基于三元概念分析的文本分类算法研究

李 贞 张 卓 王黎明
(郑州大学信息工程学院 郑州 450001)

摘 要 随着网络中三维数据的涌现,三元概念分析的优势也逐渐体现出来。三元概念分析是较新的研究领域,具有广阔的发展前景。提出基于三元概念分析的文本分类方法,该方法是一种全新的构思理念,是三元概念分析在应用上的拓展。该算法的主要思路是:首先将数据集预处理为三元背景,同时将背景中的二值关系扩展为0-1间的模糊关系,其用于表示特定条件下属性对于对象的隶属度,并基于此构建三元概念,利用三元概念表示数据集中文本、特征词与类别之间的三元关系;然后结合模糊理论中的贴近度,类比得出三元概念间的相似度,并运用相似性度量计算出训练集中三元概念与新文本的相似值。实验结果表明,文中所提模型是有效的,且在特定的数据集上相较于机器学习Support Vector Machine(SVM)算法、K-Nearest Neighbor(KNN)算法、卷积神经网络(CNN)算法以及基于形式概念分析的分类模型均有更好的分类效果。

关键词 三元概念分析,三元概念,模糊理论,文本分类,三元概念相似度

中图法分类号 TP181 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2017.08.036

Research on Text Classification Algorithm Based on Triadic Concept Analysis

LI Zhen ZHANG Zhuo WANG Li-ming
(School of Information Engineering, Zhengzhou University, Zhengzhou 450001, China)

Abstract With the emergence of three-dimensional data in the network, the advantages of triadic concept analysis (TCA) have been reflected gradually. As a relatively new field, TCA has a bright prospect. This paper proposed a text classification algorithm based on TCA, which is a novel idea and a development of TCA in application aspect. The main idea of this algorithm is firstly preprocessing the dataset so that we can convert it into triadic context, meanwhile extend the binary relation in the context to a fuzzy value between 0-1 which represents membership degree about attribute for object under certain conditions. Based on this, we can build triadic concepts and utilize it to express the ternary relation among text, term and category. Then, combined with the approach degree in fuzzy theory, we can analogize the similarity formula of triadic concepts, accordingly calculate the training set's similar value about triadic concept for a new text. Compared to support vector machine(SVM), K-nearest neighbor (KNN), convolution neural network (CNN) algorithm and classification based on formal concept analysis model, the results indicate that the proposed model in specific dataset is effective and achieves a better performance.

Keywords Triadic concept analysis, Triadic concept, Fuzzy theory, Text classification, Triadic concept similarity

1 引言

随着信息技术的发展,互联网数据及资源呈现海量特征。处理三维数据或基于条件的二维数据时,由形式概念分析扩展的三元概念分析理论可提供有效途径[1]。Wille R[2]于1982年提出了形式概念分析(Formal Concept Analysis, FCA),并将其用于概念的发现、排序和显示。它是一种基于格的分类方法,是能从一组概念层次结构中提取有象征性数据(二元数据集)的数据挖掘技术。受Peirce C的启发,Wille R等人于1995年将FCA(形式概念分析)扩展到三元情况,提出了三元概念分析(Triadic Concept Analysis, TCA)[1]。传统的基于形式概念的分析主要探讨如何从给定的对象与属性之间的二元关系中获取二元概念及其层次结构。但如今很多时候我们对问题的研究不仅仅在二元关系中,对数据所属的"条件"也更加重视,对象与属性之间的关系多数情况也是在一定的前提条件下成立的[3-4]。同时,属性包含一般属性和结构属性,在传统二元形式概念中则是一起处理这两种属性,但有时需要单独处理结构属性,这时三元概念分析就体现出了其独特的优势。三元概念对一般属性和结构属性分别处理,这样在解决实际问题时就可以缩小信息搜索范围[3-5],提高效率。

