GPS 轨迹中出行目的提取的一种智能算法

邓中伟，季民河

（华东师范大学地理信息科学教育部重点实验室，上海 200062）

摘要：从理论角度归纳和定义与出行目的识别相关的语义信息类型（时间、土地利用类、交通行为特征和受访者社会经济属性等），探讨如何通过信息集成实现GPS轨迹数据中出行目的的高效提取，在此基础上提出基于机器学习的通用建模思路。实证案例使用36位上海市居民的被动式GPS交通调查数据，利用C5.0算法分析构建出行目的提取决策树。实验结果表明：综合异源异质相关信息从GPS轨迹数据中提取行程目的具有较高精度（约87.6%），研究方法具有一定的可行性和通用性。

关键词：交通规划；居民出行调查；出行目的；GPS轨迹数据；C5.0

中图分类号：U121 **文献标识码**：A **文章编号**：1674-2850(2011)06-1064-7

Deducing trip purpose from GPS trajectory data: a machine learning approach

DENG Zhongwei, JI Minhe

(*Key Laboratory of Geographic Information Science, Ministry of Education, East China Normal University, Shanghai* 200062, *China*)

Abstract: This paper presents a machine learning approach to derive trip purpose from GPS trajectory data coupled with other relevant data sources. This approach employs a number of attributes (i. e. time stamp and land-use type of trip ends, a set of spatiotemporal indices of travel, and demographic and socioeconomic characteristics of respondents) to construct a decision tree for purpose classification. Each attribute provides partial evidence to the depiction of a given purpose, but this depiction may or may not be monotonic, and none of them can work alone towards the goal. A reasoning procedure using the adaptive boosting technique was designed to explore how these attributes could work together to achieve trip purpose derivation. This technique generated multiple decision trees to improve the classification results through a mechanism of voting from these trees. Each tree was constructed in the depth-first fashion with the root node and the split of subsequent nodes being determined on the basis of the gain-ratio computed for the relevant attributes. This procedure was implemented in the C5.0 machine learning environment with 226 GPS trip records collected from 36 respondents. The experimental results seemed rather promising: using 10 iterations for adaptive boosting, an overall classification accuracy of 87.6% was achieved.

Key words: transportation planning; household travel survey; trip purpose; GPS trajectory data; C5.0

0 引言

传统的居民出行调查方法主要包括家访/邮寄问卷调查、电话询问调查、电脑辅助电话调查（computer assisted telephone interview，CATI）、电脑辅助自我填报方法调查（computer-assisted self interview，CASI）等[1~4]。这些方法虽然成熟通用，但出行数据的收集均依赖于受访者的主观记忆和自我报告，容易导致行程漏报错报；很多受访者对所经区域不熟悉，不能提供量化甚至定性定位信息；行程距离、时间和速度等数据也不够精确；而且由于人的记忆和耐性有限，难以保障多天出行数据的

基金项目：国家自然科学基金（40771138，40671074）

作者简介：邓中伟（1984—），男，博士研究生，主要研究方向：交通地理信息智能化服务

通信联系人：季民河，教授，主要研究方向：空间统计分析、地理信息软计算、人地关系模拟. E-mail：mhji@geo.ecnu.edu.cn

收集质量。

以“智慧地球”为背景的新信息时代的到来改变了数据收集手段，各种传感器的广泛应用使得数据的收集向仪器自动记录的方向发展，数据收集的方式也由专业人员收集向广大用户参与的自发性数据收集转变。GPS仪器能够自动精确记录出行轨迹和时间，与地图数据相结合提取所需交通信息[5]，克服了上述传统交通方法的局限，自20世纪90年代中期开始逐渐在交通调查中广泛应用[6]，并有代替传统调查的趋势[7]。

出行距离、出行速度、出行时长等信息可以直接从轨迹数据中计算获得；但交通模式、出行目的等更高语义层次的信息，需要在其他信息辅助下经过逻辑推理来判定。目前的研究多是将土地利用类型信息、道路信息，人为设定简单规则提取居民出行目的，取得了一定效果[8~9]，但是人为设定规则具有主观性和不可扩展性，由于各人收集数据的方式和质量不同，处理的方式也大多从自己的数据出发，缺少通用的理论模式。

