

基于模糊-粗糙集的移动对象 k 近邻预测^{*})

袁 妍 洪晓光

(山东大学计算机科学与技术学院 济南 250101)

摘 要 已往移动对象 k 近邻预测的研究中,有各种不确定性的分析以及对对象本身预测位置不确定性的处理,而预测位置模糊不确定性导致其 k 近邻查询边界的粗糙不确定性的相关处理仍是空白。本文应用模糊-粗糙集的理论,先分析了已有方法得出的预测位置的模糊性,再用传统方法求得基于预测位置的扩展 $k+m$ 近邻集,最后借助模糊-粗糙隶属函数来最终确定所求 k 近邻集合中的各个点。实验数据表明,本方法明显提高了 k 近邻集合相对移动对象实际位置的精确度。

关键词 移动对象, k 近邻查询,模糊集,粗糙集

Fuzzy-Rough Sets for Prediction of Moving Objects' K-Nearest Neighbor

YUAN Yan HONG Xiao-Guang

(School of Computer Science and Technology, Shandong University, Jinan 250101)

Abstract There are many analyses on the diversified uncertainty and many disposals on the uncertainty of objects' predicted position in the previous study of the prediction of moving objects' k -nearest neighbor. But there have not been any measure to deal with the rough-uncertainty of moving objects' k -nearest neighbor set, which is caused by the fuzzy-uncertainty of moving objects' predicted position. In this paper, the theory of fuzzy-rough sets is employed to analyze the fuzzy position of the moving objects and its extended $k+m$ nearest neighbor set. Also, the fuzzy-rough membership function is employed to obtain the final k -nearest neighbor set. A comparison between the processed result and the initial result is made by experiments. Compared to the actual position of the moving objects, the analysis based on the theory of fuzzy-rough sets can promote the precision of its k -nearest neighbor set distinctly.

Keywords Moving object, K-nearest neighbor query, Fuzzy sets, Rough sets

1 前言

随着移动计算,无线通信及定位技术的发展,如何有效地对移动对象进行查询、管理以及提供准确的基于位置近邻服务等需求使得时空数据库研究面临着新的挑战^[1]。目前有很多移动对象索引技术的研究,其中有关未来趋势及其最近邻居预测技术由于更适合在智能交通调度、基于位置服务等领域的应用而得到了广泛的关注^[2],例如,“已知某车现在的位置、时速和运动方向,查询未来一小时后离它最近的几个加油站或者食宿供给处”。

移动对象上安装定位设备便能有效记录移动对象的运动状态,如当前所处的位置、运动速度和方向等,根据这些当前的运动特性便可以对未来进行预测。由于自身及外界诸多因素,对象的运动状态会不断地发生变化,预测过程中就会有多种不确定性出现。移动对象及其 k 近邻预测中不确定性的处理分时间上和空间上两种^[3],如:“预计什么时候到达某地方”、“预计某时刻到达什么地点”。本文主要研究移动对象未来趋势的空间不确定性。

移动对象预测位置的模糊不确定性表现在它真实位置在预测位置周边区域的模糊隶属度,这必然会引起 k 近邻集合基于“ k 近邻”不分明关系的粗糙不确定性,继而累及传统

KNN 得到的 k 近邻的精确度。已往研究没有涉及到这种现象的处理,本文借用模糊-粗糙集理论分析传统方法得到的结果,即通过移动对象预测位置的模糊隶属度,计算其扩展 $k+m$ 近邻的模糊-粗糙隶属函数,从而得出更精确的 k 近邻集合。

第 2 节和第 3 节将分别讨论移动对象及其 k 近邻查询的相关研究和模糊集、粗糙集的相关知识,第 4 节讨论了基于模糊-粗糙集理论的移动对象 k 近邻预测的具体方法和步骤,第 5 节通过实验数据得出相应参数的合适取值,以及最终近邻集的结果更精确的结论,最后总结了新方法提高精确度的同时存在的影响精确度的有待进一步研究的部分。

2 相关研究

在移动计算、位置服务等新兴应用中,需要对移动对象未来位置预测提供相关服务,这些相关的查询索引方法是目前研究的热点。

2.1 移动对象未来位置的预测

就移动对象的位置预测方面,有各种不同的索引方法。对偶转换技术(Dual transformation)将在 d 维空间中运行物体的预测问题转换到 $2d$ 维空间^[4]。Patel J. M 将此种转换由一维空间扩展到多维空间,并提出名为 STRIPES 的索引