到稿日期:2016-07-25 返修日期:2016-10-25 本文受国家青年科学基金项目(61303044)资助。

李 贞(1991-),女,硕士生,主要研究方向为三元概念分析及应用、机器学习、数据挖掘,E-mail:bright_zhenli@163.com;张 卓(1978-),男,博士,副教授,CCF会员,主要研究方向为形式概念分析及应用等,E-mail:iezhangzhuo@zzu.edu.cn(通信作者);王黎明(1963-),男,博士,教授,CCF高级会员,主要研究方向为现代软件工程技术、分布式人工智能和数据挖掘等。

三元概念分析有着极其广泛的应用前景，在数据量庞大的今天，它势必将为我们提供一个高效且实用的数据处理方法[5]。

目前，对基于形式概念分析的机器学习方法的研究成果较丰富，提出的模型比较多样，而且最终的分类和聚类结果也高效且准确。Carpineto C[6]等根据 SVM 中核函数的性质，将文本与特征词概念化成概念格中的外延与内涵，通过概念格的哈斯图计算内涵间的距离而得出相似值，进而提出了 CL-SVM 分类模型，最终对文本进行了合适的单分类处理。Belohlavek 等[7]提出了一种基于形式概念分析的新决策树归纳方法，其主要思想是用概念格即概念的层次结构推导出决策树，进而对测试数据集分类。Kang X P 等[8]提出基于概念格的多示例集成学习模型，将训练集进行处理后构成概念格，将多示例问题转变成基于概念格的多个局部多示例学习问题，通过多个局部目标特征集对训练集进行高效且准确的分类。Li S T 和 Tsai F C[9-10]通过将模糊理论与形式概念分析结合，提出基于形式概念分析的 FFCM 分类模型，并根据数据集测试得出较高准确率的分类效果。文献[11-13]总结了近年来运用形式概念分析的知识提取和机器学习算法在应用研究中所取得的成果。

三元概念分析的研究还处于起步阶段，其主要应用包括概念三元格的构造、三元关联规则挖掘、三元背景的模糊化及因子分析等，主要的成果集中在国外[5]。如 BiederMann K 根据三元概念的提出，主要研究了概念三元格的构造。Ganter B 等于 2004 年探讨了三元蕴含及关联规则挖掘。文献[14]研究出了一种新的三元二进制数据的因子分析方法，描述了对象、属性与条件三者之间的关系，提出了计算最优分解的贪心算法并评估了其性能。文献[15]研究了三元概念聚类方法，并根据实验结果的比较从 5 组三元聚类算法中展示出给定数据集“最佳模式”的三元数据形式。

尽管基于三元概念分析的研究并不完善，但是随着网络中三维数据的日益增多，其已经成为人工智能领域的研究热点之一[5]。三元概念分析是处理分析数据的新框架，本文借鉴基于形式概念分析的机器学习算法、思路以及三元概念分析本身的结构优势，将三元概念引入文本挖掘领域中，结合已有的模糊理论研究与相似度函数，建立共有的特征词空间，进而构建出基于三元概念分析的整个学习机制，并将全新的模型应用于文本挖掘领域中。该算法是对三元概念分析在应用方面的初步探索，可发展空间大，为大数据的研究奠定了基础；同时该模型的提出对于拓展机器学习的分类方法也具有深远的意义。

2 相关理论

2.1 三元概念分析的理论描述

三元概念分析是形式概念分析在三元背景上的拓展，适合处理三维数据。因此，研究三元概念分析既能进一步丰富形式概念分析理论，又能为复杂数据的知识获取提供新的理论和方法。下面给出三元概念分析中的基本定义。

定义 1[1]（三元背景） 称(K_1,K_2,K_3,Y)为一个三元背景，其中$K_1=\{g_1,\cdots,g_p\}$为对象集，每个$g_i(i\leqslant p)$称为一个对象；$K_2=\{m_1,\cdots,m_q\}$为属性集，每个$m_j(j\leqslant q)$称为一个属性；$K_3=\{b_1,\cdots,b_r\}$为条件集，每个$b_k(k\leqslant r)$称为一个条件；Y为K_1,K_2,K_3之间的三元关系，其中$Y\subseteq K_1\times K_2\times K_3$。若$(g,m,b)\in Y$，则表示对象$g$在条件$b$下具有属性$m$。

