粗糙集的模糊性度量与SVM的混合分类算法

任小康,孙正兴,郝瑞芝
REN Xiao-kang,SUN Zheng-xing,HAO Rui-zhi

西北师范大学 数学与信息科学学院,兰州 730070
College of Mathematics & Information Technology,Northwest Normal University,Lanzhou 730070,china
E-mail:273839182@qq.com

REN Xiao-kang,SUN Zheng-xing,HAO Rui-zhi.Measure of rough sets's fuzziness and SVM hybrid classification algorithm.Computer Engineering and Applications,2010,46(7):46-48.

Abstract: This paper uses the information entropy method to measure the rough set's fuzziness,and makes the pretreatment before the reduction of rough's decision attribute with eliminating the difference which due to the redundance of decision attribute. Combination of SVM in solving the small sample,nonlinear and high dimensional pattern recognition problem has a lot of unique performance advantages.In this paper,an improved algorithm is given and the classification results are tested.
Key words: information entropy;rough set;fuzzy degree;reduction;Support Vector Machine(SVM)

摘 要: 采用信息熵的方法来度量粗糙集的模糊性可以在约简之前对粗糙的决策属性进行预处理,从而消除因决策属性的冗余而带来的分类决策的偏差。结合 SVM 在解决小样本、非线性及高维模式识别问题中表现出许多特有的优势。对该类别的方法进行了改进,并对分类的结果进行了测试。
关键词: 信息熵;粗糙集;模糊度;约简;支持向量机
DOI: 10.3778/j.issn.1002-8331.2010.07.014 **文章编号:** 1002-8331(2010)07-0046-03 **文献标识码:** A **中图分类号:** TP301.6

1 引言

Rough 集理论[1]为研究不确定知识和数据的表达、学习归纳提供了一种重要的理论方法，已经广泛应用于模式识别、知识发现、问题求解以及不确定推理等领域。粗糙集理论是处理不完全和不精确信息的一种有效的数学工具[2],分类的目的是为了找到不确定信息的规律,从而使机器能够模拟人类进行自主的认知和学习。所以粗糙集的不确定性度量应是关注的焦点。

在 Vapnik[3]提出支持向量机(SVM)后,支持向量分类机得到快速的发展。在分类过程中 SVM 是目前研究的重点,原始的 SVM 算法的速度都比较慢，原因在于其利用传统的标准二型优化技术来解决对偶问题。因为 SVM 在训练时需要计算和存储核函数矩阵,当样本点的维数较大时,在寻优过程中需要进行大量矩阵运算,因此浪费了大量的内存和时间。影响 SVM 算法计算量的主要因素是训练样本数，尤其是样本的支持向量数，大量的计算则导致计算机因内存容量限制而无法计算,无法实现对信息的有效分类。

结合粗糙集对不精确、不确定和不完善的数据的模糊性度量来去除冗余属性的方法和它不需要任何先验知识,可分析出信息间的相互关系，在知识不受影响的前提下对属性进行约简，具有处理大容量高维信息的强大功能，可以有效地降低 SVM 训练样本维数。该方法通过去除冗余和约简后的属性作为 SVM 的训练样本可以大幅度减少训练样本数。结合两者的优点,提出新的分类方法对数据进行分类。结果表明,该算法有更好的可行性与优越性。

2 基于信息熵粗糙集的模糊度与属性约简

目前,度量粗糙集不确定的方法主要有粗糙度、粗糙熵、模糊度和模糊熵。由文[4-6]研究工作发现,粗糙集的粗糙度随着知识粒度的减小而单调递减。这符合人们的认知直觉。但是,很多实例表明,当属于一个集合的正域或负域中的知识颗粒被细分时,它的粗糙度可能不发生变化,这与人们的认知直觉不吻合。为了克服这个问题,有的研究者提出粗糙熵,如 Liang[7]等人定义了一种粗糙熵，它是集合 X 的粗糙度与近似空间中的知识粒度之积,并得到结论:这种粗糙熵随着知识颗粒的细分严格单调递减。这个结论在一定程度上弥补了用粗糙度度量粗糙集不确定性的不足。但是分析发现,如果对集合 X 负域的知识颗粒(与 X 无关)进行细分,粗糙度将不变(符合人们的认知规律),但粗糙熵却严格递减(不符合人们的认知规律)。这说明与集合 X 无关的知识颗粒的变化也会导致 X 的粗糙熵的变化,这与人们对不确定性问题的认知不符。

