

基于正态分布对模糊概念自动计算的 FPDA 应用设计

刘浩舸 管建和

(中国地质大学(北京)信息工程学院 北京 100083)

摘 要 计算机控制系统具有自动机的特征,可以用有限自动机理论来对其进行描述,有限自动机是计算机科学各个方面的重要基石。但除确定性的有限自动机理论外,还有许多模糊事件应该由模糊自动机的隶属函数来解决,文中将重点放在具有正态分布特征的事件上。提出了一种实现自动运算的模糊自动机模型,若要在正态分布中得到“可能发生”和“很可能发生”这类模糊型事件的概率,可以只用这种模型实现自动计算。

关键词 模糊数,正态分布,形式语言,下推自动机

中图法分类号 TP181 文献标识码 A

FPDA of Fuzzy Concepts Automatic Calculation PDA Designing Based on Normal Distribution

LIU Hao-ge GUAN Jian-he

(Faculty of Information Engineering, China University of Geosciences (Beijing), Beijing 100083, China)

Abstract The controlling program of the computer system has the characteristics of FA, which can be described by the finite automaton theory. Finite automaton is an important cornerstone of every aspect of computer science. But besides deterministic finite automaton theory, there are multiply fuzzy events which should be operated by fuzzy automaton according to their membership function. In this paper we focused on the events based on normal distribution. In order to realize the automatic calculation of formulations and improve the effectiveness of the statistics of normal distribution events, this paper presented a fuzzy automaton model to achieve automatic operation, and we gave examples when wanting to get the probability of fuzzy events like “may-occur”, “very-likely-occur” or “rarely-occur” in normal distribution.

Keywords Fuzzy number, Normal distribution, Formal language, Pushdown automaton

1 引言

形式文法是一类语言生成的规则,按照一种文法转化规则可以生成具有相同形式的某种语言,它是一种用来产生形式语言的数学模型。自动机是一种具有离散输入/输出系统的数学模型,它具有有限的内部状态,它的内部状态随着字符的输入而发生变化,它是识别形式语言的一种数学模型。我们使用自动机来设计各种处理器,以在大量的文本编辑器和编译器中处理字符串。

但是随着科学的发展,特别是计算机科学和人工智能的发展,面临的一个重要的课题就是如何让计算机去模拟人脑思维。所以在 1956 年模糊数学诞生之后,又很迅速地产生了模糊语言这个分支。所谓模糊语言,实际就是用模糊数学的方法和观点将自然语言定量化和数学化,以适应计算机程序的需要。语言的模糊性正是语言的精髓,有了模糊性才能表达丰富的情感。同时,自动机作为识别语言的工具,也必须结合模糊数学这一种工具才能让计算机在语句识别中变得越来越智能,才能在一定程度上对人脑进行模拟。通过验证,模糊有限状态自动机的推导可以转化成为传统的确定性有限状态自动机的推导。

模糊自动机可以识别很多有穷自动机不能识别的语言和指令。近年来,各式各样关于模糊型自动机的算法也层出不穷。A Blanco 将模糊自动机的理论应用于人工神经网络,并展示了双层智能神经网络能够与基于模糊语言集合的正则文

法相互等价^[2]。雷红轩等人^[7]于 2007 年提出了取之于非负实数 R^+ 的模糊自动机的概念并将此类自动机应用到动态规划求解最短路线的问题中,得到了求最短路线的算法。朱文倩^[8]于 2009 年提出了一种基于整数加群的模糊自动机的数学模型,并将对策论中的多人博弈问题转化为自动机模型,给出了求最优解的算法。

Krzysztof Wawryn^[10]于 1996 年发表了形式语言在逻辑门电路设计中的应用,并将实验结果应用到了 CMOS 运算放大器的设计方法中。EAHMED^[3]还提出了细胞自动机模型,该自动机模拟了在免疫系统反应较弱时肿瘤细胞的快速增长。Steven B^[9]也将自动机运行应用在实时系统中。文献^[15]总结了自动机史上最重要的一些成就。

正态分布在概率论和数理统计中占有重要的地位,在理论和实践上都有重要的应用。它广泛存在于自然、生产和科技等各个领域。模糊自动机的原理同样可以应用到正态分布中,传统的确定型正态分布只能计算确定的指标在总体中所占的比重,比如“某日正午 12 点温度是 40℃ 的天数的比率”、“一个班中体重是 120 斤的男生所占的比重”或“一个社区中年龄是 45 岁的人的比率”。但如果要计算不确定概念的数据,如某日的“正常”温度、一个班里“中等体重”的男生或一个社区里“中年人”的比率,则其就无法运行了。因此需要结合自动机的设计与模糊数学,才能按照隶属函数将区间数输入到正态分布公式中,从而计算模糊概念在正态分布中的概率。