定义 2[1]（三元概念） 定义三元背景(G,M,B,Y)的三元概念是三元组(K_1,K_2,K_3)，满足$K_1\subseteq G,K_2\subseteq M,K_3\subseteq B$，同时对于每一个$\{i,j,k\}=\{1,2,3\}$且$j<k$，有$(K_j\times K_k)^{(i)}=K_i$。其中，称$K_1$为三元概念中的外延，$K_2$为内涵，$K_3$为条件。

表 1 是一个三元背景的交叉表，构成的所有三元概念可以由集合的形式表示：$\zeta(\mathcal{R})=\{(\varnothing,123456,abcd),(C,126,abcd),(C,123456,ab),(B,15,cd),(A,124,d),(AC,12,abcd),(AC,126,c),(BC,6,b),(ABC,123456,\varnothing),(ABC,1,acd),(ABC,\varnothing,abcd)\}$。

表 1 三元背景(K_1,K_2,K_3,Y)

	a						b						c						d					
	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6	1	2	3	4	5	6
A	×	×						×	×					×	×					×	×	×		
B	×										×	×				×		×		×			×	
C	×	×	×	×	×	×		×	×	×	×	×		×	×			×		×	×			×

从表 1 中可以清晰地看到，三元背景中的关系仅表示在一定条件下某个属性属于或者不属于某个对象，每一个“×”代表属于/存在关系。

2.2 模糊集与三元概念分析的结合理论

模糊理论由 Zadeh 提出。为了能更好地表示模糊信息，Tadrat J 等将模糊理论与形式概念分析相结合，提出了模糊形式概念分析(FFCA)[16]。本文提出了模糊理论与三元概念分析相结合的文本分类模型，即模糊三元概念分析(FTCA)的分类模型，其将文本概念化为更抽象的三元概念，继而对新文本进行分类。

传统的三元概念分析使用二值形式来表达对象、属性与类别之间存在或者不存在的关系，但是它并没有考虑到不确定或不精确的信息。文中提出的 FTCA 模型将这种关系的二值形式延伸至 0-1 区间的实数，即隶属度值，这样不但可以表示属于或不属于的关系，还能用隶属度的值来表示属于的程度。在一个类别中一个属性对于一个对象的隶属度值代表这个属性的重要程度，同时也可根据隶属度值来计算两个概念间的相似度值。这些概念将在下面逐一进行解释定义。

定义 3（模糊三元背景） 一个模糊三元背景表示为一个四元组$K=(G,M,B,Y=\varphi(G\times M\times B))$，其中$G$是对象集，$M$是属性集，$B$是条件集，$Y$是$G,M$和$B$三者之间的模糊关系集，对于任意的$g\in G,m\in M,b\in B$，每一个模糊关系$(g,m,b)\in Y$都有一个在$[0,1]$之间的隶属度$\mu(g,m,b)$。

定义 4（模糊三元概念） 在模糊三元背景(G,M,B,Y)下，一个模糊三元概念定义为一个三元组$(\varphi(K_1),K_2,K_3)$，其中$K_1\subseteq G,K_2\subseteq M,K_3\subseteq B$，并且对于任意一个对象$g\in\varphi(K_1)$，隶属度$\mu_g=\min\limits_{m\in K_2,b\in K_3}\mu(g,m,b)$，$\mu(g,m,b)$是对象$g$在条件$b$下具有属性$m$的隶属度。

定义 5（模糊三元概念基数） 三元概念中对象的隶属度表示模糊三元概念的模糊值，因此模糊三元概念$(\varphi(K_1),K_2,K_3)$的基数定义为$|\varphi(K_1)|$，其中$|\varphi(K_1)|$为模糊集$\varphi(K_1)$的基数。

