 密级：

硕 士 学 位 论 文



论文题目 混合模糊语义细胞的学习及其应用

作者姓名 盛守波

指导教师 汤永川

学科(专业) 计算机应用技术

所在学院 计算机科学与技术学院

提交日期 2017-03-08

A Dissertation Submitted to Zhejiang University for the Degree of

Master of Engineering



TITLE: The Study and Application of Hybrid Fuzzy Semantic Cell

Author: Shoubo Sheng

Supervisor: Yongchuan Tang

Subject: Computer Application Technology

College: Computer Science and Technology

Submitted Date: 2017-03-08

研究生学位论文独创性声明

本人声明所呈交的学位论文是本人在导师指导下进行的研究工作及取得的研究成果。除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，也不包含为获得 **浙江大学** 或其他教育机构的学位或证书而使用过的材料。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示谢意。

学位论文作者签名： 签字日期： 年 月 日

学位论文版权使用授权书

本学位论文作者完全了解 **浙江大学** 有权保留并向国家有关部门或机构送交本论文的复印件和磁盘，允许论文被查阅和借阅。本人授权 **浙江大学** 可以将学位论文的全部或部分内容编入有关数据库进行检索和传播，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

（保密的学位论文在解密后适用本授权书）

学位论文作者签名： 导师签名：

签字日期： 年 月 日 签字日期： 年 月 日

摘要

概念实体的表达往往具有一定的模糊性，这种模糊性是蕴含在在语义中出现的。使用合适的概念模型来表达模糊语义具有重要的意义。模糊语义细胞作为最小的模糊概念的表示单元，在数据挖掘、机器学习以及知识发现中具有重要的作用。在概念空间（论域）上，模糊语义细胞被称为“关于”，“类似”以及“和接近”的语义标签，其中代表概念的原型，是定义在论域上的距离函数，则是概念空间中定义在上其他点和的距离的概率密度函数。在模糊语义细胞的学习中我们需要关注语义的覆盖程度、描述的清晰程度以及描述的模糊性这三个因素，因此模糊语义细胞的学习原则很自然地就联系到最大覆盖率、最具典型性和最大模糊熵这三个指标之上。

本文中混合模糊语义细胞是建立在模糊语义细胞的学习基础之上，模糊语义细胞学习的最终目标是要寻找最佳的来刻画具有某个概念的数据集，而混合模糊语义细胞则在此基础上做了更深一层的拓展，考虑具有若干个相关的概念的集合,其中每个概念都对应使用模糊语义细胞来描述第个概念的数据集，混合模糊语义细胞的学习是为了能够寻找到一组最合适的权重参数来刻画某个概念在此概念集合（主题）中的影响程度或者是重要程度，借鉴之前的模糊语义细胞的学习原则，我们需要重新定义并计算语义细胞的两个数字特征：期望粒度和模糊熵。最终将学习混合模糊语义细胞的问题转化为了非线性约束优化问题。

**关键词：**　概念，模糊语义细胞，覆盖率，典型性，模糊熵，混合模糊语义细胞，权重，期望粒度

Abstract

The expression of conceptual entities often comes with some ambiguity, which is implied in semantic expression. The use of appropriate conceptual model to express fuzzy semantics is of grate importance. Fuzzy semantic cell, as the representation unit of the smallest fuzzy concept, plays an important role in data mining, machine learning and knowledge discovery. In the conceptual space(discourse), fuzzy semantic cell is called ‘about ’, ‘similar to ’, ‘close to ’ semantic label, where ,, are the prototype of label, a distance function on , and a probability density function on . In fuzzy semantic cell learning, we need to pay attention to three factors: semantic coverage, clarity of description, and ambiguity of description. Therefore, the ambiguous semantic cell learning goal is naturally related to the maximum coverage, the most typical and the maximum fuzzy entropy of the three indicators above.

In this paper, mixed fuzzy semantic cell is built on the basis of fuzzy semantic cell learning. The ultimate goal of fuzzy semantic cell learning is to find the best to describe the data set of a certain concept while the mixed fuzzy semantic cell is further expanded on this basis. Consider a collection of several related concepts, each of which corresponds to fuzzy semantic cell which describe the data set of concept . Mixed fuzzy cell learning is to find a set of the most appropriate weight parameters to describe the impact or the degree of importance of a concept in the set. Referring to the principle of the learning of fuzzy semantic cell, we need to redefine and compute the two numeric characteristics of semantic cell: expectation granularity and fuzzy entropy . Finally, the learning of mixed fuzzy semantic cell is transformed into a nonlinear constrained optimization problem.

**Keywords：**concept, fuzzy semantic cell, coverage, specificity, fuzzy entropy, mixed fuzzy semantic cell, weight, expectation granularity

目录

[摘要 i](#_Toc167852796)

[Abstract ii](#_Toc167852797)

[第1章 绪论 1](#_Toc167852798)

[1.1 课题背景 1](#_Toc167852799)

[1.1.1 1](#_Toc167852800)

[1.2 本章小结 1](#_Toc167852801)

[1.2.1 1](#_Toc167852802)

[第2章 1](#_Toc167852803)

[2.1 第一节 1](#_Toc167852804)

[2.1.1 1](#_Toc167852805)

[2.2 本章小结 1](#_Toc167852806)

[2.2.1 1](#_Toc167852807)

[第3章 1](#_Toc167852808)

[3.1 第一节 1](#_Toc167852809)

[3.1.1 1](#_Toc167852810)

[3.2 本章小结 1](#_Toc167852811)

[3.2.1 1](#_Toc167852812)

[第4章 1](#_Toc167852813)

[4.1 第一节 1](#_Toc167852814)

[4.1.1 1](#_Toc167852815)

[4.2 本章小结 1](#_Toc167852816)

[4.2.1 1](#_Toc167852817)

[第5章 1](#_Toc167852818)

[5.1 第一节 1](#_Toc167852819)

[5.1.1 1](#_Toc167852820)

[5.2 本章小结 1](#_Toc167852821)

[5.2.1 1](#_Toc167852822)

[第6章 1](#_Toc167852823)

[6.1 第一节 1](#_Toc167852824)

[6.1.1 1](#_Toc167852825)

[6.2 本章小结 1](#_Toc167852826)

[6.2.1 1](#_Toc167852827)

[第7章 1](#_Toc167852828)

[7.1 第一节 1](#_Toc167852829)

[7.1.1 1](#_Toc167852830)

[7.2 本章小结 1](#_Toc167852831)

[7.2.1 1](#_Toc167852832)

[第8章 1](#_Toc167852833)

[8.1 第一节 1](#_Toc167852834)

[8.1.1 1](#_Toc167852835)

[8.2 本章小结 1](#_Toc167852836)

[8.2.1 1](#_Toc167852837)

[参考文献 1](#_Toc167852838)

[攻读硕士学位期间主要的研究成果 1](#_Toc167852839)

[致谢 1](#_Toc167852840)

图目录

[图 1.1简单的语音信号多描述编解码过程 1](#_Toc164669029)

[图 3.1 流程图 1](#_Toc164669030)

表目录

[表 2.1简单的多描述分配表 1](#_Toc164669160)

# 绪论

## 课题背景

我们对于事物的认知是不断从最基本的实体所赋予的概念的学习中得到的，但通过语言所表达出来的概念往往是具有模糊性的。这种模糊性对于人类的学习和思考具有重要的意义，尤其是在模式识别、信息交流以及抽象学习等方面。控制论的创始人维纳提到：人具有模糊概念的能力。人的思考和判断具有模糊性，这使得我们在作出决策的时候并不需要精确地了解事物的特征属性，比如我们在判断一个类似杯子形状物体是否是杯子时，我们并不需要精确知道杯子的具体的数据比如高度，口半径和底半径，以及是否具备把手等数据，而只要把看到的实物与我们大脑中的样本进行对比就可以知道；当我们开车转弯的时候也并不需要确切知道到底要将方向盘转多少角度，以多大的速度过弯，而只需要根据当前的位置就能以适当的操作过去，而针对这种问题，对计算机来说就是一个比较复杂繁琐的事情，需要将汽车的位置、速度、角度作为状态变量列出方程，求出汽车的轨迹。人脑具有存储和处理模糊知识、模糊信息的能力，这正是人脑的优越性所在。我们习惯使用高矮、大小、胖瘦、美丑等词来描述事物，但我们主观是无法定量去描述怎么样的才算高，怎么样的才算矮等。因此对于这种边界模糊的概念，经典的集合论的思想以及现有的分类算法都会显得无所适从。1965年L.A.zadeh率先提出使用模糊集[1] 理论来描述具有模糊概念的问题，与经典集合不同，经典集合描述的是一种确切的拥有（包含）关系，论域中的元素要么属于某个集合，要么不属于这个集合，对应的特征函数要没是1要么是0，而模糊集合打破了这种绝对的隶属关系。在模糊集中隶属函数是整个模糊集理论的基础， 隶属函数为定量分析提供了依据，也为其他的很多的模糊控制、模糊语言和模糊推理打下了基础。

计算机科学发展的一个方向就是让机器来模拟人类的思维，让计算机能够理解人类的自然语言，从而能够让计算机具备更高的智能，而模糊语言就是朝着这个方向在前进。在一个有限的字符集序列的集合中，一个字符串序列是否是合法的是很容易判定的，但是即便不合法的字符序列有时候也能表示出语句的基本含义，如何能够通过一种形式化的语言来描述字符序列的合法程度是计算机理解自然语言的关键。于是很容易联想到可以使用隶属函数来表示它们属于合法序列的程度，把合法语句成分看成是定义在所有可能字符串上的一个模糊集合，而对于各种语言成分的语义可以使用定义在不同论域上的模糊集来表示，这样模糊语言被表示成[2]的四元组，根据模糊集T和命名关系N的隶属函数的计算不同还分为非结构化模糊语言和结构化模糊语言等，但模糊语言的表示并不是一成不变的，现阶段对模糊语言的理论研究还处于初步阶段，但基于模糊概念的模糊程序设计语言仍取得了一定的进展，尤其是在模糊控制领域，模糊语言有了较为成熟的发展和运用，利用有经验的专家领域和操作者带有模糊性的语言总结出的语言控制规则，加以定量的描述而设计出模糊控制器，在实际的运用中取得了很好的效果，英国学者E.H.mamdani在1974年首次提出模糊控制到1987年首次在日本地铁电力机车自动运输系统投入运行，是模糊控制领域发展最为迅猛的时代，模糊技术迅速风靡日本和欧美，1985年世界上第一块模糊逻辑芯片在美国贝尔实验室诞生，成为了模糊技术进展的又一里程碑，世界上很多知名的公司都在积极地从事这方面的研究，给模糊技术的应用注入了新的活力，开辟了诱人的前景，所以研究模糊语言对模糊技术是非常必要的。