机器学习采用归纳、综合的方法改善自身的性能，自动发现训练样本中的定理、定律和规则等。决策树是应用最为广泛的机器学习方法的一种，它在背景知识缺乏的情况下，能够从一组无次序、无规则的实例中推导出决策树表示形式的分类规则，形成分类器和预测模型，可以对未知数据进行分类或预测、数据挖掘等[10]。研究首先归纳和定义与出行目的相关的信息类型，以及如何从GPS轨迹数据中获得这些信息，提出了基于机器学习的智能化提取出行目的的通用建模思路，最后通过分析数据的特点，选择使用C5.0算法，对上海居民出行目的智能提取进行实证研究，验证研究方法的科学性和可行性。

1 方法论

1.1 建模思想

在考虑基于GPS轨迹数据智能提取居民出行目的之前，思考在日常生活中对出行目的逻辑推定的过程。

情形1：“张三从家到了学校”

根据常识，居民的出行活动往往发生在特定的场所，由于“家”和“学校”的界定，可以推断该出行活动是“上学、上班、送孩子上学、业务办公、观光休闲”中的一种，出行端点的土地利用类型信息与出行目的密切相关，为出行目的识别提供了重要依据。

情形2：“他在周一的早上7：30出发，8：00到达”

增加了时间信息后，候选出行目的“旅游休闲”的可能性降低，出行活动具有特定的时间结构，“观光休闲”的行为极少发生在工作日的早晨，时间信息获得使候选出行目的进一步缩小范围，成为出行目的识别的另一重要依据。

情形3：“在学校逗留了5 min”

“逗留5 min”的事实与“上学、工作、业务办公”推断相矛盾，至此，可以将出行目的判定为“送孩子上学”。不同的出行活动往往表现出不同的交通行为特征，逗留时间、出行时长、出行速度、出行距离、交通模式的差异同样是出行目的推定的重要依据。

情形4：“张三是位45岁的律师，他是一个12岁孩子的父亲”

在了解张三的职业和家庭情况后，甚至在情形1～情形3部分信息缺失的情况下也可以轻易判定这是一个“送孩子上学”的出行。将此类信息定义为受访者的家庭社会经济属性，这些信息在调查时常常作为静态信息存入数据库中。

综上所述，出行目的的推定可以看作一个基于一定的先验知识，在获得特定信息情况下的逻辑推理过程，推理结果的可靠性取决于时间、土地利用类型、交通行为特征、受访者社会经济属性等信息，它们不必非常完备，相互之间可以补充和佐证。

1.2 机器学习算法：C5.0

决策树算法是以实例为基础的归纳学习算法，以其易于提取显示规则、计算量相对较小、可以显示重要决策属性和较高的分类准确率等优点而广泛应用。决策树的构建是一种自上而下，分而治之的归纳过程，即从根节点开始，对每个非叶节点，找出其中对应样本集中的一个属性对样本进行测试，根据不同的测试结果将训练样本集划分为若干个子样集，每个子样本构成一个新叶节点，对新节点再重复上述划分过程，这样不断循环，直到达到特定的终止条件。其中，测试属性的选择和如何划分样本集是构成决策树的关键环节，不同的决策树算法在此使用的技术不尽相同。研究采用了 C5.0 算法，它同其前身 C4.5 一样，是用信息增益比率测试样本属性，改进之处包括计算和内存占用性能上有极大提高（～90%），更加适用于大数据集分类；采用 Boosting 方法改进了精度，可使误差率降低一半；增加支持日期变量等新的变量类型[11]。

一个属性的信息增益就是由于使用这个属性分割样例而导致的期望熵降低。属性 A 相对样例集合 S 的信息增益 Gain(S，A)被定义为

$$\mathrm{Gain}(S,A)\equiv \mathrm{Entropy}(S)-\sum_{v\in \mathrm{Values}(A)}\frac{|S_v|}{|S|}\mathrm{Entropy}(S_v), \tag{1}$$

其中，Values(A) 为属性 A 所有可能值的集合；S_v 为 S 中属性 A 的值为 v 的子集（即 $S_v=\{s\in S\mid A(s)=v\}$）. 式（1）中第一项是集合 S 的熵，第二项是用 A 分类 S 后熵的期望值，因此，Gain(S，A)为由于知道属性 A 的值而导致的期望熵减少。但 Gain(S，A) 有偏向多值的属性，为克服这个局限，可用增益比率（gain ratio）取代。增益比率通过加入分裂信息（split information）项惩罚多值属性，用以衡量属性分裂数据的广度和均匀性：

$$\mathrm{SplitInformation}(S,A)\equiv -\sum_{i=1}^{c}\frac{|S_i|}{|S|}\log_2\frac{|S_i|}{|S|}, \tag{2}$$

