基于数据挖掘的企业合作伙伴的寻求方法研究①

魏宏业[1]，吕永波[1]，何 崴[2]，张仲义[1]
(1. 北京交通大学，北京 100044；2. 中国航天科技集团十院，北京 100854)

摘要：企业竞争日趋激烈，如何在竞争中立于不败之地成为企业必须要解决的问题. 通过构建企业联盟，有效组织企业的业务流程，达到资源共享，充分发挥企业的核心竞争力，将各企业的资源有效整合，从而迅速占领市场，是企业的成功之道. 寻找理想的合作伙伴是企业联盟成败的关键因素. 文章介绍了基于支持向量机的数据挖掘方法，建立了基于支持向量机的合作伙伴寻求模型，挖掘潜在的理想合作伙伴，研究表明，用支持向量机方法来发现理想合作伙伴是可行的.

关键词：企业联盟；合作伙伴；数据挖掘；支持向量机
中图分类号：TP311 **文献标识码**：A **文章编号**：1007 - 9807(2004)01 - 0060 - 05

0 引 言

市场竞争日益激烈，市场快速多变且难以预测，市场机遇转瞬即逝. 越来越多的企业意识到单凭自身内部的资源整合已经难以把握快速变化的市场机遇，于是它们开始将眼睛转向企业外部. 作为企业外部资源优化整合的一种手段——动态联盟[1](virtual corporations，VCs)，正伴随着经济全球化趋势和信息技术的不断发展，成为许多企业的现实选择.

动态联盟又称“虚拟企业”，它是一些相互独立的商业过程或企业组成的暂时联合，这些商业过程或企业称之为“伙伴(partner)”. 各个伙伴在诸如设计、制造、分销等领域分别为该联盟贡献出自己的核心能力[2]，并相互联合起来实现技能共享和成本分担，以把握快速变化的市场机遇. 因此，选择“正确”的合作伙伴是决定合作成败的关键因素，Zhan Su[3]曾提出了一些伙伴选择的基本原则，在定量方法方面，为减少可能的伙伴组合数，大多采用分阶段法，进行层层筛选. 如 Tulluri 和 Baker 提出了一种设计有效虚拟企业的两阶段定量方法[4]，钱碧波等提出了敏捷虚拟企业伙伴选择的三阶段结构化进程等[5].

企业合作伙伴的寻找是一个复杂的问题，想要寻找“最优解”是很难的，只能寻找“满意解”，理想的合作伙伴是由诸多因素决定的，很难用一种确定性的方法来表达. 在 20 世纪末出现的多学科相互交融和相互促进的新兴边缘学科——数据挖掘(data mining，DM)，可以从海量的数据中提取有效的、潜在的知识. 基于此，本文提出了一种基于支持向量机(support vector machines，SVMs)的数据挖掘方法从海量的企业信息库中搜索动态联盟伙伴，并给出了一个典型算例，以说明该模型及算法的有效性.

1 数据挖掘

数据挖掘是从大量的数据中提取或“挖掘”知识，也有人把数据挖掘视为另一个常用的术语数据库中的知识发现(knowledge discovery in databases，KDD)的同义词[6,7]. 进行数据挖掘的主要过程是：根据相应数据的特点来选取规则模板，对数据

① 收稿日期：2002 - 09 - 23；修订日期：2003 - 08 - 14.
基金项目：国家自然科学基金资助项目(601720623).
作者简介：魏宏业(1973—)，男，满族，辽宁丹东人，博士生.

进行选取、清洗和转换,然后根据需要应用归纳学习方法、决策树方法、粗集方法、最邻近方法、人工神经网络技术、遗传算法等来进行数据挖掘,挖掘得到的结果可通过解释成为知识,经过筛选后加入知识库,帮助进行决策.

从数据挖掘方法上,数据挖掘通常分为两类:统计模型和机器学习技术,其中机器学习与数据挖掘关系最密切.统计模型应用于数据挖掘主要是进行评估,常用的统计技术有概率分布、相关分析、回归、聚类分析和判别分析等;机器学习是人工智能的一个分支,也称为归纳推理,通过学习训练数据集,发现模型的参数,并找出隐含的规则[8].常用的机器学习方法如人工神经元网络、决策树和遗传算法在数据挖掘中的应用都很广泛.