前面提到即便是不合法的字符序列也能够表达语句的基本含义，这是由语言的字符串与其意义的对应关系决定的，这种关系称之为语义。可见语义是模糊语言的核心。1971年zadeh给模糊语义做了定量的理论，并在此理论上完备了语言变量的概念，使得具有程度性的词以及以实数集为论域的词有了定量的语义描述。有了这些词的原子语义，再加上一组规定的语义规则算法，就能计算出合成词的含义，合成词又可以和合成词链接形成更加复杂的合成词，然而给它们都一一定义其语义（即隶属函数）是很困难的，于是研究人员定义了一系列的语义算子来解决这个问题，使用语义算子和由它们组合的合成词的语义计算规则来表示模糊语义。语义算子在一定程度上简化了模糊概念的建模，但是还是避免不了隶属程度计算的复杂性，所以对模糊概念的语义建立一套简洁、直观、易计算的模型是非常有必要的，对模糊推理和模糊控制有着重要的意义。

综上所述，模糊语义在表示模糊概念方面是很简单直观的，因此模糊语义的建模就成为了研究模糊语义的重点，我们要做的就是如何在一个独立的数据集中学习出一个模糊语义，有了这部分的理论支持，我们就可以将问题的研究范围拓展到如何针对一个具有相同主题的多个模糊概念的学习，也就是混合模糊语义的学习，混合模糊语义的建模具有很强的应用价值，比如对于某一个针对新闻事件，具有多个方面的报道，每个报道可能侧重不同的领域（我们可以称之为原型），例如对美国大选的网页和媒体报道，有的侧重经济、有的侧重军事、移民等等，针对这些网页数据，利用混合语义来表示这个主题就可以在一定程度上分析出舆论的演变趋势，这种演变趋势是通过相应的语义在混合语义中的影响程度确定的，根据这个特性，可以利用混合模糊语义来做热点追踪和舆情控制等具有价值的应用。因此我们将研究模糊语义和混合模糊语义的学习。

## 国内外研究现状

（1）在模糊集中，隶属函数涵盖了信息的模糊性或者语义的不确定性，与传统的集合论不同的是，集合的运算及性质有较大的差异，并且集合的运算是体现在隶属函数的运算，因此正确地构造模糊隶属函数在模糊控制及概念学习等领域至关重要。基于模糊集和隶属函数的性质，之后的研究根据特定的问题以及研究人员根据自身对于模糊概念的理解提出了多种不同形式的隶属函数，例如模糊统计法、例证法、专家经验法以及二元对比排序法（按照对比的测度又可以划分为比较法、对比平均法、优先关系定序法以及相似优先对比法），二元对比排序法是目前一种较为实用的方法。从本质上来说，隶属函数应当是客观的，但由于人们认知的局限性，不同的人看法也不尽相同使用的方法也可能会不同，但在反应模糊概念的角度上是具有一致性的，它们都是客观实际的一个近似，而且应该在实践中不断加以修正和完善，不断地接近客观实际。

以上隶属函数的确定都存在一些问题，例如模糊统计法是基于频率的，因此计算量相当大；例证法在遇到例证数据较少时，计算的结果不能反映真实的情况；专家经验法存在一定的主观性；二元对比排序法适用于低维的数据集，在高维数据上面又可能面临“维度灾难”。由于目前还没有一套成熟有效的方法，大多数系统的确立还停留在经验和实验的基础上，因此对模糊的度量不具有通用性。但分析已有方法会发现大部分的隶属函数都只是利用了“局部”数据特征，并没有从整体的集合数据的特性去考虑；另外对于隶属函数的确定缺乏一个统一的“参照物”。之后的研究开始考虑数据整体特性，提出了分析法和参照已知函数法（模糊分布法），分析法对所考虑的模糊集所特有的性质进行分析，利用分析所得出的结论建立隶属函数（一般是基于一定的先验知识），因为不同类型的模糊集，具有不同的特殊属性，因此采用分析法建立隶属函数的模式也各有不同。参照已知函数法在论域为实数时，选择已知的带有参数的函数作为所要构造的模糊集的隶属函数的参照函数，然后确定出参数就得到了模糊集的隶属函数。

（2）直觉模糊集是模糊集理论的发展中最有影响力的拓展，它们都是在遵循Zadeh模糊集理论中模糊概念的基础上增加了一个新的参数—非隶属函数，隶属函数用来刻画“亦此亦彼”的概念，而非隶属函数正好相反，它刻画的是“非此非彼”的模糊概念。Atanassov[3] 在1986年给出了很完备的定义，并给出了系统性的运算和定理，之后直觉模糊集被证明了是一种的特殊的L-模糊集，这样就可以使用L-模糊集来研究直觉模糊逻辑的问题，和L-模糊集一样，隶属函数和非隶属函数仍然是直觉模糊集的理论基础，隶属函数和非隶属函数决定了直觉模糊集的研究和应用。关于隶属函数和非隶属函数的确定方法，刑清华和刘付显在广义基本概率分配（BPA）函数、广义信任函数和广义似然函数这三个改进定义的基础上建立直觉模糊集隶属度函数和非隶属函数，并且论证了的正确性[4]。对于两个直觉模糊集的所表示的模糊概念的接近程度的度量称为匹配度[5]，目前常用的计算匹配度的方法有语义距离、贴近度和相似度，通常常用的是用相似度来衡量，计算相似度的方法有最大最小法、算术平均最小、几何平均最小、相关系数法和指数法等。

（3）随后推动模糊概念的学习往前进一步发展的是原型理论的提出，原型理论和前面提到的隶属函数的确立不同，原型理论给出隶属某个概念的“参照物”，元素和这个原型之间的距离度量用来衡量这个元素在多大程度上属于这个概念，原型理论的中心趋势模型被概念化为表示一组样本的平均值。在机器学习领域，无论是聚类还是分类都能看到原型理论的影子。之后Lawry和Tang将原型理论和模糊集理论创造性地结合在了一起，通过这种方式，使得具有原型集合P的模糊概念L可以由定义在论域上的嵌套的随机集邻域来建模。用L描述论域中的元素x成立的前提是有邻域包含x的概率作为基础的。从这种方法导出的适当性度量的计算与标签语义具有一致性[14,15]。在此理论的基础上，随后他们又提出了认知结构的概念来描述模糊概念的最小的语义单元，也就是本文中的语义细胞。

语义细胞用来表示以“关于”、“类似”和“邻近”的模糊概念，其中是的原型集合。d是论域上的距离的度量（一般使用的欧氏空间中的距离），而则表示的是其他的点和原型之间的临近大小（粒度）的概率密度函数，这里的概率密度函数的确定会有两种情形，在没有先验知识的情况下，一般会假定一些已有的比较符合实际数据分布的函数，例如对数正态分布或密率分布。而在知道数据的分布的前提下，我们就可以直接利用这个带参数的函数，之后的工作就需要根据已有的数据对密度函数中的参数进行估计。这种建模方式在知识表示和机器学习中具有重要的意义。

## 本文的主要研究内容与贡献

本文的主要工作旨在研究从数据集中学习模糊语义细胞及混合模糊语义细胞，由于模糊语义细胞是建立在模糊集理论和原型理论基础上，所以学习的目标包含两个部分：原型P的学习和隶属度函数的学习。在学习的过程中需要考虑两个影响最终学习结果的指标：期望隶属度和模糊熵，我们首先计算三个数字特征：隶属度、期望粒度、模糊熵。期望隶属度反映的是整个语义细胞的平均半径，或者叫做邻近度，半径越大，涵盖论域中的元素也就越多，所表述的语义就越模糊；半径越小，包含的中的元素就越少，表述的语义就越具体。模糊熵借鉴了信息论中信息的不确定性度量，也就是熵，用来描述潜在的概念的不确定程度。这两个数字特征不仅提供了语义细胞的直观的表征，也构成了模糊语义细胞学习的两个重要的原则:

1.适当信息粒度[27]原则

2.最大模糊熵原则

适当信息粒度原则作为粒度计算的基本原理之一，是在实验数据的基础上形成一个信息颗粒，这个信息粒度能够很好地对实验数据覆盖并且使得信息有较大的确定性[7,37]，但是很明显这是相互冲突的，更高程度地对实验数据进行覆盖意味着要包含的数据越多，就必然会使得信息的含义变得不明确，因此在与语义细胞的学习过程中我们在遵循适当信息粒度的基础上，加入了一个粒度调节因子用来满足覆盖率和确定性。其次还需要遵循最大模糊熵原则，因为在学习的过程中还需要考虑模糊概念的边界不确定的因素，这使得模糊语义细胞的学习具有很好的鲁棒性。这样就将模糊语义细胞的学习问题转化为了非线性优化问题，优化中同时考虑覆盖率、信息确定性以及边界的模糊性，最后将这些因素归一化到目标函数中学习。

有了模糊语义细胞学习的支撑，我们就可以对基于多概念的混合模糊语义细胞进行学习，在模糊语义细胞的基础上，我们需要重新定义隶属度、期望粒度和模糊熵的计算。其次我们需要考虑的是每个模糊概念在整个概念集合中被描述的程度是不一样的，也就是构成概念集合的成分比重是不一样的。比如我们在日常生活中谈到：盛食物的器皿，那么我们往往想到的是碗、碟、钵等这些日常生活中常见的，而像那些诸如瓶子、杯子等虽然也可以用来盛食物，但可能用法上面就不如上面的器皿那样直接。所以我们考虑在学习的过程中为每个模糊语义细胞分配权重系数。这样概念的集合DB就可以通过形式化表示。最后优化的目标函数依然是和模糊语义细胞的优化目标类似，使得平均覆盖率尽量高、语义尽量明确、具有较好的鲁棒性。在模糊语义细胞中对于隶属度的学习包含两个部分：原型P和数据分布参数，我们使用拟牛顿法来学习，在这里也采用这种方式，关于最终的目标函数是一个多参数的非线性优化。我们使用遗传算法来找到一组最佳的权重向量，完成了对混合模糊语义细胞的学习过程。

混合模糊语义细胞的学习对具有多个模糊概念集合的表示具有较强的应用价值，尤其是在处理具有时间序列的概念演化方面具有很好的效果，例如针对某一新闻事件如美国大选，关于这一新闻事件的众多媒体报道、网页，它们所描述的侧重点各有不同（如政治、经济、军事、移民等），我们可以将之称之为概念，我们可以将这些文档使用DocumentToVector工具对其向量化，每个向量就是这个概念下的一个数据。同一时间段，新闻媒体会侧重某一个或几个概念对它们报道的比较多，相应的文档就比较多，而每个时间段都会有关于这些概念不同的报道，这样通过对比某一个时间段中各个概念的报道程度就可以分析出现在媒体舆论偏向，通过参观某个概念的时间演化的过程中报道程度（权重）的变化，就可以知道接下来这个概念是不是主流，利用这样的特征，我们就可以做诸如舆情分析和热点追踪等有价值的应用。

## 文章的组织结构

如何对模糊概念的语义进行建模是文章讨论的核心，本文先从模糊语义细胞的学习入手，然后在模糊语义细胞的基础上学习混合模糊语义细胞。并且分析了它们在表示具有多个模糊概念的集合上的研究，在一些标准的文本数据集上都取得了较好的表征。本文的主要集中在对在一些生成的数据集上完成语义细胞的学习和在文本数据集上的运用，一共5个章节，每个章节的安排如下：