其中，S_1 到 S_c 为 c 个值的属性 A 分割 S 而形成的 c 个样例子集。增益比率度量由 Gain(S，A) 和分裂信息共同定义，即

$$\mathrm{GainRatio}(S,A)\equiv\frac{\mathrm{Gain}(S,A)}{\mathrm{SplitInformation}(S,A)}. \tag{3}$$

研究在具体分析中使用了 C5.0 的 Boosting 方法，通过对同一数据集进行多次迭代计算提高分类精度。其过程是首先采用传统方法生成一个用于分类的决策树，在保证上次错分的样例在本次迭代分类正确的前提下重新生成新的决策树，通过逐次迭代生成多个决策树，决策树的数目取决于迭代的次数，然后，根据分类的精度对这些树赋予不同的权重形成一个综合的用于分类的决策树。

1.3 属性变量的定义和提取

依据 1.1 节论述，出行目的识别可以转化为基于受访者社会经济属性、时间、交通行为特征、土地利用类型等部分信息进行分类的问题。如图 1 所示，这些信息可以通过以下方式获得。

社会经济属性：居民出行交通调查涵盖了受访者家庭和个人的社会、经济属性内容。GPS 仪器记录的是时空信息，无法收集此类信息，实践中仍通过传统调查方法进行静态信息收集。近年来，数据的收集方式有所改变，如本研究团队利用互联网为载体，先后设计了 2 种不同的调查方式[12～13]。在分析中，社会经济属性表通过受访者的 ID 号与受访者的 GPS 轨迹数据表实现唯一关联。

时间信息：为 GPS 仪器记录的、由定位卫星精确时钟确定的时间信息。应用中根据常识将连续记录的时间分为工作日和非工作日。由于 C5.0 算法能够处理连续的时间数据类型，未对一天中的时间进行分段处理。

交通行为特征信息：GPS 轨迹数据以秒为间隔单位记录的点位信息，因此每个点位具有空间坐标（精度：3～15 m）和时间戳记。轨迹是行程提取的基础数据。提取过程为：先根据前后点的时间和距

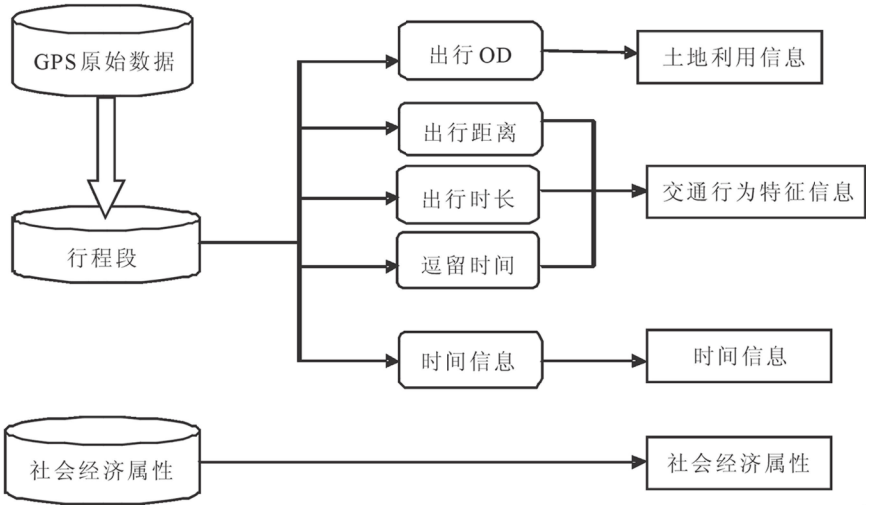

图1 属性变量的定义与提取

Fig. 1 Attribute generation and deducing

离计算每个点位的速度，然后将速度小于5 km/h且连续超过120 s的点集综合简化为一个“停靠点”，并根据“停靠点”将连续的轨迹线划分为不同的行程段，最后计算各个行程段的出行时长、距离、平均速度以及不同行程段间的逗留时间，作为该行程段的交通行为特征属性。

