软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2014,25(9):2026−2036 [doi: 10.13328/j.cnki.jos.004632]
©中国科学院软件研究所版权所有.
E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

海量信息融合方法及其在状态评价中的应用*

李嘉菲[1,2], 周 斌[1,2], 刘大有[1,2], 胡 亮[1,2], 王 峰[1,2]

[1](吉林大学 计算机科学与技术学院,吉林 长春 130012)

[2](符号计算与知识工程教育部重点实验室(吉林大学),吉林 长春 130012)

通讯作者: 胡亮, E-mail: hul@jlu.edu.cn

摘 要: 针对证据理论无法有效处理海量信息融合的不足,提出一种结合聚类和凸函数证据理论的海量信息融合方法,旨在解决状态评价等普遍而重要的应用问题.该方法首先基于聚类算法 BIRCH 对采集的海量信息进行预处理,形成多个簇;然后,针对状态评估类问题所用数据大多为数值数据和序数数据这一特点,计算每个簇的质心,并将其作为该簇的代表信息,基于广义三角模糊隶属函数对每个质心信息进行基本概率指派形成证据;最后,基于凸函数证据理论完成各证据的组合,从而完成了海量信息的融合.仿真实验结果表明:该方法既高效又合理地融合了海量信息,为海量信息融合技术的发展提供了一条探索途径.

关键词: 证据理论;聚类;信息融合;海量信息;状态评价

中图法分类号: TP18

中文引用格式: 李嘉菲,周斌,刘大有,胡亮,王峰.海量信息融合方法及其在状态评价中的应用.软件学报,2014,25(9): 2026−2036. http://www.jos.org.cn/1000-9825/4632.htm

英文引用格式: Li JF, Zhou B, Liu DY, Hu L, Wang F. Massive information fusion algorithm and its application in status evaluation. Ruan Jian Xue Bao/Journal of Software, 2014,25(9):2026−2036 (in Chinese). http://www.jos.org.cn/1000-9825/4632.htm

Massive Information Fusion Algorithm and Its Application in Status Evaluation

LI Jia-Fei[1,2], ZHOU Bin[1,2], LIU Da-You[1,2], HU Liang[1,2], WANG Feng[1,2]

[1](College of Computer Science and Technology, Jilin University, Changchun 130012, China)

[2](Key Laboratory of Symbolic Computation and Knowledge Engineering of Ministry of Education (Jilin University), Changchun 130012, China)

Corresponding author: HU Liang, E-mail: hul@jlu.edu.cn

Abstract: To solve the problem that the evidence theory can’t efficiently deal with the fusion of massive information, a new method combining clustering and the convex evidence theory is put forward. The method aims to solve the common and important application problems of the status evaluation. First, the famous clustering algorithm BIRCH is performed to pre-process the data, generating multiple clusters. Second, the centroid of each cluster is calculated as the representation of the cluster pertaining to the fact that most data used for status evaluation have numeric attribute or ordinal attribute. Then, to form the evidence provided by the information in each cluster, the centroid information is given a basic probability assignment value based on the generalized triangular fuzzy membership function. Finally, evidences are combined according to the combination rule of the convex evidence theory. As a result, the massive information fusion is achieved. The results of simulation experiment show that the presented method can efficiently and reasonably perform the massive information fusion, providing a new way to improve the massive information fusion techniques.

Key words: evidence theory; clustering; information fusion; massive information; status evaluation

* 基金项目: 国家自然科学基金(61133011, 61170092, 60973088, 61202308); 国家高技术研究发展计划(863)(2011AA010101); 吉林省重点科技攻关项目(20130206046GX)

收稿时间: 2014-01-20; 定稿时间: 2014-04-09

© 中国科学院软件研究所 http://www.jos.org.cn

状态评价是一类相当普遍的重要问题,它是一种典型的有序命题类问题[1].在日常生活中,评价类问题随处可见,如:环境监测领域,把室内环境舒适度分为很舒适、较舒适和不舒适 3 类;评估桥梁状况时,将其分为潜在危险、严重破坏、轻微破坏和状态极佳 4 类;农业生产领域,将耕作土地的肥沃程度分为高肥力、中上肥力、中肥力、中下肥力、低肥力 5 类.为了提高状态评价系统的可靠性和性能,现在普遍采用了多传感器信息融合技术.以室内环境舒适度评估为例,舒适度的评价非常复杂,不仅与温度、湿度、气流、太阳辐射等物理条件有关,而且随人体热平衡、个人活动量和着装多少而变化[2].由于人类的感觉具有模糊性、不确定性和主观性而很难量化[3],所以,舒适度目前是一种不可直接测量的物理量,只能借助于信息融合技术对影响环境舒适度的多种属性进行推理,以评价环境的舒适程度.

近年来,世界范围的信息化变革,使得几乎每个行业都面临着大数据问题,这些数据具有海量、动态、不确定和多源异构等特点,上述特点使得数据的存储、传输与及时响应和处理面临巨大的挑战[3−5].信息融合作为有效整合和管理数据的重要工具之一,其研究正吸引着越来越多的研究者的关注[3,4].Dempster-Shafer 证据理论既能处理随机性所导致的不确定性,又能处理模糊性所导致的不确定性,在信息融合领域中获得了广泛应用[3,4].Basir 等人采用传统证据理论对多传感器采集的证据进行融合来诊断汽车引擎的故障,并提出两种新方法提高 mass 函数建模和证据组合的有效性[6].Panigrahi 等人使用当前证据和历史行为,提出结合证据理论和贝叶斯学习的方法来检测信用卡欺诈行为[7].Leung 等人将群体决策与传统证据理论相结合,提出一种能够自动识别并处理不可靠证据的集成信息融合方法,可以有效地处理高冲突证据[8].Mora 等人使用传统证据理论对空间和卫星采集的数据进行融合,以判定森林中的主要树种[9].传统证据理论模型不适于求解有序命题类问题,凸函数证据理论模型[10]是传统证据理论的一种重要改进,它给出了适合于有序命题类问题的组合函数,有效地解决了有序命题类问题的不确定性处理问题,拓展了证据理论模型的应用范围[11].但是在融合海量信息时,证据理论具有潜在指数级复杂度的缺点会更加突出.如何合理高效地利用海量信息、使用证据理论进行融合,是一个亟待解决的问题.

