest->state[i][j]]=i;
            ydest[dest->state[i][j]]=j;
        }
    }
}

```

(上接第 31 页)

当的优化及处理,可以进一步发挥 SVM 本身处理问题的优越性。

表 1 数据实验结果表

序号	训练集预处理前			训练集预处理后		
	实际结果	预测结果	总耗时(s)	实际结果	预测结果	总耗时(s)
1	+1	+1	302	+1	+1	233
2	+1	+1		+1	+1	
3	+1	+1		+1	+1	
4	+1	+1		+1	+1	
5	+1	+1		+1	+1	
6	-1	+1*		-1	-1	
7	-1	+1*		-1	-1	
8	-1	-1		-1	-1	
9	-1	-1		-1	-1	
10	-1	-1		-1	-1	

#### 4 结束语

将 RS 和 SVM 相结合,利用 RS 对 SVM 处理的训练

```

}
//计算当前状态的每个将牌的与目标状态之间的距离的总和。
for(i=1;i<9;i++)
{
    result=result+abs(xcur[i]-xdest[i])+abs(ycur[i]-ydest[i]);
}
return result;
}

```

再将估价函数最小的结点从 Open 表中删除,加入到 Close 表中去。然后对该结点进行扩展,分 4 种情况(空格左移,右移,上移,下移)讨论,再对 4 种移动所得到的结点重复上述操作过程直到找到目标结点为止。

#### 4 结 论

本程序主要是用 A\* 算法来搜索八数码问题的最优解。通过输入大量的初始状态和目标状态发现,在一般情况下都可以找到最优的动作序列,但对某些复杂的初始状态虽能得到正确解却不能完全得到最短的搜索路径。这是有待改进的地方。

#### 参考文献:

- [1] 林尧瑞,马少平.人工智能导论[M].北京:清华大学出版社,1989.
- [2] 马少平,朱小燕,人工智能[M].北京:清华大学出版社,2004.
- [3] 尼尔逊 N.J. 人工智能原理[M].北京:科学出版社,1983.
- [4] Ansari N,Hou E. 用于最优化的计算智能[M].李 军,边肇祺译.北京:清华大学出版社,1999.
- [5] 王万森.人工智能原理及其应用[M].北京:电子工业出版社,2000.

样本进行预处理,可以缩短 SVM 的训练时间,并在不影响 SVM 预测系统预测精度的前提下提高 SVM 预测系统的实时性,为实际预测问题的处理提供了一个很好的解决方案。

#### 参考文献:

- [1] 张 辉,张 浩,陆剑峰.SVM 在数据挖掘中的应用[J].计算机工程,2004,30(6):7-8.
- [2] 张文修,吴伟志,梁吉业,等.粗糙集理论与方法[M].北京:科学出版社,2001.
- [3] Pawlak Z. Rough sets[J]. International Journal of Information and Computer Science,1982,11:241-256.
- [4] 李孟歆,吴成东,夏兴华.粗糙集理论及其应用[J].沈阳建筑工程学院学报,2001,17(4):296-299.
- [5] 邓乃扬,田英杰.数据挖掘中的新方法-支持向量机[M].北京:科学出版社,2004.