2.1 信息熵

信息熵是一个非常广泛的概念,1948 年 Shannon 信息熵[8]的提出为信息的不确定度量奠定了理论基础,Klir 基于 Shannon 熵提出了一种度量不确定性的信息熵[9]:

$$H(F_X^B)=-\frac{2}{n}\sum_{i=1}^{n}\mu_X^B(x_i)\mathrm{lb}\mu_X^B(x_i) \quad (1)$$

作者简介: 任小康(1963-),男,教授,研究领域为多媒体信息处理;孙正兴(1985-),男,计算机应用专业研究生,研究方向为多媒体信息处理。
收稿日期: 2009-04-27 **修回日期:** 2009-07-02

容易验证,该信息熵不满足模糊度定义,不是模糊度。

2.2 集合 X 的隶属函数

设 U 是非空对象集,对象子集 $X\subseteq U$,则对于任意的 $x(x\in U)$,x 属于集合 X 的隶属函数[10]为:

$$\mu_X^B(x)=\frac{|X\cap[x]_B|}{|[x]_B|}$$

显然,$0\leqslant\mu_X^B(x)\leqslant 1$,它表示任意一个元素属于集合 X 的程度。

令 $F_X^B=\{\mu_X^B(x_1),\mu_X^B(x_2),\cdots,\mu_X^B(x_n)\}$,则 F_X^B 是集合 U 上的一个模糊集(即 $F_X^B\in F(U)$)。

2.3 基于信息熵的粗糙集模糊度

为了应用信息熵测量粗糙集的模糊度[11],进一步分析发现:粗糙集的模糊性来自边界域的两个部分,一部分是边界域中属于集合 X 的元素,一部分是边界域中不属于集合 X 的元素,而式(1)的信息熵只考虑了前面一部分,没有涉及第二部分。为此,提出一种新的基于信息熵的粗糙集的模糊度度量方法:

$$dz(F_X^B)=-\frac{1}{n\ln 2}\sum_{i=1}^{n}[\mu_X^B(x_i)\ln\mu_X^B(x_i)+(1-\mu_X^B(x_i))\ln(1-\mu_X^B(x_i))] \quad (2)$$

直观上讲,式(2)由 $\mu_X^B(x_i)\ln\mu_X^B(x_i)$ 和 $(1-\mu_X^B(x_i))\ln(1-\mu_X^B(x_i))$ 两部分信息熵构成,前者主要反映属于集合 X 的元素"贡献"的不确定性,后者主要反映不属于集合 X 的元素"贡献"的不确定性,这两部分同时考虑才能更精确地刻画粗糙集的不确定性。

定理 1 设格 $\langle P(A),\in\rangle$ 中任意一条链为 $\varnothing=B_0\subset B_1\subset B_2\subset\cdots\subset B_m=A$,如果 $U/B_{i+1}<U/B_i$ 则对于任意的 $X\subseteq U$,都有 $dz(F_X^{B_{i+1}})\leqslant dz(F_X^{B_i})$。

定理 1 更能很好地描述近似空间中知识粒度减小的变化趋势。式(2)依赖两部分信息熵,既利用了度量不确定性的 Shannon 熵,又结合了粗糙集的特点,同时构造集合 X 的边界域中属于 X 的那部分元素"贡献"的不确定性和不属于 X 的那部分元素"贡献"的不确定性,非常直观。随着集合 X 的边界域上的知识颗粒的"不成比例"的细分,粗糙集的模糊度将严格递减;而集合 X 边界域上的知识颗粒被"成比例"细分时,粗糙集的模糊度不变。这更加准确地刻画出人们对不确定性问题的认知规律。

