基于多目标数据挖掘的城市交通仿真算法研究*

张 昕[1], 关志超[1], 杨东援[2]
(1. 深圳市城市交通规划研究中心, 广东 深圳 518034;
2. 同济大学, 上海 200092)

摘 要: 城市交通仿真是智能交通系统领域内的核心技术之一, 其基础在于实时交通数据的采集和分析整理。而城市实时交通数据包括静态信息和动态数据, 需要将多源数据进行融合并对数据进行分析和挖掘, 提取交通特征。提出了一种基于聚类集成的多目标聚类分析框架。同时在此框架下, 提出了一个启发式的聚类算法 k-WANMI, 进行快速有效的聚类分析。实验结果表明, 提出的方法有效的满足多数据源的应用需求, 提出的框架和算法能够处理混合数据、处理具有不同权重的属性并且能够进行多目标分析。

关键词: 城市交通仿真; 多目标; 数据挖掘; 聚类分析算法
中图分类号: TP319 文献标识码: A 文章编号: 0529-6579 (2007) S2-0210-05

进入新世纪, 经济全球化和信息网络化进程的不断加快, 深圳作为中国最大的改革开放经济特区, 城市经济持续快速增长, 交通运输业得到了迅猛的发展, 交通基础设施的兴建对城市经济的"瓶颈"制约进行了一定的缓解, 但与城市经济发展、社会进步和市民公众的服务要求仍有很大差距。在大规模新建交通基础设施的同时, 已有基础设施也迫切需要采用现代科学与先进技术进行信息化、智能化改造挖潜, 充分发挥既有基础设施的作用。为了适应城市交通运输工程建设和运营管理的需要, 大量复杂、基础性及共性的高新科学技术问题亟待解决, 必须采用先进的科技手段进行研究[1]。

城市交通仿真是智能交通系统领域内的核心技术之一[2]。其目标在于运用利用计算机及其网络技术, 充分结合和发挥地理信息、图像识别, 微波遥感等相关数据检测和采集技术, 实时采集各种交通数据和信息, 并对数据进行合理的分析、加工和整合, 从微观、中观、宏观等角度分析, 找出导致城市道路交通拥挤问题的症结, 并进一步预见各种对策和外部因素变化对系统的影响。同时, 将历史数据和未来预测数据与交通仿真有机融合, 最终对深圳市城市交通系统规划设计与改造方案实行优化, 为解决这些问题提供辅助决策支持手段。由此可见, 城市交通仿真的基础在于实时交通数据的采集和分析整理。而城市实时交通数据包括路网、土地、建筑物等静态信息和视频、线圈、红外、雷达以及浮动车等动态数据, 如何将多源数据进行融合并对数据进行分析和挖掘, 提取交通特征, 是本文研究的出发点。

在数据分类与融合方法的研究大多数集中在聚类分析技术。聚类分析是按照特定标准把数据集分割成不同的簇, 使簇内的相似性尽可能大; 而簇间的区别性也尽可能大。从数据分类分析的角度而言, 每个簇实际上是一个数据类别。国内外目前对多源数据融合的研究还有所不足, 主要集中在使用已有分类算法对数据进行细分, 大多数研究局限于特定应用, 没有形成对问题本质的抽象和概括。虽然数据分类问题从本质上可以看作一个聚类分析问题并利用现有的聚类算法进行处理。然而, 由数据分类问题自身的特点所决定, 对用于数据分类的聚类算法提出了一些新的要求, 主要包括:

(1) 由于涉及不同类型的数据属性变量, 因此, 新的聚类算法要能够有效处理混合属性的数据(既包含范畴属性, 又包含数值属性)。

(2) 目前针对数据分类的大部分聚类分析算法中, 属性都赋予了同样的权重; 而在实际应用中, 尤其是数据分析中, 不同属性可能具有不同的权重。因此, 新的聚类算法要能够处理具有不同权重属性的情况。

* 收稿日期: 2007-07-09
基金项目: 国家"十一·五"科技支撑计划重点资助项目 (2006BAG02B02); 建设部科技计划项目 (建科验字2006-096号); 深圳市政府投资计划重点工程项目 (深发改 2004-481、2005-814号)
作者简介: 张昕 (1977年生), 男, 工学博士; E-mail: zx@supec.com

(3) 很多情况下，决策人员仅依据某个属性变量进行数据分类，这很难用传统的聚类分析概念进行解释。所以数据分类本质上是一个多目标优化(多目标聚类分析)问题，即聚类结果要尽可能的满足每个数据类属性的属性值分布，而各个属性的属性值分布常常是矛盾的。每个数据类属性变量对应于一个优化准则，如果针对每个数据类属性变量定义目标函数，从多目标优化的角度，就可以将仅依据某个数据类属性变量进行数据分类的结果解释为多目标聚类分析的最优解。

