文章编号:1001 - 9081(2006)05 - 1109 - 02

一种面向态势估计中分群问题的聚类方法

黄 雷,郭 雷
(西北工业大学 自动化学院,陕西 西安 710072)
(harry@ mail. nwpu. edu. cn)

摘 要:对目标分群技术问题进行了描述,分群或聚类问题是态势估计需要实现的一个重要功能,主要根据底层融合的结果应用聚类分析法实现战场目标分群。目标分群的结果有助于确定态势元素之间的相互关系,从而解释问题领域的各种行为,辅助指挥决策。提出使用 Chameleon 算法对战场目标或群进行划分,根据相对互连性 RI 和相对近似性 RC 所表征的相似度把它们形成更高层次的群。

关键词:数据融合;态势估计;分群;Chameleon 算法
中图分类号: TP182 **文献标识码**:A

Method of clustering about situation assessment

HUANG Lei, GUO Lei
(*College of Automation, Northwest Polytechnical University, Xi'an Shaanxi* 710072, *China*)

Abstract: The issue of object clustering about situation assessment and was described that object clustering is one of important function which situation assessment need achieve was pointed out. Object clustering mainly uses cluster analysis based on the information from lower level data fusion. The output of object clustering is helpful to determine the relationship among situation elements thus interpreting the actions related to problem field and supporting command decisions. Chameleon algorithms were used to clustering based on the RI and RC.

Key words: data fusion; situation assessment; clustering; Chameleon algorithms

0 引言

美国联合领导实验室 JDL(Joint Directions of Laboratories)数据融合小组(DFS)建立数据融合处理模型,把数据融合分为四层:第一层处理包括数据和图像的配准、关联、跟踪和识别;第二层处理包括态势提取、态势分析和态势预测,统称为态势估计;第三层处理是敌兵力威胁估计;第四层处理是优化融和处理。态势估计接受低层融合的结果,从中提取对当前军事态势尽可能准确、完整的感知。在军事应用领域中,态势估计至今没有统一的定义,JDL 数据融合处理模型对态势估计的描述为:作为二级数据融合,态势估计是建立关于作战活动、事件、时间、位置和兵力要素组成的一张视图,将所观测到的战斗力量分布与活动和战场周围环境、敌作战意图及敌机动性有机地联系起来,分析并确定事件发生的原因,得到关于敌方兵力结构,使用特点的估计,最终形成战场综合态势图。态势觉察中需要利用提取的态势特征元素把平台按照空间和功能进行分群从而帮助指挥员做出正确的决策。在不同的战场态势中的战场目标具有不同的组织和空间结构,结构中不同的组成部分起着不同的作用。形成这一结构的过程称为目标分群,是态势估计要完成的主要任务之一。分群中使用的传统聚类方法对于各战场目标相互间的关系挖掘的不一定深入,所以,本文分析了态势估计的目标分群问题,提出将基于动态模型的 Chameleon 聚类算法引入到目标分群来实现目标向空间群的聚类,在此基础上,按照最近邻原则和一定的规则逐级形成目标编群的体系结构。

1 分群问题描述

目标分群将关于目标对象的可用数据按空间、功能及相互作用等属性逐级分群,以揭示目标之间的相互联系,确定相互合作的功能,从而解释问题领域的各种行为。战场目标分群的形成过程是以数据驱动的前向推理过程,即将规则应用于有效数据以产生一个可推理的假设结构。因此,基于一定的规则是目标分群的主要特征。分群的基本思想是对有用的数据进行分组,以便后续评估确定态势元素之间的相互关系,并能够据此从各个层次解释战场态势的行为特性。目标或群之间的空间距离是用于分群的重要的一个属性。分群从低到高可以分为:空间群是按照位置关系对空间上的目标进行集合划分;功能群是执行类似功能的在战术上相关的平台组集合;敌/我/中立方群则根据平台组的敌我友属性分群。群的形成过程就是目标分群或聚类问题的求解过程。战场目标分群主要基于聚类分析法,本文就是利用一种通用聚类算法 Chameleon 来实现战场目标分群。