^{*})基金项目:教育部重点科技计划项目(03102);国家自然科学基金(60673130);省重大科技专项项目(2004GG4201022);山东省中青年科学家奖励基金(2005BS01002)。袁 妍 硕士研究生,主要研究方向:数据库高维索引,粗糙集理论、模糊集理论在数据库中的应用等;洪晓光 教授,硕士生导师,主要研究方向:数据库理论与应用,非经典数据库。

结构,用规则的分级网络划分实现对偶转换空间的索引^[5]。TPR-TREE(Time Parameterized R-tree)用以时间为参数的边界矩形实现移动对象的分类索引^[6]。针对 TPR-TREE 随时间的推移其查询性能降低的特点,STAR-TREE (Spatio-temporal Self Adjusting R-TREE)当查询性能降低到设定的限度时,将自动调整索引结构^[7]。Saltenis 对 TPR-TREE 进行扩展,引入移动对象的失效时间 (expiration time)^[8]。最近,Jensen 提出了基于 B^+ -TREE 的索引结构 B^x -TREE^[9]。 B^x -TREE 仍然用线性函数 $f(t)$ 来描述移动对象的运动状态,但移动对象的分类索引方式不再使用 TPBR,而是按移动对象的更新时刻的时间轴分类。

以上的算法都是仅有当前速度影响预测。而预测本身的不确定性更强,移动物体的实际位置常处于不确定体积之外,导致频繁的速度更新,因此指数滤波方法被较多地用于时间序列预测中^[10]。但这仍是有其不确定性存在的,本文将在预测位置模糊不确定性方面对其分析,即通过一个函数来表现它真实位置在预测位置周边区域的模糊隶属度。这主要为下一步获得更精确的 k 近邻集合服务。

2.2 移动对象 k 近邻的预测

移动对象的 k 近邻(K-Nearest Neighbor,KNN) 查询用于找出相对于某个查询对象,被查询对象中距离它最近的一个或多个最近邻对象。目前,已经提出多种 KNN 查询算法,但大多都是基于 R 树的,适用于静态的对象^[11, 12]。第一个涉及移动对象的最近邻查询算法采用取样的方法,然而这无法保证查询结果的精确性^[13]。Benetis 给出了一种通过数学方法来支持移动对象的连续最近邻查询的算法^[14],但该算法每次查询只能回答一个最近邻查询对象。所以 Tao 在 DF 和 BF 查询算法的基础上,提出用于移动对象的 TP 查询算法^[15]。

以上方法都没有涉及到预测移动对象位置的模糊不确定性会影响 k 近邻精确度,以及 k 近邻集合基于“ k 近邻”不明确关系的粗糙不确定性。本文借用模糊-粗糙集理论,通过移动对象预测位置的模糊隶属度,计算其扩展 $k+m$ 近邻的模糊-粗糙隶属函数,最终得出更精确的 k 近邻集合中的各个点。

3 模糊-粗糙集相关知识

模糊集与粗糙集是两种应用最为广泛的处理不确定性的方法,两者既有区别,又相互补充。

3.1 模糊集

模糊集理论以隶属度作为基石,模糊子集由其隶属函数来描述^[16]。给定论域 U 上的一个模糊子集 A 。

定义 1 对于任意 $x \in U$ 都确定了一个 $\mu_A(x)$, $\mu_A(x)$ 称为 x 对 A 的隶属度。映射 $x \rightarrow \mu_A(x)$ 称为 A 的隶属函数。 A 的隶属函数是一个 A 的连续特征函数。隶属函数的值域为 $\mu_A(x) \in [0, 1]$, 当 $\mu_A(x) = \{0, 1\}$ 时,模糊子集 A 就是一个普通集合,在模糊集中,我们不能简单地某元素属于不属于某个集合,而只能说多大程度属于某个集合。

3.2 粗糙集

粗糙集理论所有的概念和运算都是以不可分辨关系(一种等价关系)为基础,通过引入上近似集和下近似集在集合运算上进行的,这通常称为粗糙集理论的代数描述^[17]。设 U 是有限论域, R 是 U 上的等价关系, $[x]_R$ 是 U 上的 R 等价类,一个知识库就是一个关系系统 $K = (U, R)$ 。

