

加权模糊聚集算子研究^{*}

李爱平 贾 焰 吴泉源

(国防科技大学计算机学院 长沙 410073)

摘 要 在数据挖掘、模糊专家系统和多 Agent 协同决策过程中,要经常面对信息聚集技术和对多个模糊数据来源进行聚集运算,一般用得最多的是合取、析取及加权平均等算子,但是不同的领域有着不同的需求,本文着重对加权聚集算子进行研究。首先,提出了加权平均关系与析取关系结合后的析取-加权平均算子,该算子解决了加权平均算子不能区分析取与合取的关系。然后,提出了一种最大加权平均算子和最小加权平均算子,该算子将最大最小值算子与加权平均算子进行了泛化,解决了同时考虑信息局部性特征与信息整体性特征的问题。理论分析表明,本文提出的加权模糊聚集算子对于模糊信息源的聚集运算起到了很好的补充和完善的功能。

关键词 模糊聚集算子,析取-加权平均算子,加权算子,最大加权平均算子,最小加权平均算子

A Research on Weighted Fuzzy Aggregation Operators

LI Ai-Ping JIA Yan WU Quan-Yuan

(Computer School of National University of Defense Technology, Changsha 410073)

Abstract Information fusion techniques and fuzzy aggregation operators are commonly applied into several fields of data mining, fuzzy expert system and multi-agent cooperated decision. Because different fields imply different requirements, a large number of aggregation operators exist today. The mostly used is Maximum operators, Minimum operations, and weighted mean operators. In particular, this article pays more attention on weighted fuzzy operators. First, we put forward the disjunction-weighted average operators, which solve the problem that weighted average operator cannot distinguish the relationship between disjunction and conjunction. Secondly, this article generalizes the maximum/minimum and weighted average operators, puts forward maximum weighted average operators and minimum weighted average operators. The fuzzy information sources local character and whole character are both considered in these operators. The theory results show these fuzzy aggregation operators are good complements for fuzzy information fusion.

Keywords Fuzzy aggregation operators, Disjunction-weighted average operators, Weighted operators, Maximum weighted average operators, Minimum average operators

1 前言

信息聚集技术在数据库技术、模糊专家系统和多 Agent 协同决策过程中应用得越来越广泛,需要对多个模糊数据来源进行聚集运算,一般用到最多的是合取、析取及加权平均等算子^[1,2]。但是不同的领域有着不同的需求,需要的模糊聚集算子也是不一样的。在数据库中,模糊聚集算子在模糊查询、数据挖掘、数据聚集运算中有着重要应用^[3]。在专家系统中,模糊聚集算子一方面可以为复合模糊命题具有不同模糊逻辑关系的子命题之间提供不同的证据逻辑组合计算方法;另一方面可以满足不同规则推理出同一结果时(并行规则模式),提供不同的逻辑真值聚合运算方法。因此,将不同的模糊聚集算子组合起来可以提供更多的不确定推理模型^[4]。另外,模糊聚集算子在将多个专家系统针对同一问题决策的结果进行聚合时,也是重要的技术之一。

本文第 2 节介绍相关工作;第 3 节给出析取-加权平均算子;第 4 节给出最大最小加权平均算子;第 5 节给出变权算

子;最后总结全文并介绍未来的研究工作。

2 相关工作

在模糊聚集算子方面,国外主要以 Yager 教授、Torra 教授等人为聚集运算、不同的加权模糊聚集方面进行了深入研究^[1,2,5,6]。其中,Yager 教授提出了有序加权聚集算子(Ordered Weighted Aggregation operators, OWA)的概念^[6],对权重进行了新的解释和定义,并进一步提出了几何有序加权聚集算子(OGWA),区间值有序加权算子等。Torra 教授等人在其基础上对如何构造加权模糊聚集算子进行了研究^[2]。但是,他们的研究主要以理论为主,对实际应用过程中碰到的情况考虑不是很足,在实际应用过程中需要对其进行改进。

国内研究方面,何新贵教授系统介绍了加权平均算子的理论,并给出了其若干应用模式^[7],在很多应用领域中获得了成功应用。由于应用情况非常多,也碰到了一些问题^[4,8]。在变权算子方面,汪培庄教授与李洪兴教授提出了“惩罚性”变权分析的方法,并得到了很好的应用^[9]。但是,由于应用的