2 Chameleon 算法描述

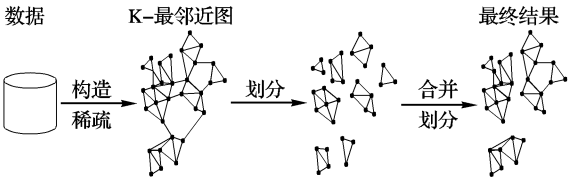

图 1 Chameleon:基于 K-最邻近图和动态建模的层次聚类

Chameleon 是一个在层次聚类中采用动态模型的通用聚类算法。Chameleon 算法是一个二阶段算法:第一阶段用图划分算法(如超图分割算法 hMetis)把 K-最邻近图划分为较小的相对独立的群;第二阶段用一个凝聚的层次聚类算法通过

收稿日期:2005 - 11 - 09;**修订日期**:2006 - 02 - 13

作者简介:黄雷(1977 -),男,湖北麻城人,讲师,硕士研究生,主要研究方向:多传感器数据融合、本体、语义网; 郭雷,男,教授,博士生导师,主要研究方向:数字图像处理、数据融合.

反复合并,可以实现群的聚类。聚类过程中,如果两个群间的互连性和近似度与群内部对象间的互连性和近似度相关,则合并两个群。由于动态模型的合并过程,所以有利于自然的和同构的聚类的发现,且只要定义了相似度函数就可以应用于所有类型的数据。

Chameleon 算法进行聚类分析的过程(如图 1),Chameleon 算法基于通常采用的 K-最邻近图(K-Nearest Neighbor Graph)方法描述它的对象。K-最邻近图中的每一个节点代表一个对象,如果一个对象是另一个对象的 K 个最类似的对象之一,在这两个对象之间存在一条边,边的权重用两个对象间的相似度表示。这样的好处是,距离很远的对象完全不相连,边的权重代表了潜在的空间密度信息。Chameleon 通过两个对象的相对互连性(Relative Interconnectivity) $RI(C_i,C_j)$ 和相对近似性(Relative Closeness) $RC(C_i,C_j)$ 来决定对象间的相似度。

两个对象 C_i 和 C_j 之间的相对互联性 $RI(C_i,C_j)$,把一个群做最小截断时需要去掉的边的权重之和定义为该群的互连性,就可以用 C_i 和 C_j 合并后形成的群的互连性与 C_i 和 C_j 的平均互连性的比率定义为 $RI(C_i,C_j)$。它的定义如下:

$$RI(C_i,C_j) = \frac{|EC_{\{C_i,C_j\}}|}{\frac{1}{2}\left(|EC_{C_i}| + |EC_{C_j}|\right)}$$

其中,$EC_{\{C_i,C_j\}}$ 是连接 C_i 和 C_j 的所有边的权重和;类似,EC_{C_i}(或 EC_{C_j})是把群 C_i(C_{C_j})划分为两个大致相等部分的最小等分线切断的所有边的权重和(即将图分为两个大致相等部分需要切断的边的加权和)。相对互连性可以处理群间形状不同和互连程度不同的问题。

两个对象 C_i 和 C_j 之间的相对近似性 $RC(C_i,C_j)$,用 C_i 和 C_j 合并后形成的类的近似度与 C_i 和 C_j 的平均内部近似度的比率定义相对近似性 $RC(C_i,C_j)$。它的定义如下:

$$RC(C_i,C_j) = \frac{\overline{S}_{EC_{\{C_i,C_j\}}}}{\frac{|C_i|}{|C_i|+|C_j|}\overline{S}_{EC_{C_i}} + \frac{|C_j|}{|C_i|+|C_j|}\overline{S}_{EC_{C_j}}}$$