为此,本文提出一种结合聚类和凸函数证据理论的海量信息融合方法,旨在解决状态评价等普遍而重要的应用问题.该方法首先基于一种综合的层次聚类方法 BIRCH[12]对采集的海量信息进行预处理,形成多个簇;然后,针对状态评价类问题所用数据大多为数值数据和序数数据这一特点计算每个簇的质心,将其作为该簇的代表信息;接着,基于广义三角模糊隶属函数[16]对每个质心信息进行基本概率指派(basic probability assignment,简称 BPA);最后,基于凸函数证据理论完成各证据的组合,从而完成了海量信息的融合.本文在一个室内环境舒适度判定系统中验证了该方法的有效性,该方法也可以很方便地推广到其他状态评价系统中.

1 预备知识

1.1 凸函数证据理论

凸函数证据理论模型[1,10]比传统的证据理论模型更适合对有序命题类问题进行不确定性推理[10,11].它构造了适合有序命题类问题和具有凸性质的证据组合函数,是非常适合求解有序命题类问题的不确定性推理模型[10].

定义 1. 简单命题 $P_1,P_2,...,P_n$ 是一组有序命题,如果它们满足:

- 对 $i=1,2,...,n$,命题 P_i 的主词项均为 S,谓词项为 s_i;
- 对 $i=1,2,...,n$,s_i 均描述 S 的同一性质或特征;
- 谓词项 $s_1,s_2,\ldots,s_n$ 描述 S 同一性质的程度依次增强或减弱.

定义一组有序命题间的小于等于关系,记为≤.

定义 2. 一组有序命题 $P_1,P_2,...,P_n$ 的真值$|P_1|,|P_2|,...,|P_n|$应呈现出凸的性质,即对任意 $P_i\leqslant P_j\leqslant P_k$,都有$|P_j|\geqslant \min\{|P_i|,|P_k|\}$成立.

文献[10]给出针对有序命题的新的综合函数 f 的定义.

© 中国科学院软件研究所 http://www.jos.org.cn

定义 3. 设概念 $S=\{s_1,s_2,\ldots,s_n\}$,$s_1,s_2,\ldots,s_n$ 为一组有序命题. $\Re=\{\mu\mid\mu$是$2^S\cup\{\bar{S}\}$上的基本支持函数$\}$ 表示 $2^S\cup\{\bar{S}\}$ 上的基本支持函数空间,$f:\Re\times\Re\to\Re$为其不确定性值的综合函数.对于$\forall\mu_1,\mu_2\in\Re$,有:

① 当$\mu_1=\mu_0$,有 $f(\mu_0,\mu_2)=f(\mu_2,\mu_0)=\mu_2$;

② 当$\mu_1\neq\mu_0$ 且$\mu_2\neq\mu_0$ 时,有:

$$f(\mu_1,\mu_2)(s_i)=\begin{cases}\sum_{1\leqslant k\leqslant i}\left\{\frac{1}{2}\mu_1(s_k)[1+\mu_1(\bar{S})]+\frac{1}{2}\mu_2(s_k)[1+\mu_2(\bar{S})]\right\}\Big/(g-k+1), & \text{if } i<g\\ \sum_{i\leqslant k\leqslant n}\left\{\frac{1}{2}\mu_1(s_k)[1+\mu_1(\bar{S})]+\frac{1}{2}\mu_2(s_k)[1+\mu_2(\bar{S})]\right\}\Big/(k-g+1), & \text{if } i>g\\ \sum_{1\leqslant k\leqslant g}\left\{\frac{1}{2}\mu_1(s_k)[1+\mu_1(\bar{S})]+\frac{1}{2}\mu_2(s_k)[1+\mu_2(\bar{S})]\right\}\Big/(g-k+1)+ & \\ \sum_{g+1\leqslant k\leqslant n}\left\{\frac{1}{2}\mu_1(s_k)[1+\mu_1(\bar{S})]+\frac{1}{2}\mu_2(s_k)[1+\mu_2(\bar{S})]\right\}\Big/(k-g+1), & \text{if } i=g\end{cases}\tag{1}$$

其中,

- $g=\left\lfloor\sum_{1\leqslant i\leqslant n}\left\{\frac{1}{2}\left[\mu_1(s_i)+\frac{\mu_1(s_i)}{1-\mu_1(\bar{S})}\mu_1(\bar{S})\right]+\frac{1}{2}\left[\mu_2(s_i)+\frac{\mu_2(s_i)}{1-\mu_2(\bar{S})}\mu_2(\bar{S})\right]\right\}\times i\right\rfloor$;
- $\mu_0\in\Re$,满足$\mu_0(s_i)=0$,对 $i=1,2,\ldots,n$.

定义 3 确定的证据理论模型被称作凸函数证据理论模型.可以证明:如上定义的函数 $w=f(\mu_1,\mu_2)$是基本支持函数,并且是一个凸函数,对于$\forall\mu_1,\mu_2\in\Re$[10].$g\in[1,n]$,g 表示最有可能为真之命题的序号.