第一章绪论部分主要介绍了目前关于模糊概念的研究背景和现状，并介绍了目前几种比较常见的隶属函数的确定方法并且指出了它们各自的局限性，之后引入了关于推动模糊集理论发展的直觉模糊集以及原型理论，然后阐述了本文的主旨聚焦在原型理论和模糊集理论结合的基础上，探索学习表示模糊概念的语义建模，分为模糊语义细胞的学习和混合模糊语义细胞的学习。最后论述了混合模糊语义细胞模型在实际中的运用，表明这种学习多模糊概念的模型具有一定的应用价值。

第二章节主要介绍了本文中的一些涉及模糊语义细胞的定义、推论以及关于模糊语义细胞的数字特征。我们首先给出了模糊语义细胞的一般定义，接着又给出了在模糊语义细胞关于隶属度的度量，分别给出了在语义细胞下和隶属函数的定义。紧接着给出语义细胞的两个数字特征：期望粒度和模糊熵，由于期望粒度和模糊熵需要借助数据粒度的概率密度，文中给出了一个常用的概率密度函数并论证了它的合理性。最后给出在给定概率密度函数下的期望粒度和模糊熵。

第三章给出了语义细胞学习的方法，首先给出了覆盖率的概念，随后给出了模糊语义细胞学习需要满足的三个约束：覆盖率尽量高、语义尽量明确（期望粒度较小）、最大化模糊熵。前两个约束可以看成同一个，即具有合适的隶属度，第三个约束是为了保证模型具有很好的鲁棒性。然后给出了最终所要优化的目标函数的一般形式，根据目标函数的矩阵的正定性，目标函数在定义域内一定存在全局最优解，我们对原型、分布参数利用拟牛顿法进行更新找到最优解，最后我们给出了优化的算法。

第四章在第三章的基础上引入了混合模糊语义细胞的概念，在论域中有多个概念的模糊集中，重新定义了隶属函数、期望粒度和模糊熵，这些数字特征都是基于多个模糊语义细胞的数字特征的加权形式所得到的，因此模糊语义细胞的学习和之前的相同，在混合模糊语义细胞中，我们还需要确定的就是一组权向量。最终需要优化的目标函数形式也和模糊语义细胞类似，最后我们使用GA遗传算法来确定权重向量。

第五章是实验部分，分别使用了两种不同的数据集，第一种是合成的数据集，用来模拟混合模糊语义细胞的学习，并在实验的基础上讨论了参数对实验结果的影响，并得出了一些重要的结论。第二种是TDT2语料库中数据量排名前30的语义种类的数据，在其中选取了若干的主题概念以及和这些概念相关的词汇作为要研究的对象，通过WordToVector工具将其向量化后作为我们讨论的数据集，然后运用以上方法学习到一组权重向量。

第六章对文章的主要工作作了总结，并且指出了其在研究具有时间序列的模糊概念时的优势，展望了其在诸如热点追踪和舆情分析方面的应用前景。

## 本章小结

这一章节提出了文章所要研究的问题的背景，国内外研究现状，根据现有的在模糊语义建模方面的问题，提出文章的研究内容，并在实验的基础上论证了在模糊语义的表征方面优势所在，并且对文章的工作中的前景做了展望。

# 模糊语义细胞

概念是心理学当中最基本的认知单元，并且在学习、推理等认知过程中蕴含了某些重要的信息[38,39,18]。对概念进行建模在认知心理学和人工智能方面具有很重要的价值。对概念的建模主要是如何表征蕴含在概念中的模糊语义，目前使用最为广泛的是通过语义算子来连接一些具有清晰语义的词，比如我们可以先预定义一些列原子词，由原子词加上一些连接词“或”、“且”、“非”；或者可以由原子词加上一些描述性的程度副词“很”、“略”、“极”、“较”，还可以使用一些原子词和一些表示模糊化的修饰词“大约”、“近似”，还可以加入一些具有判定意义的修饰词“偏向”、“倾向于”、“多半是”等。原子词可以使用多种修饰词来组合，形成的合成词还可以使用这些修饰词进行连接，形成更佳复杂的合成词。这些合成词的隶属函数的表示非常复杂，其计算和处理也需要较多的存储空间。

为了简化语义表达，扎德和其他研究人员使用语义算子的方法来解决这一问题。按照以上修饰词的类型分为以下四种类型：语气算子、模糊化算子、判定化算子、连接与否定算子，并给出了在各语气算子的隶属函数的定义。这是目前比较常用的模糊语义的表示。另外一种的简化是通过在原型理论的基础上引入了隶属度的概率密度函数建立一种认知结构。以下将首先介绍本文需要用到的一些概念和定义。

## 语义标签与模糊语义细胞

在论域上，称为一语义标签集合，对每一个语义标签，有集合，是属于概念的原型集合，被称为“关于”、“近似”，“接近”的语义表达。不失一般性，本文中的用中的一个元素表示。这种认知结构我们把它称为模糊语义细胞，用来表示模糊语义的概念。这种表示的认知结构有三个组成部分：的原型、定义在上的距离的度量，以及定义在其他元素与原型之间的距离的概率密度函数，以下本文将介绍模糊语义细胞的定义及相关概念。

### 模糊语义细胞

**定义1：（模糊语义细胞）**在论域上，模糊概念的模糊语义细胞被表示成三元组，其中是的原型，是定在上的用来测量与距离的度量，是的隶属度分布的概率密度函数。

模糊语义细胞非常直观地体现了语义的细胞结构，它包含两个部分：语义细胞核和语义细胞膜。模糊概念的原型代表语义细胞核，的模糊边界代表了语义细胞膜，这种模糊性是通过的隶属度分布的概率密度函数而确定的。关于距离的度量，这里采用使用比较广泛的欧氏距离，且对于，，。上面对模糊语义细胞做了定性的描述，以下将通过引入-邻域以及在其基础上导出的隶属函数的概念对模糊语义细胞做定量的分析。为了统一，以后我们将语义标签、概念、语义细胞统一使用符号表示。

**定义2：（-邻域）**对任意的模糊语义细胞且，-邻域被定义为：

(1)

定义了在以原型为圆心，为半径的球形邻域内的所有点的集合，是给定任意的粒度的的延伸。在这个概念中，的语义是明确的，的模糊性来自于它的边界是模糊的，这个模型也符合了人类的学习模式：即基于样例的归纳学习。这说明对于，与原型之间的距离是一个定义在随机变量，它的的概率密度函数是。根据以上的关于模糊语义细胞-隶属度的定义，的半径是由粒度表示的。是服从概率分布的随机变量。根据这些我们就可以计算定义在上对的积分来表示论域中的任意一个点在多大程度上隶属于这个模糊语义细胞。即这个的取值在上时，这样模糊语义细胞一定会包含这个点。我们把它称之为隶属度。下面我们将给出语义标签的隶属函数的定义。

**定义3:（隶属函数）**在论域上，模糊语义细胞的隶属函数被定义为：

(2)

由于，是定义在[0,+]的概率密度函数是关于的单调递减函数，且=1。所以从概率上来说，这个定义既解释了对于论域中任意的点在多大程度上属于这个模糊语义细胞，也可以解释了用这样的模糊语义细胞来刻画这个点的置信度有多少[17]。从模糊集角度来说，如果如果把看成是一个模糊集，那么就是这个模糊集的隶属函数。从原型理论角度来说反映的是和原型之间的相似性。图1给出了在二维空间下的模糊语义细胞和其粒度服从对数正态分布的概率密度函数和隶属函数的例子。后面实验部分会论证这种分布的合理性。



**图1.**（a）二维情形下的模糊语义细胞的，这里的退化为一个点。（b）模糊语义细胞粒度的概率密度函数。（c）模糊语义细胞的隶属函数

### 语义细胞的数字特征

前面讨论了语义细胞的概念及隶属度和隶属函数的定义，这一小节将对模糊语义细胞进行全面的量化分析。以下将会从模糊语义细胞的三个数字：原型、期望粒度和模糊熵三个方面进行介绍。根据模糊语义细胞的定义我们知道它分为两个部分：原型和模糊边界，原型指定了模糊语义细胞的典型的元素（集合），而模糊边界则引出了另外两个数字特征，期望粒度和模糊熵。

**定义4.（期望粒度）**模糊语义细胞的粒度服从概率密度函数为的分布，则

的期望粒度为：

(3)

期望粒度反应的是模糊语义细胞的半径（边界）的平均取值。这样就可以用一个确定的半径来度量具有模糊边界的语义细胞的邻近程度。

**定义5.（模糊熵）**模糊语义细胞的粒度服从概率密度函数为的分布，则

的模糊熵为：

(4)

注意：在，的取值有可能为负值，但具有单调递增、极值、凸的特点。模糊熵不仅能表示模糊语义粒度的模糊性，而且根据它的定义我们可以看出，它与整个论域是独立的，只依赖于数据的分布，这方便了以后的模糊语义细胞的学习。这里的是连续的情况，对于论域中的元素是有限离散的情况，在模糊集理论中有两种定义模糊熵的方法。第一种方法是由Zadeh[41]提出：对于一模糊集和概率分布，则

(5)

第二种是由Deluca和Termini[4]给出的，与上面的不同的是他们并没有使用概率分布，而是使用的香农的信息熵的形式：

(6)

其中是归一化常数，这种形式的熵是最常用的用来评估一个系统的模糊性，它满足Ebanks对模糊熵需要满足五个属性的定义：Sharpness、Maximality、Resolution、Symmetry和Valuation[6]。对于模糊语义细胞来说，这种方式显然不合适，因为模糊语义细胞的论域是一个无限连续的集合。模糊熵的引入保证了信息的不确定性，在这个基础上学习出的模糊语义细胞具有很好的鲁棒性。

## 粒度概率密度分布

关于如何假设模糊语义细胞粒度的分布，分为两种情况：如果在我们已经提前知道模糊语义细胞中的元素的分布模型，就直接使用；而在没有先验知识的情况下，我们就必须为这种概率密度假设出一个比较合理的模型，概率密度的设定不是一成不变的，也没有一个统一的标准。这里给出两种最常用的分布模型：幂律分布和对数正态分布，这两种分布在社会和自然、生物领域运用的比较广泛。

### 幂律分布

幂律分布[Barabasi 2011]被证明是自然界许多方面都遵循的一种分布，它具有以下的形式化表示方式：

(7)

这里是随机变量，是常量，这种分布是的递减函数，为了判断某种随机变量是否是服从幂律分布，通常是对(7)取对数后做如下的线形拟合：

(8)

然后取两边的残差

(9)