其中,$\overline{S}_{EC_{\{C_i,C_j\}}}$ 是连接 C_i 和 C_j 的边的平均权重,$\overline{S}_{EC_{\{C_i\}}}$(或 $\overline{S}_{EC_{\{C_j\}}}$)是把 C_i(或 C_j)划分为两个大致相等部分的最小等分线切断的所有边的平均权重。近似度是指所有做最小截断时需要去掉的边的平均权重。在合并过程中,优先合并群间近似度与群内近似度相近的群。

3 群形成过程

3.1 初始分群的形成

这一阶段用分割式算法把 K-最邻近图划分为较小的相对独立的子群,分为三部分:

第一步:利用低级别融合结果数据集和距离度量来构建一个相似度的矩阵:

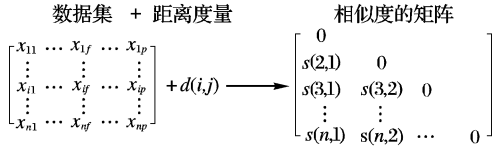

图 2 构建相似度矩阵

距离度量通常用来定量的描述观测—观测对或观测—航迹对之间的相似性。距离度量可采用广泛使用的加权欧氏距离:

$$d_{ij}^2 = \tilde{y}_{ij} S^{-1} \tilde{y}_{ij}^T$$

第二步:利用相似度的矩阵构建 K-最邻近图

相似度矩阵是整个算法进一步处理的基础,为了便于理解和简化计算,在接下来的计算中采用的是基于图的模型。首先是把相似度矩阵转换成基于相似度的 K-最邻近图,即把数据对象看成是图的顶点,对象之间的相似度作为链接(边)的权重。

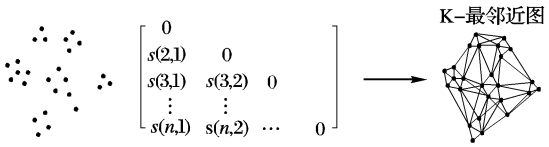

图 3 构建 K-最邻近图

第三步:以 K-最邻近图及分割式算法求得初始的分群

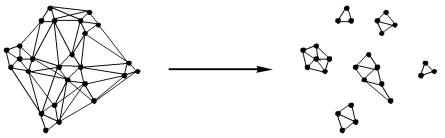

图 4 求解初始分群

依据递归水平,在 K-最邻近图上,做最小截断,用图分区算法(如:hMetis)将 K-最邻近图反复划分成小的无连接子图。hMetis 算法根据最小化截断的边的权重和来分割 K-最近邻居图,图上的一个最小截断是指把 K-NN Graph 分区成两个近似的、等大小的部分,使被分区的总权重最小。然后把每一个子图看成一个初始子群,重复该算法直至达到标准。

3.2 群的合并

Chameleon 通过两个群 C_i 和 C_j 的相对互连性 $RI(C_i,C_j)$ 和相对近似性 $RC(C_i,C_j)$ 来决定两个群之间的相似度。访问每个群,计算它与临近群的 RI 和 RC。合并 RI 和 RC 分别超过 T_{RI} 和 T_{RC} 的群对。若满足条件的临近群多于一个,合并具有最高绝对互连性的群。重复上两步,直到没有可合并的群。

Chameleon 使用两种方法来合并相邻的群:

1) C_i 和 C_j 的相对互连性和相对近似性必须满足用户指定的阈值 T_{RI} and T_{RC}。

Chameleon 计算每一个群 C_i 和它邻近的群 C_j 的相对互连性和相对近似性,看是否满足下列条件:

$$RI(C_i,C_j) \geqslant T_{RI} \text{ and } RC(C_i,C_j) \geqslant T_{RC}$$

若一个以上的相邻群满足上述条件,Chameleon 选择与 C_i 相对互连性较大的群如 C_j 合并。重复这个过程直到没有满足条件的群为止。

2) Chameleon 定义相似度函数,若 C_i 和它的相邻群 C_j 的函数值最大则合并。函数定义为:

$$F(C_i,C_j) = RC(C_i,C_j) * RI(C_i,C_j)^{\alpha}$$

其中:α 是一个从 0 到 1 之间的用户指定的参数,减少 α 表示 $RI(C_i,C_j)$ 更重要。若 $\alpha > 1$ 则表示相对近似性更重要。重复上述过程直到没有满足条件的子群为止。