在实际应用中，通常要在属性集上进行操作，通过属性的

隶属度进一步阐述与对象和类别之间的关系。上述定义 2 中通过外延的隶属度来描述模糊三元概念，下面将要提到的分类过程中，经过对概念结构的分析，由于最终要确定训练样例与新文本的相似度，需在共有的特征空间中操作，因此从内涵的角度考虑定义三元概念更为合适。

根据三元序集上的定义和三元伽罗瓦连接的含义，三元概念中的外延、内涵和条件都是一个闭包系统；同时三元概念中的预序关系也是建立在外延、内涵和条件关系上的。当增大外延集合时，内涵和条件集合会随之减少；反之亦然。因此依据概念三元格的特点，我们可以从外延角度和内涵角度定义三元概念。

定义 6（从内涵角度定义的模糊三元概念）　在模糊三元背景(G,M,B,Y)下，一个模糊三元概念定义为一个三元组$(\varphi(K_2),K_1,K_3)$，$K_2\subseteq M,K_1\subseteq G,K_3\subseteq B$，对于任意一个属性$m\in\varphi(K_2)$，均有隶属度$\mu_{K_2}(m)=\min\limits_{g\in K_1,b\in K_3}\mu(m,g,b)$，其中$\mu(m,g,b)$是模糊关系$(m,g,b)\in Y$中的一个隶属度。

2.3　三元概念间的相似性函数

在模糊理论中贴近度表示两个模糊集合之间的彼此接近程度，在模糊模式识别方法中采用贴近度的大小识别待判别模糊子集的模式类别[17]。为衡量待识别子集的类别，需要判别测试模糊集与训练模糊集合之间的相对贴近程度。我们选用其中一种最小最大贴近度计算公式，对模糊集合间的关系推导及公式原理可参考文献[18]。

$$\sigma(A,B)=\frac{\sum\limits_{k=1}^{n}(A(x_k)\cap B(x_k))}{\sum\limits_{k=1}^{n}(A(x_k)\cup B(x_k))} \tag{1}$$

其中，$\sigma(A,B)$表示模糊集合 A 与模糊集合 B 之间的贴近度；$\cap$和$\cup$为 Zadeh 算子，$\cap$取小算子，$\cup$取大算子；k 表示第 k 个特征指标，共有 n 个。$a\cap b=\min(a,b)$，$a\cup b=\max(a,b)$。

相应地，可以类比式(1)中描述模糊集之间的贴近度来定义模糊三元概念间的相似性函数。下面给出三元概念间相似函数的定义。

定义 7（三元概念间最大最小相似度）　模糊三元概念$(\varphi(K_2),K_1,K_3)$与另一个模糊三元概念$(\varphi(K_2'),K_1',K_3')$间的最大最小相似性函数定义为式(2)：

$$FuzzySim((\varphi(K_2),K_1,K_3),(\varphi(K_2'),K_1',K_3'))=FuzzySim(\varphi(K_2),\varphi(K_2'))=\frac{|\varphi(K_2)\cap\varphi(K_2')|}{|\varphi(K_2)\cup\varphi(K_2')|} \tag{2}$$

模糊集中有各种定义$\cap$和$\cup$的操作方式，本文选择上述最普遍的使用方式，$\cap$取最小值，$\cup$取最大值。即 $\varphi(K_2)\cap\varphi(K_2')$ 定义为 $\mu_{K_2\cap K_2'}=\min(\mu_{K_2}(m),\mu_{K_2'}(m))$，且 $\varphi(K_2)\cup\varphi(K_2')$定义为 $\mu_{K_2\cup K_2'}=\max(\mu_{K_2}(m),\mu_{K_2'}(m))$。