当(9)的残差小于某个给定的阈值，则可以认为近似服从幂律分布。

在没有先验知识的情况下，假定语义细胞的粒度服从幂律分布是合理的，并且由Bianconi[2008]和Zipf定理为这种合理性提出了解释[6]：（1）幂律分布反映了社会、技术特别是生物网络总是向着一种有序的方向趋近。（2）同时Zipf法则（幂律分布的另一种形式）在众多的语料库中也被发现。很多学者也尝试去解释这种现象出现的原因，Mandelbrot [1953]在理论上证明了在这种分布下，能够最小化单位信息所需要的代价，也就是说在模糊语义细胞中的元素的分布服从幂律分布的时候，人们理解这个概念的语义所要付出的代价最小。而我们所要研究的问题是概念的语义表示，离不开文本语料库的支持，所以这种假设元素的分布服从幂律分布是合理和必要的。图二给出了实际语料中关键词的权重排序曲线以及对它做幂律分布拟合之后的直线。



**图2.**实际文档语料中关键词的权重排序曲线及对其做幂律分布拟合之后的直线（语料来源：福岛核电站事故；2011年3月14日；1300个关键词）[5]。

值得注意的是，幂律分布在的时，在上才有确定的期望，在时，方差有上界，一般幂律分布带有指数，这样才能确定其期望。幂律函数多种形式，如果考虑的取值具有下限的情况，幂律分布还具有以下的形式[Wikipedia]：

(10)

以及加入了指数项形式的幂律分布[Wikipedia]：

(11)

由于幂律分布的期望的收敛性是依赖于随机变量的取值下限以及的取值决定的，会给计算带来不便，因此接下来介绍另外一种常见的分布：对数正态分布。

### 对数正态分布

本文中假设模糊语义细胞的粒度分布服从对数正态分布，并从UCI数据集的分布测试结果来说明假设的合理性。对数正态分布具有以下的形式：

(12)

对于分布未知的连续型随机变量，一般情况我们直接假设它服从高斯分布，这在现实的自然和社会科学中是很常见的分布，并且在UCI数据集上的测试的结果说明假设的合理性[1]。表一是对这些数据正态分布拟合和对数正态分布拟合的似然比较，图三是各数据集中将统计平均值作为原型，以数据和原型之间的欧氏距离作为的粒度分布直方图，从实验的结果来看，很容易看出对数正态分布的拟合效果要优于正态分布。

**表1.**对UCI各数据集的正态分布拟合和对数正态分布拟合的似然比较



**图3.**四个数据集上在确定原型和粒度的度量之后，粒度分布的直方图

确定了粒度的分布之后，模糊语义细胞具有以下两个数字特征的推论：

**推论1.** 设模糊语义细胞的粒度服从对数正态分布，其中分布参数和未知，为的期望粒度，则有：

(13)

证明：由（3）与（12）可得

**推论2.** 设模糊语义细胞的粒度服从对数正态分布，其中分布参数和未知，为的模糊熵，则有：

(14)

证明：由（4）与（12）可得

s

通过（13）与（14）发现，期望粒度和模糊熵最终都和粒度分布的参数相关，而与原型无关，但这三个参数的学习是同步的。以下将给出模糊语义细胞学习的两大原则：

## 模糊语义细胞的学习

前面提到了模糊语义细胞的学习分为原型的学习和隶属函数的学习，隶属函数部分主要是确定粒度分布中的未知参数，原型和分布参数的学习是同时进行的，并且需要在一定的规则上进行。以下就介绍模糊语义细胞在学习的过程中需要遵循的两大原则：合理粒度和最大化模糊熵。

### 合理粒度原则

如何评价一个模糊语义细胞的优劣，最直观的描述有两个方面：一是学习出的语义细胞要能最大限度地涵盖论域中的元素，也就是模糊语义细胞要有足够的覆盖率，这是一个模糊语义细胞最直观的表征，较高的覆盖率说明学习出的模糊语义细胞能够比较全面地反应这个概念集合；另外一方面当随着覆盖率的提升（在统一的原型维度上），相应的语义细胞的模糊边界（半径）也变得越来越远，这样导致语义细胞所要表达语义越来越模糊，变的没有意义。比如如何去描述全体浙江大学计算机学院在读学生的身份，我们从尽量涵盖全面且语义比较明确的角度可以说浙江大学的学生，而如果描述成杭州的大学生、浙江的大学生、大学生等这样概念就太泛了。因此这里合理的粒度不仅具有较高的覆盖率，还必须具有较明确的语义，这是两个冲突和协调的方面。Pedrycz在确定有意义的信息粒度方面做出了卓越的贡献，提出了合理粒度原则[27]。接下来将围绕合理粒度介绍覆盖率和语义清晰度。

**定义6.（覆盖率）**给定集合，和语义细胞（固定，也称为清晰语义细胞），则对的覆盖率为：

(15)

其中表示中元素落在邻域中个数，，，越多的元素落入这个邻域，覆盖率越高。考虑连续分布的模糊语义细胞时，对于覆盖是定义在概率的基础上，下面将给出模糊语义细胞覆盖率的定义：

**定义7.（模糊语义细胞的覆盖率）**给定集合和模糊语义细胞，则对的覆盖率为：

(16)

**定理1.** 给定集合和模糊语义细胞，则：

(17)

证明：根据（16）的定义，不失一般性，我们对中的元素和原型之间距离度量排序，对，，则：

因此（17）成立。从（17）可以看出，模糊语义细胞对集合的覆盖率等于中的每个元素的隶属度的平均值。与（15）不同的是，清晰语义细胞中中的元素的被覆盖的取值情况是二值{0,1}的，而模糊语义细胞使用（0，1）之间的值表示被覆盖的可能性。

以上讨论了合理粒度原则的第一个方面，合理粒度要求模糊语义细胞的边界不能太远，使得语义尽量明确。而期望粒度反映了模糊语义细胞边界的平均取值情况，因此使用期望粒度作为语义明确度的度量是合理的，并且期望粒度越小，说明针对性越强，语义越明确，期望粒度越大，针对性越弱，语义越含糊。

### 最大化模糊熵

最大熵模型指出，对任意一个随机事件的概率密度进行预测的时候，应当满足以下条件：（1）满足所有已知的全部条件，即要利用已知的条件，最大化减少不确定性。（2）对未知的条件不作任何假设，保证容错率。最大熵模型在很多领域都有运用[20,11,38,39,2,29]。在本文中，模糊语义细胞的粒度分布模型假设为对数正态分布模型，分布的参数未知，并且模糊熵定义在概率密度之上，因此最大化模糊熵模型满足模糊语义细胞的学习目标。

综上所述，学习模糊语义细胞的目标是要使得覆盖率尽可能较高、语义尽量明确、模糊熵尽量较大。根据（14）的结论我们发现，模糊熵越大，对应的分布的期望和方差越大，而正是期望粒度，所以模糊熵和期望粒度是相互冲突的因素；越大，则覆盖率越高，所以模糊熵和覆盖率是相互协同的因素；同样地，当覆盖率越大，语义越模糊，所以覆盖率和期望粒度相互冲突，而语义越模糊，则模糊熵越大，所以覆盖率和模糊熵相互协同；同理，期望粒度和覆盖率相互冲突，和模糊熵相互冲突。后面实验部分也会证明这一点。下面将根据以上的原则及结论给出具体学习策略。

## 模糊语义细胞的学习算法

前面论证了模糊语义细胞学习所需要遵循的三大因素：最大化覆盖率、语义最明确、最大化模糊熵，并且讨论了它们之间相互协同和制约的关系，基于这些，模糊语义细胞学习被表示成如下：

**定义8.（目标函数）**设有概念集合，则用来表示这个概念集合的模糊语义细胞可以通过最小化目标函数学习得到：

(18)其中为一调节因子，用来平衡和两项，因为,而值的变动范围会比前面一项大，因此调节因子是为了保证前后两项的大小处于同一个量级，而不至于出现前面一项被后面一项“吞噬”的情况。需要注意的是因为模糊语义细胞的原型的学习是隐含在中的，而隶属函数是由原型和粒度的分布决定的，根据本文假设的对数正态分布则要学习的参数有、、，结合（13）、（14）、（17）有：

(19)

以下证明在这三个参数各自的定义域内，目标函数有下界。

证明：

因为，只需证明有下界即可。

令，分别对、求导得：

根据梯度下降法，令，，得，。

而，带入、，得，故在定义域内是凹函数，且有全局最小值。因此目标函数有下界。由于所需要优化的是非线性无约束的优化问题，可以采用最常用的拟牛顿法或者是启发式的遗传算法来解决，由于这一部分的算法与混合语义细胞的学习算法类似，算法流程可以参考混合语义细胞的学习。

## 本章小结

本章开头介绍了目前比较流行的模糊语义表示方法：语义连接算子，并指出了它的局限行，随后提出了一种基于原型理论的模糊语义表示方法：模糊语义细胞。首先给出了模糊语义细胞的一般性定义，在模糊语义细胞中最为关键的三个部分：原型、距离度量、概率密度函数。然后在此基础上探讨了-邻域和隶属函数这两个基本概念。在模糊语义细胞中只和粒度分布相关的两个数字特征：期望粒度和模糊熵，随后我们讨论了两个典型的分布：幂律分布和对数正态分布，并研究了在对数正态分布下的两个数字特征的简化形式。最后我们给出了模糊语义细胞学习所要遵循的三大原则：覆盖率最大、语义明确、最大化模糊熵，并在此基础上提出了最终学习的目标函数，将问题转化为了非线性无约束优化问题。

# 混合模糊语义细胞

前面介绍模糊语义细胞的相关定义以及学习的原则和算法，本章将介绍在具有多个模糊概念的集合中，如何寻找一个统一的比较全面的刻画这些概念的模型。我们知道对概念的建模使用模糊语义细胞是很直观的，这是因为一个语义细胞正好能够涵盖这个概念的数据集。而对多个相关的概念集合用一个语义细胞去表示是很困难的（不能用一个粒度的分布去概括其他概念的粒度分布），而且最终的学习结果可能不敛性。这里我们我们借鉴高斯混合模型（GMM）的思想，使用多个独立的模型来刻画每个概念的粒度分布，然后根据每个概念在整个概念集合中的“贡献度”不同，引入权重系数，最后利用若干模糊语义细胞以及一组权重向量就表示了这个模糊概念集合。为了简单起见，这里我们将这个模糊概念集合定义为主题。

使用多个的模糊语义细胞来刻画同一个主题是必要的。例如媒体在报道美国大选这一主题时，某一时间段的新闻、网页、视频网站等媒体会从各个方面去报道这一事件，但都有各自的侧重的概念，比如有的聚焦于美国的政局，有的则针对经济，有的关注军事领域，有的可能更对移民更感兴趣等等。如果我们抛开概念仅仅使用“美国总统大选”来刻画这一主题，这样的话不仅仅让语义变的更为模糊，而且不能聚焦和跟踪我们关心的地方。