定义 2 给定知识库 $K = (U, R)$,对于每个子集 $S \subseteq U$ 和 U 上的一个等价关系 R ,如果 $R_*(S) = \{x \subseteq U | [x]_R \subseteq S\}$,则称 $R_*(S)$ 是 S 的 R 下近似;如果 $R^*(S) = \{x \subseteq U | [x]_R \cap S \neq \emptyset\}$,则称 $R^*(S)$ 是 S 的 R 上近似。此外,称 $(R_*(S), R^*(S))$ 为 S 的粗糙集,称 R 为 U 上的不分明关系。

3.3 模糊-粗糙集

模糊集对集合中子类边界的不清楚定义进行了模型化,反映了由类别之间的重叠体现出的隶属边界的模糊性;而粗糙集则体现了由于属性不足引起的集合中对象间的不可区分性,即由于知识的粒度而导致的粗糙性。它们处理的是两种不同类别的模糊和不确定性,两者的有机结合能更好地处理不完全知识。由此 D. Dubois, H. Prade 在文[18]中提出了模糊-粗糙集(Fuzzy-Rough Sets)的概念。模糊-粗糙集的主要思想是当等价关系使模糊集合的论域变得粗糙时,定义此模糊集合的相应上近似和下近似;或者把等价关系弱化为模糊相似关系,从而得到一个更具表达力的粗糙模型。

定义 3 设 (U, R) 是 Pawlak 近似空间,即 R 是论域 U 上的一个等价关系。当等价类的元素所属的类别不明确时,等价类便表示为模糊集的形式 $F = \{F_1, F_2, \dots, F_H\}$, F_j 是模糊集, $j \in \{1, 2, \dots, H\}$ 。给定 R 上的一个模糊划分 θ ,利用上近似 θ^* 和下近似 θ_* 的形式表达任一模糊集合 F ,称 $\theta^*(F)$ 和 $\theta_*(F)$ 为模糊-粗糙集。定义模糊上、下近似如下:

$$M_i = \mu_{\theta^*(F)}(F_i) = \sup_x (\min(\mu_{F_i}(x), \mu_F(x))) \forall i$$

$$m_i = \mu_{\theta_*(F)}(F_i) = \inf_x (\max(1 - \mu_{F_i}(x), \mu_F(x))) \forall i$$

上式表示了模糊事件 F 的可能性和必然性程度。 M_i 是 F 中的 F_i 的可能性隶属度, m_i 是 F 中的 F_i 的必然性隶属度。

4 基于模糊-粗糙集的移动对象 k 近邻预测

本方法以 TPR-TREE 为基本结构预测移动对象位置和其 k 近邻,应用模糊-粗糙集理论分析其预测位置的模糊不确定性、相关 k 近邻集合的粗糙不确定性。

4.1 移动对象预测位置的不确定性

设需要预测的移动对象为 P ,其位置的 n 维参考坐标是 $S(p_1, p_2, \dots, p_n)$ 。移动过程中, P 的速度会受其它多种因素影响而有相应变动,所以它的实际位置不一定在预测位置 S ,并且在其附近每个位置点都有可能性。

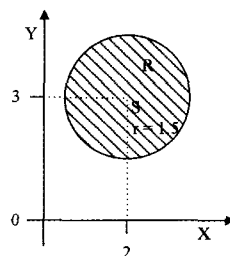


图 1

以 S 为中心,设定一模糊因子 r ,从而有了预测位置的模糊空间 $R([p_{1-r}, p_{1+r}], [p_{2-r}, p_{2+r}], \dots, [p_{n-r}, p_{n+r}])$ 。从当前时间到预测时间的差额为 Δt 。 r 与 Δt 成正比,可根据 Δt 大小来拟订。图 1 的阴影区演示了二维状态下, P 以 $S(2, 3)$ 为参考坐标的模糊空间 R 。

移动对象 P 实际位置相对模糊空间 R 的隶属函数 μ 为:

$$\mu = [1 + (\frac{|P-S|}{\alpha})^2]^{-1}$$

其中 $|P-S|$ 为 n 维空间中 P 和 S 间的距离, α 是相对于 r 的一个参数因子, $\alpha > r$,有多种取法,如取 $\alpha = 2r$ 。这样得出的隶属模糊空间曲线如图2。