文献[1]对上述组合函数进行了深入细致地分析,并给出了改进方法,将定义 3 中的公式(1)改写为

$$f(\mu_1,\mu_2)(s_i)=\begin{cases}\sum_{1\leqslant k\leqslant i}\{w_1\mu_1(s_k)[1+\mu_1(\bar{S})]+w_2\mu_2(s_k)[1+\mu_2(\bar{S})]\}\Big/(g-k+1), & \text{if } i<g\\ \sum_{i\leqslant k\leqslant n}\{w_1\mu_1(s_k)[1+\mu_1(\bar{S})]+w_2\mu_2(s_k)[1+\mu_2(\bar{S})]\}\Big/(k-g+1), & \text{if } i>g\\ \sum_{1\leqslant k\leqslant g}\{w_1\mu_1(s_k)[1+\mu_1(\bar{S})]+w_2\mu_2(s_k)[1+\mu_2(\bar{S})]\}\Big/(g-k+1)+ & \\ \sum_{g+1\leqslant k\leqslant n}\{w_1\mu_1(s_k)[1+\mu_1(\bar{S})]+w_2\mu_2(s_k)[1+\mu_2(\bar{S})]\}\Big/(k-g+1), & \text{if } i=g\end{cases}\tag{2}$$

其中,

- $gd=\sum_{1\leqslant i\leqslant n}\{w_1\times\mu_1(s_i)[1+\mu_1(\bar{S})/(1-\mu_1(\bar{S}))]+w_2\times\mu_2(s_i)[1+\mu_2(\bar{S})/(1-\mu_2(\bar{S}))]\}\times i$;
- $g=\begin{cases}\lceil gd\rceil, & \text{当}gd-\lfloor gd\rfloor\geqslant\Delta_2\text{时}\\ \lfloor gd\rfloor, & \text{当}gd-\lfloor gd\rfloor\leqslant\Delta_1\text{时}\\ \text{把}g=\lfloor gd\rfloor\text{和}g=\lceil gd\rceil\text{得到的结果合成一个,当}\Delta_1<gd-\lfloor gd\rfloor<\Delta_2\text{时} & \end{cases}$.

其中,$w_1+w_2=1$,$\Delta_1,\Delta_2>0$(两个待定常数),一般可选$\Delta_1=0.2$,$\Delta_2=0.8$.

当被组合的证据的个数 $m>2$ 时,公式(2)变成公式(3):

令 $A_j(k)=\mu_j(s_k)[1+\mu_j(\bar{S})/(1-\mu_j(\bar{S}))]$:

$$f(\mu_1,\mu_2,\ldots,\mu_m)(s_i)=\begin{cases}\sum_{1\leqslant k\leqslant i}\left(\sum_{1\leqslant j\leqslant m}w_j\times A_j(k)\Big/(g-k+1)\right), & \text{if } i<g\\ \sum_{i\leqslant k\leqslant n}\left(\sum_{1\leqslant j\leqslant m}w_j\times A_j(k)\Big/(k-g+1)\right), & \text{if } i>g\\ \sum_{1\leqslant k\leqslant g}\left(\sum_{1\leqslant j\leqslant m}w_j\times A_j(k)\Big/(g-k+1)\right)+ & \\ \sum_{g+1\leqslant k\leqslant n}\left(\sum_{1\leqslant j\leqslant m}w_j\times A_j(k)\Big/(k-g+1)\right), & \text{if } i=g\end{cases}\tag{3}$$

其中,

© 中国科学院软件研究所 http://www.jos.org.cn

- $gd=\sum_{1\leqslant i\leqslant n}\left\{\sum_{1\leqslant j\leqslant m}w_j\times\mu_j(s_i)[1+\mu_j(\bar{S})/(1-\mu_j(\bar{S}))]\right\}\times i$;
- $g=\begin{cases}\lceil gd\rceil, & \text{当}gd-\lfloor gd\rfloor\geqslant\Delta_2\text{时}\\ \lfloor gd\rfloor, & \text{当}gd-\lfloor gd\rfloor\leqslant\Delta_1\text{时}\\ \text{把}g=\lfloor gd\rfloor\text{和}g=\lceil gd\rceil\text{得到的结果合成一个,当}\Delta_1<gd-\lfloor gd\rfloor<\Delta_2\text{时}\end{cases}$.

其中, $\sum_{1\leqslant j\leqslant m}w_j=1,\Delta_1,\Delta_2>0$. 结果合成方法参见文献[1].

1.2 聚类方法BIRCH

现有的传统聚类分析方法大多数仅局限于处理数据规模较小、包含连续属性或者分类属性的数据聚类问题,无法有效处理现实生活中包含混合属性的超大规模甚至海量数据的聚类问题[14].

BIRCH[8,9]是为大量数值数据聚类设计的,它将层次聚类与迭代划分等其他聚类方法集成在一起.BIRCH 使用聚类特征(clustering feature,简称 CF)概括一个簇,使用聚类特征树(CF-树)表示聚类的层次结构.这些结构帮助聚类方法在大型数据库、大规模数据集甚至在流数据库中取得高的速度和可伸缩性,还使得 BIRCH 方法对新对象增量或动态聚类也非常有效[13].基于 BIRCH 的上述优点,本文选取其对状态评估类应用中的海量数据进行预处理,将相似的数据聚集到同一簇内.

簇的聚类特征[13]是一个 3 维向量,汇总了对象簇的信息,定义如下:

$$CF=\langle n,LS,SS\rangle,$$

其中,LS 是 n 个点的线性和,而 SS 是数据点的平方和.

聚类特征本质上是给定簇的统计汇总,使用其概括簇可以避免存储个体对象或点的详细信息,只需固定大小的空间来存放聚类特征.这是空间中 BIRCH 有效性的关键.