2　支持向量机的基本理论

统计学习理论从 20 世纪 70 年代末诞生,到 20 世纪 90 年代之前都处在初级研究和理论准备阶段,近几年才逐渐得到重视,其本身也趋向完善,并产生了支持向量机这一将这种理论付诸实现的有效的机器学习方法.目前,SVM 算法在模式识别、回归估计、概率密度函数估计等方面都有应用.例如,在模式识别方面,对于手写数字识别、语音识别、人脸图像识别、文章分类等问题,SVM 算法在精度上已经超过传统的学习算法或与之不相上下.

2.1　支持向量机的基本思想

支持向量机的基本思想是通过用内积函数定义的非线性变换将输入空间变换到一个高维空间,在这个高维空间中寻找输入变量和输出变量之间的一种非线性关系,其基本结构见图 1.支持向量机有着严格的理论基础,采用结构风险最小化原则,具有很好的推广能力;支持向量机算法是一个凸二次优化问题,保证找到的解是全局最优解;能较好的解决小样本、非线性、高维数等实际问题,因此,支持向量机是当今研究的热点问题[9].

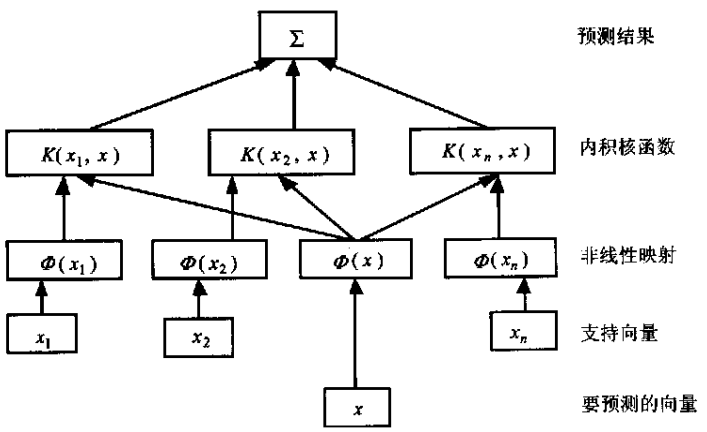

图 1　支持向量机结构示意图

2.2　支持向量机分类(support vector classification, SVC)算法

训练样本集 D_n 由 n 个样本$(\vec{x}_i, y)$ $(i = 1,2,\cdots, n)$ 组成,这里 $\vec{x}_i \in X$, $y \in \{-1,1\}$.在两类样本线性可分的情况下,SVC 通过寻找"最优分类超平面"实现结构风险最小化,这一超平面不仅准确区分两类样本而且最大化分类间隔.这等同于要求超平面 $w^{\mathrm{T}}\vec{x}_* + b = 0$ 中的 w 是以下凸二次规划的解:

$$\left.\begin{aligned} &\min \quad \frac{1}{2} w^{\mathrm{T}} w \\ &\text{s.t.} \quad y(w^{\mathrm{T}}\vec{x}_i + b) \geq 1, i = 1, \cdots, n \end{aligned}\right\} \tag{1}$$

引入拉格朗日函数

$$L(w, b, \alpha) = \frac{1}{2} w^{\mathrm{T}} w - \sum_{i=1}^{n} \alpha_i [\, y(w^{\mathrm{T}}\vec{x}_i + b) - 1]$$

可以获得(1) 的对偶规划(2):

$$\left.\begin{aligned}&\max \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}\alpha_i\alpha_j y_i y_j \vec{x}_i^{\mathrm{T}}\vec{x}_j\\&\text{s.t.}\quad \alpha_i \geq 0, i = 1,2,\cdots,n\\&\sum_{i=1}^{n}\alpha_i y_i = 0\end{aligned}\right\} \tag{2}$$

设 $\tilde{\alpha}_i (i = 1,\cdots,n)$ 是(2) 的解,则 SVM 求得的最优分类超平面为

$$\tilde{w}^{\mathrm{T}}\vec{x}_* + b = 0$$

其中

$$\tilde{w} = \sum_{i=1}^{n}\tilde{\alpha}_i y_i \vec{x}_i$$

而 b 可由约束条件

$$\alpha_i[y_i(w^{\mathrm{T}}\vec{x}_i + b) - 1] = 0 \quad i = 1,\cdots,n$$

求得. 若 $\tilde{\alpha}_k \neq 0$,则

$$y_i(\tilde{w}^{\mathrm{T}}\vec{x}_k + b) - 1 = 0$$

此时

$$b = y_i - \tilde{w}^{\mathrm{T}}\vec{x}_k$$

$\tilde{\alpha}_k \neq 0$ 时对应的 $\vec{x}_k$ 称为支撑向量.