混合模糊语义细胞的学习是建立在模糊语义细胞学习的基础上，同样遵循着合理粒度原则和最大化模糊熵原则，学习的参数在模糊语义细胞的基础上加上了权重系数，即整个主题被描述成，。需要注意的是关于混合模糊语义细胞的学习是同时进行的，即多个模糊语义细胞参数（、、）的学习和权重的学习是在一个目标函数中进行的，并非先确定模糊语义细胞，后确定权重。另外在混合模糊语义细胞的学习中，三大数字特征也和模糊语义细胞有所不同，最终所要优化的目标函数与模糊语义细胞也有所区别：模糊语义细胞的优化是非约束的，而混合模糊语义细胞的学习是要满足，即最后需要求解的是约束优化的问题。在接来下的小节中将会给出，最后会给出详细的学习算法。

## 混合模糊语义细胞概念

**定义9.（混合模糊语义细胞）**在论域上，主题的混合模糊语义细胞被表示成，，其中是模糊概念的原型，是定在上的用来测量与距离的度量，是的隶属度分布的概率密度函数，是第个模糊语义细胞对主题的贡献度。

混合模糊语义细胞类似于高斯混合模型，每个模糊概念的粒度都遵循相同的分布，但是各自的分布参数不同。且在重构混合模糊语义细胞的时候使用了不同的权重。在反映主题方面，混合模糊语义细胞不如模糊语义细胞描述概念那么直观直观，但是它也具有和模糊语义细胞对应的数字特征。以下将先引入混合模糊语义细胞的-邻域以及在其基础上的隶属函数。

**定义10.（-邻域）**对任意的混合模糊语义细胞，且，-邻域被定义为：

(20)

-邻域给出了一种在论域上衡量哪些概念是属于某个给定粒度的主题的度量。由于使用模糊概念的析取范式来表示混合模糊语义细胞，即属于的元素也是一定属于对应主题的，所以使用来表示落在主题中的元素是合理的。由-邻域的定义，的半径是由粒度表示的。又因为是服从概率分布的随机变量。根据这些我们就可以计算定义在上对的积分来表示论域中的任意一个点在多大程度上隶属于这个混合模糊语义细胞。即这个的取值在上时，这样一定会包含这个点。我们也把它称之为隶属度。另外考虑到让隶属于这个主题是由多个贡献度不同的共同完成的，因此需要对进行加权形式之后的积分。下面我们将给出主题的隶属函数的定义。

**定义11.（隶属函数）**在论域上，混合模糊语义细胞的隶属函数被定义为：

(21)

距离的度量仍然使用欧氏空间中的距离，并且规定当且仅当时，，，即属于某个概念的数据恰好就是这个概念的原型时，它就一定是属于这个概念，而不属于其他的概念，或者说该数据和其他的概念没有关联。容易证明与具有类似的性质：

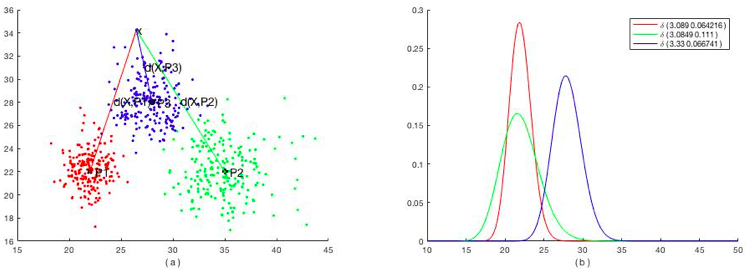
，，当且仅当，。

**证明：**

，等号成立的条件是当且仅当，此时，

。 原命题得证。

图4给出了在二维空间下的混合模糊语义细胞中各概念的粒度分布的情形。

**图4.**混合模糊语义细胞模型(a)三种分布参数不同的对数正态分布的随机点集(b)三种不同的对数正态分布的概率密度函数

## 混合模糊语义细胞的数字特征

同模糊语义细胞一样，混合模糊语义细胞同样具有期望粒度和模糊熵这两个数字特征，所不同的是在模糊语义细胞中，研究的对象是隶属于概念的元素集合，而在混合语义细胞中，研究的对象是隶属于主题的元素集合，这两个数字特征在迁移到主题上面时，其形式也有所不同，因为对于混合模糊语义细胞来说，它的“原型”是泛化之后的原型，相当于多个模糊语义细胞的原型集合加上一组权重因子，所以泛化之后的期望粒度可以看成是期望粒度的期望，泛化后的模糊熵是加权模糊熵，以下将给出混合模糊语义细胞的两个数字特征的定义：

**定义12.（期望粒度）**混合模糊语义细胞中的粒度服从概率密度函数为的分布，则的期望粒度为：

(22)

同样期望粒度是混合模糊语义细胞的半径（泛化边界）的平均取值。我们可以使用这样的数字特征来度量某个主题的内容边界的大小。当的粒度服从期望为方差为的对数正态分布时，有以下推论：

**推论3.** 混合模糊语义细胞中的粒度服从概率密度函数为的对数正态分布，则的期望粒度为：

(23)

证明参考推论（1）的证明。

**定义13.（模糊熵）**混合模糊语义细胞中的粒度服从概率密度函数为的分布，则的模糊熵为：

(24)模糊熵也称为差分熵，在，的取值有可能为负值[Wikipedia]，所以模糊熵并非是Shannon离散熵的延伸，需要相应的矫正方法来限制离散点的密度。但是在本文中，模糊熵作为数字特征的引入并非是要严格刻画模糊语义细胞（混合模糊语义细胞）粒度的模糊性，它在后面的学习中是起着约束的角色，并且具有单调递增、极值、凸的特点，可以用来简化计算。

**推论4.** 混合模糊语义细胞中的粒度服从概率密度函数为的对数正态分布，则的模糊熵为：

(23)

证明参考推论（2）的证明。

## 混合模糊语义细胞的学习

在混合模糊语义细胞的表示中，我们是通过结合多个概念的模糊语义细胞来表示不确定信息的一种新的形式，它的最直观的理论原型就是证据理论[40]（DS），实际上混合模糊语义细胞就是证据理论中识别框架的元素个数（模糊概念的种类数）被固定的一种特殊的形式，而每个模糊语义细胞的权重就相当于通过基本信任分配函数mass计算后的基本概率数[41]。但是由于在证据理论中多值变量的基本信任分配函数的框架的量级是2重指数运算（），计算起来相当复杂。因此不太适合用来处理模糊语义细胞的权重计算。基于此我们依旧采用和模糊语义细胞类似的方式去优化求解待定参数，以下讨论混合模糊语义细胞的学习的原则。

### 合理粒度原则

混合模糊语义细胞中，合理粒度要求主题能够尽可能地全面覆给定的集合，同时这种覆盖是具有针对性的覆盖，反映在量化数字特征上就是覆盖率较高、期望粒度较小。以下给出混合模糊语义下的覆盖率和期望粒度的概念：

**定义14.（覆盖率）**给定集合，和混合模糊语义细胞则对的覆盖率为：

(25)其中表示中元素落在邻域中个数，，，在连续情况下的混合模糊语义时，覆盖率被定义为：

**定义15.（混合模糊语义细胞的覆盖率）**给定集合和混合模糊语义细胞，则对的覆盖率为：

(26)

**定理1.** 给定集合和混合模糊语义细胞，则：

(27)

证明：根据（16）的定义，不失一般性，我们对中的元素和主题之间距离度量进行排序，对，，则：

,原命题得证。

观察（26）我们发现混合模糊语义细胞的覆盖率和模糊语义细胞的覆盖率具有相同的形式，进一步地，如果我们将（25）中的看成是模糊语义细胞泛化之后的粒度概率密度函数，则混合模糊语义细胞就可以被看成是一种特殊的模糊语义细胞，它的原型就是各概念的原型经过泛化之后的原型，也可以称之为伪原型（实际上不存在这样的实体原型）。合理粒度不仅要求具有较高的覆盖率，同时还需要有比较明确的语义，而根据前面的知识，我们使用泛化的期望粒度来衡量语义的明确程度，越小，说明这个主题要表达的意思就越具体，反之就越抽象。

### 最大化模糊熵

在混合模糊语义细胞中，最大化模糊熵和模糊语义细胞中的略有不同，它是建立在泛化的粒度概率密度的基础上。的形式比较复杂，因为泛化的概率密度可能不是对数正态分布的概率密度，所以可能不具有推论（2）的形式，但是作为学习的一个限制原则，在这里我们依然使用作为限制项。

模糊语义细胞的学习目标就是要使得主题对数据的覆盖率较高、主题更具有针对性，模糊熵较大。反映到数字特征上就是尽量大、较小、较大。同模糊语义细胞一样，以下将给出具体的学习算法和策略。

## 混合模糊语义细胞的学习算法

我们知道模糊语义细胞的学习是一个非线性的无约束优化，而在混合模型下时，我们需要对权重加以的限制，因此混合模糊语义细胞的学习是一个约束优化的问题，以下给出具体的目标函数的形式。

**定义16.（目标函数）**给定多概念集合和混合模糊语义细胞，可以通过最小化目标函数学习得到：

(28)

对于第一项很显然是有界的，接下来将证明同样有界。

证明：

令，分别为组成模糊语义细胞的多个概念原型向量、均值向量、方差向量、权重向量。由（23）、（24）可得

在模糊语义细胞的学习中我们知道是有下界的，故有下界。

综上所述，目标函数有下界，并且可以将目标函数转化为非线性的无约束优化问题，本文采用拟牛顿法求解约束优化问题，以下将给出在MATLAB中使用拟牛顿法（和BFGS算法）求解的算法流程：

**Algorithm** Mix fuzzy semantic cell learning

**Require:** Data set

**Ensure:** Mix fuzzy semantic cell

1：**function**

：the prototype vector of mix fuzzy semantic cell

：the parameters vector of probability density function

：the weight vector of fuzzy semantic cell which satisfied

2：**Begin**

3：**Initialize**：

4：**for** each data **do**

5： **Compute：**

6：**end for**

7：**Compute：**

8：**Compute：**

9：**return**

10：**end function**

11：

12：**function** BFGS()

13： **Begin**

14： **Initialize：**,, // B0 is a positive definite matrix and is the first-order optimality measure

15：

16： **Compute：**the gradient vector of functionat ,

17： **while () do**

18： **Compute：**search direction

19： **Update：** //is the step size which is obtained by line search

20： **Compute：**the gradient vector of function at ,

21： **Set：**,

22： **Update： //**

23：

24： **end while**

25： **return**

26：**end function**

## 本章小结

本章节首先引入了如何借鉴GMM的思想来表示多个模糊概念的集合,随后我们参照模糊语义细胞的学习思路给出了混合模型下语义细胞的一些基本的定义及数字特征。在混合模糊语义细胞中，-邻域和隶属函数的概念和模糊语义细胞中的略有差别，在混合模糊语义细胞中，中的元素到“原型”的距离被泛化成了这个元素到距离这个元素最近的某个概念的原型的距离，在这种度量方式下，我们得出了和模糊语义细胞一些类似的结论，最后我们利用模糊语义细胞的学习原则得出了最后需要优化的目标函数：