4 算例

设在某一时刻 t 经过一级融合采集到了 5 个空中威胁目标,以每个目标的位置(X,Y,Z),速度(V)和雷达截面为主要状态要素来聚类,如表 1 所示。 (下转第 1129 页)

局域计算,并改进了最大邻域半径的设置方法,在降低 CABDET 算法计算开销的同时提高了算法对不同数据分布的适应能力。图 5 给出了 F-CABDET 算法对数据集 DB1、DB2、DB3 的聚类结果。

(a) Radius=3　　(b) Radius=2.5　　(c) Radius=3.5

图 5　F-CABDET 算法的聚类结果

从图 5 中可以看出,测试集 DB1、DB2 具有如下特征:具有明显的簇边界;同一簇内密度分布较为均匀;能采用全局参数统一描述各个簇;簇具有任意形状;DB2 中存在噪音数据。对于具有此类特征的数据集,F-CABDET 算法取得了与 DBSCAN 算法和 CABDET 算法相似的聚类结果[6]。测试集 DB3 中簇间的密度分布与另一个簇的密度分布较为接近,全局参数无法刻画数据集的几何特征。F-CABDET 算法采用了可变邻域半径,具备了较强的发现潜在的不同密度簇的聚类能力,因而得到了与 CABDET 算法相同的聚类效果,并优于 DBSCAN 算法。F-CABDET 算法在输入参数的设置上类似于 CABDET 算法,具有弱的参数敏感性。

2.2　执行效率的比较

实验分别对 F-CABDET、CABDET 和 DBSCAN 算法在不同测试集上的聚类时间进行了比较,这三个算法均采用 Matlab 编写,在数据的预处理阶段,CABDET 算法和 DBSCAN 算法通过建立距离矩阵 $Dist(i,j)$ 实现的,未采取优化技术;F-CABDET 算法使用了基于窗口的计算方法,并将该方法应用到 DBSCAN 算法上以考察其在不同算法中的应用能力。表 1 给出了各个算法在上述三个测试集上的运行时间,其中 DBSCAN* 采用了基于窗口的优化方法。从表中可以看出,F-CABDET算法大幅度降低了在不同测试集上的执行时间,约为 CABDET 算法执行时间的 1/3 ~ 1/5,对 DBSCAN 算法的优化也取得了相似的结果,证明了基于窗口的优化技术对此类计算具有相同的优化能力。

表 1　执行时间比较(单位:s)

算法	测试集		
	DB1	DB2	DB3
CABDET	0.658 4	0.308 3	40.641
F-CABDET	0.160 5	0.104 0	8.013
DBSCAN	0.523 3	0.263 0	34.805
DBSCAN*	0.091 6	0.056 2	6.530

3　结语

真实的数据集合通常数据量大并且密度分布不均,客观上要求算法具有较短的执行时间和可变的参数设置。F-CABDET算法通过采用基于窗口的优化技术大幅度降低了算法的执行时间,能更好的满足不同类型的应用需求,该算法只需要一个输入参数作为根节点的初始邻域半径,在聚类过程中动态改变邻域半径,不仅可以发现任意形状的簇,而且具有处理异常数据的能力。实验结果表明 F-CABDET 算法具有快速的聚类能力和较好的聚类结果。F-CABDET 算法采用不同的距离定义作为相似性度量,能处理多种类型的数据集,今后的研究任务是将该算法应用于实际数据类型。

参考文献:

[1] HAN JW, KAMBER M. Data mining concepts and techniques[M]. Beijing: Higher Education Press, 2001.346 - 381.

[2] HINNEBURG A, KEIM DA. An efficient approach to clustering in large multimedia databases with noise[A]. KDD'98[C]. New York, 1998.58 - 65.