在两类样本线性不可分,即一个超平面不能把两类完全分开的情况下,引入松弛变量

$$\xi_i \geq 0, \quad i = 1,\cdots,n$$

使超平面 $w^{\mathrm{T}}\vec{x}_* + b = 0$ 满足:

$$w^{\mathrm{T}}\vec{x}_i + b \geq 1 - \xi_i \quad \text{当 } y_i = +1$$

$$w^{\mathrm{T}}\vec{x}_i + b \leq -1 + \xi_i \quad \text{当 } y_i = -1$$

不难发现, $\sum_{i=1}^{n}\xi_i$ 是错分样本数的一个上界,因此引入以下目标函数

$$\Psi(w,\xi) = \frac{1}{2}w^{\mathrm{T}}w + C\sum_{i=1}^{N}\xi_i$$

其中: C 是一个正常数,可以称为惩罚因子,如果 C 很大,说明对于错分的惩罚大. 此时 SVC 通过求解以下二次规划(3) 来实现

$$\left.\begin{aligned}&\min \Psi(w,\xi)\\&\text{s.t.}\quad \sum_{i=1}^{n}\alpha_i y_i = 0\\&0 \leq \alpha_i \leq C, i = 1,\cdots,n\end{aligned}\right\} \tag{3}$$

与线性情形类似,可以把(3) 转化为它的对偶规划(4)

$$\left.\begin{aligned}&\max \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \vec{x}_i^{\mathrm{T}}\vec{x}_j\\&\text{s.t. } 0 \leq \alpha_i \leq C, i = 1,\cdots,n\\&\sum_{i=1}^{n}\alpha_i y_i = 0\end{aligned}\right\} \tag{4}$$

关于二次规划问题,目前已有多种方法求解,支持向量机问题中常用的方法有内点算法[10]、SMO 算法[11]、分解方法等.

为了更好地处理非线性可分问题,SVC 采用核函数代替样本间的点积,因此可以获得规划问题(5), (2) 和(4) 可以看作它的特例:

$$\left.\begin{aligned}&\max \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j k(\vec{x}_i,\vec{x}_j)\\&\text{s.t. } 0 \leq \alpha_i \leq C, i = 1,\cdots,n\\&\sum_{i=1}^{n}\alpha_i y_i = 0\end{aligned}\right\} \tag{5}$$

上述问题获解时,对所有的 i, Karush-Kuhn-Tucker (KKT) 条件成立:

$$\alpha_i = 0 \Rightarrow y_i f(\vec{x}_i) \geq 1$$

$$0 < \alpha_i < C \Rightarrow y_i f(\vec{x}_i) = 1,$$

$$\alpha_i = C \Rightarrow y_i f(\vec{x}_i) \leq 1$$

其中

$$f(\vec{x}_*) = \operatorname{sgn}\Big(\sum_{\text{与支撑向量对应的 } i} y_i\alpha_i k(\vec{x}_i,\vec{x}_*) + b\Big) \tag{6}$$

是 SVC 的训练结果,它表示了一个分类曲面.

目前常用的核函数主要有:

1) 多项式核函数

$$K(x_i, y) = (x_i y + 1)^d \quad (d = 1,2,\cdots,n) \tag{7}$$

2) 径向基函数核函数

$$K(x_i, y) = \exp\left[-\frac{|x_i - y|^2}{\sigma^2}\right] \tag{8}$$

3) Sigmoid 核函数

$$K(x_i, y) = \operatorname{th}[\Phi(x_i y) + \theta] \tag{9}$$

3 基于支持向量机的企业伙伴寻找模型

动态联盟的盟主必须最先抓住市场机遇,对整个产品概念及其相应的关键技术进行创新,并利用虚拟制造等技术完成新产品的设计开发. 盟主要求盟友与之进行的战略配合分为两类:产品生产和市场开拓. 按照优势互补的原则,盟主企业在分析自身的核心资源后,根据不同的战略配合目标选择盟友[12~14]. 图 2 为合作伙伴评价指标体系.