根据定义 7 中的概述，两个三元模糊概念 $Con1$ 和 $Con2$ 之间的最大最小相似度计算公式如式(3)所示：

$$E(Con1,Con2)=\frac{|Con1\cap Con2|}{|Con1\cup Con2|} \tag{3}$$

利用式(3)的相似度计算公式即可计算出每个新文本构建的概念与每个类别中训练模型之间概念的模糊相似度。

举例说明基于三元概念分析的文本分类算法对新文本的分类操作过程。该算法的主要思想是以三元概念作为机器学习算法中的学习模型，即首先将训练数据集形式化为三元背景，再根据三元背景生成三元概念，此时三元概念表示为文本、特征词、类别以及三者之间的三元关系。表 2 是一个模糊三元背景的交叉表，显示了对象集、属性集和条件集之间的三元关系，其中所列举的对象集有 $K_1=\{D1,D2,D3,D4,D5,D6,D7,D8\}$，$D$ 表示文本；属性集 $K_2=\{T1,T2,T3,T4,T5\}$，T 表示文本中提取的特征关键词；条件集 $K_3=\{C1,C2\}$，C 表示文本所属类别。

表 2　模糊三元背景

	C1					C2				
	T1	T2	T3	T4	T5	T1	T2	T3	T4	T5
D1	0.87	0.61	0	0	0	0	0	0	0	0
D2	0	0	0	0	0	0	0.81	0	0.65	0
D3	0.76	0	0.75	0	0	0	0	0	0	0
D4	0.54	0	0.84	0	0.83	0	0	0	0	0
D5	0	0	0	0	0	0.66	0.77	0	0.90	0.88
D6	0	0	0	0	0	0	0.75	0	0.59	0.81
D7	0	0	0	0	0	0	0	0.51	0	0

在表 2 中，经过计算得到词语的 TF-IDF 值代表隶属度，表示对象、属性和条件之间的模糊三元关系。

基于上表中的三元背景，根据定义 4 可以生成所有的模糊三元概念，这些概念以〈文本 D(隶属度)，词语 T，类别 C〉的方式表示，即〈外延(隶属度)，内涵，条件〉，列举如下：

$Con1=\langle\{D1(0.87),D3(0.76),D4(0.54)\},\{T1\},\{C1\}\rangle$

$Con2=\langle\{D3(0.75),D4(0.54)\},\{T1,T3\},\{C1\}\rangle$

$Con3=\langle\{D1(0.61)\},\{T1,T2\},\{C1\}\rangle$

$Con4=\langle\{D4(0.54)\},\{T1,T3,T5\},\{C1\}\rangle$

$Con5=\langle\{D2(0.65),D5(0.77),D6(0.59)\},\{T2,T4\},\{C2\}\rangle$

$Con6=\langle\{D5(0.77),D6(0.59)\},\{T2,T4,T5\},\{C2\}\rangle$

$Con7=\langle\{D5(0.66)\},\{T1,T2,T4,T5\},\{C2\}\rangle$

$Con8=\langle\{D7(0.51)\},\{T3\},\{C2\}\rangle$

根据定义 6，从概念的内涵角度考虑，可以将所有的三元概念以〈词语 T(隶属度)，文本 D，类别 C〉的方式表示。例如，表 2 模糊三元背景中生成的概念 $Con2$ 可由如下形式表示：

$Con2=\langle\{T1(0.54),T3(0.75)\},\{D3,D4\},\{C1\}\rangle$

$Con2$ 中 $T1$ 的隶属度 $\mu_{Con2}(T1)$可由下式计算得到：

$$\mu_{Con2}(T1)=\min[\mu_{D3}(T1),\mu_{D4}(T1)]=\min[0.76,0.54]=0.54$$

同理可以计算出所有概念与相关联词语的隶属值，由此根据模糊三元概念，利用其特性建立起概念-词 $R_{Con\text{-}T}$ 的重要关系，如表 3 所列。我们将模糊三元概念作为学习样例对新文本进行分类。三元概念中已经包含了文本与类别的关系，

在上述例子中 $Con1$ 属于 $C1$ 类，而 $Con8$ 属于 $C2$ 类，因此不需要再对其关系进行计算，与二元概念相比省去了计算文本与类别关系的过程，更直观、简便，提高了效率。