CF-树是一棵高度平衡树,它存储了层次聚类的聚类特征.CF-树有两个参数:分支因子 B 和阈值 T.分支因子定义了每个非叶节点子女的最大数目,而阈值参数给出了存储在树的叶节点中子簇的最大直径.这两个参数影响结果树的大小[13].

BIRCH 方法包括两个阶段[13]:

- 阶段 1:BIRCH 扫描数据库,建立一个初始存放于内存的 CF 树,它可以被看作数据的多层压缩,试图保留数据内在的聚类结构;
- 阶段 2:BIRCH 采用某个聚类算法对 CF 树的叶节点进行聚类,如典型的划分方法.把稀疏的簇当做离群点删除,而把稠密的簇合并为更大的簇.

1.3 证据的基本概率指派方法

正确获得证据理论中基本概率赋值,是应用证据理论的基础和关键,也是实际应用中最难的一步[15,16].BPA 设置得是否合理,直接关系到融合结果是否正确.BPA 生成可以分为两大类:一类是专家根据主观经验加以设定,一类是系统根据一些已知条件自动生成 BPA.文献[16]提出的 BPA 属于 BPA 自动生成方法,指在系统具有一定样本数据的前提下,根据传感器报告自动生成 BPA 函数.该方法首先基于广义三角模糊数描述模型数据库中已知状态的特征属性(模糊模型标记),然后确定传感器测量值与模型库中模型标记的似然度,该似然度表示在采集的模糊信息下确定为某一目标的可能性,在数值上表示了传感器信息对某一命题支持的程度,利用似然度确定传感器输出的基本概率指派.该方法在目标识别中的应用实例说明:它可以较好地反映传感器报告对各个目标的隶属程度,具有较好的通用性,且计算复杂度低.为此,本文采用该 BPA 指派方法对传感器的测量信息进行基本概率赋值以产生证据.

2 基于凸函数证据理论的海量信息融合方法

随着大数据时代的来临,物联网、面向复杂应用背景的多传感器系统的大量涌现将产生大量数据,这些数

© 中国科学院软件研究所 http://www.jos.org.cn

据具有海量、动态、不确定和多源异构等特点,使得数据的存储、传输与及时响应和处理面临巨大的挑战[3−5].

在实际应用中,所有融合方法都会面临处理不确定信息的挑战,而证据理论因其不确定性的表示、量度和组合方面的优势受到广泛地重视,是一种有前景的信息融合方法[3,4].但是,现有基于证据理论的信息融合方法还不具有处理海量信息(证据)的能力,这一缺点将严重阻碍证据理论在信息化变革的发展和应用.为此,本文将数据挖掘领域擅长处理大规模数据的综合聚类算法 BIRCH[11]引入凸函数证据理论,提出一种新的基于凸函数证据理论的海量信息融合方法.算法的主要思想是:首先,基于聚类方法 BIRCH 对采集的海量信息进行预处理,形成多个簇;然后,计算每个簇的质心,将质心作为该簇的代表;接着,使用本文第 1.3 节介绍的方法对每个质心信息中与融合目标相关的各个属性进行基本概率指派,组合后形成每个簇的代表性证据;最后,基于凸函数证据理论完成各簇代表性证据的组合,从而完成了海量信息的融合.由于本算法主要目标是处理状态估计应用中的数据,而这类数据主要是数值数据和序数数据,序数数据又可以转化为数值数据来处理[13],因此在聚类数值数据后,某一个簇的质心可以作为该簇的代表完成后续运算.算法的主要步骤如下:

1) 根据采集数据的特点,选择与融合目标相关的所有属性 $A_1,A_2,...,A_n$;
2) 根据 $A_1,A_2,...,A_n$,采集数据的规模和特点确定 BIRCH 算法的分支因子 B、阈值 T 和叶节点中子簇数的最大值 L.设定初始值 B=10,T=0,L=15;
3) 根据步骤 1)和步骤 2),BIRCH 算法对采集的 N 个海量数据进行聚类形成 r 个簇 $C_1,C_2,...,C_r$;
4) 根据步骤 3)的聚类结果,计算每个簇的质心数据 $Q_1,Q_2,...,Q_r$;
5) 根据应用问题,确定辨识框架Θ:$\{H_1,H_2,...,H_k\}$;
6) 建立模型库的模糊模型标记.根据给定的 $H_1,H_2,...,H_k$ 的一定量样本数据,针对样本的某个属性 A_i,可以确定该属性的最小值、最大值和平均值,基于这 3 个属性值,可以建立一个三角形模糊数来描述命题 H_j.据此,建立对应的隶属函数 $\mu_{A_iH_j}(x), i=1,2,...,n, j=1,2,...,k$;
7) 确定传感器采集数据的观测函数.对属性 A_i,首先计算出模型库中所有模型标记属性 A_i 的平均方差,根据计算所得平均方差和传感器当前的测量值,将该测量值扩展成可表示的三角模糊数,进而获得与其对应的传感器观测函数 $g_{A_i}(x), i=1,2,...,n$;
8) 确定传感器采集数据与模型库中模型标记的似然度.传感器观测函数 $g_{A_i}(x)$ 与目标模糊模型标记 $\mu_{A_iH_j}(x), j=1,2,...,k$ 两条曲线相交部分纵坐标的最大值即为传感器报告与模型库中模型标记的似然度 $\mu'_{A_iH_j}(x), j=1,2,...,k$;
9) 生成基本概率指派:
   a) 初始化 BPA.令 $m'_{A_i}(H_j)=\mu'_{A_iH_j}(x), j=1,2,...,k$;
   b) 令 $U_n=\max\limits_{j=1,2,...,k}\mu'_{A_iH_j}(x)$;
   c) 设置全集Θ的初始 BPA: $m'_{A_i}(\Theta)=1-U_n$;
   d) 将 $m'_{A_i}(H_j), j=1,2,...,k$ 和 $m'_{A_i}(\Theta)$ 归一化处理后,获得此时传感器测量生成的属性 A_i 的数据对应的基本概率指派 $m_{A_i}(H_j), j=1,2,...,k$ 和 $m_{A_i}(\Theta)$;
10) 对于步骤 4)中得到的每个簇的质心信息,根据选定的属性 $A_1,A_2,...,A_n$,重复步骤 6)~步骤 9),生成每个簇的质心对应的 n 条证据;
11) 使用经典证据理论的组合公式,融合步骤 10)产生的 n 条证据,形成能反映簇 C_i 对融合目标支持程度的合成证据 $cm_i(H_j)$,i=1,2,…,r,j=1,2,…,k;
12) 令 $cm_i(H_j)$,i=1,2,…,r,j=1,2,…,k 的权重 $w_i=\dfrac{\text{第}i\text{个簇包含的数据量}}{r\text{个簇包含的数据总量}}, i=1,2,...,r$;
13) 利用凸函数证据理论的证据组合公式(3),利用步骤 12)分配的权重对步骤 11)生成的 r 条合成证据进行融合,确定最终融合结果.