土地利用类型：“停靠点”同时也是各个行程的起点或终点（orgin-destination，OD）。在ARCGIS环境下，将各个行程端点提出生成点图层，与土地利用类型图做“面内点”（point-in-polygon）叠置分析，将土地利用类型信息赋值给行程的起点和终点。

1.4 C5.0建模

GPS轨迹通过上节描述的处理被划分为行程，且每条行程赋有受访者社会经济、时间、交通行为特征、土地利用类型等不同的属性变量。所有行程及其属性组织成一个数据库的表。至此，出行目的识别即可转为依据这些多源异质属性变量的分类。C5.0分类建模算法比较适合这种训练实例用“属性-值”对（pair）表示，目标函数为离散值输出，可能需要析取描述的、而训练数据又包含错误和属性值存在缺失的情况[14]。表1列出了研究对行程进行目的分类的机器学习的输出和输入变量。

表1 决策树的输入属性变量和输出目标变量

Table. 1 Input and target variables for trip purpose identification

因素	类型	变量
输入	土地利用	土地利用类型
	社会经济属性	职业，收入水平，家庭结构，年龄
	时间	工作日，非工作日，一天中不同的时段
	交通行为特征	平均速度，交通模式，出行距离，逗留时长
输出	出行目的	出行目的

2 实例研究

2.1 数据的收集和预处理

实例研究选定了50位拥有私家车的华东师范大学教职工进行了为期3 d的调查。受访者除了佩戴GPS收集出行的时空数据，还需填写传统的出行调查表格，收集家庭成员、职业、年龄、性别等调查者的社会经济属性信息，并需额外填写每个行程的目的和地点，用于构造训练样本；具体调查流程参见文献[12]。调查使用了12台GPS轨迹数据记录仪，历时2.5个月。在征集阶段，有12位受访者因故无法接受调查，故总体回应率为76%．调查过程中出现一些操作失误和硬件故障，导致少数样本数据收集不完整，最后仅有36位受访者的数据用于分析，部分受访者轨迹数据可视化如图2所示。