著名的语言学家乔姆斯基也将文法分成 4 类:0 型文法、

4 模糊自动机的设计

对正态分布密度函数进行识别时需要设计一台下推自动机 (Pushdown automaton)。下推自动机 PDA 在 FA 的基础上增加了堆栈功能,它是一个七元组: $M = \langle Q, \Sigma, \Gamma, \delta, q_0, Z, F \rangle$, 其中:

- (1) Q 是自动机非空的有穷状态集合, $\forall q \in Q, q$ 成为 M 的一个状态。
- (2) Σ 是输入字母表, 正态分布公式应作为字符串输入。
- (3) Γ 是栈符号表。
- (4) Z 被称为开始符号, 是 M 开始时存储在堆栈中的唯一符号。
- (5) q_0 是初始状态, $q_0 \in Q$, 也叫启动状态。
- (6) F 是 M 的最终状态的集合, $F \in Q$, 在本文中 q_3 是唯一的终止状态。
- (7) δ 是转换函数, 它表明自动机状态的变化和栈顶符号的变化。

这里给出用于识别正态分布密度函数的自动机 M_N :

$$M_N = (\{q_0, q_1, q_2, q_3, q_4, q_5\}, \{/, *, -, \sigma, \mu, 2, 1, \text{Pi}, \text{exp}, \text{sqrt}, (,), x\}, \{E, A, B, C\}, \delta, q_0, E, \{q_3\})$$

公式的识别过程如下:

```

E → expE
  → exp(AE)
  → exp(-1 * (BAE)
  → exp(-1 * ((CBAE)
  → exp(-1 * ((x-μ)CBAE)
  → exp(-1 * ((x-μ) * BAE)
  → exp(-1 * ((x-μ) * (x-μ)BAE)
  → exp(-1 * ((x-μ) * (x-μ))AE)
  → exp(-1 * ((x-μ) * (x-μ)) / AE)
  → exp(-1 * ((x-μ) * (x-μ)) / (2 * σ * σBAE)
  → exp(-1 * ((x-μ) * (x-μ)) / (2 * σ * σ)AE)
  → exp(-1 * ((x-μ) * (x-μ)) / (2 * σ * σ))E)
  → exp(-1 * ((x-μ) * (x-μ)) / (2 * σ * σ)) / E)
  → exp(-1 * ((x-μ) * (x-μ)) / (2 * σ * σ)) / (A)
  → exp(-1 * ((x-μ) * (x-μ)) / (2 * σ * σ)) / (sqrt(C)
  → exp(-1 * ((x-μ) * (x-μ)) / (2 * σ * σ)) / (sqrt(2 * Pi) * σE)
  → exp(-1 * ((x-μ) * (x-μ)) / (2 * σ * σ)) / (sqrt(2 * Pi) * σ)

```

转换规则 δ 的原理如图 2 所示。

状态 q_1 接收到字符 x 转换到状态 q_2 时, 就要按照之前定义的隶属函数 $g(x)$ 来进行转换, 此条模糊状态转换语句的效果是: 读入的数字 x 会乘上相应的隶属函数计算结果 $g(x)$, 即将区间 $[A, B]$ 中所有的随机变量 x 放入函数 $f(x) * g(x)$ 中进行计算。将整个区间的函数进行积分, 其结果就是模糊数的概率值。

$$H(x) = \int_{-\infty}^{+\infty} h(x) dx = \int_{-\infty}^{+\infty} f(x) * g(x) dx$$

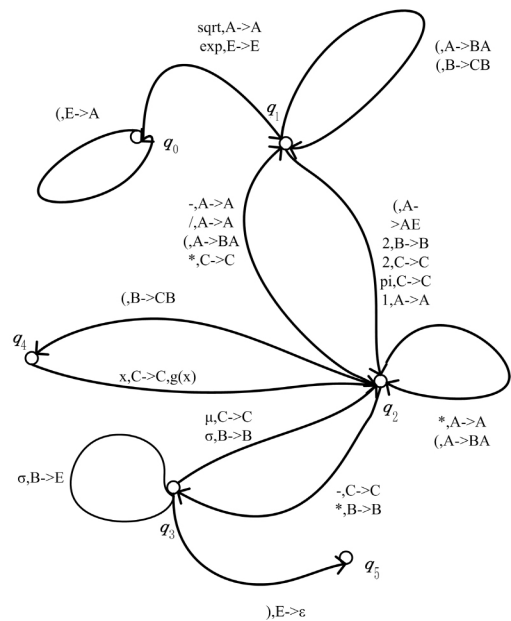


图 2

结束语 本文在广泛应用于各个领域的正态分布公式的基础上将模糊数 x 加入到公式中, 并定义“ x ”的隶属函数为 $g(x)$, 给出了 3 个可能的例子, 并给出了它们的隶属函数。基于形式语言的语法规则, 设计了一个 CFG 来模仿这个公式的生成过程, 最后设计了下推自动机对公式进行识别处理。同时, 利用模糊数 $g(x)$ 来加强自动机在正态分布事件的模糊概念计算中的作用。