最后使用拟牛顿法求解，并给出了具体的算法流程。

# 混合模糊语义细胞的学习及应用实验

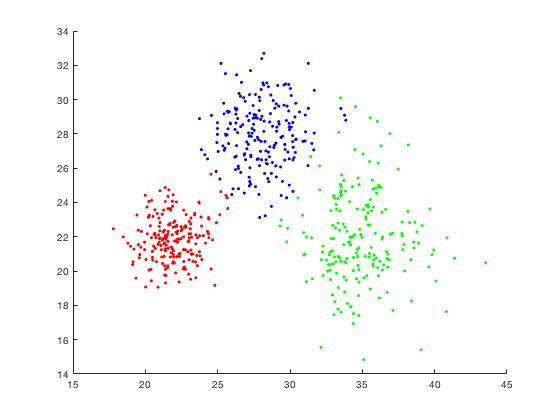
第三章在第二章的基础上详细介绍了混合模糊语义细胞具体的学习参数和数字特征，并给出了最后要优化的目标函数的形式及算法。本章将会在不同的数据集中去模拟混合模糊语义细胞的学习过程，并在学习的过程中对混合模糊语义细胞的各项参数及数字特征进行分析和讨论，我们主要关注在每次迭代的过程中目标函数值、原型、以及分布的参数、，权重向量这些与分布相关的参数随着的变化趋势；此外研究混合模糊语义细胞的三个数字特征：覆盖率、期望粒度、模糊熵随着的变化同样是必要的，通过它们的变化能够直观地看出混合模糊语义细胞学习的方向是越来越趋向于一种综合的最优水平。

模拟混合模糊语义细胞的数据集分为两类：第一类是人工合成的数据集，另一类是选自TDT2[7]中数据量排名前30的不同的语义类别的文档并且经过TF-IDF之后的数据集。以下将详细介绍这两种数据集。

## 实验数据集简介

### 合成数据集

在本实验中，第一类研究的数据集为二维空间中的点，我们按照给定的三个对数正态分布（分布参数分别为、、，、、，、、）随机生成了3种不同的数据集，每种数据集包含200个数据（见图5），学习的目标就是在这600个数据中学习到混合模糊语义细胞。



**图5.**在二维情形下，三种不同分布的数据集

### 真实文本语料数据集

第二类数据集是来自TDT2语料库[web]，TDT2是Nist主题检测和跟踪语料库，它收录了1998年上半年的6个不同媒体（两个新闻节目APW、NYT;两个音频节目VOA、PRI;两个电视节目CNN、ABC）所报道的11201个话题文档，这些文档按照内容的主题的语义分成了96个类别，这里选取了数据量排名前30的种类，一共包含了9394篇文档，得到这些文档之后，根据已经给定的向量将每篇文档变成向量，每个向量的长度都是1，每篇文档都被表示成了36771维的术语向量，所以最后的文档-术语的共生矩阵就是9394\*36771的矩阵，它是一个高度稀疏的矩阵，为了能够计算，我们选取了其中5个语义种类的1645篇文档，并且去除了和所选文档语义不相关的24218个术语，最后得到了1645\*12553大小的共生矩阵，这里它是高维空间中的点集，需要先进行降维之处理，我们使用PCA降维并根据“滚石法”选取了这个矩阵的前100个特征向量和共生矩阵做内积将共生矩阵转换为1645\*100的矩阵，最后将其进行中心化处理。最终学习的目标就是要在这一主题中寻找到一组最佳的权重、各语义种类的原型和分布参数来最佳刻画这一文档语料。

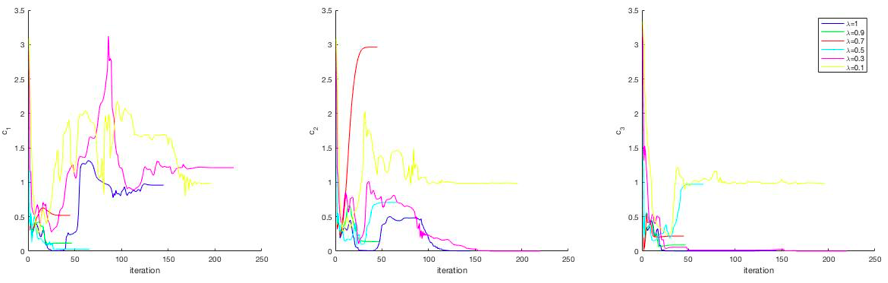
## 合成数据集实验及结果

在数据是二维情形下的合成数据集中，在学习之前我们需要设定混合模糊语义细胞的初始参数，这些参数分为两类：一类是是与数据集分布相关的参数（、、、），另一类是混合模糊语义细胞模型的调节参数（）。实验首先会给出若干组调节参数，对于每一组固定的调节参数带入模型最终求得最优的一组分布参数（、、、）和目标函数的最小值，同时会给出每一次的迭代的之后的分布参数和模糊语义细胞的三个数字特征、、的结果来表示学习的过程，最后我们对比在不同的调节参数下对最终实验结果的影响，这里我们选择、、、、、，并对结果进行分析。

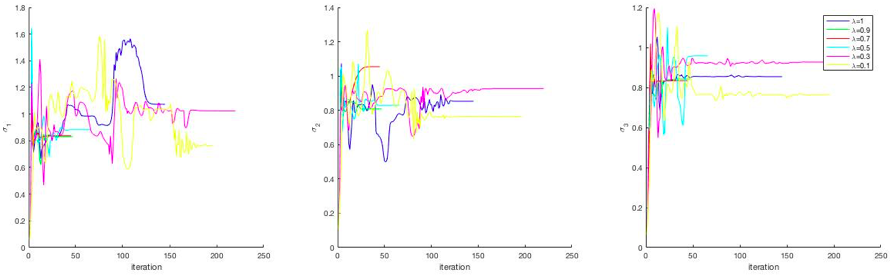
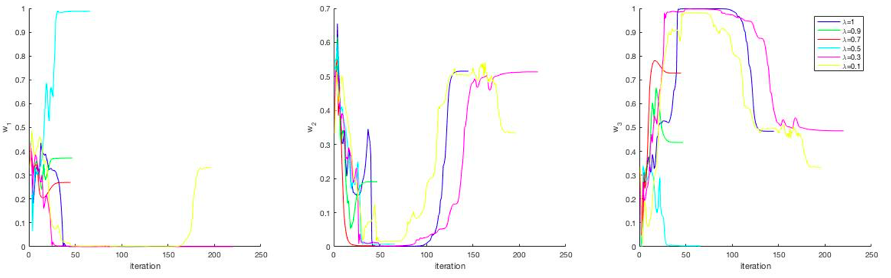
表2给出实验的初始参数，图6、图7、图8、图9给出了在在取不同值的情况下的，混合模糊语义细胞的分布参数和目标函数随每一次迭代学习之后的变化。表3给出了在在取不同值时三个概念集合的分布参数（、、、）以及目标函数的最终的学习结果。实验结果发现，随着取值的不同，除的取值最终都比较稳定外，和的取值都会有不同程度的变化，但是我们发现最终目标函数的取值都是很稳定的。

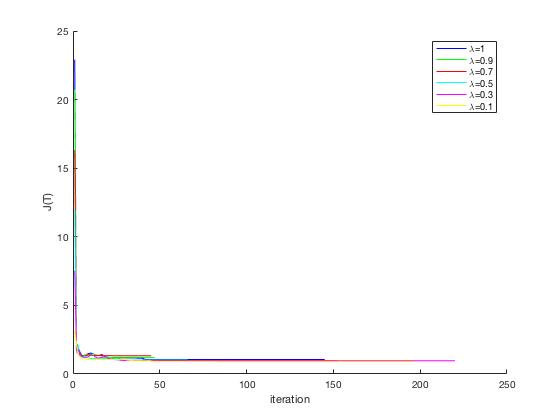
**表2.**混合模糊语义细胞学习的初始参数

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
| DataSet1 | (1,1) | 0.1 | 0.01 | 1/3 |
| DataSet2 | (2,2) | 0.1 | 0.01 | 1/3 |
| DataSet3 | (3,3) | 0.1 | 0.01 | 1/3 |

****

**图6.**取不同值时三种概念数据集的分布参数随迭代次数的变化

**图7.**取不同值时三种概念数据集的分布参数随迭代次数的变化**图8.**取不同值时三种概念数据集的权重参数随迭代次数的变化

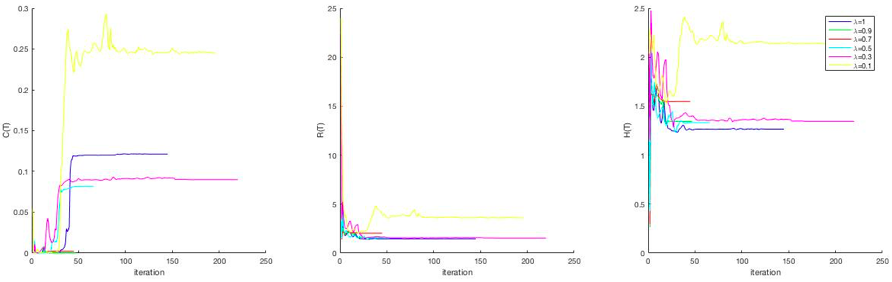


**图9.**取不同值时混合模糊语义细胞学习的目标函数值随迭代次数的变化

**表3.**不同取值下从概念集中学习的混合模糊语义细胞分布参数和目标函数值

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  |  |  |  |  |
|  | (33.2746,33.2746)  (21.7140,21.8596)  (21.7430,21.8406) | 0.9589  0.0048  0.0051 | 1.0741  0.8534  0.8539 | 0.0002  0.5155  0.4834 | 1.0603 |
|  | (61.8560,61.8555)  (42.3039,42.3031)  (32.0622,32.0598) | 0.1166  0.1379  0.0896 | 0.8316  0.8082  0.8422 | 0.3724  0.1903  0.4373 | 1.2092 |
|  | (15.0618,15.0617)  (11.2655,11.2654)  (9.6914,9.6911) | 0.5208  2.9636  0.2195 | 0.8378  1.0544  0.8339 | 0.2695  0.0028  0.7277 | 1.3295 |
|  | (27.5959,27.9051)  (19.8658,19,8659)  (15.8668,15.8665) | 0.0270  0.7075  0.9721 | 0.8844  0.8297  0.9583 | 0.9885  0.0082  0.0032 | 1.0217 |
|  | (37.6680,37.6726)  (27.5474,27.9635)  (27.5474,27.9635) | 1.2133  0.0013  0.0013 | 1.0243  0.9271  0.9271 | 0.00002  0.5136  0.4864 | 0.9684 |
|  | (27.8548,27.3724)  (27.8548,27.3725)  (27.8548,27.3725) | 0.9856  0.9856  0.9856 | 0.7651  0.7651  0.7651 | 0.3333  0.3338  0.3330 | 0.9004 |