本文提出算法的时间复杂度与 BIRCH 算法和凸函数证据理论融合证据时间紧密相关.BIRCH 算法的时间

© 中国科学院软件研究所 http://www.jos.org.cn

复杂度为 $O(N)$,其中,N 为被聚类的对象数[13].而针对辨识框架Θ:$\{H_1,H_2,...,H_k\}$,融合 r 个簇的质心对应的 n 条证据的时间为 $O(r\times n\times k)$.故,整个算法的计算时间为 $\max\{O(N),O(r\times n\times k)\}$.在处理实际应用问题时,$n$ 与 k 的值是确定的,因此,本算法具有线性的处理时间,明显优于传统的凸函数证据理论算法.

3 仿真实验

海量数据的规模通常是 TB 级甚至 PB 级的,在现实生活中较难采集到.为了验证本文提出算法的有效性,我们选择室内环境舒适度这一典型的状态评价类问题中的大规模数据进行仿真实验.由于舒适度的评价非常复杂,本文选取温度、湿度这两个可直接测控并对人类舒适感影响最大的两个变量为代表进行判断,在两个房间 Room 1 和 Room 2 内,分别部署 3 个和 4 个温湿度传感器,我们每隔 10s 采集一次数据.这 7 个传感器 24 小时采集的数据量为 24 000 多条,10 天可以采集 2.5×10^5 条,30 天约采集 7.5×10^5 条数据.由于判定房间内 24 小时内某些时间段的环境舒适度会更有助于指导执行改进措施,因此选择在这两个房间里各传感器在 3 个连续时间段(15 点~19 点、7 点~20 点和 0 点~24 点)采集的大规模数据进行舒适度评价的仿真实验,实验过程和结论也适用于实际评估中更多传感器更长时间采集的规模为 TB 级的海量数据.我们分别使用经典凸函数证据理论方法(简称 CET 算法)与本文算法对这 6 组数据进行融合,根据融合结果,评价某时间段内、某房间的环境舒适度.

3.1 证据的生成和组合

设室内环境温度为 0°C~40°C,空气湿度为 0~100%,舒适度论域Θ={不舒适,较舒适,很舒适}.如前所述,采用广义三角模糊数来定义温度隶属函数 $\mu_{t_i}(x)$ 和湿度隶属函数 $\mu_{h_i}(x)$,其中,i=1,2,…,n:

$$\mu_{t_i}(x)=\begin{cases}0, & x<a\\ \dfrac{x-a}{b-a}, & a\leqslant x\leqslant b\\ \dfrac{c-x}{c-b}, & b\leqslant x\leqslant c\\ 0, & x>c\end{cases} \tag{4}$$

$$\mu_{h_i}(x)=\begin{cases}0, & x<A\\ \dfrac{x-A}{B-A}, & A\leqslant x\leqslant B\\ \dfrac{C-x}{C-B}, & B\leqslant x\leqslant C\\ 0, & x>C\end{cases} \tag{5}$$

根据相关文献的研究结果[2,17],可以得到不同舒适度对应的温度和湿度区间.

Table 1 Relationship between different comfort degree and temperature and humidity

表 1 不同舒适度与温、湿度的对应关系

	不舒适	较舒适	很舒适
温度	0°C~13°C, 33°C~40°C	13°C~22°C, 29°C~33°C	22°C~29°C
湿度	0%~28%, 94%~100%	28%~47%, 86%~94%	47%~86%

由表 1 中的数据,可以确定温度和湿度属性在某一区间上的最小值、最大值和平均值.基于这 3 个属性值,可以建立相应的三角模糊函数来描述舒适度的状态.特别地,对同一舒适度状态对应的温度和湿度的各个区间,要分别建立隶属度函数,在确定传感器采集数据与模型库中模型标记的似然度时,这些区间上的隶属度函数都要与传感器观测函数比较.以不舒适状态对应的温度为例,由于不舒适对应温度的两个区间,我们分别建立两个区间上的三角模糊函数 $\mu_{t_1}(x)$ 和 $\mu'_{t_1}(x)$:

© 中国科学院软件研究所 http://www.jos.org.cn

$$\mu_{t_1}(x)=\begin{cases}0, & x<0\\ \dfrac{x}{6.5}, & 0\leqslant x\leqslant 6.5\\ \dfrac{13-x}{6.5}, & 6.5\leqslant x\leqslant 13\\ 0, & x>13\end{cases},$$

$$\mu'_{t_1}(x)=\begin{cases}0, & x<33\\ \dfrac{x-33}{3.5}, & 33\leqslant x\leqslant 36.5\\ \dfrac{40-x}{3.5}, & 36.5\leqslant x\leqslant 40\\ 0, & x>40\end{cases}.$$

其他舒适度状态对应的温度、湿度隶属函数的构建方式类似,不再赘述.下面我们以 Room 1 的 1 号传感器采集的数据(1,52.75,19.25,2012/4/13 17:01:08)为例,说明根据其温度和湿度的测量值以及上述隶属函数进行证据 BPA 指派的过程.此数据的格式为(传感器编号,湿度,温度,采集时间).