2.4 属性约简

信息系统表示为一个决策表 (U,A),$A=C\cup D$,$C\cap D=\Phi$,C 是条件属性集,D 是决策属性集,$y\in D$ 是整体决策而不是对于"决策子集" $W\in U/y$ 的一个局部决策。决策属性 $y\in D$ 关于条件属性 $X\in C$ 的支持子集是子集 $S_X(y)=\bigcup_{w\in U/y}W^{(U/X)^-}=\bigcup_{W\in U/y}(\bigcup_{V\in U/X,V\in W}V)$,$spt_X(y)=\left|\bigcup_{w\in U/y}W^{(U/X)^-}\right|/|U|$,称为 y 关于 X 的支持度。

令 $Y\subseteq D$ 是决策属性子集,Y 关于 X 的支持子集是 $S_X(Y)=\bigcup_{W\in U/Y}S_X(W)$,$Y$ 关于 X 的支持度是 $spt_X(Y)=\left|\bigcup_{W\in U/Y}spt_X(W)\right|/|U|$。从支持子集与支持度可以评价属性的重要性,并显示出属性之间的依赖关系。例如,已知 $X_1\supseteq X_2$ 蕴涵 $S_{X_1}(Y)\supseteq S_{X_2}(Y)$,其中 $X_1,X_2\subseteq C,Y\subseteq D$,则 $x\subseteq X$ 在 X 中的重要性为 $sig_{X-\{x\}}^Y=(|S_X(Y)|-|S_{X-\{x\}}(Y)|)/|U|$;如果 $S_X(Y)=S_{X-\{x\}}(Y)$,则称 x 在 X 中是不重要的,否则称 x 在 X 中是重要的。由所有在 X 中是重要属性 x 组成的集合称为 X 的核(相对 Y 而言),表示为 $C_X^Y=\{x\in X|sig_{X-\{x\}}^Y>0\}$。

属性约简[1]就是要找到 $X\subseteq C$ 的一个极小子集 X_0 使得 $S_{X_0}(Y)=S_X(Y)$,其中 $X\supset\cdots\supset X_0$。由于约简不是惟一的,所以依据核属性与对决策属性的支持度可以得出较合理的相对属性约简。

在约简过程中不能删除决策表中的重要属性与核属性,从约简中选择一个较合理的,相对约简时应遵循一下几个原则[1]:(1)首选包含核属性的约简;(2)选择包含重要属性多的约简;(3)属性重要性会发生变化的属性,应尽量包含在约简中;(4)相容性与近似精度应满足要求。

3 支持向量机

SVM 的原理[13]:SVM 就是通过在原空间或经投影后在高维空间构造最优分类面。将给定的属于两个类别的训练样本分开,构造超平面的依据是两类样本距离超平面距离的最大化。

设线性可分样本集 $(x_i,y_i)_{1\leqslant i\leqslant N}$,$x_i\in R^d$,$y_i\in\{-1,1\}$,是类别标号,$d$ 维空间中线性判别函数的一般形式为 $g(x)=w\cdot x+b$,相应的分类面方程为 $w\cdot x+b=0$。将 $g(x)$ 进行归一化,使所有的 x_i 都满足 $|g(x)|\geqslant 1$,即离分类面最近的样本 $|g(x)|=1$,这样分类间隔就等于 $2/\|w\|$。求解最优分类面就等效于最小化 $\|w\|$,原问题为:

$$\begin{aligned}&\min\Phi(w)=\frac{1}{2}\|w\|^2=\frac{1}{2}(w\cdot w)\\&\text{s.t. } y_i(w\cdot x_i+b)\geqslant 1\quad i=1,2,\cdots,l\end{aligned} \quad (3)$$