3.1　企业合作伙伴寻找模型的建立

假设企业合作伙伴寻求实例($\vec{x}_i$, y)组成,这里 $\vec{x}_i \in R^n$,为影响合作伙伴选择的属性, $y \in \{-1,1\}$ 为企业伙伴选择的结果 -1 为不适合,1 为适合,建立基于支持向量机的企业合作伙伴寻求模型,就是寻找 $\vec{x}_i$, y之间的关系:

$$f: R^n \rightarrow \{-1,1\} \qquad (i = 1,2,\cdots,k)$$
$$y = f(\vec{x}_i)$$

式中: R^n 为影响企业合作伙伴选择的 n 个因素.

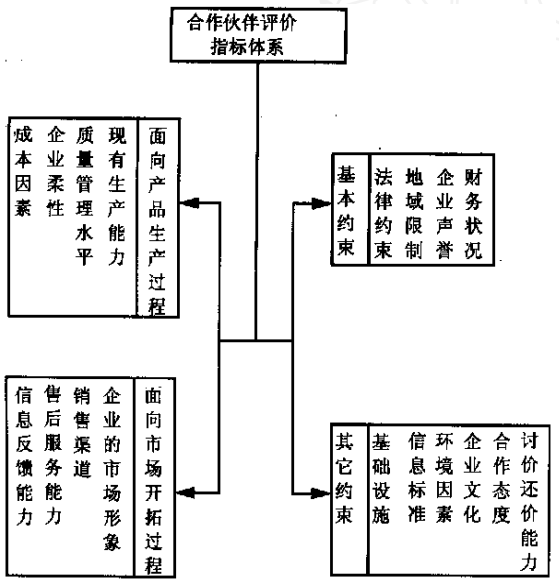

图 2　合作伙伴评价指标体系

理想的企业合作伙伴是由多因素决定的,各因素的关系也非常复杂,因此作者认为,选择企业理想合作伙伴的问题是非线性不可分的问题,所以模型采用式 (5);根据支持向量机理论,企业合作伙伴寻找模型的建立,也即寻求式 (6),式中 $\vec{x}_*$ 为要预测的企业样本; $\vec{x}_i$ 为 k 个样本中的第 i 个样本; $k(\vec{x}_i, \vec{x}_*)$ 为核函数,可在式 (7), (8), (9)中选择. α_i, b 可通过求解二次规划问题获得.

3.2　实验算例

下面以实例说明企业合作伙伴的寻求过程,盟主企业根据企业的实际情况以及合作的目的,选用的评价指标有:备选企业对指定产品的生产能力、对生产指定产品的报价、参与联合开发新产品的能力、盟主企业与备选企业在企业文化方面是否一致或相容. 本文用 VC6.0 开发了相应的支持向量机程序,从企业信息库中取出 78 个企业样本,随机的选取了 64 个样本,针对某企业,由专家评价选出了 32 个理想伙伴样本,32 个不能合作的企业样本,作为寻求模型的学习样本. 其余的 14 个企业样本用于测试,如表 1,

表 1　测试样本

序号	指定产品的生产能力(件)	指定产品报价(万元)	联合开发新产品的能力	企业文化	是否理想
1	4 000	5	差	有冲突	否
2	3 000	4	一般	一致	否
3	2 000	2	强	可兼容	是
4	3 800	4.2	差	有冲突	否
5*	1 800	2	强	一致	是
6	4 200	4.9	差	有冲突	否
7*	2 200	4.4	一般	可兼容	否
8	4 200	2.2	强	一致	是
9	3 600	4.5	一般	有冲突	否
10	3 800	4.5	一般	有冲突	否
11	2 800	4.2	强	可兼容	否
12	3 000	2.5	强	可兼容	是
13	2 000	4.3	一般	可兼容	否
14	4 000	4.5	差	一致	否

本文选用径向基函数作为模型的核函数,采用了分解法进行二次规划, $\sigma^2 = 0.29$. 对于联合开发新产品的能力(差 = 0.3,一般 = 0.6,强 = 0.9) 和企业文化(有冲突 = 0.3,可兼容 = 0.6,一致 = 0.9) 的属性值分别进行处理. 训练结果是理想和不理想两类的支持向量数分别为 28 和 30 个, $b = -0.061\,187\,4$. 对 14 个样本进行测试,结果是错误率为 14.28 %(表 1 中标“ * ”的为出错的样本),此错误率在应用中是可以接受的,由实验结果可知,基于支持向量机的企业合作伙伴寻求模型具有很好的分类预测性能,可以在企业信息库中应用,从海量的数据中发现潜在的、理想的企业合作伙伴,从而为企业的决策提供有力的支持.