^{*} 基金项目:国家“八六三”高技术研究发展计划基金项目(2004AA112020,2005AA112030),武器装备预研基金项目(51415010304KG0175)。李爱平 博士,助研,主要研究方向为专家系统、数据库与分布计算;贾 焰 博士,教授,博士生导师,主要研究方向为数据库与分布计算;吴泉源 博士生导师,教授,主要方向为专家系统,分布计算。

复杂性,在有些情况不能满足,本文针对不同情况进行了分析,并对最大最小值算子与加权平均算子进行了泛化,同时提出了一些新的加权模糊聚集算子。

3 析取-加权平均算子

加权平均算子提出以后,由于它的简单性、易理解性和一定程度上的合理性,很快得到了大量的应用^[7]。但是,在实际应用过程中,发现要进行聚集运算的模糊信息源之间并不仅仅是简单加权关系,有时是析取关系与加权关系的融合。此时,简单的加权平均算子便显出处理能力的不足^[4,8]。

由于加权平均关系各信息源之间的关系只有“ $\wedge$ ”型一种表述符号^[7],因此在实际应用过程中为便于用户理解和使用,在描述信息源之间关系时一般使用“AND”来描述运算“ $\wedge$ ”符号,在实际计算时采用的却不是合取关系中的三角模算子(如 Min 算子),而是加权平均算子。但是,却没有一种与之相对应加权平均算子可以代替子命题之间有“OR”关系的算子。如果一定要用加权平均算子来代替,便会得到加权平均算子代替“AND”和“OR”之后运算结果一样的后果^[8]。这便造成一种误解,认为加权平均算子不能区分析取与合取的关系。造成这种现象的原因是,三角模算子利用摩根定律后的对偶算子是余三角模算子,这便可与各信息源之间的合取和析取的关系对应起来,而加权平均算子利用摩根定律计算的对偶算子还是它自身。例如,在专家系统知识库应用中会碰到这样的情况:

如果(叶片为褐色 或者 叶片为黑色 或者 叶片为黄褐色 并且有黑色斑点)并且(生育期为幼苗期 或者 生育期为成苗期)

则 小麦可能患有黑穗病

如果单纯考虑症状之间关系,专家认为叶片的颜色与生育期之间存在加权平均关系,而且叶片的颜色的权重为 0.7,生育期的权重为 0.3。但是叶片为褐色与叶片为黑色之间并没有加权平均关系,它们之间是互斥的。我们将上述的这种关系称为析取-加权平均关系。

最一般的解决该问题的方法是将其拆分为 6 条规则,但是当知识库中的规则数目增加时,给知识库的管理维护和推理机的效率都会造成很大影响。根据文^[9]中的定义,任何一个模糊逻辑公式 P 都可表示为一个合取范式,即 $P = \bigwedge_{i=1}^m P_i$ 其中, $P_i (i=1, \dots, m)$ 为子句,即 $P_i = \bigvee_{j=1}^{n_i} p_j$, p_j 为原子公式,给出如下定义:

定义 1(析取-加权平均逻辑公式) 符合如下形式的公式:

$$\mathcal{Q} = w_1 * \mathcal{Q}_1 \wedge w_2 * \mathcal{Q}_2 \wedge \dots \wedge w_n * \mathcal{Q}_n, (n \geq 1) \quad (1)$$

其中, $w_i \in [0, 1]$ 且 $\sum_{i=1}^n w_i = 1$, 并且 $\mathcal{Q}_i = (P_{i1} \vee P_{i2} \vee \dots \vee P_{im_i})$, $(m_i \geq 1)$, P_{m_i} 为原子逻辑公式, $\mathcal{Q}_i$ 为析取公式, $\mathcal{Q}$ 的真度为:

$$T(\mathcal{Q}) = \sum_{i=1}^n w_i * \max(T(P_{i1}), T(P_{i2}), \dots, T(P_{im_i})) \quad (2)$$

则称式(1)中的 $\mathcal{Q}$ 为析取-加权平均逻辑公式,式(2)中的运算关系为析取-加权平均算子。

显然,当 $m_i = 1, (i=1, \dots, n)$ 时, (1) 式变为普通的加权平均逻辑公式。

类似地,还可以定义另外一种析取-加权平均关系的算子。

定义 2 $\mathcal{Q}, w_i, \mathcal{Q}_i, P_{m_i}$ 的定义同定义 1, 定义 $\mathcal{Q}$ 真度的计

算方式:

$$T(\mathcal{Q}) = \sum_{i=1}^n w_i * (\min(1, \sum_{j=1}^{m_i} T(P_{ij}))) \quad (3)$$