目前，针对上述第一个要求，即混合属性数据的聚类分析，存在一些有效的算法[3-4]，其中较有代表性的是 k-prototypes算法。也存在一些能够处理属性具有不同权重的聚类算法，如文献[4]中给出的算法。此外，多目标聚类分析问题也受到了一定的重视[5-6]。然而，目前还没有能同时满足上述三个要求的聚类算法。因此，本文针对数据分类问题对聚类分析的三个需求，提出新的基于聚类集成的多目标聚类分析算法，以满足多目标数据源的应用需求。

1　多目标聚类分析框架

聚类集成是一种将样本集上运行的多个聚类算法的输出合成为一个划分的方法，其目的是对多个结果进行优化，最终得到该样本集的一个较好的划分[7-8]。

设样本集 $D=\{X_1, X_2, \cdots, X_n\}$，D中每个样本都具有 r个属性，样本的属性集为 $A=\{A_1, \cdots, A_r\}$。在 D的 n个样本上的一个聚类划分可以表示为 k个簇的集合 $\{C_l \mid l=1, \cdots, k\}$ 或一个标签向量 $\lambda \in N^n$。其中 λ用于标识每个样本所属的簇，因此 λ中的元素的取值范围为 1, …, k，在 λ中，若第 i个元素的值为 j，则表示样本 X_i属于第 j个簇。一个聚类算法可以看作一个输出标签向量的函数 Φ。

该框架包括两个步骤：单一属性空间聚类和聚类集成。

第一步，单一属性空间聚类，是针对样本集的每个属性分别对样本集进行聚类分析(不考虑该属性外的其他属性)。按照每个属性进行聚类完成以后，便得到了该样本集的 r个聚类结果，即 r个类别标签向量 $\lambda^{(1)}, \lambda^{(2)}, \cdots, \lambda^{(r)}$。

由于我们对每个属性进行单独的处理，针对不同的属性类型，可以使用不同的聚类算法；由于不同的属性变量在数据类型和数据分布上可能存在较大差异，所以这种灵活性在实际的数据挖掘应用中是非常重要的。例如，如果输入数据集既有离散属性又有连续属性，则对不同的属性可以有不同的选择。对于连续属性而言，可以选择比较常用且有效的数值聚类算法如 k-means等，或者进行离散化都可以得到一个标签向量作为下一阶段的输入。

同样，对于离散属性，如果令该属性上取值相等的所有样本属于同一个类别，则得到了针对该属性的最优划分(离散属性的属性值是无序的，所以相同属性值之间的距离是 0，而不相同的属性值之间的距离是 1；只需定义优化目标为簇内距离之和，则上述划分达到该目标函数的最小值)。更确切的讲，对样本集 $D=\{X_1, X_2, \cdots, X_n\}$ 和属性集 $A=\{A_1, \cdots, A_r\}$，不妨设 V_i是离散属性 A_i的属性值的集合。令 $\Phi^{(i)}$是一个将 V_i中的属性值映射为自然数的函数，则可以得到该属性的最优划分的类别标签向量 $\lambda^{(i)}$：

$$\lambda^{(i)}=\{\Phi^{(i)}(X_j, A_i) \mid X_j \in D, A_i \in V_i\}。$$

因此，该算法框架能够有效满足数据分类的第一个需求，即有效的处理混合属性的数据。由于这一步相对简单，所以在本文后续的讨论中，假设这一步工作已经完成，即样本集 D已被转换为一个类标签向量组成的离散数据集合。

在第二步，通过集成函数 Γ对这 r个聚类结果进行集成，变换为单一的类别标签向量 λ。由于最终聚类结果 λ需要尽可能的满足每个属性确定的结果 $\lambda^{(1)}, \lambda^{(2)}, \cdots, \lambda^{(r)}$，而各个属性确定的划分常常是矛盾的，所以这一步面临的是一个多目标优化问题(以最终划分和单个属性变量确定的划分为变量，可以定义一个优化目标函数。定义目标函数的方法很多，本文从信息论的公共互信息的角度进行定义，采用最终划分和单个属性变量确定的划分之间的公共互信息为目标函数)。与单目标优化问题的本质区别在于，多目标优化问题的解非唯一，而是存在一个最优解集合，集合中元素称为 Pareto最优或非劣最优(non-dominated)。所谓 Pareto最优，就是不存在比其中至少一个目标好而比其它目标不坏的解，即不可能通过优化其中部分目标而其它目标非最坏。因此，前面提到的仅仅依据某个属性变量进行数据分类得到结果，就可以归结为多目标聚类分析的一个最优解，在理论上给出了一个很好的解释。所以，该算法框架有效的满足了数据分类的第三个需求。