采用 Lagrange 乘子 $\alpha=(\alpha_i,\alpha_{i+1},\cdots,\alpha_n)$ 求解该二次规划问题,可以得到最优分类面,其中 $w=\sum\alpha y_i x_i$,x_i 是位于分类间隔面上的样本,这些训练样本被称为支持向量,分类函数为:

$$f(x)=sign\left(\sum_i\alpha y_i x_i\cdot x+b\right) \quad (4)$$

对于线性不可分的情况,SVM 引入了松弛变量 ξ 和惩罚因子 C,使目标函数变为:

$$\Phi(w,\xi)=\frac{1}{2}(w\cdot w)+C\left(\sum_{i=1}^{N}\xi_i\right) \quad (5)$$

另一方面,SVM 通过核函数将输入的低维空间的非线形问题映射到高维特征空间线性问题,在新空间上求解最优分类面,线性可分的核函数为 $K(x,x_i)=(x\cdot x_i)$,这样得到的分类函数为:

$$f(x)=sign(\sum_i\alpha y_i\cdot k(x_i\cdot x)+b) \quad (6)$$

总之,SVM 建立在统计学理论的基础上,在解决小样本、非线性及高维模式识别问题中表现出很多优势,并在许多应用中取得了很好的结果。

4 基于信息熵粗糙集的模糊性度量与 SVM 的混合分类算法

4.1 算法步骤

步骤 1 用户对前 N 个属性样本进行标记得到:相关属性集 I^+,无关属性集 I^-;

步骤 2 通过相关属性集 I^+,无关属性集 I^- 的约简找到属性

集的核,根据核计算各属性隶属该核的 $dz(F_X^B)$并分别进行排序,消去 $dz(F_X^B)$中模糊度相同的属性(保留一个),并通过属性约简,得到新的 I_1^+、I_1^-;

步骤 3 用 SVM 训练样本集 (x_i,y_i),$x_i \in I_1^+ \cup I_1^-$,$y_i=\begin{cases}+1,\text{if } x_i \in I_1^+\\-1,\text{if } x_i \in I_1^-\end{cases}$;

步骤 4 用 SVM 对训练样本进行学习并进行分类:$f(x)=sign(\sum_i \alpha y_i \cdot k(x_i \cdot x)+b)$;

步骤 5 如果分类正确则显示结果,否则将该样本进行标记加入前 N 个标记的样本中返回步骤 1 重新进行训练。

由于检索的时候,用户标记的样本是在特征空间中离分类面最近的样本,因此这些样本很适合构造 SVM 分类器。又因为支持向量是位于分类面间隔上的样本,而距离分类间隔面远的样本,对 SVM 分类器的构造没有影响。通过对分类不正确的样本进行标记并进行训练可以进一步地优化支持向量,使新得到的支持向量更加接近最优分类面。以上方法主要借助信息熵的方法来度量粗糙集的模糊性,去除冗余属性,并通过相关属性约简得到作为分类面的支持向量。借助构造的分类面将未标记的属性样本进行分类。

4.2 试验结果

为了验证该算法的实用性,选用了疾病诊断数据 D[14],由于该属性样本为线性不可分的情况。试验是环境是 MatLab7.0,所用的分类软件是 LIBSVM,台湾大学林智仁(Lin Chih-Jen)副教授等开发设计的一个简单、易于使用和快速有效的 SVM 模式识别与回归的软件包。

目标函数变为:

$$\Phi(w,\xi)=\frac{1}{2}(w\cdot w)+C\left(\sum_{i=1}^{N}\xi_i\right)$$

对应的核函数选用:

$$K(x_i,x)=\exp(-\|x-x_j\|^2)(\delta=1)$$

先对诊断数据进行模糊度计算并排序去除冗余得到 D',数据 D'有 14 个属性 C_1~C_{14},属性约简后保留 5 个属性$\{C_1,C_3,C_5,C_8,C_{11}\}$,分别使用未去冗余的算法和去冗余的算法对剩余样本进行训练。