[3] ESTER M, KRIEGEL HP. A density-based algorithm for discovering clusters in large spatial databases with noise[A]. KDD'96[C]. Portland OR, 1996.226 - 231.

[4] ANKERST M, BREUNIG MM. Optics: Ordering points to identify the clustering structure[A]. Proc. ACM SIGMOD'99 Int. Conf. On Management of Data[C]. Philadelphia. PA, 1999.49 - 60.

[5] 赵艳广,谢帆,宋俊德.一种新的聚类算法:等密度线算法[J].北京邮电大学学报,2002,25(2):8 - 13.

[6] DAI WD, HOU YX, HE PL. A clustering algorithm based on building a density-tree[A]. ICMLC2005[C]. Guangzhou, 2005.1999 - 2004.

(上接第 1110 页)

表 1　一级融合输入的目标数据

目标编号	X/m	Y/m	Z/m	V/mps	雷达截面 RCS/m^2
U_1	1 105	1 608	1 012	799	9.01
U_2	877	1 549	991	962	12.80
U_3	433	629	1 199	800	8.10
U_4	499	781	1 201	227	4.0
U_5	1 107	1 609	970	2 000	0.10

对表 1 的数据进行计算后,首先可得到相似性矩阵,再依据相似性矩阵,构造出 K-最邻近图(取 k = 3)。使用图分区算法(hMetis)将 K-最邻近图反复分区,形成许多小的无连接子图。再采用第二种合并群(见 3.2 节群的合并)的方法,取 α = 2,可以得到目标分群结果,如图 5。

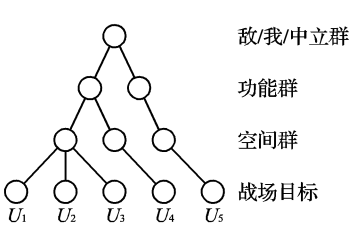

图 5　目标编群结果

5　结语

Chameleon 算法综合考虑了互连性和近似性,将互连性和近似性都大的群合并。改变了其他算法只重视某一方面的情况,从而可以获得高质量的聚类。仿真结果表明,该算法效果优于仅仅使用 KNN 算法。但是,在聚类过程中,需要注意 K-最近邻居图中 k 值的选取,最小二等分的选取和用户指定方式中阈值的选取等情况,这些都是下一步工作中需要继续研究的问题。

参考文献:

[1] KARYPIS G, HAN EH, KUMAR V. CHAMELEON: A Hierarchical clustering algorithm Using Dynamic Modeling[J]. COMPUTER, 1999,32(8): 68 - 75.

[2] KARYPIS G, AGGARWAL R, KUMAR V. *et al.* Multilevel hypergraph partitioning: application in VLSI domain[A]. Proceedings of the 34th Conference on Design Automation. Anaheim, CA: ACM Press[C], 1997. 526 - 529.

[3] MCMICHAEL D . A Statistical Approach to Situation Assessment [A]. Proc. 2nd Inter. Conf. on Information Fusion[C], 1999.

[4] HALL DL, LLINAS J. An Introduction to Multisensor Data fusion[A]. Special Issue on Data Fusion[C]. Proceedings of the IEEE, 1997.

[5] STEPHEN S, PETER S, DALE K. Data fusion a conceptual approach to level 2 fusion (situational assessment)[A]. Proceedings of SPIE, Aerosense03[C]. Orlando, FL, 2003.

[6] 杨万海.多传感器数据融合及其应用[M].西安:西安电子科技大学出版社,2004.

[7] 何友,王国宏,陆大金,等.多传感器信息融合及应用[M].北京:电子工业出版社,2000.

[8] 熊伟,何友.信息融合系统中的态势预测技术[J].火力与指挥控制,2001,26(4):47 - 50.

[9] 马云.数据融合中态势评估技术研究与实现[D].西安 西安电子科技大学,2003.

[10] 程岳.数据融合中态势估计技术研究[D].西安:西安电子科技大学,2003.