4　结束语

企业合作伙伴选择是由多因素决定的,很难建立一种确定的数学模型. 利用企业以往合作的经验,通过对合作实例的分析,寻找理想合作伙伴和各种因素之间的关系是非常重要的. 基于支持向量机的数据挖掘的方法作为一种优秀的学习和搜索算法,具有很好的推广能力和较强的非线性动态数据处理能力,通过对学习样本的学习,预测样本,从而在海量的企业信息库中发现潜在的、有效的理想合作伙伴. 但是,在选择支持向量机参数时还存在很大的盲目性,比如本文中 σ^2 参数的选择,只是根据作者的经

验试取,因此,作者将在以后的研究中着力进行支持向量机参数选择优化问题的研究. 随着研究的进一步深入和对模型不断应用完善,必将会对企业再造过程提供有力的支持.

参 考 文 献:

[1]Byrne A J. The virtual corporation[J]. Business Week, 1993, (2): 98—103.

[2]Prahalad C K, Hamel Gary. The core competence of the corporation[J]. Harvard Business Review, 1990, (56): 79—91.

[3]Zhan S, Poulin D. Partnership management within the virtual enterprise in a network[J]. IEMC, 1996, 645—650.

[4]Talluri S, Baker R C. A quantitative framework for designing efficient business process alliances[A]. International Conference on Engineering Management and Control[C]. 1996. 656—661.

[5]钱碧波, 潘晓弘等. 敏捷虚拟企业合作伙伴选择的方法研究[J]. 机电工程, 1999, 6: 41—44.

[6]Chen M S, Han J, Yu P S. Data mining: An overview from a database perspective[J]. IEEE Transactions on Knowledge and Data Engineering, 1996, 8(6): 866—883.

[7] Fayyad U M, Piatetsky-Shapiro G, Smyth P. From Data Mining to Knowledge Discovery and Data Mining[M]. Cambridge: AAAI Press, 1996. 1—34.

[8]Chen L D, Toru S. Data mining methods, applications, and tools[J]. Information System Management, 2000, 17(1): 65—70.

[9]Burge C J C. A tutorial on support vector machines for pattern recognition[J]. Data Mining and Knowledge Discovery, 1998, (2): 121—167.

[10]Alex J Smola, Bernhard Schoelkopf. A Tutorial on Support Vector Regression[R]. NeuroCOLT2 Technical Report Series NC2-TR-1998030, 1998.

[11]Vladimir N Vapnik. 统计学习理论的本质[M]. 北京:清华大学出版社, 2000.

[12]郑文军等. 虚拟企业合作伙伴评价体系及优化决策[J]. 计算机集成制造系统, 2000, 6(5): 63—67.

[13]冯蔚东, 陈 剑. 虚拟企业中的风险管理与控制研究[J]. 管理科学学报, 2001, 4(3): 1—9.

[14]黄梯云, 冯玉强, 周宽久. 决策支持系统中的建模知识表示研究[J]. 管理科学学报, 2001, 4(1): 45—51.

Analysis of method for enterprise to search partners base on data mining

WEI Hong-ye[1], *LÜ Yong-bo*[1], *HE Wei*[2], *ZHANG Zhong-ye*[1]

1. Beijing Jiaotong University, Beijing 100044, China;

2. 10th Institute of China Aerospace, Beijing 100854, China

Abstract: It is very important for an enterprise to construct its business alliance with its business partners. In this case, selection of optimal partners is crucial to the success of the alliance. This article describes a technology of data mining, i.e. support vector machines, then constructs an optimal partner searching model based on SVMs(support vector machines). Practical effectiveness of SVMs for searching optimal partner in huge enterprise database is proven in the paper.

Key words: enterprise alliance; cooperative partners; data mining; support vector machines