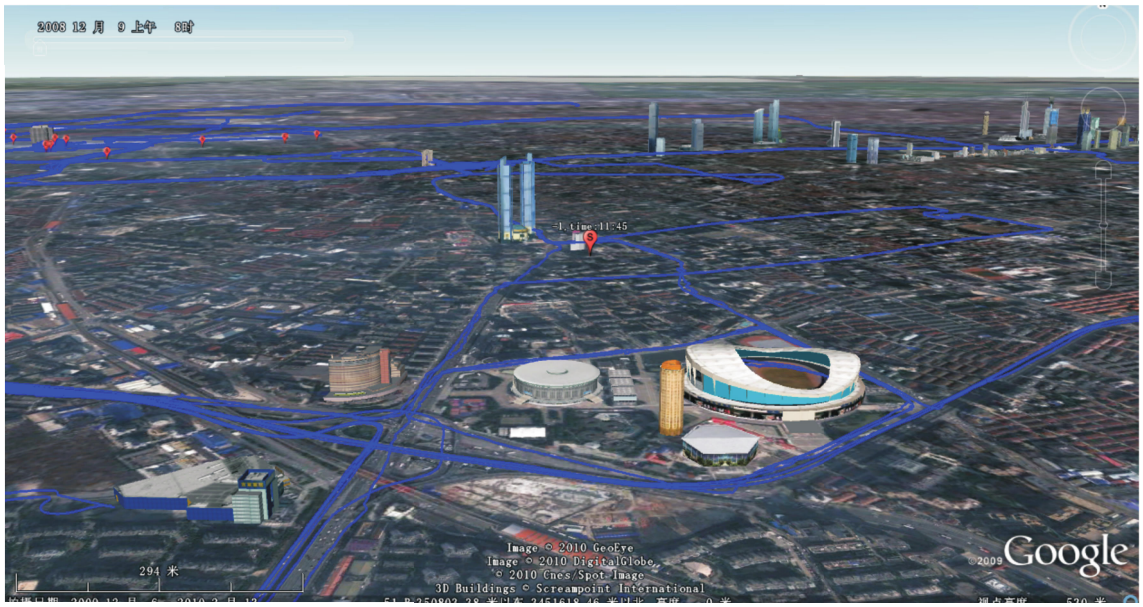

图 2　基于 Google Earth 的受访者轨迹可视化示意图

Fig. 2　Visualization of respondents' trajectory based on Google Earth

按照 1.3 节基于 GPS 轨迹数据和受访者的静态属性信息进行了变量的定义和建模，经过数据预处理得到 36 人、共 226 条行程段，这些行程段具备有了行程端点的土地利用类型、交通行为特征、受访者社会经济属性、行程发生时间等信息，同时还通过传统的调查方法获得了这些行程目的，从而构造了机器学习的训练样本，训练样例示例如表 2 所示。

表 2　训练样例示例表

Tab. 2　Instance of training sample for machine learning

行程 ID	起点土地利用类型	终点土地利用类型	工作日/非工作日	起点时间	终点时间	逗留时长	家庭规模/人	受访者年龄/岁	受访者职业	行程目的
1	住宅	学校	工作日	06：51：17	07：46：42	09：03：05	3	42	教师	上班
2	学校	商场店铺	工作日	16：49：57	17：33：00	00：37：23	3	42	教师	购物
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

2.2　结果分析

在 C5.0 软件环境下，对训练样本进行分类。为提高分类的精度，采用了迭代 10 次的 Boosting 方法，分类结果如表 3 所示。在未采用 Boosting 方法之前，错分比率为 19.5%. 在选择了 Boosting 进行 10 次迭代分类后，分类精度明显提高，错误率降为 12.4%，分类精度达到了 87.6%.

表 4 为分类错误矩阵表。由表 4 中可以看出："回家"类出行目的识别精度最高（～95.6%），其次为"上班"、"购物休闲"、"接送人"等活动，识别率在 90%左右；分类精度较低的是"其他活动"（～73.3%）与"业务出行"（80.8%）。其主要原因是 36 位受访者均为高校教师，他们生活相对规律、简单，"回家"、"上班"、"接送人"、"购物休闲"是生活的主要内容，

表 3　Boosting 迭代分类精度表

Tab. 3　Errors percentage of Boosting 10 trials

迭代次数/次	决策树大小	错分比率/%
0	25	44 (19.5)
1	24	66 (29.2)
2	32	61 (27.0)
3	24	73 (32.3)
4	25	59 (26.1)
5	26	65 (28.8)
6	32	67 (29.6)
7	31	59 (26.1)
8	23	60 (26.5)
9	28	61 (27.0)
Boosting		28 (12.4)

表4 出行目的分类结错误矩阵

Tab.4 Error matrix for trip purpose classification

回家（*a*）	上班（*b*）	接送人（*c*）	其他活动（*d*）	购物休闲（*e*）	业务出行（*f*）	出行目的（*s*）	精度/%
66		2			1	（*a*）：回家	95.6
2	61	3	2			（*b*）：上班	89.7
1	1	30		1		（*c*）：接送人	90.9
3	6	1	23	1	1	（*d*）：其他活动	73.3
	1				9	（*e*）：购物与休闲	90.0
	1		1	9		（*f*）：业务出行	80.8

这些出行在训练样本中相对充足；另外这些活动的时间、地点相对固定，容易识别。相反，“其他活动”和“业务出行”的活动时间和活动地点的土地利用类型多样化，加上样本量本身就很小，识别错误率偏高，但是在样本充足的情况下，这部分的分类精度可望得到进一步提高。

3 结论与建议

被动式GPS交通出行调查是未来交通调查的重要趋势，居民出行目的是居民出行交通调查不可或缺的内容，如何从GPS轨迹数据提取交通信息是该方法成功的关键。通过对与出行目的识别相关的信息种类归纳，探讨了如何在交通调查和GPS轨迹数据中定义和提取这些信息，在此基础上将出行目的提取转化为依据出行行程属性进行分类的问题，为基于机器学习的建模提供了基础和思路。研究的实例分析根据数据特点，选择了C5.0机器学习-决策树算法对样本进行了行程目的提取，结论与建议如下：

1）导入的异源语义数据在行程目的智能化提取中起到很大作用，在对6种出行目的的划分中获得了较高的整体识别精度（～87.6%）。GPS轨迹的时空特征结合具有语义形态的活动时段（时长）、土地利用类型、交通行为特征、受访者社会经济属性信息，可以有效地增加划分力度。因此，在GPS仪器辅助的交通调查设计中，可以此4类信息的收集为导向。