结合模糊数学和下推自动机理论, 可以预测不确定事件发生的可能性, 并实现其自动计算。这样的学科交叉的组合有望进一步的发展, 针对不同的实际情况也可以提出更多不同的模糊数。如果能对人类情感进行更加细致和科学的划分, 这种模型将在人工智能方面得到广泛的应用。

参考文献

- [1] ABRIAL J R. The B-Book[M]. Cambridge University Press, Cambridge, 1996.
- [2] BLANCO A, DELGADO M. MC Pegalajar Fuzzy automaton induction using neural networks[J]. International Journal of Approximate Reasoning, 2001, 27(1): 1-26.
- [3] AHMED E. Fuzzy Cellular Automata Models in Immunology [J]. Journal of Statistical Physics, 1996, 85(1): 291-294.
- [4] STEIN C M. Estimation of the Mean of a Multivariate Normal Distribution[J]. Annals of Statistics, 1981, 9(6): 1135-1151.
- [5] ASTESIANO E, et al. Fundamental Approaches to Software Engineering (FASE 98)[M]//Lecture Notes in Computer Science, vol. 1382, Springer, Berlin, 1998.
- [6] WU Q E, PANG X M, HAN Z Y. Fuzzy automata system with application to target recognition based on image processing[J]. Computer and Mathematics with Applications, 2011(61): 1267-1277.
- [7] 雷红轩, 彭家寅, 牟廉明. 取值于 R^+ 的模糊自动机及其在动态规划中的应用[J]. 内江师范学院学报, 2007, 22(6): 33-35.
- [8] 朱文倩, 贺巧, 昌春艳, 等. 基于整数加群模糊自动机及其在对策论中的应用[J]. 内江师范学院学报, 2009, 24(8): 41-43.
- [9] BRADLEY A S, HENDERSON B W, KENDALL B D, et al. A formal design language for real-time systems with data[J]. Science of Computer Programming, 2001(40): 3-29.

地性调度平均降低 14.12%，最高降低 17.01%，说明了集成抢占式的本地性调度算法的平均完成时间得到降低。同时比较集成抢占式非本地性调度算法可以发现，其完成时间也有降低，说明这种算法在降低完成时间的同时，保证了数据本地性。

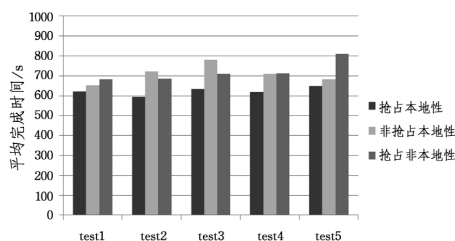


图 4 平均完成时间的比较

接下来测试不同的 Map 任务个数对于两种算法的影响，结果如图 5 所示。从集成了抢占式的调度算法后与非抢占式的本地性调度算法的对比可以看出，当 Map 任务个数较少时，其并没有明显的差别；而伴随 Map 任务个数的增加，两种算法的平均完成时间差值在逐渐增大。由实验数据可知，对于集成了可抢占式的本地性调度算法，Map 任务数越大，调度算法就越优。

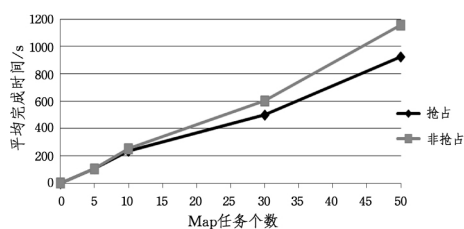


图 5 不同 Map 任务数的平均完成时间比较

结束语 本文分析了 Hadoop 集群下对数据本地性调度的改进，并指出在保证原算法的数据本地性的前提下可以通过抢占的方式减少作业的平均完成时间。通过分析 MapReduce 计算框架，找出抢占与重调度时机，从而达到抢占空闲 Reduce 任务的目的。通过实验可以发现，集成了可抢占式的本地性作业调度的平均等待时间、平均完成时间都有了不同程度的降低。