表 3 模糊三元概念的向量表示模型

	$T1$	$T2$	$T3$	$T4$	$T5$
$Con1$	0.54	—	—	—	—
$Con2$	0.54	—	0.75	—	—
$Con3$	0.87	0.61	—	—	—
$Con4$	0.54	—	0.84	—	0.83
$Con5$	—	0.75	—	0.59	—
$Con6$	—	0.75	—	0.59	0.81
$Con7$	0.66	0.77	—	0.90	0.88
$Con8$	—	—	0.51	—	—

如现有新文本 Document 以及对应的关键词隶属度为(0.52,0.85,0.6,0.72,0.9)，类别设为 C_k，则 Document 所生成的概念与表 3 所给关系 $R_{Con\text{-}T}$ 中的 $Con4$ 模糊最大最小相似度的计算可由下式得出。

$$FuzzySim2(Con_{Doc},Con_4)=\frac{|Con_{Doc}\cap Con_4|}{|Con_{Doc}\cup Con_4|}=0.5064$$

以上是以简单的文本为例来说明三元概念分析方法在分类中的可行性，以下将详细阐述分类模型。

3 基于三元概念分析的文本分类模型

FTCA 以更为广义的角度定义了概念，更清晰地揭示了对象、属性与条件之间的关系。相较于模糊二元形式分析，在文本分类中三元概念能够直接显示出概念与类别之间的关系，不再用模糊集中复杂的合成计算得到。这个模型清晰明了，简单直观。本文的主要思路是：以 FTCA 为工具从语料库中提取所需内容，将文本抽象成概念的形式，用概念而非文本作为训练样例。具体做法为：首先，文本通过自然语言处理过程，采用预处理和特征提取的方法找出文本中对应的关键词；然后，通过分析文本、特征词、类别之间的内在联系形成模糊三元背景，继而根据已有的概念和模糊集的生成算法得到模糊三元概念，为了计算概念与文本(测试样例)间的相似度，需要找出每个概念与关键词的关系；最后，对测试文本分类，根据计算出的每一类中的所有概念与文本的相似度，求出文本具有最高相关度的类别，即为最终类别。本部分将描述该分类模型的主要步骤。

基于模糊三元概念分析的文本分类的整体框架如图 1 所示。

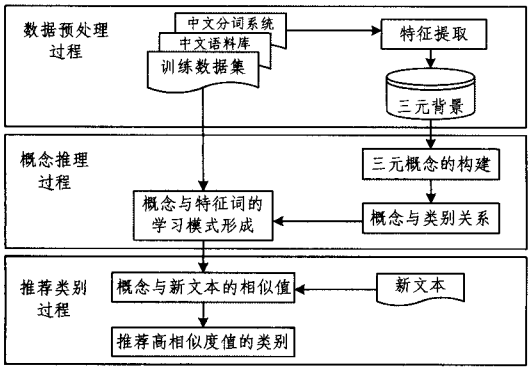

图 1 基于模糊三元概念分析的分类框架

3.1 文本预处理过程

文本预处理的目的是将每一个文本处理成下面步骤中需要使用的特征空间的形式。数据预处理的过程实际上是处理自然语言的过程。首先要对文本分词，过滤停用词并移除标点。在对中文的文本进行分类时，需要将句子中的词语进行分割，以便提取重要信息。本文将使用严格的单个词语的索引，不采用任何词干提取的方法。

数据分词和删除不相关信息后，对文本进行特征提取也是一个很重要的步骤。从语料库中检索出重要信息时，特征提取便可用来获取文本中的关键词。为方便处理三元概念模型中的相似度计算，采用经典的 TF-IDF(Term Frequency-Inverse Document Frequency)方法计算每个词在每篇文本中的权重。TF-IDF 的定义如式(4)所示，其中 $tf_i(d)$ 表示词语 i 在文本 d 中出现的频率，N 表示文本集中的文本总数，df_i 是含有词语 i 的所有文本数目。词 i 在文本 d 中的权重 $w_i(d)$ 也可以被简单认为是词频，$\log\frac{N}{df_i}$ 是逆向文件频率，可以作为词语普遍性的度量。