图10展示了取不同值时每次迭代学习的过程中，混合模糊语义细胞的三个数字特征、、的变化趋势，表4给出了不同情况下的最终的数字特征。

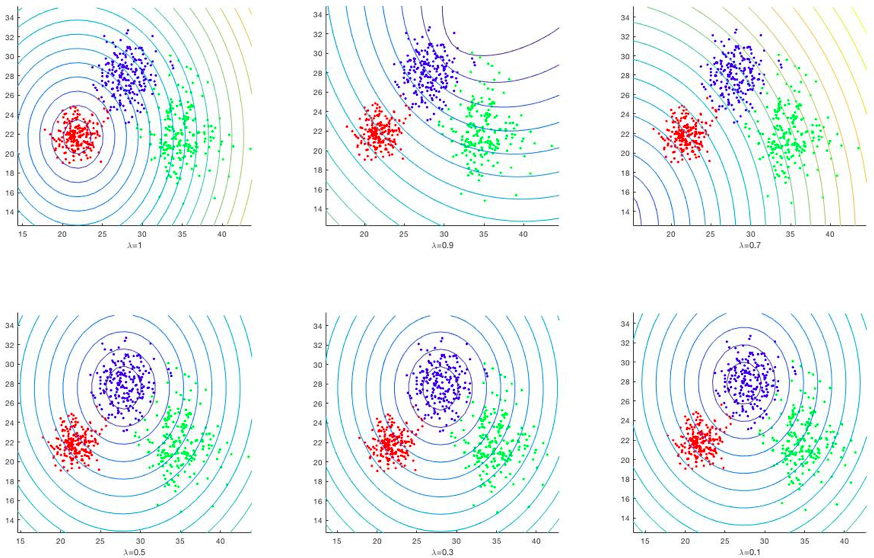


**图10.**取不同值时混合模糊语义细胞三个数字特征、、随迭代次数的变化

**表4.**不同取值下学习到的混合模糊语义细胞的三大数字特征、、

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  | 0.1212 | 1.4474 | 1.2659 |
|  | 0.00007 | 1.5760 | 1.3435 |
|  | 0.0023 | 2.0215 | 1.5476 |
|  | 0.0819 | 1.5387 | 1.3314 |
|  | 0.0899 | 1.5390 | 1.3446 |
|  | 0.2449 | 3.5906 | 2.1368 |

实验结果发现，当增大时，覆盖率、期望粒度、模糊熵都会减小，而且发现覆盖率、期望粒度、模糊熵的变动都相对比较平滑，这也从侧面说明了模糊熵的引入使得模型不容易受的影响，另外还发现除某些特殊的取值，这三个数字特征的取值都比较集中，说明模型具有很好的鲁棒性。图11给出在取不同的值的时候，混合模糊语义细胞泛化“原型”到平面各点的加权距离的等高线。可以发现当、、时，泛化“原型”（等高线簇的中心）变动比较大，而、、时泛化的“原型”靠近第二概念集合的中心且趋于稳定。



**图11.**取不同值时混合模糊语义细胞的泛化原型到平面各点的加权距离的等高线

综上所述，的取值对混合模糊语义细胞的学习有很大的影响，我们把称之为粒度控制因子，也就是说，当随着粒度控制因子增加时，主题的潜在的概念变的越来越清晰，并且潜在模糊性也越来越小，所以也可以成为确定性控制因子或者称之为模糊控制因子。综合来看当、、时学习得到的混合模糊语义细胞的泛化“原型”趋于稳定，、、时混合模糊语义细胞的三个数字特征取之都比较集中，所以是一个比较好的折中选择，后面的实验部分还会继续讨论关于对混合模糊语义细胞学习的影响。

## TDT2数据集实验及结果

前面介绍了在TDT2降维处理之后的1645\*100文档向量数据集中，根据语义它们被分为5大种类，每个种类的文档数量分别为160、1221、73、122、65，为了能够学习到表征这五个语义集合的混合模糊语义细胞，考虑到学习的“原型”维度较高，泛化的“原型”由五个模糊语义细胞的原型（每个100维）组成，这样光原型学习就有500个参数，计算非常耗时并且复杂，所以这里为了简化计算过程我们将混合模糊语义细胞的学习分为两步：

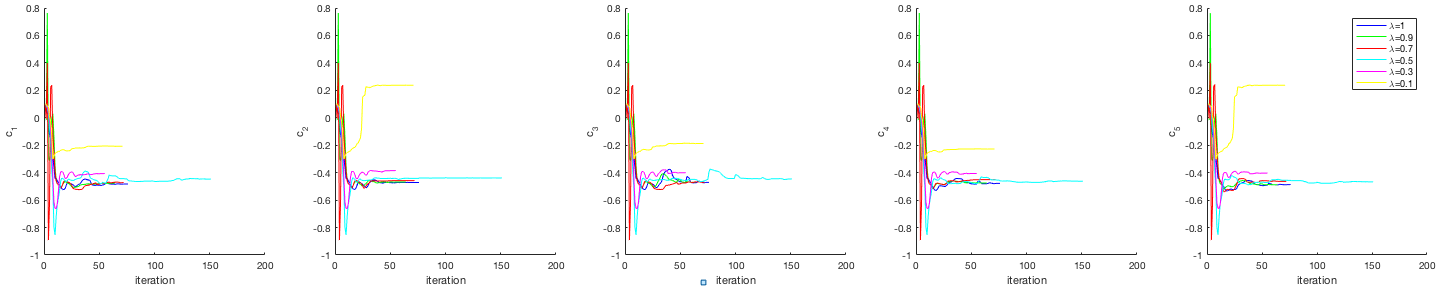
1. 各模糊语义细胞原型的学习
2. 混合模糊语义细胞的其他分布参数（、、）的学习

原型的学习包含在模糊语义细胞的学习之中，关于模糊语义细胞的学习第二章的2.4节给出了具体的目标函数，具体的算法流程和混合情况下的算法类似，唯一不同的是它是无约束的非线性优化，需要使用MATLAB中的非线性无约束优化工具来求解；其他的15个分布参数按照非线性约束优化来求解，同样使用MATLAB工具箱来求解。在之后的讨论部分我们会说明这样分两步学习的结果和直接的混合模糊语义细胞的学习是近似的。

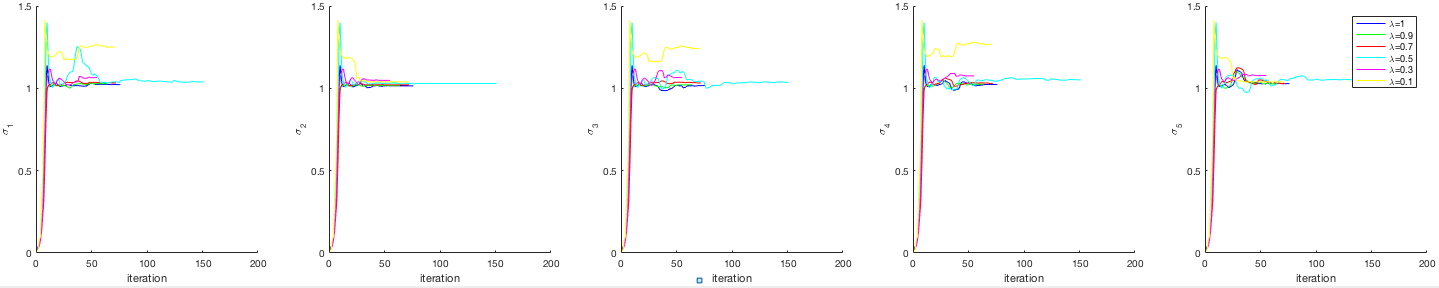
表5给出了文档向量的混合模糊语义细胞学习的分布参数的初始值,其中初始的原型都是100维（每一位都是1）的向量。图11、图12、图13、图14展示在粒度控制因子在取不同值时分布参数和目标函数的变化。附录中表6给出了在在取不同值时三个概念集合的分布参数（、、）以及目标函数的最终的学习结果。实验结果发现除时偏离比较大之外，三种分布参数收敛的比较集中和稳定，说明模型具有很好的鲁棒性。

**表5.**文档向量混合模糊语义细胞学习的初始参数

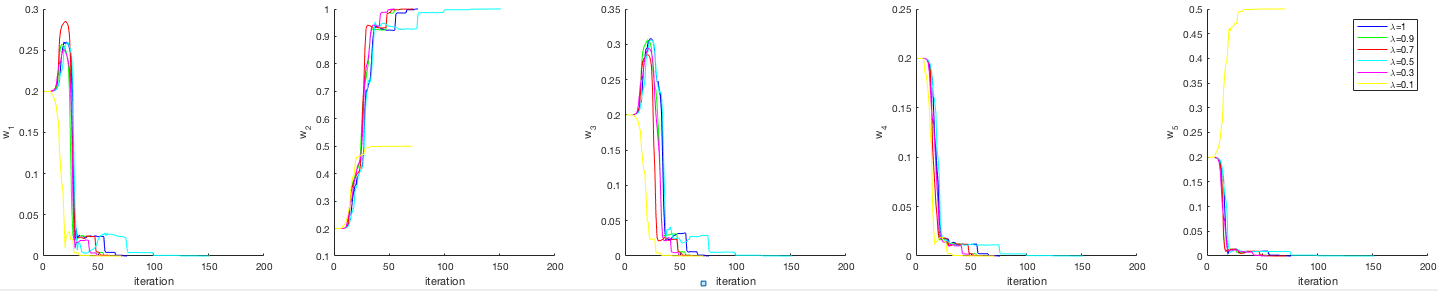
|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
| DocumentSet1 | (1,…,1) | 0.1 | 0.01 | 1/5 |
| DocumentSet2 | (1,…,1) | 0.1 | 0.01 | 1/5 |
| DocumentSet3 | (1,…,1) | 0.1 | 0.01 | 1/5 |
| DocumentSet4 | (1,…,1) | 0.1 | 0.01 | 1/5 |
| DocumentSet5 | (1,…,1) | 0.1 | 0.01 | 1/5 |