测试 1 选用分类参数 C=5,两种方法的比较分别为,当分类结果完全正确时,对应的允许误差 ξ_i 和迭代次数 N_i 如图 1 所示。

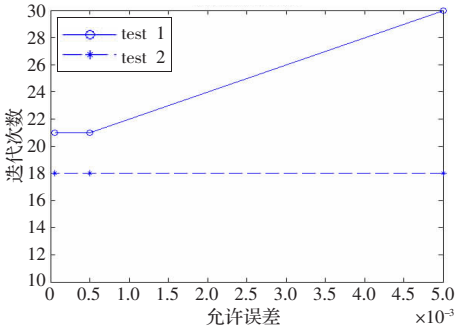

图 1 测试对象的分类结果

注:Test1 表示未数据预处理算法,Test2 表示数据预处理算法

测试 2 当 ξ=0.005 时,改变分类参数 C=$\{1,10\}$和 δ=$\{0.01,10\}$,两种分类算法的结果中对应的支持向量 S_i 和迭代次数 N 如表 1 所示:

表 1 不同参数的分类结果比较

算法	δ	C=1			C=10		
		支持向量	迭代次数	正确率/(%)	支持向量	迭代次数	正确率/(%)
1	0.01	7	9	100	51	60	100
	10.00	7	9	100	51	79	100
2	0.01	5	8	100	20	30	100
	10.00	5	8	100	20	30	100

注:算法 1 表示未进行数据预处理,算法 2 表示进行数据预处理

测试 1 表明,在不同的允许误差下,达到相同的分类结果,经过数据预处理的分类算法拥有较少的支持向量和迭代次数。测试 2 表明平滑因子 δ 和分类参数 C 变化较大时两种算法的分类结果都比较稳定。

5 结语

基于信息熵的粗糙集的模糊性度量和粗糙集的约简避免了 SVM 算法的维数灾难。可以由较少的支持向量和较少次的迭代得到比较稳定和更优的分类结果。由于没有引入好的反馈策略,算法的性能还有待进一步提高。

参考文献:

[1] Pawlak Z.Rough sets[J].Internal J Inform Comput Sci,1982,11(5):341-356.

[2] Wang Guo-Yin.Rough set theory and knowledge discovery[M].Xi'an:Xi'an Jiaotong University Press,2001.

[3] Vapnik V.The nature of statistical learning theory[M].Berlin:Springer-Verlag,1995.

[4] Chakrabarty K,Biswas R,Nanda S.Fuzziness in rough sets[J].Fuzzy Sets and Systems,2000,110:247-251.

[5] Wang Guo-yin,Zhao Jun,An Jiu-jiang,et al.A comparative study of algebra viewpoint and information viewpoint in at-tribute reduction[J].Fundamenta Informaticae,2005,68(6):289-301.

[6] Zhao Jun,Wang Guo-Yin.Research on system uncertaintymeasures based on rough set theory[C]//Proceedings of the RSKT2006,Chongqing,2006:227-232.

[7] Liang J Y,Shi Z Z.The information entropy,rough entropyand knowledge granulation in rough set theory[J].International Journal of Uncertainty,Fuzziness and Knowledge-Based Systems,2004,12(1):37-46.

[8] Shannon C E.A mathematical theory of communication[J].The Bell System Technical Journal,1948,27:379-423.

[9] Klir G J,Wierman M J.Uncertainty based information[M].New York:Physica-Verlag,1998.

[10] Yang Lun-Biao,Gao Yi-Ying.Fuzzy mathematics:Theoryand application[M].Guangzhou:South China University of Technology Press,2004.

[11] 王国胤,张清华.不同知识粒度下粗糙集的不确定性研究[J].计算机学报,2008(9).

[12] 张文修,吴伟志,梁吉业.粗糙集理论与方法[M].北京:科学出版社,2001:24-34.

[13] Burges C J C.A tutorial on support vector machines for pattern recognition[J].Data Mining and Knowledge Discovery,1998,2(2).

[14] Arthur A,David N.The UC I machinelearning repository[EB/OL].http://archive,ics.uc.iedu/m l/machine2learning2databases/.