则 $T(\mathcal{Q})$ 是一个析取-加权平均算子。

定义 1, 2 扩展了普通的加权平均关系,使之可以支持含有析取关系的加权平均聚集的运算。在赋权重时,将具有析取关系的复合子命题作为一个模式赋一个权重即可。

4 最大、最小加权平均算子

加权平均算子过分注重各模糊信息源之间的整体性,没有考虑各模糊信息源的局部性之间的差异,而最小和最大算子又过分注重各模糊信息源的局部性,没有考虑各模糊信息源之间的整体性。

定义 3 向量 $v = (w_1, w_2, \dots, w_n)$ 称作 n 维权重向量当且仅当 $w_i \in [0, 1]$ 且 $\sum_{i=1}^n w_i = 1$ 。

首先给出有序加权聚集 OWA 算子的定义。

定义 4^[6] 令 w 为一 n 维权重向量,映射: $OWA_w: \mathcal{R}^n \rightarrow \mathcal{R}$ 是一个 n 维 OWA 算子,当它符合如下形式时:

$$OWA_w(x_1, \dots, x_n) = \sum_{i=1}^n w_i x_{\sigma(i)} \quad (4)$$

其中, $\{\sigma(1), \dots, \sigma(n)\}$ 是 $\{1, \dots, n\}$ 满足 $x_{\sigma(i-1)} \geq x_{\sigma(i)}, i = \{2, \dots, n\}$ 的一个排序(即 $x_{\sigma(i)}$ 在集合 $\{x_1, \dots, x_n\}$ 里是第 i 大的)。

显然,当权向量 $w = (1, \dots, 0)$ 时, OWA_w 算子即为 Max 算子;当权向量 $w = (0, \dots, 1)$ 时, OWA_w 算子即为 Min 算子。 OWA_w 算子通过改变权向量的不同将 Max 算子和 Min 算子进行了泛化。(注: OWA_w 算子中权向量的含义与定义 3 中的权向量的含义不同, OWA_w 算子中的权向量表达的是各变量取值大小的重要性,而定义 3 中的加权平均中权向量表达的含义是不同信息来源的重要性或可信性等。)

下面给出最大最小加权平均算子定义(注:下面定义中的权向量 w 是加权平均含义的权向量)。

定义 5(最大加权平均算子) 令 w 为一 n 维权重向量,映射: $W_w^{\max}: [0, 1]^n \rightarrow [0, 1]$ 是一个 n 维最大加权平均算子,当它符合如下形式时:

$$W_w^{\max}(x_1, \dots, x_n) = \sum_{i=1}^n \max(\min(na_{ik(i)}, 1 - \sum_{j=1}^{\sigma(i)-1} na_{kj}), 0) * x_{\sigma(i)} \quad (5)$$

其中, $\{\sigma(1), \dots, \sigma(n)\}$ 是 $\{1, \dots, n\}$ 满足 $x_{\sigma(i-1)} \geq x_{\sigma(i)}, i = \{2, \dots, n\}$ 的一个排序(即 $x_{\sigma(i)}$ 在集合 $\{x_1, \dots, x_n\}$ 里是第 i 大的)。

类似地,给出最小加权平均算子的定义。

定义 6(最小加权平均算子) 令 w 为一 n 维权重向量,映射: $W_w^{\min}: [0, 1]^n \rightarrow [0, 1]$ 是一个 n 维最小加权平均算子,当它符合如下形式时:

$$W_w^{\min}(x_1, \dots, x_n) = \sum_{i=1}^n \max(\min(na_{ik(i)}, 1 - \sum_{j=\sigma(i)+1}^n na_{kj}), 0) * x_{\sigma(i)} \quad (6)$$

其中, $\{\sigma(1), \dots, \sigma(n)\}$ 是 $\{1, \dots, n\}$ 满足 $x_{\sigma(i-1)} \geq x_{\sigma(i)}, i = \{2, \dots, n\}$ 的一个排序(即 $x_{\sigma(i)}$ 在集合 $\{x_1, \dots, x_n\}$ 里是第 i 大的)。

性质 1(连续性) 最大最小加权平均算子是连续的。

性质 2(泛化性) 当算子 $W_w^{\max}$ 最大值的权重不小于 $1/n$ 时,算子 $W_w^{\max}$ 变为算子 Max, 当算子 $W_w^{\min}$ 最小值的权重不小于 $1/n$ 时,算子 $W_w^{\min}$ 变为最小值算子 Min; 当算子 $W_w^{\max}$ 的 n 维权向量中除了最小值的权重外其它值权重都为 $1/n^2$ 时,则算子 $W_w^{\max}$ 变为等权平均算子; 同样,当算子 $W_w^{\min}$ 的 n 维权向量(下转第 203 页)