$$w_i(d)=tf_i(d)\times\log\frac{N}{df_i} \tag{4}$$

分类模型的第一步即数据预处理的算法如过程 1 所示。

过程 1 Preprocess Dataset

输入：文本数据集

输出：预处理后的模糊三元背景 Fuzzy TriadicContext

1. 搜索将预处理的文本，生成所有类别文本的集合 preText(i)，i 为文本类别总数，并对其进行分词 partition
2. For m=1:i 遍历所有类别
   For n=1:doc_n 遍历每一类别中选取的文本数 doc_n
   获取文本 preTemp
   提取词项 Term=strcat(term,preTemp)
   End for
   End for
3. Term=rid(Term,stopwords) 对词项进行停用词过滤 stopwords 以及标点移除 delimiter 等处理
4. 构建 doc_term_class 三维零矩阵
5. For l=1:i
   For j=1:preTemp.size
   For k=1:Term.size
   统计词频 doc_term_class(l,j,k)
   End for
   End for
   End for
6. 计算 Tfidf(doc_term_class(l,j,k))，构建出三维多值矩阵
7. 根据交叉验证 1:4 分成测试数据与训练数据
8. Fuzzy TriadicContext=full(Tfidf) 得到训练数据对应的模糊三元背景

3.2 模糊概念构建与推理过程

模糊三元概念是在已知模糊三元背景的基础上构建的，一旦计算出归一化的 TF-IDF 值，则可根据文本、关键词以及文本类别这三者之间的关系形式化为模糊三元背景。文本集作为模糊三元背景中的对象集，特征提取后得到的关键词作为模糊三元背景中的属性集，而作为学习样例已知的文本类

别则作为模糊三元背景的条件集合。本文中将归一化 TF-IDF 值作为隶属度值。

从三元背景出发，根据三元概念和模糊集的定义构建所有的模糊三元概念集。本文的研究只关注单分类的问题，也就是说所有的学习样例和测试样例都只属于一个类别，不存在多类别的情况，这也是非常特殊的三元背景。因此在构建三元概念时可根据二维的思想构建出三元概念集[19]，也可以从三维数据构建概念集合。本文参考了文献[20]和文献[21]中形式概念分析对数据的提取方法，运用 h 生成算子，直接从三元背景中构建三元生成算子，继而提取出三元概念。

分类模型第二步即构建模糊三元概念算法的伪码如下所示。

过程 2　Produce Triadic-concepts

输入：模糊三元背景 Fuzzy TriadicContext

输出：模糊三元概念 Fuzzy TriadicConcepts

1. 根据 Fuzzy TriadicContext 构建三元生成算子，得到 tri-generators

　For each generator∈ tri-generators

　　进行隶属度值的计算，运用公式

　　$\mu_{K_2}(m)=\min_{g\in K_1,b\in K_3}\mu(m,g,b)$

　End for

2. 针对 tri-generators 进行 h 闭包算子的运算，即在同一类别中对象和类别不变，扩充属性，同时计算隶属度值，得到 tri-sets

3. For each settri-sets

　扩充类别部分

　删除重复的集合，得到 Tri-concepts

　End for

4. 得到 Fuzzy TriadicConcepts

本文将构建出的模糊三元概念作为训练样例，其与测试样例（未分类的文本）拥有相同的特征空间，此时对象、属性与条件的关系就变成了概念与属性间的关系。在上述例子中，表 2 的三元概念模型便转变成了表 3 的基于词向量的学习形式。在有监督的学习中，一个新文本将根据训练数据提供的学习机制进行分类。概念构建与推理过程的详细伪码如下所示。

过程 3　Reasoning Triadic-concepts

输入：构建好的三元概念 Fuzzy TriadicConcepts

输出：基于词向量的三元概念 $Concept_j$-Term

1. 构建完成的三元概念 TriadicConcepts 为从内涵角度定义的带有隶属度的三元概念 TriadicConcepts-Membership

2. 从而得到概念与类别的关系 $R_{Con\text{-}C}$

　For each $Concept_j$∈ TriadicConcepts

　　For each $Concept_j$ 属性

　　For each $Concept_j$ 同一个属性对应不同对象

　　Con(membership)=min(obj_m-membership)