在数据挖掘中，经常使用的多目标优化技术有三类：

1) 目标聚集法：用聚集操作符(aggregation operator)将多目标转化为单目标，进而用单目标

优化技术求解；

2) 字典序法：将所有的目标函数按照重要性进行排序，然后根据已排序的目标函数从最重要的目标函数开始逐级进行优化，最终得到优化解；

3) Pareto最优法：利用非劣最优的概念，求出所有的 Pareto解。其中，目标聚集法中的求和操作符是最常用的选择。因此，本文也采用求和操作符对目标函数进行聚集。

由于在数据分类应用中，不同属性具有不同的权重。因此，目标函数要能够处理属性具有不同权重的情况。设 $\omega = \{\omega_1, \omega_2, \ldots, \omega_r\}$ 为不同属性的权重的集合，令 λ为样本集 D的任意一个划分，则其与 D的 r个划分（分别针对 r个属性进行聚类得到的结果）的平均加权标准化共有信息量（Weighted Average Normalized Mutual Information，WANMI）定义为：

$$\phi^{(WANMI)}(\Lambda\lambda) = \frac{1}{r}\sum_{q=1}^{r}\phi^{(NMI)}(\lambda, \lambda^{(q)}) \quad (1)$$

设指定聚类的个数为 k，用 $\lambda^{(k-opt)}$表示样本集 D的最优划分，其应满足的条件是：

$$\lambda^{(k-opt)} = \arg\max_{\lambda}\sum_{q=1}^{r}\omega_q\phi^{(NMI)}(\lambda, \lambda^{(q)}) \quad (2)$$

其中，λ遍历所有可能的 k—划分。

样本集的最优划分应满足与样本集的 r个划分的平均共有信息程度最高，由此可以用公式（1）作为聚类分析的目标函数。从（1）可以看出，不同属性的权重优先级得到了很好的处理，所以满足了数据分类应用处理不同权重的属性的需求。

目标函数（1）定义了一个复杂的组合优化问题。如果采用枚举遍历的方法，n个样本的所有可能的 k—划分的个数为：$\frac{1}{k!}\sum_{i=1}^{k}\binom{k}{i}(-1)^{k-i}i^n$。

在 $n >> k$时，其个数近似的表示为：$k^n/k!$。例如将 16个样本划分为 4个簇，则有 171，798，901种可能。因此，本文设计了一个直接优化目标函数的聚类算法 k-WANMI，进行快速有效的聚类分析。

2 聚类分析算法

k-WANMI算法以待聚类的数据集合、期望生成的聚类个数以及属性权重作为输入，算法执行之前首先为集合中的数据设定初始所属类别，然后遍历整个数据集，如果发现更改某条数据所属的类别以后能提高目标函数值，则更改该条数据所属的类别，继续考察下一条数据，如此多次遍历整个集合，直到不存在某条数据需要改变其类别为止，这样便得到了一个局部最优解。

k-WANMI算法的流程图如下所列。数据存储在磁盘上，算法对数据进行顺序的扫描。

```
//k-WANMI聚类算法
/输入：样本集 D，期望聚类的个数 k，以及属性权重集合
/输出：样本集 D的聚类结果
/*初始化 */
  Begin
    对于 D中每个样本数据 t
    i++
    更新属性 A的 AHi
    if I<= k then
        将 t的类别属性设成 i
    else
        将 t放到其距离最近的类中
    write<t, i>
/*聚类 */
  Repeat
    not_moved= true
    while没有到达 D中的最后一条记录 do
    读下一条记录 <t, Ci>
    分别假设 t属于其他的 k-1个类别，并计算相应的 ANMI，找到使目标函数取值最大的类 Cj
    ifCi!= Cj then
        write<t, j>
    not_moved= false
  Until not_moved
  End
```

在初始化阶段，我们首先选择前 k条数据初始化每个簇的直方图。后续的数据被放到其距离最近的簇并更新簇的直方图。同时存储每条数据的类别标签并构造每个属性的直方图。

在聚类阶段，迭代遍历整个数据集，如果发现更改某条数据所属的类别能提高目标函数的取值，则更改该条数据所属类别，继续考察下一条数据，如此多次遍历整个集合，直到不存在某条数据需要改变其类别为止，这样就得到一个局部最优解。其中最为关键的步骤是计算 WANMI的值。