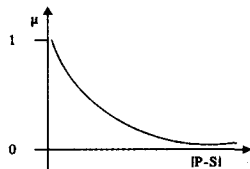


图2

显然,随着 P 与 S 间距离的增大, μ 越来越小,逐渐趋近于零。也就是说 P 位置距离 S 越远,可能性越低。

4.2 移动对象预测近邻集的不确定性

移动对象的预测位置本身有模糊不确定性,这必然会会影响其 k 近邻查询的精确度,即产生 k 近邻集合基于“ k 近邻”不分明关系的粗糙不确定性。有了移动对象 P 的预测位置 S 后,为减少 P 位置的模糊性对最终近邻集结果精确度的影响,当要求查询 k 近邻集合 U 时,暂时查询 $k+m$ 近邻集合 $W(d_1, d_2, \dots, d_{k+m})$, $m < k$,从 d_1 至 d_{k+m} 到 S 的距离依次递增。 D_i 表示点 d_i 到位置 S 的距离。根据模糊-粗糙集的理论对 W 进行如下分析,从而得出更为精确的 k 近邻集。

首先,在模糊空间 R 中取 $2n$ 个 P 的可能位置点 $(x_1, x_2, \dots, x_{2n})$ 。 W 中各点到这 $2n$ 个取样点间的距离可以表示为:

$$D_{ij} = \mu^{-1} |d_i - x_j|, (i=1, 2, \dots, k+m; j=1, 2, \dots, 2n)$$

$|d_i - x_j|$ 表示点 d_i 和点 x_j 间的距离。

第二步,有了 D_{ij} ,我们可以得出 d_i 属于 k 近邻集合 U 的模糊度为:

$$\mu(d_i) = \begin{cases} 0.51 + 0.49 \left(\frac{\sum_{j=1}^{2n} D_{ij}}{2n \cdot D_i} \right)^{-1}, & i \leq k, \\ 0.49 \left(\frac{\sum_{j=1}^{2n} D_{ij}}{2n \cdot D_i} \right)^{-1}, & i > k \end{cases}$$

$(i=1, 2, \dots, k+m)$

此外,就所求 k 近邻集合 U 而言,移动对象 P 空间位置的模糊不确定性导致了集合 U 的边界是粗糙不确定的。又由于此处更为关注集合 U 内的各点到移动对象 P 实际位置的距离,而集合 U 内部或许存在距离很近的不分明点,因此应对这些不分明点进行分析。

第三步,考虑集合 U 内各个近邻点之间的模糊相似度:

$$\tilde{\mu}(d_i) = \sum_{j=1}^{k+m} (1 + \beta |d_i - d_j|^{2/(k+m-1)})^{-1},$$

$(j=1, 2, \dots, k+m)$

其中的指数 $2/(k+m-1)$ 是为了调节距离对函数值的影响; β 为调节参数,用于控制函数值的变换范围,有多种取法,可以固定或者根据距离来定。如取:

$$\beta = \left(\frac{1}{k+m} \sum_{j=1}^{k+m} |d_i - d_j|^{2/(k+m-1)} \right)^{-1},$$

$(j=1, 2, \dots, k+m)$

第四步,对于移动对象 P 的实际位置和其 k 近邻集合 U 出现的这两类不确定性,用模糊-粗糙隶属函数来综合两者的影响:

$$T(d_i) = \frac{1}{k+m} \tilde{\mu}(d_i) \cdot \mu(d_i), (i=1, 2, \dots, k+m)$$

最后,移动对象的预测 k 近邻集合 U 由此模糊-粗糙隶属

函数来确定,分别计算 W 中 $k+m$ 个点的 $T(d_i)$,然后由大到小排序成数组 $Y(q_1, q_2, \dots, q_{k+m})$,数值较大的前 k 个(q_1 到 q_k)即组成所求 k 近邻集合 U 的成员。

5 精确度评估

在我们数值分析中,比较不同维数下不同的 k, r 取值时,模糊-粗糙集方法得出的结果相对于实际结果的精确度也不同。图3描述了 r 取中间值时维数和 k 对精确度的影响,可以看出,中维数时,精确度较高,并且同维数下 k 越大,得到的移动对象 k 近邻集合的精确度越理想。

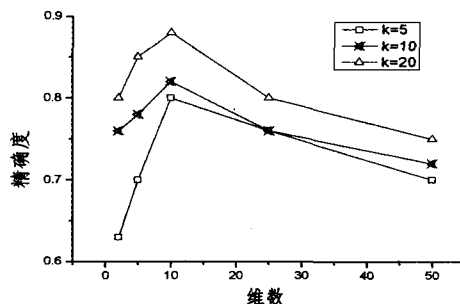


图3 普通 r 时维数和 k 对精确度的影响

取一般情况下的 $k=10$,维数和 r 大小对精确度的影响在图4中的曲线可以体现。 r 取值较大或较小时都会影响最后 k 近邻结果的精确度。特别当高维较大 r 时,精确度最低。