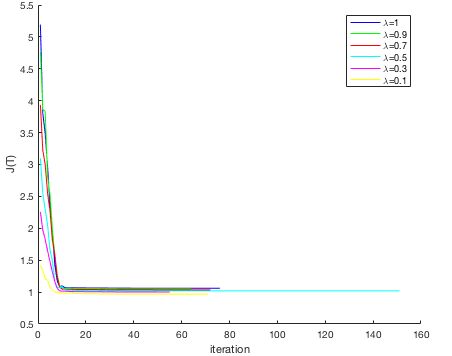
**图11.**取不同值时五种语义的文档向量分布参数随迭代次数的变化



**图12.**取不同值时五种语义的文档向量分布参数随迭代次数的变化

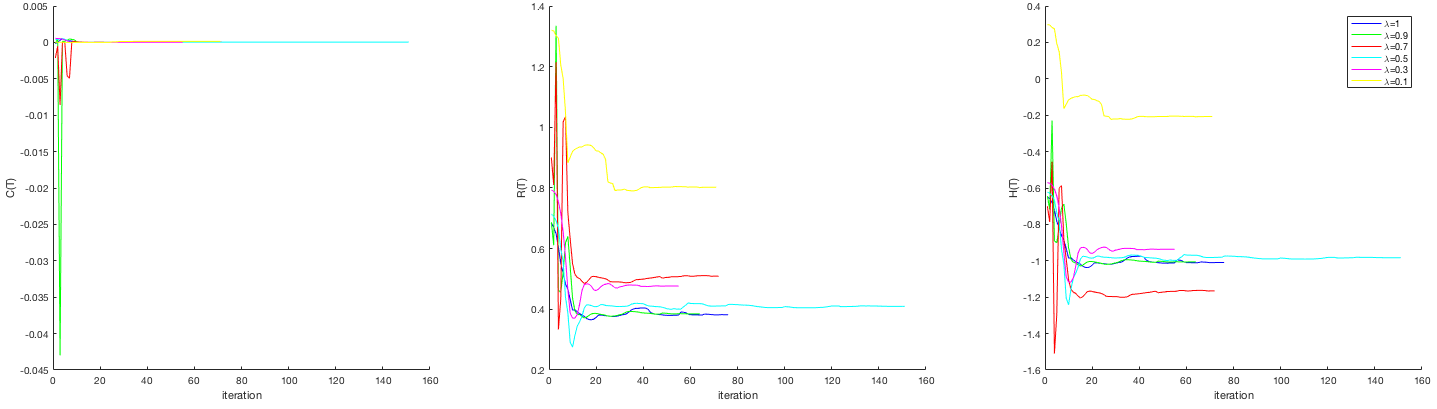


**图13.**取不同值时五种语义的文档向量权重参数随迭代次数的变化

****

**图14.**取不同值时文档向量的混合模糊语义细胞学习的目标函数值随迭代次数的变化

通过表6的结果不难发现，在学习到的5个权重中，第二类语义文档集合所占的权重几乎为1（时第二类和第五类权重相等），这也可以从它们各自的文档向量可以看出来，在五类数据集中，第二类所占的数据量最多且更加集中。图14给出了在取不同值时文档向量的混合模糊语义细胞的三个数字特征、、在迭代学习的过程中变化趋势。表7给出了取不同值时的数字特征的最终的结果。

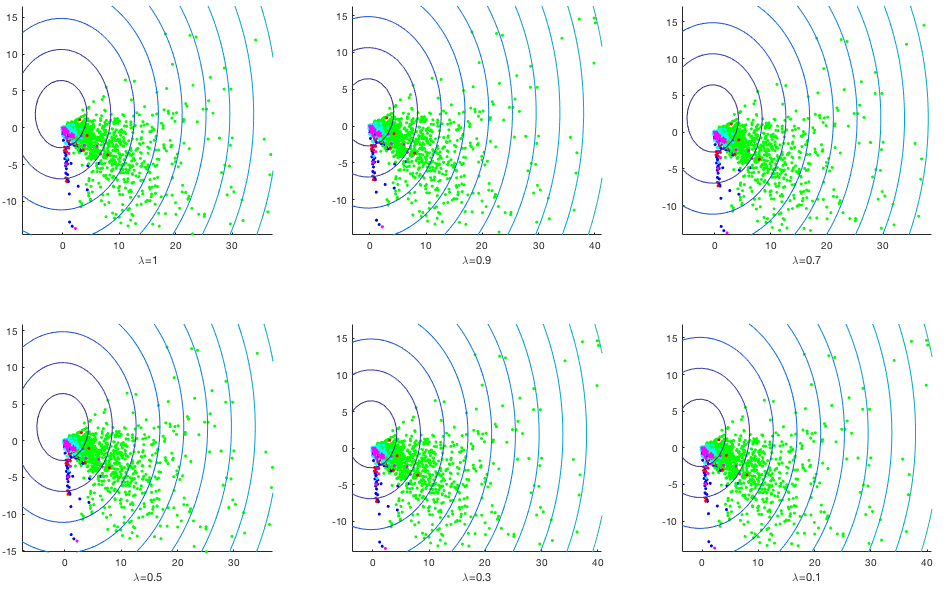


**图14.**取不同值文档向量三个数字特征、、随迭代次数的变化

**表4.**不同取值下学习到的混合模糊语义细胞的三大数字特征、、

|  |  |  |  |
| --- | --- | --- | --- |
|  |  |  |  |
|  | 0.0439 | 0.3823 | -1.0092 |
|  | 0.0000 | 0.3852 | -1.0062 |
|  | 0.0001 | 0.5092 | -1.1655 |
|  | 0.0417 | 0.4098 | -0.9833 |
|  | 0.0221 | 0.4769 | -0.9364 |
|  | 0.1404 | 0.8025 | -0.2074 |

实验结果表明，当减小时，、、不是严格意义上的单调递增，但总体上呈增大的趋势。其中随的变化比较集中，、除特殊的外都比较集中。图15给出了取不同的值的时候，文档向量的泛化“原型”到平面各点的加权距离的等高线（文档向量已经经过中心化处理并且被投影到二维平面）。可以发现无论当取何值时，文档向量最后的泛化“原型”都稳定在同一个位置附近，这也是很容易理解的，因为这五种语义的文档向量彼此融合在一起，并不像合成数据集那样分散，它们大部分数据都是重叠的，因此最后学习出来的泛化“原型”既是某一个语义种类的原型，也是混合模糊语义细胞的原型，因此这个原型几乎是不变，和本身没有联系。之后的分析讨论部分将对这一现象作进一步解释。



**图11.**取不同值时文档向量泛化原型到平面各点的加权距离的等高线

## 实验结果分析与讨论

前面的实验部分分别展示了两种通常情况下的数据分布的情况，一种是每种概念的数据集分布相对比较分散且数据量相对一致，另外一种是每种数据分布相对比较集中并且数据量是偏斜的。

通过以上合成数据集和真实的文档-向量数据集的实验结果可以发现，首先粒度控制因子的选择对实验的结果影响很大，因此选取合适的对学习最佳的混合模糊语义细胞有重要的意义，通过合成数据集的实验结果（图11）发现，当、、的时候学习到的泛化的“原型”变动比较大，且、时覆盖率相对较低，而当、、的时候泛化的“原型”比较集中并且较稳定；而在数字特征(图10)中，除时数字特征偏离其他的取值时较大之外，其他情况下的取值都比较集中和稳定，而且当时，通过表3最终学习到的权重分布来看，三个概念集合所占的权重一样，不仅如此，学习到的、、也是一致的，即最后学习到的三个原型最后退化为同一个具体的原型。进一步分析我们会发现，这个退化的具体的“原型”非常接近生成模拟数据集的时候设定的分布参数中第三个数据集的原型，同时伴随着原型的收敛，权重越来越趋于相等，期望粒度、模糊熵 变大，即最后的主题涉及的范围变大，语义变的越来越模糊。综上，当取时学习到混合模糊语义细胞最佳。而观察文档-向量数据集的实验结果发现，图11和图12都表现出与合成数据集相似的性质，当的时候、以及数字特征偏离其他的取值时较大，并且通过表4可以发现取其他值时权重集中在第二个数据集，而这个情况从数据本身的分布角度来说也是合理的,通过图11我们发现，由于第二类语义的文档的数量较多，并且分布比较集中，造成了对其他数据集的很大程度上的覆盖，因此可以认为整个文档向量数据集的主题是集中在描述第二种语义文档的概念。而当是权重开始慢慢向其他的语义文档集合开始倾斜，学习的准确性有所下降。同时根据表4的结果我们还发现当、时，混合模糊语义细胞的覆盖率相对而言比较低，因此综合来看，的取值在之间时，混合模糊语义细胞学习的效果最佳。

实验的第二关键的地方在于模型中模糊熵的引入，

**表6.**不同取值下文档向量的混合模糊语义细胞分布参数和目标函数值

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  |  |  |  |  |
|  | -0.4829  -0.4716  -0.4710  -0.4783  -0.4874 | 1.0229 1.0147 1.0162 1.0229 1.0284 | 0.0000  1.0000  0.0000  0.0000  0.0000 | 1.063 |
|  | -0.4775  -0.4682  -0.4699  -0.4750  -0.4879 | 1.0286 1.0164 1.0211 1.0245 1.0274 | 0.0000  0.9999  0.0000  0.0000  0.0000 | 1.0521 |
|  | -0.4729  -0.4580  -0.4707  -0.4511  -0.4636 | 1.0324  1.0211  1.0323  1.0299  1.0301 | 0.0000  0.9999  0.0000  0.0000  0.0000 | 1.0355 |
|  | -0.4472  -0.4384  -0.4461  -0.4627  -0.4669 | 1.0389 1.0295 1.0361 1.0516 1.0536 | 0.0000  0.9999  0.0000  0.0000  0.0000 | 1.0186 |
|  | -0.4074  -0.3850  -0.4005  -0.4058  -0.4031 | 1.0648 1.0482 1.0657 1.0758 1.0776 | 0.0000  1.0000  0.0000  0.0000  0.0000 | 1.0006 |
|  | -0.2073 0.2371  -0.1878  -0.2272 0.2371 | -0.2073 0.2371  -0.1878  -0.2272 0.2371 | 0.0000  0.5000  0.0000  0.0000  0.5000 | 0.9690 |

参考文献

[1] D.Spinellis, K.Raptis. Component mining: a process and its pattern language[J]. Information and Software Technology, 2000(42):609~617

[2] 杨芙清, 梅宏, 李克勤. 软件复用与软件构件技术. 电子学报, 1999, 27(2):68~75

[3] 杨芙清. 软件复用及相关技术. 计算机科学，1999, 26(5):1~4

[4] Nenad Medvidovic, Richard Taylor. A Classification and Comparison Framework for Software Architecture Description Languages[J] . IEEE Transactions on Software Engineering, 2000, 25(1):70~93

[5]窦郁宏, 陈松乔. 程序挖掘中需求描述的研究. 计算机工程与应用, 2002, 10:53~56

[6] 杨瑞林, 李力军. 新型低合金高强韧性耐磨钢的研究. 钢铁， 1999（7）：41~45

[7] 贾名字. 工程硕士论文撰写规范. 硕士学位论文,上海交通大学,2000

[8] 胡海洋, 杨玫. Cogent后组装技术研究与实现. 电子学报, Dec 2002, 30(12):1823~1827

[9]任洪敏, 钱乐秋. 构件组装及其形式化推导研究. 软件学报, 2003, 14(6):1066~1074

注意：

参考文献的排列按照学位论文中所引用的文献顺序排列，论文中参考文献引用需用上标。

文献数量合理，不太少也不滥用，文后列出的参考文献在正文中必须有对应的引用。

文献来源正宗权威，是学术文献，出典可查。

攻读硕士学位期间主要的研究成果

[1] D.Spinellis, K.Raptis. Component mining: a process and its pattern language[J]. Information and Software Technology, 2000(42):609~617

[2] Schinstock, D.E., Cuttino, J.F. Real time kinematic solutions of a non-contacting, three dimensional metrology frame[J]. Precision Engineering, 2000, 24(1):70-76

致谢

署名

当前日期