该算法的时间复杂性依赖于样本的个数 n，样本属性的个数 r，直方图的数目、聚类数目 k和重复遍历样本的次数（迭代次数 i）。假设样本集的每个属性都有同样多的属性值个数（p个），则在最坏的情况下，样本初始化的时间复杂性为 O（nkrp），遍历样本集、产生最终划分的时间复杂性为 O（$i$nk^2 rp^2），故整个算法的时间复杂性为 O（$i$nk^2 rp^2）。实际上参数 p一般较小（每个属性确定的划分中簇的个数一般不会太大），所以可以看

作一个较小的常数值。同时，哈希表数据结构的使用也降低了参数 p的影响对运行时间的影响。所以，在实际情况下，算法的期望时间复杂性为 $O(tkn^2+tn)$。

算法需要在存储 $(k+1)\times r$个直方图和数据集，所以空间复杂性为 $O(tkp+nr)$。

从复杂性分析可以看出，k-WANMI算法的复杂性和数据集的大小以及属性个数成线性关系。因此，该算法具有良好的可扩展性。

3　算法实验结果及评价

这一节通过在实际采集到的实时交通数据上的实验，比较 k-WANMI算法相对其他算法的在聚类精度上的优越性，以及 k-WANMI算法的可扩展性。

聚类结果的评价一直比较困难。本文所采用的数据集中，由于每个样本真实类标签的存在，可以得到聚类准确率一个较为客观的计算方法。假设目标数据集中的类标签的取值有 g个（其集合分别用 $s_1, \cdots, s_g$表示），聚类结束后簇的个数为 k（k个簇的集合分别用 $c_1, \cdots, c_k$表示），则聚类准确率定义为：$\text{accuracy}=\dfrac{\sum_{i=1}^{k} a_{ij}(|s_j|)}{g}$

其中 a_i是第 i个簇中类标签为 j的样本数目，且类标签为 j的样本是在第 i个簇中相对于其全局分布所占比例最大的样本，即对任意的 $l(\neq i\leqslant g)$，有 $\left(\dfrac{a_{ij}}{|c_i|}\right)\Big/\left(\dfrac{|s_j|}{n}\right)\geqslant\left(\dfrac{a_{lj}}{|c_l|}\right)\Big/\left(\dfrac{|s_j|}{n}\right)$

进一步简化上述条件公式，有 $\left(\dfrac{a_{ij}}{|s_j|}\right)\Big/\left(\dfrac{a_{lj}}{|s_j|}\right)$

相应的，聚类的错误率可以定义为 $\text{error}=1-\text{accuracy}$。

3.1　算法正确率评价

由于交通仿真数据集合既含有连续属性，又含有离散属性，目前混合属性数据的聚类分析算法中，较有代表性的是 k-prototypes算法。因此，我们选择 k-prototypes算法与 k-WANMI算法进行比较。

由于所有实验数据集都包含连续属性，所以 k-WANMI算法按照如下方式处理连续属性：设最终簇的数目为 k，在 k-WANMI算法的第一阶段，利用 k-means算法分别对每个连续属性进行划分，得到的 k划分作为第二阶段的输入。在所有的实验中，k-prototypes算法采用随机的初始化方法，其聚类结果用 10次运行结果的平均值来表示。此外，k-prototypes算法中需要的离散属性权重始终设置为 1。并且在实验中，所有的属性设定具有相同的权重。

由于 k-WANMI算法和 k-prototypes算法都需要以最终的簇的数目为输入参数，故在实验中令簇的数目从 2变化到 9进行聚类错误率的比较。图 1给出了两个算法数据集上错误率的比较。和 k-prototypes算法相比，k-WANMI算法在数据集上具有较低的平均错误率。

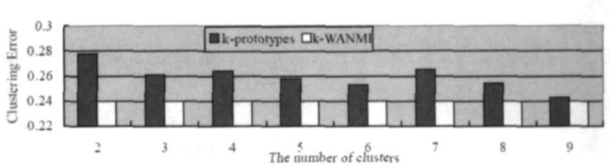

图 1　在不同簇数目下的聚类错误率

Fig. 1　Clustering error percentage in different number of clusters

根据图 1，二者的相对性能可以总结为表 1。

表 1　不同聚类算法的相对性能比较

Tab. 1　Performance comparison of different clustering algorithms

算法	平均聚类错误率
k-prototypes	0.26
k-WANMI	0.24

实验结果表明，k-WANMI算法相对已有的算法，在数据集的聚类精度上有着一定的优越性。

3.2　算法可扩展性评价

为了测试 k-WANMI算法在数据分类分析时的可扩展性，仍然以交通采集数据集为测试集，测试其运行时间。k-WANMI算法用 Java实现，实验在一台 HP MI370服务器上运行，配置为 Pentium4 XEON 3.6G×2，8G内存，操作系统为 Windows 2003 Server。