首先,计算出模型库中所有模型标记温度的平均方差为 4.根据计算所得平均方差,将测得的温度值扩展成可表示的三角模糊数(15.25,19.25,23.25),其对应的传感器观测函数 $g_{t_1}(x)$ 为

$$g_{t_1}(x)=\begin{cases}0, & x<15.25\\ \dfrac{x-15.25}{4}, & 15.25\leqslant x\leqslant 19.25\\ \dfrac{23.25-x}{4}, & 19.25\leqslant x\leqslant 23.25\\ 0, & x>23.25\end{cases}.$$

温度传感器观测函数 $g_{t_1}(x)$ 分别与 3 种舒适度状态隶属函数 $\mu_{t_1}(x)$和$\mu'_{t_1}(x)$、$\mu_{t_2}(x)$和$\mu'_{t_2}(x)$、$\mu_{t_3}(x)$ 曲线相交部分纵坐标的最大值即为温度传感器报告与模型库中 3 种舒适度状态的似然度,则温度传感器报告各舒适度状态的似然度为

$$\mu_t(\text{不舒适})=0,$$
$$\mu_t(\text{较舒适})=0.7941,$$
$$\mu_t(\text{很舒适})=0.1667,$$
$$\mu_t(\Theta)=0.2059.$$

归一化处理后,对温度而言,这条传感器数据生成的基本概率指派为

$$m_t(\text{不舒适})=0,$$
$$m_t(\text{较舒适})=\frac{0.7941}{0+0.7941+0.1667+0.2059}=0.6807$$
$$m_t(\text{很舒适})=\frac{1667}{0+0.7941+0.1667+0.2059}=0.1429$$
$$m_t(\Theta)=\frac{0.2059}{0+0.7941+0.1667+0.2059}=0.1764$$

这样,就完成了使用温度属性生成基本概率指派的全部过程.同理,可使用湿度属性进行 BPA 指派,得到如下结果:

$$m_h(\text{不舒适})=0,$$
$$m_h(\text{较舒适})=0.1789,$$
$$m_h(\text{很舒适})=0.4384,$$
$$m_h(\Theta)=0.3827.$$

© 中国科学院软件研究所 http://www.jos.org.cn

将新生成的证据 m_t 和 m_h 使用证据理论的组合规则进行融合,形成新证据为

$$m(\text{不舒适}) = 0,$$
$$m(\text{较舒适}) = 0.5102,$$
$$m(\text{很舒适}) = 0.2400,$$
$$m(\Theta) = 0.2498.$$

利用所有新生成的证据,使用凸函数证据理论进行组合得到的融合结果,就可以判断该房间在某一时间段内的环境舒适度.

3.2 BIRCH算法中参数的选择

BIRCH 算法执行的关键步骤是构造 CF-树,CF-树构建时需要确定 3 个参数:分支因子 B、阈值 T 和叶节点中子簇数的最大值 L.这 3 个参数的取值,直接影响结果树的大小.根据文献[12]的实验结果,我们选定 B 的初始值为 10、T 的初始值为 0、L 的初始值为 15.我们根据所用的温度和湿度数据的特点进行了多次实验,在 B 和 L 值固定的前提下,以每次增加 0.05 的幅度调整 T 值,再观察最终融合结果的支持度值.具体的实验结果见表 2.

Table 2 Results when B is 10, L is 15 and T is set to different value

表 2 B=10,L=15,T 在不同取值时的实验结果

阈值 T	算法执行时间(ms)	生成叶节点的数量	融合结果	较舒适的 BPA
0	1 280	2 779	较舒适	0.810 2
0.05	686	928	较舒适	0.792 5
0.1	593	560	较舒适	0.805 5
0.15	562	451	较舒适	0.810 6
0.2	530	462	较舒适	0.813 5
0.25	515	313	较舒适	0.814 0
0.3	499	221	较舒适	0.813 1
0.35	484	212	较舒适	0.806 0
0.4	468	120	较舒适	0.802 2
0.45	468	96	较舒适	0.796 7
0.5	452	71	较舒适	0.789 9

从表 2 中不难看出:随着 T 的增加,生成的 CF-树中叶节点的数量逐渐减少,算法的执行时间也随之减少.在各种取值情况下,融合的结果都是“较舒适”,只是证据对“较舒适”这一结论的支持程度不同.在 T 为 0.25 时,“较舒适”的 mass 函数值取得最大值 0.814 0,因此我们选定 T 的取值为 0.25.类似地,在 T 和 L 值固定的前提下,以每次增加 5 的幅度调整 B 值,然后观察最终融合结果的支持度值.具体的实验结果由于文章篇幅限制略去,我们最终选定 B 的取值为 30,因为此时生成的叶节点数由原来的 313 降为 259,时间几乎不变,而“较舒适”的 mass 函数值仍取得最大值 0.814 0.进而,我们在保持 T=0.25,B=30 的前提下,以每次增加 5 的幅度来调整 L 值,仍然以“较舒适”的 mass 函数值为目标函数,确定此时使该函数取得最大值 0.814 4 的 L 为 40,此时的运算时间为 530ms,而生成的叶节点数降为 100.因此,我们通过实验最终选定 BIRCH 算法的参数 B=30,T=0.25,L=40.