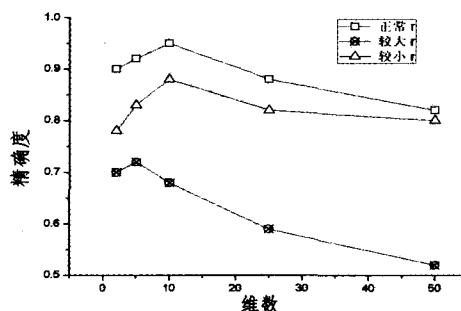


图4 $k=10$ 时维数和 r 对精确度的影响

最后,我们比较一下使用模糊-粗糙集方法分析后的结果和原始TPR-TREE结果。

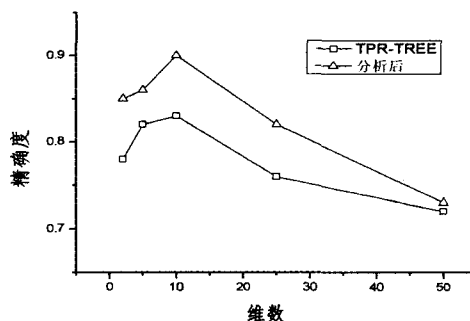


图5 分析前后精确度的变化

图5显示了普通 r 和 $k=10$ 情况下分析方法前后精确度的比较,显然,模糊-粗糙集理论方法有效地提高了移动对象 k 近邻集合预测的精确度。

结论 在移动计算、智能交通调度、基于位置服务、天气预报等领域的新兴应用中,需要对移动对象当前以及未来位

置的预测提供相关服务。而移动对象预测问题的不确定性是普遍存在并且急需解决的问题,不确定性处理的好还是差直接影响到查询、管理结果的精确度,进而影响到用户需求。其中,空间不确定性中的移动对象预测位置的模糊不确定性必然会累及传统 KNN 得到的 k 近邻的精确度,同时会引起 k 近邻集合基于“ k 近邻”不分明关系的粗糙不确定性。针对这种已往研究没有涉及到的不确定性的处理,本文通过传统方法扩大 k 近邻查询范围,应用模糊-粗糙集理论分析预测得出的 $k+m$ 近邻集合,借助移动对象预测位置的模糊隶属度和模糊-粗糙隶属函数,最终确定所求 k 近邻集合中的各个点。尽管计算量略微增大,但数值分析可以表明本方法很大程度上提高了移动对象 k 近邻预测的精确度。

另外,本方法的模糊-粗糙隶属函数中用到了 $k+m$ 近邻集合 W 中各点到模糊空间 R 中 $2n$ 个移动对象 P 可能位置取样点的距离。这种取样的方法会一定程度上影响精确度。所以,未来的工作将致力于基于非取样方法的移动对象 k 近邻预测的精确度的进一步提高。

参考文献

- Konig-Ries B, Makki K. Research Direction for Developing an Infrastructure for Mobile & Wireless Systems. Consensus Report of the NSF Workshop, Scottsdale, AZ, October 2001
- Chon H D, Agrawal D, Abbadi A E. Data Management for Moving Objects. In: Proceeding of International Telematics and LBS Workshop, 2003
- Trajcevski G. Managing Uncertainty in Moving Objects Databases [J]. ACM Transactions on Database Systems, 2004, 29(3): 463~507
- Agarwal P K, Arge L, Erickson J. Indexing Moving Points. In: Proceedings of PODS, 2000. 175~186
- Patel J M, Chen Y, Chakka V P. STRIPES: An Efficient Index for Predicted Trajectories. In: Proceedings of SIGMOD, 2004.