量有很大的影响。当聚类数目大于实际聚类数, k -way 谱聚类方法的效果差^[11]。因此如何自动地确定聚类数目是一个关键的问题, 是未来研究的方向。

3. 如何解决模糊聚类的问题: 尽管在文档聚类中, 谱聚类取得了很好的效果。但是在文档聚类中, 单个词可能属于多个类, 单个文档可能是多主题的文档。这就需要我们用模糊聚类的方法解决。如何确定基于模糊聚类与谱方法的联合: 如建立模糊标准的图划分的目标函数等, 是我们的研究方向。

4. 运用到海量数据中去: 当我们用谱聚类, 不可避免地要计算矩阵的特征值与特征向量。通常这种计算的代价很大, 求解非稀疏矩阵的所有特征向量的标准解法需要 $O(n^3)$ 。同时当应用到海量数据时, 相似度矩阵也很大, 可能会超出计算机的内存。Dhillon^[12]在研究了谱聚类和核 k -means 聚类方法的关系后, 将谱聚类问题用核 k -means 算法求解, 并且运用到核 k -means 算法的优化技巧^[13,14], 来解决海量数据的计算。但该方法只用于非二分图的谱聚类方法, 对二分图的聚类问题仍然是我们未来的研究方向。

参考文献

- Hagen L, Kahng A B. New spectral methods for ratio cut partitioning and clustering. IEEE Trans. Computer-Aided Design, 1992, 11(9): 1074~1085
- Shi J, Malik J. Normalized cuts and image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2000, 22(8): 888~905
- Ng A, Jordan M, Weiss Y. On spectral clustering: Analysis and an algorithm. Advances in Neural Information Processing Systems 14. MIT Press, 2001
- Meila M, Shi J. A random walks view of spectral segmentation. In: 8th International Workshop on Artificial Intelligence and Statistics, 2001
- Zha H, He X, Ding C, Gu M, Simon H D. Bipartite Graph Partitioning and Data clustering. In: Proc. of ACM 10th Int'l Conf. Information and Knowledge Management (CIKM 2001), Atlanta, 2001. 25~31
- Dhillon I S. Co-clustering documents and words using Bipartite Spectral Graph Partitioning. KDD 2001 San Francisco, California, USA
- Ding H Q. Unsupervised feature selection via two-way ordering in gene expression analysis. <http://bioinformatics.oupjournals.org/cgi/reprint/19/10/1259>
- Ding C, He X, Zha H, Gu Ming, Simon H. Spectral Min-Max Cut for Graph Partitioning and Data Clustering. In: Proc. of 1st IEEE Int'l Conf. Data Mining, San Jose, CA, 2001
- Fiedler M. Algebraic connectivity of graphs. Czechoslovak Math. J. 1973, 22: 298~305
- Ding C, He X, Zha H. A spectral method to separate disconnected and nearly-disconnected Web graph components. KDD '01, San Francisco, California, USA
- Verma D, Meila M. A Comparison of Spectral Clustering Algorithms. [UTCS Technical Report # TR-04-25]
- Dhillon I, Guan Y, Kulis B. A Unified View of Kernel k -means, Spectral Clustering and Graph Cuts. [UTCS Technical Report # TR-04-25]
- Zhang R, Rudnick A. A large scale clustering scheme for kernel k -means. In ICPR02, 2002. 289~292
- Dhillon I S, Guan Y, Kogan J. Iterative clustering of high dimensional text data augmented by local search. In: Proceedings of The 2002 IEEE International Conference on Data Mining, 2002
- Cu M, Zha H, Ding C, He X, Simon H, Xia J. Spectral relaxation models and structure analysis for K -way graph clustering and bi-clustering. <http://citeseer.st.psu.edu/context/2380189/639394>
- Weiss Y. Segmentation using eigenvectors: a unifying view. <http://www.cs.huji.ac.il/~yweiss/iccv99.pdf>
- Meila, Xu L. Multiway Cuts and SPectral Clustering. <http://www.stat.washington.edu/www/research/reports/>
- Belkin M, Niyogi P. Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering. Advances in Neural Information Processing Systems 14 (NIPS 2001)
- Yu S U, Shi J. Multiclass spectral clustering. In: Int'l Conf. on Computer Vision, 2003
- Kannan R, Vempala S, Vetta A. On clusterings - good, bad, and spectral. In: Proceedings of the 41st Annual Symposium on Foundations of Computer Science, 2000
- Chan P K, Schlag M, Zien J Y. Spectral k -way ratio-cut partitioning and clustering. IEEE Trans. CAD-Integrated Circuits and Systems, 13: 1088~1096
- Bach F R, Jordan M I. Learning spectral clustering. Neural Info. Processing Systems 16 (NIPS 2003), MIT Press, 2004
- Brand M, Huang K. A unifying theorem for spectral embedding and clustering. In: Int'l Workshop on AI & Stat (AI-STAT 2003), 2003
- Ding C, He X, Zha H, Simon H. Unsupervised learning: self-aggregation in scaled principal component space. In: Proc. 6th European Conf. Principles of Data Mining and Knowledge Discovery (PDMK 2002), 2002. 112~124
- Ding C. Document Retrieval and Clustering: from Principal Component Analysis to Self-aggregation Networks. In: Int'l Workshop on AI & Stat (AI-STAT 2003), 2003
- Kamvar S D, Klein D, Manning C D. Spectral Learning. IJCAI-03, 2003
- Chung F R K. Spectral Graph Theory. Amer. Math. Soc, 1997
- Xing E P, Jordan M I. On semidefinite relaxation for normalized k -cut and connections to spectral clustering. <http://www.cs.berkeley.edu/~epxing/Paper/TR-SDP5.ps>
- Zha H, He X, Ding C, Gu M, Simon H. Spectral relaxation for K -means clustering. <http://www.cs.psu.edu/~zha/papers/ming.ps>
- Brin S, Page L. The Anatomy of a Large-Scale Hypertextual Web Search Engine. In: Proc. 7th International World Wide Web Conference, 1998