2）使用机器学习方法处理行程提取问题的尝试获得良好效果，表明机器学习能够较好地从GPS样本数据和先验知识的结合中发现潜在规则和规律，使得出行目的的提取不必依赖于研究者的个人经验，为基于GPS轨迹数据的居民出行目的提取提供了通用的数据分析方法。另外，在数据样本较小且存在属性值缺失的情况下，学习算法的选择也非常重要。这里使用的C5.0算法便表现出相当好的稳健性和适用性。

3）通过数据预处理将连续记录的轨迹点转化为离散的行程段，更便于属性变量的定义、提取和关联。离散化处理往往是机器学习的首要条件之一，而预处理的质量直接影响分类质量，因此在处理方法的选择和参数指定上需要更深入的研究。

说明

该研究初期成果曾在第7届交通运输国际会议上宣读。

[参考文献] (References)

[1] van EVERT H, WERNER B, EEHARD E. Survey design: the past, the present and the future[R]. Netherlands: AVV Transport Research Center, 01081132, 2004.

[2] LAVRAKAS D P J. Telephone survey methods: sampling, selecting, and supervision[M]. Inc.: Sage Publications, 2005.

[3] 广州市交通规划研究所，赛思达公司（MVA）. 2005广州市居民出行调查总报告[R]. 广州：广州市交通规划研究所，MVA，2006.
Guangzhou Academy of Vrban Planning and Design Q MVA. The household travel survey report of Guangzhou, 2005[R].

Guangzhou: Guangzhou Academy of Vrban Planning and Design & MVA, 2006. (in Chinese)

[4] 张卫化，陆化普. 城市交通规划中居民出行调查常见问题及对策[J]. 城市规划学刊，2005 (5)：90-94.
ZHANG W H, LU H P. Some problems in investigating urban citizen travels and countermeasure[J]. Urban Planning Forum, 2005(5): 90-94. (in Chinese)

[5] WAGNER D B. Lexington area travel data collection test: GPS for personal travel surveys[R]. Final Report for OHIM, OTA, and FHWA, 1997.

[6] STOPHER P R, PHILIP B, STEPHEN G. Using passive GPS as a means to improve spatial travel data: further findings[A]. Paper Presented to the 23rd Conference of Australian Institutes of Transport Research[C]. Australia: 2001. 1-14.

[7] WOLF J, LEE M. Synthesis of and statistics for recent GPS-enhanced travel surveys[A]. Submitted to the International Conference on Survey Methods in Transport: Harmonization and Data Comparability[C]. France: 2008.

[8] WOLF J, GUENSLER R, BACHMAN W. Elimination of the travel diary: experiment to derive trip purpose from GPS travel data[J]. Transportation Research Record, 2001, 1768: 125-134.

[9] STOPHER P R, CLIFFORD E, ZHANG J, et al. Deducing mode and purpose from GPS data[R]. Sydney: Paper report to Institute of Transport and Logistics Studies, 2008.

[10] QUINLAN J R. Bagging, boosting, and C4.5[A]. Proceedings of the 13th National conference of Artificial Intelligence[C]. Portland, OR: 1996. 725-730.

[11] FREUND Y, SCHAPIRE R E. Experiments with a new boosting algorithm[A]. Machine Learning: Proceeding of the Thirteenth International Conference[C]. 1996. 148-156.

[12] 邓中伟，季民河，陈雯，等. 耦合被动式GPS与网络调查的居民出行调查[J]. 交通运输系统工程与信息，2010，10 (2)：178-183.
ZHENG Z W, JI M H, CHEN W, et al. Coupling passive GPS tracking and web-based travel surveys[J]. Journal of Transportation Systems Engineering and Information Technology, 2010, 10(2): 178-183. (in Chinese)

[13] CHEN W, JI M. A prompted recall interview platform for GPS-based HTS: design and development[A]. Paper Presented to the International Workshop on GIS for Transportation[C]. Wuhan: 2009. 157-163.

[14] WU X D, KUMAR V, QUINLAN J R, et al. Top 10 algorithms in data mining[J]. Knowledge Inference Systems, 2008, 5(4): 1-37.