3.3 大规模数据融合结果分析

基于第 3.1 节和第 3.2 节的方法和结论,我们采集了 Room 1 和 Room 2 在 3 个连续时间段的 6 组数据,数据规模从 4 000~24 000 不等,分别使用 CET 算法和本文算法进行融合.实验结果如图 1 和表 2 所示.

图 1 比较了两种算法在融合相同规模数据时的运行时间,通过图 1 的实验结果我们可以看出:本文算法在运行时间上有明显的优势,随着数据量的增加,本文算法融合时间的增长速度比 CET 算法慢得多,当数据集的规模超过 15 000 条时,CET 算法的时间接近于本文算法的两倍.由此仿真实验结果可以推断:当数据规模为海量时,本文算法的运行时间将比 CET 算法少得多,在时间方面有更大的优势.

© 中国科学院软件研究所 http://www.jos.org.cn

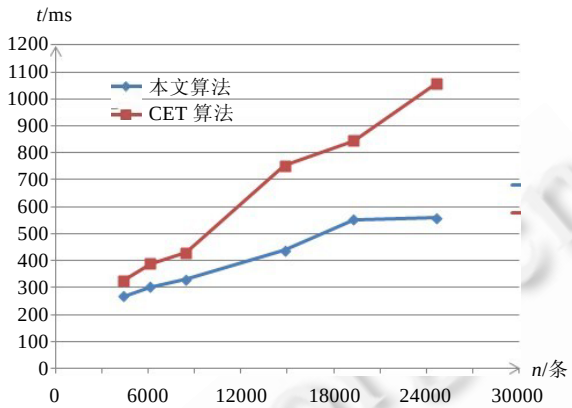


Fig.1 Time comparison between CET and our algorithm

图 1 CET 算法与本文算法运行时间比较

表 3 对 CET 算法和本文算法的最终融合结果进行了比较,不难看出:两种方法对同一数据的融合处理得到的结论是一致的.就“较舒适”的 BPA 值而言,两种算法的差别不大,在区间[0.01,0.1]内,BPA 值间的差别与原始数据的分布有关,如果原始数据的值分布范围比较广,则两种算法的融合结果的 BPA 值差别较大;反之,差别较小.

Table 3 Final fusion results comparison between CET and our algorithm

表 3 CET 算法与本文算法最终融合结果的比较

房间号(数据量)	使用算法	不舒适的 BPA	较舒适的 BPA	很舒适的 BPA	融合结果
Room 1 (4 385)	CET 算法	0.000 4	0.768 9	0.230 7	较舒适
	本文算法	0.000 1	0.686 1	0.313 8	较舒适
Room 2 (6 128)	CET 算法	0.110 9	0.705 4	0.183 7	较舒适
	本文算法	0.089 4	0.678 4	0.232 2	较舒适
Room 1 (8 431)	CET 算法	0.000 1	0.705 3	0.294 6	较舒适
	本文算法	0.000 3	0.624 9	0.374 8	较舒适
Room 2 (19 280)	CET 算法	0.168 6	0.684 9	0.146 5	较舒适
	本文算法	0.170 8	0.695 6	0.113 6	较舒适
Room 1 (14 840)	CET 算法	0.003 9	0.786 6	0.209 5	较舒适
	本文算法	0.001 6	0.788 7	0.209 7	较舒适
Room 2 (24 661)	CET 算法	0.056 6	0.808 2	0.135 2	较舒适
	本文算法	0.171 8	0.718 2	0.110 0	较舒适

综上,我们不难看出:与传统的凸函数证据理论方法相比,本文提出的算法在处理大规模或海量数据的融合问题时有明显的时间优势,且融合结果与传统方法一致.因此,从整体效果上看,本文算法既高效又合理地处理海量信息融合问题,为海量信息融合技术的发展提供了一条新的探索途径.

4 结 论

凸函数证据理论模型能够有效处理有序命题类问题中的不确定性,拓展了证据理论模型的应用范围.但随着大数据时代的来临,在融合海量信息以解决状态评估等有序命题类问题时,传统的 CET 方法将具有指数级的时间复杂度,这是研究者无法接受的.针对这一问题,必须对传统的 CET 模型进行改进,才能完成海量信息的融合.本文正是以此为切入点,通过使用大数据聚类方法 BIRCH 对采集的海量数据进行预处理,形成多个簇.然后计算每个簇的质心,将其作为该簇的代表,使用基于广义三角模糊数的方法对每个质心信息进行基本概率指派.最后,基于凸函数证据理论完成各证据的组合,从而完成了海量信息的融合.仿真实验结果证明:本文提出的算法能在保证融合结果正确的前提下大幅度减少海量信息的融合时间,是一种合理高效的海量信息融合新方法.下一步工作计划获得其他应用领域规模更大的数据,如桥梁健康状况数据、土地肥沃程度数据等,利用本文算

© 中国科学院软件研究所 http://www.jos.org.cn

法进行测试,针对测试结果改进本文算法.

致谢 在此,我们由衷地感谢给本文工作提出宝贵建议的评审老师们.

References:

[1] Liu DY, Yang K, Tang HY, Chen JZ, Yu QY, Chen GF. A convex evidence theory model. Journal of Computer Research & Development, 2000,37(2):175−181 (in Chinese with English abstract).

[2] Han JF. Research on the synthesis of comfort degree for fuzzy sensors based on temperature and humidity. Journal of Transducer Technology, 2002,21(6):19−24 (in Chinese with English abstract).