(上接第 139 页)

TCP 连接记录组成,每个记录有 34 个连续属性,共有 5 个类别。在本实验中,主要参数设置如下:窗口大小 100,结点更新阈值 90%,结点添加阈值 90%,类最大生存期 10。实验中采用共 5 组数据,每组数据为 4M,表 1 显示了 InnoDB2 算法和传统算法的分类准确率和时间效率的平均值。可以看出 InnoDB2 算法和传统算法有相似的分类准确率,但是 InnoDB2 有较好的时间运行效率。

表 1 算法分类准确率和训练时间比较

	分类准确率(%)		训练时间(s)	
	InnoDB2	IDB2	InnoDB2	IDB2
1	98.93	98.00	0.2528	0.3029
2	98.36	98.47	0.5950	0.6394
3	93.61	93.94	0.7491	0.8434
4	96.28	95.34	0.6010	0.6805
5	95.67	95.98	0.4511	0.5416

结束语 本文提出了一个能实时检测局部概念漂移并随之自适应调整的数据流分类算法 InnoDB2。当局部概念漂移出现时,该算法不是重新构造一个全新的 DB2 层次分类树,而是仅更新漂移所影响的局部结点,使得该算法具有较好的时间运行效率。实验结果证实该算法可以较好解决数据流分类中局部概念漂移的问题。

637~646

- Saltenis S, Jensen C S, Leutenegger S T, et al. Indexing the Positions of Continuously Moving Objects. In: Proceedings of SIGMOD, 2000. 331~342
- Procopiu C M, Agarwal P K, Har-Peled S. STAR-TREE: An Efficient Self-Adjusting Index for Moving Objects. In: Proceedings of the Intl Workshop on Algorithm Engineering and Experiments, 2002. 178~193
- Tao Y, Papadias D, Sun J. The TPR*-TREE: An Optimized Spatio-Temporal Access Method for Predictive Queries. In: Proceedings VLDB, 2003. 790~801
- Jensen C S, Lin D, Ooi B C. Query and Update Efficient B+-TREE Based Indexing of Moving Objects. In: Proceedings of VLDB, 2004. 768~779
- Xu B, Wolfson O. Time-Series Prediction with Applications to Traffic and Moving Objects Database [c]. In: The Third International ACM Workshop on Data Engineering for Wireless and Mobile Access(MobiDE'03), San Diego, California, 2003. 56~60
- Cheung K L, Fu A W. Enhanced Nearest Neighbour Search on the R-TREE [J]. ACM SIGMOD Record, 1998, 27(3): 16~21
- Roussopoulos N, Kelley S, Vincent F. Nearest Neighbor Queries [C]. In: Proceedings of the ACM SIGMOD International Conference on Management of Data, 1995. 71~79
- Song Z, Roussopoulos N. K-Nearest Neighbor Search for Moving Query Point [C]. In: Proceedings of the 7th International Symposium on Spatial and Temporal Databases, 2001. 79~96
- Benetis R, Jensen C S, Karciuskas G, et al. Nearest Neighbor and Reverse Nearest Neighbor Queries for Moving Objects [C]. In: Proceedings of the International Data Engineering and Applications Symposium, 2002. 44~53
- Tao Y, Papadias D. Time-parameterized Queries in Spatio-temporal Databases [C]. In: Proceedings of the ACM SIGMOD Conference, 2002. 334~345
- Zadeh L A. Fuzzy Sets. Information and Control, 1965. 338~353
- Pawlak Z. Rough Sets. International Journal of Computer and Information Sciences, 1982. 341~356
- Dubois D, Prade H. Rough fuzzy sets and fuzzy rough sets [J]. International Journal of General Systems, 1990. 191~209

参考文献

- Babcock B, Babu S, Datar M, et al. Models and Issues in Data Stream Systems. In: Proc. of the 21st ACM Symposium on Principles of Database Systems (PODS2002), 2002. 1~16
- 金澈清, 钱卫宁, 周傲英. 流数据分析与管理综述. 软件学报, 2004, 15(8): 1172~1181
- 孙玉芬, 卢炎生. 流数据挖掘综述. 计算机科学, 2007, 34(1): 1~5
- Gama J, Castillo G. Learning with Local Drift Detection. In: ADMA, 2006. 42~55
- Tsymbal A, Pechenizkiy M, Cunningham P, et al. Handling Local Concept Drift with Dynamic Integration of Classifiers. Domain of Antibiotic Resistance in Nosocomial Infections. In: CBMS, 2006. 679~684
- Volkan V, Jennifer G D. A Hierarchical Method for Multi-class Support Vector Machines. In: ICML, 2004. 105~113
- Vapnik V. Statistical Learning Theory. Wiley, New York, 1998
- Kreel U. Pairwise Classification and Support Vector Machines. Advances in Kernel Methods-Support Vector Learning, Cambridge, MIT Press, 1999. 255~268
- Newman D J, Hettich S, Blake C L, et al. UCI Repository of Machine Learning Databases, 1998