(上接第 197 页)

量中除了最大值的权重外其它值权重都为 $1/n^2$ 时, 则最小加权平均算子变为等权平均算子。

性质 3(归一性) n 维最大、最小加权平均算子的新权系数都是 n 维权重向量。

定理 1 最大最小加权平均算子与加权平均算子 WM , 最小值算子 Min , 最大值算子 Max 有如下关系:

$$Min \leq W_{\min}^{\max} \leq WM \leq W_{\max}^{\min} \leq Max \quad (7)$$

定理 1 说明算子 $W_{\max}^{\min}$ 和 $W_{\min}^{\max}$ 是介于 Max 算子, 加权平均算子 WM 和 Min 算子之间的算子。算子 $W_{\max}^{\min}$ 和 $W_{\min}^{\max}$ 将最大最小值算子与加权平均算子进行了泛化, 综合了最大最小值算子与加权平均算子的优点。

结束语 本文提出的加权模糊聚集算子适合于所有的模糊集合与模糊信息源的聚集运算, 在数据挖掘、模糊专家系统、模糊控制系统、多 Agent 系统以及数据集成领域等均有广泛应用前景。我们将其应用到了开发的专家系统平台中, 对复合模糊命题的计算与多结果的聚集运算在实践过程中得到了良好的应用。进一步工作是根据本文的理论研究成果, 针对不同的需求研究合适的具体变权算子形式。

参考文献

- Yager R R. On weighted median aggregation. International Journal of Uncertainty, Fuzziness and Knowledge-based Systems, 1994. 101~113
- Domingo J, Torra V. Median-Based Aggregation Operators for Prototype Construction in Ordinal Scales. International Journal of Intelligent Systems, 2003, 18: 633~655
- Han J W, Kamber M. Data Mining: Concepts and Techniques. San Mateo, CA: Morgan Kaufmann, 2000
- Li A P, Wu Q Y. On Harmonic Triangular Norm Aggregation Operators in Multicriteria Decision. In: Proceedings of the 8th World Multiconference on Systemics, Cybernetics and Informatics, Orlando, 2004. 66~71
- Yager R R. Induced aggregation operators. Fuzzy Sets Systems, 2003
- Yager R R. On ordered weighted averaging aggregation operators in multicriteria decision making. IEEE Trans Syst Man Cybernet, 1988, 18: 183~190
- 何新贵. 模糊知识处理的理论与技术(第 2 版). 北京: 国防工业出版社, 1999
- 陆志峰. 模糊逻辑的研究, 计算机工程与应用, 1999, 8: 15~19
- 汪培庄, 李洪兴. 模糊系统理论与模糊计算机. 北京: 科学出版社, 1996
- Dubois D, Prade H. Weighted minimum and maximum operations in fuzzy sets theory. Information Sciences, 1986, 39: 205~210