图 2给出了 k-WANMI算法在簇数目固定为 2时，随着数据量的增长运行时间的情况。可以看到，执行时间随着数据量的增长而增加，而且基本上是线性增长。

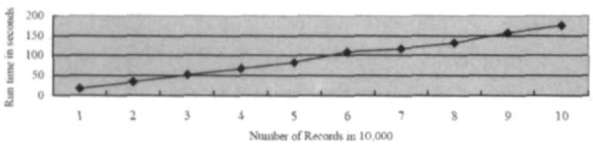

图 2 K-WANMI算法对数据量的可扩展性

Fig. 2 Scalability of k-WANMI to the number of objects

图 3给出了 k-WANMI算法随着簇的数目的增长，运行时间的变化情况。尽管执行时间随着簇的数目的增长而很快增加，其增长幅度还是可以接受的。

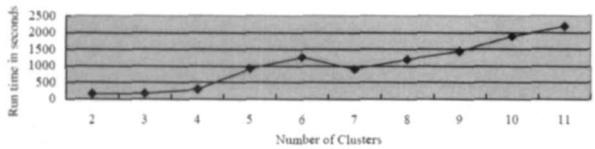

图 3 K-WANMI算法对簇的数目的可扩展性

Fig. 3 Scalability of k-WANMI to the number of clusters

以上的实验结果表明，k-WANMI算法具有良好的可扩展性，能够高效的处理大规模数据集。

4 结 论

本文针对城市交通仿真的多数据源问题对聚类分析的特殊需求，从聚类集成的角度定义了聚类问题的优化目标函数，提出了一种基于聚类集成的多目标聚类分析框架。同时，在此框架下，提出了一个启发式的聚类算法，进行快速有效的聚类分析。理论分析表明，提出的框架和算法能够处理混合数据、处理具有不同权重的属性并且能够进行多目标分析。实验结果表明，本文提出的方法有效的满足多数据源的应用需求，弥补了现有的聚类分析算法的不足。

参考文献：

[1] 朱照宏，杨东援，吴 兵. 城市群交通规划[M]. 上海：同济大学出版社. 2007.

[2] 杨东援. 交通规划决策支持系统[M]. 上海：同济大学出版社，1997.

[3] HUANG Z. Extensions to the K-Means Algorithm for Clustering Large Data Sets with Categorical Values[J]. Data Mining and Knowledge Discovery, 1998, 2(3): 283—304.

[4] FRIGUI H, NASRAOUI O. Unsupervised learning of prototypes and attribute weights[J]. Pattern Recognition, 2004, 37(3): 567—581.

[5] LAW M H C, TOPCHY A P, JAIN A K. Multiobjective Data Clustering[C]. Proc. of 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04), 2004, 2: 424—443.

[6] HANDL J, KNOWLES J. An Evolutionary Approach to Multiobjective Clustering[C]. IEEE Transactions on Evolutionary Computation, 2006.

[7] STREHL A, GHOSH J. Cluster Ensembles—A Knowledge Reuse Framework for Combining Partitions[C]. Proc. of the 8th National Conference on Artificial Intelligence and 4th Conference on Innovative Applications of Artificial Intelligence, 2002: 93—99.

[8] FROSSYNIOTIS D, PERTSELAKIS M, STAFYLOPATIS A. A Multi-Clustering Fusion Algorithm[C]. Proc. of the Second Hellenic Conference on AI, 2002: 225—236.

Research on Urban Traffic Simulation Algorithm Based on Multi-objective Data Mining

ZHANG Xin1, GUAN Zhi-chao2, YANG Dong-yuan2

(1. Shenzhen Urban Transport Planning Center, Shenzhen 518040, China;
2. Tongji University, Shanghai 200092, China)

Abstract: Urban traffic simulation is one of the core technologies of Intelligent Transport System. Its basis is to collect and analyze the real-time traffic information, which include static and dynamic data. How to fuse data from multi-sources and extract traffic characteristics is the key research point. In this paper, we propose a new cluster ensemble based multi-objective cluster analysis framework, and correspondingly bring forward a heuristic clustering algorithm named k-WANMI. The experimental result proves our method can handle mixed categorical, numeric data, and attributes with different weight. Its also able to perform multi-objective cluster analysis. It effectively satisfies the requirement of multi-sources data.

Key words: urban traffic simulation; multi-objective; data mining; cluster analysis algorithm