这种集成了抢占式的本地性调度算法依然存在一些不

足，例如添加抢占和重调度机制后增加了系统开销。测试的作业功能和数据类型不全面，在大数据情况下的性能测试还不是很多，实验在普遍性上还有所不足。接下来的工作重点 是研究如何降低系统开销以及当开销为何值时可被接受等问题，同时将在更大的实验数据集上进行改进和验证。

参考文献

- [1] Hadoop[EB/OL]. [2014-2-01]. <http://hadoop.apache.org>.
- [2] SHVACHKO K, KUANG H, RADIA S, et al. The hadoop distributed file system[C]// Proceedings of the 26th IEEE Symposium on Mass Storage Systems and Technologies. IEEE, 2010: 1-10.
- [3] DEAN J, GHEMAWAT S. MapReduce. Simplified data processing on large clusters[J]. Communications of the ACM, 2008, 51(1): 107-113.
- [4] ZAHARIA M, BORTHAKU D, SARMA J S, et al. Job Scheduling for Multi-user Mapreduce Clusters[R]. EECS Department, University of California, Berkeley, Tech. 2009.
- [5] 董西成. Hadoop 技术内幕: 深入解析 MapReduce 架构设计与实现原理[M]. 北京: 机械工业出版社, 2013.
- [6] 胡丹, 于炯. Hadoop 平台下改进的 LATE 调度算法[J]. 计算机工程与应用, 2014, 50(4): 86-89.
- [7] 何文峰. 基于任务特征与公平策略的 Hadoop 作业调度算法研究[D]. 武汉: 华中科技大学, 2013.
- [8] 燕明磊. Hadoop 集群中作业调度研究[J]. 软件导刊, 2015, 14(4): 1-2.
- [9] 储雅, 马廷淮. 云计算资源调度: 策略与算法[J]. 计算机科学, 2013, 40(11): 8-13.
- [10] 陶昌俊. Hadoop 平台的作业调度算法[D]. 合肥: 中国科学技术大学, 2015.
- [11] PALANISAMY B, SINGH A, LIU L, et al. Purlieus: locality-aware resource allocation for MapReduce in a cloud[C]// Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis. 2011.
- [12] HAMMOUD M, SAKR M F. Locality-Aware Reduce Task Scheduling for MapReduce[C]// Proceedings of International Conference on Cloud Computing Technology & Science. Beijing, 2011: 570-576.
- [13] BIRKEDAL L, ROTHWELL N, TOFTE M, et al. The ML kit (version 1)[R]. Tech. Report DIKU-Report 93 = 14, Department of Computer Science, University of Copenhagen, Universitetsparken 1, DK-2100 Copenhagen, 1993.
- [14] MIKL N A. Real Zeros and Normal Distribution for Statistics on Stirling Permutations Defined by Gessel and Stanley[J]. Siam Journal on Discrete Mathematics, 2007, 23(1): 401-406.
- [15] KARNEY C F F. Sampling Exactly from the Normal Distribution[J]. ACM Transactions on Mathematical Software, 2016, 42(1): 3.
- [16] ABRIAL J R. The B-Book[M]// Cambridge University Press. Cambridge, 1996.
- [17] ALUR R, COURCOUBETIS C, DILL D. Model-checking for real-time systems[C]// IEEE 5th Annual Symp. on Logic In Computer Science. Philadelphia, 1990: 414-425.
- [18] ALUR R, DILL D. Automata for modeling real-time systems[C]// 17th Internat. Colloquium on Automata, Languages and Programming (ICALP 90). Lecture Notes in Computer Science, Springer, Berlin, 1990.

(上接第 559 页)

- [10] WAWRYN K. A Formal Language Description and Inference Strategy for Analog Circuit Design[J]. Circuits Systems Signal Processing, 1996, 15(6): 771-805.
- [11] BUTHAYNA H, EILOUTI A. A Formal Language for Palladian Palazzo Façades Represented by a String Recognition Device[J]. Nexus Network Journal, 2008, 10(2): 245-268.
- [12] ALUR R, DILL D. Automata for modeling real-time systems[C]// 17th Internat. Colloquium on Automata, Languages and Programming (ICALP 90). Lecture Notes in Computer Science, Springer, Berlin, 1990.
- [13] ZHANG X W, LI Y M. Intuitionistic fuzzy recognizers and intuitionistic fuzzy finite automata[M]. Springer-Verlag, 2009.
- [14] KRITHIVASAN K, SHARDA K. Fuzzy w-automata[J]. Information Science, 2001, 138(2001): 257-281.
- [15] WANG J, YIN M, GU W. Fuzzy multiset finite automata and their languages[J]. Soft Comput, 2013, 17(3): 381-390.
- [16] KRITHIVASAN K, SHARDA K. Fuzzy w-automata[J]. Information Science, 2001(138): 257-281.