[3] Nakamura EF, Loureiro AAF, Frery AC. Information fusion for wireless sensor networks: Methods, models, and classifications. ACM Computer Survey, 2007,39(3):9. [doi: 10.1145/1267070.1267073]

[4] Bahador K, Alaa K, Fakhreddine OK, Saiedeh NR. Multi-Sensor data fusion: A review of the state-of-the-art. Information Fusion, 2013,14:28−44. [doi: 10.1016/j.inffus.2011.08.001]

[5] Debasis B, Jaydip S. Internet of things: Applications and challenges in technology and standardization. Wireless Personal Communications, 2011,58:49−69. [doi: 10.1007/s11277-011-0288-5]

[6] Otman B, Yuan XH. Engine fault diagnosis based on multi-sensor information fusion using Dempster-Shafer evidence theory. Information Fusion, 2007,8(4):379−386. [doi: 10.1016/j.inffus.2005.07.003]

[7] Suvasini P, Amlan K, Shamik S, Majumdar AK. Credit card fraud detection: A fusion approach using Dempster-Shafer theory and Bayesian learning. Information Fusion, 2009,10(4):354−363. [doi: 10.1016/j.inffus.2008.04.001]

[8] Yee L, Ji NN, Ma JH. An integrated information fusion approach based on the theory of evidence and group decision-making. Information Fusion, 2013,14(4):410−422. [doi: 10.1016/j.inffus.2012.08.002]

[9] Brice M, Michael AW, Joanne CW. An approach using Dempster-Shafer theory to fuse spatial data and satellite image derived crown metrics for estimation of forest stand leading species. Information Fusion, 2013,14(4):384−395. [doi: 10.1016/j.inffus.2012.05.004]

[10] Yang Y, Liu DY, Wu LZ, Wang WY. An important extension of an evidence-theory based inexact reasoning model. Chinese Journal of Computers, 1990,13(10):772−778 (in Chinese with English abstract).

[11] Liu DY, Ouyang JH, Tang HY, Chen JZ, Yu QY. Research on a simplified evidence theory model. Journal of Computer Research & Development, 1999,36(2):134−138 (in Chinese with English abstract).

[12] Zhang T, Ramakrishnan R, Livny M. BIRCH: An efficient data clustering method for very large database. In: Proc. of the ACM SIGMOD Int'l Conf. on Management of Data. 1996. 103−114. [doi: 10.1145/233269.233324]

[13] Han JW, Kamber M, Pei J. Data Mining: Concepts and Techniques. 3rd ed., Beijing: China Machine Press, 2012 (in Chinese).

[14] Chiu T, Fang DP, Chen J. A robust and scalable clustering algorithm for mixed type attributes in large database environment. In: Proc. of the 7th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining. 2001. 263−268. [doi: 10.1145/502512.502549]

[15] Xiao JY, Tong MM, Zhu CJ, Wang XL. Basic probability assignment construction method based on generalized triangular fuzzy number. Chinese Journal of Scientific Instrument, 2012,33(2):429−434 (in Chinese with English abstract).

[16] Deng Y, Zhu ZF, Zhong S. Fuzzy information fusion based on evidence theory and its application in target recognition. Acta Aeronautica et Astronautica Sinica, 2005,26(6):754−758 (in Chinese with English abstract).

[17] Xie DP, Qin HF, Yu CB. Research of monitoring system for indoor comfort degree based on fuzzy theory. China Measurement & Testing Technology, 2008,34(4):126−128 (in Chinese with English abstract).

附中文参考文献:

[1] 刘大有,杨鲲,唐海鹰,陈建中,虞强源,陈桂芬.凸函数证据理论模型.计算机研究与发展,2000,37(2):175−181.

[2] 韩峻峰.基于温湿度的模糊传感器舒适度合成法研究.传感器技术,2002,21(6):19−24.

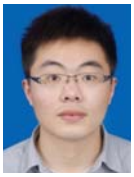
[10] 杨莹,刘大有,吴立真,王伟元.对一种基于证据理论的不确定性处理模型的重要扩充.计算机学报,1990,13(10):772−778.

[11] 刘大有,欧阳继红,唐海鹰,陈建中,虞强源.一种简化证据理论模型的研究.计算机研究与发展,1999,36(2):134−138.

© 中国科学院软件研究所 http://www.jos.org.cn

[13] 韩家炜,Kamber M,裘健.数据挖掘:概念与技术.第 3 版,北京:机械工业出版社,2012.
[15] 肖建于,童敏明,朱昌杰,王小蕾.基于广义三角模糊数的基本概率赋值构造方法.仪器仪表学报,2012,33(2):429−434.
[16] 邓勇,朱振福,钟山.基于证据理论的模糊信息融合及其在目标识别中的应用.航空学报,2005,26(6):754−758.
[17] 谢东坡,秦华锋,余成波.基于模糊理论的室内环境舒适度监测系统研究.中国测试技术,2008,34(4):126−128.

李嘉菲(1976－),女,吉林长春人,博士,副教授,CCF 高级会员,主要研究领域为信息融合,数据挖掘,智能信息处理.
E-mail: jiafei@jlu.edu.cn



周斌(1989－),男,硕士生,主要研究领域为信息融合,数据挖掘.
E-mail: 164908156@qq.com

刘大有(1942－),男,博士,教授,博士生导师,主要研究领域为知识工程与专家系统,分布式 AI,不确定性推理.
E-mail: dyliu@jlu.edu.cn

胡亮(1968－),男,博士,教授,博士生导师,主要研究领域为分布式计算,物理信息系统,网络与信息安全.
E-mail: hul@jlu.edu.cn

王峰(1987－),男,博士生,CCF 学生会员,主要研究领域为计算机易失性数据取证分析,物联网异构信息融合,物理信息系统.
E-mail: wangfeng12@mails.jlu.edu.cn

© 中国科学院软件研究所 http://www.jos.org.cn