　　End for

　　End for

　End for

　求出所有 $Concept_j$(membership)

　计算出训练样例中所有三元概念与词 T 的关系 $R_{Con\text{-}T}$

3. 得到 $Concept_j$-Term

3.3　文本分类过程

本文的思想是在文本分类的过程中用更抽象定义的三元概念代替文本作为学习样例，每一个学习样例的类别信息对新文本的类别分类都起着至关重要的作用。我们在构建模糊三元概念的同时，已经将类别信息与文本、关键词融为一体，从开始 $R_{D\text{-}C}$ 文本与类的关系成功转变为 $R_{Con\text{-}C}$ 概念与类的关系。在求出的模糊三元概念集中，每一个模糊概念都仅属于一个类别，因此在 $R_{Con\text{-}C}$ 中，可以用 0 和 1 代表概念属于或不属于某一类。$R_{Con\text{-}C}$ 为新文本提供了关键的类别信息。

FTCA 分类方法的主要目的是基于格空间已经训练好的三元概念对新文本进行分类。对于目标文本 d，首先可以经过 TF-IDF 预处理计算得到与训练样例中相同的词空间向量，然后将未知类的目标文本 d 设成 C_k 类。此时，我们也同样对测试样例构建出模糊三元概念，d 是概念中的外延，词空间向量中的 *term* 表示概念中的内涵，而 C_k 是概念中的条件，TF-IDF 作为隶属度值。以共同的词向量为中介，通过定义 7 最大最小相似性函数提出的相似性操作可分别计算出新文本与每一个类的模糊三元概念之间的模糊相似性度量。新文档与类别 *Category* 的相关度计算公式如式(5)所示：

$$Corr_{Catg_i}(d)=\sum_{j=1}^{n}FuzzySim(d,concept_j)/N \tag{5}$$

其中，$FuzzySim(d,concept_j)$ 为文本 d 生成的三元概念与类别 i 中的概念 $concept_j$ 的模糊相似度，N 为类别 i 中全部模糊三元概念的个数。文本分类过程的详细伪码如下所示。

过程 4　Classification

输入：新文本 d 和训练样例（三元概念与词间关系 $Concept_j$-Term）

输出：新文本的类别 Category(d)

1. 由过程 1 得到测试数据集，并构建相应的三元概念 Con_{Doc}

2. 在具有相同特征空间 Term 的 Con_{Doc} 与 Con_j 条件下分类

3. 对于训练样例中的每一个类别 $Catg_i$

　计算出 $Catg_i$ 中模糊三元概念总数 N

　For each $Catg_i$

　　For each $Concept_j$∈$Catg_i$

　　计算测试集的三元概念与该类每一个概念的模糊相似度 FuzzySim，并求出该类相似度之和 $Sum_{Catgi(d)}$

　　$Sum_{Catgi(d)}$ += E(Con_{Doc}, $Concept_j$)

　　End for

　　求出 d 在该类 $Catg_i$ 的相关联程度

　　$Corr_i(d)=Sum_{Catgi(d)}/N$

　End for

　计算出 d 与所有类别相关度的最大值

　　$Category(d)=\arg\max_{C_i} Corr_i(d)$

4. 得到文本类别 Category(d)

基于模糊三元概念分析的分类模型，FTCA 算法的整体描述如算法 1 所示。

算法 1　FTCA

输入：文本数据集

输出：测试文本的类别

1. 预处理数据得到三元概念

　　Produce Triadic-concepts

　根据模糊三元概念得到 $R_{Con\text{-}T}$

　　Reasoning Triadic-concepts

2. Classification

3.4 算法时间复杂度分析

根据以上不同过程的伪码表述，对每个模块进行时间复杂度分析，从对数据的预处理到最终的分类，算法的总时间复杂度分析及其形式化证明如下。

命题 1 FTCA 算法的时间复杂度为 $O(\|L_1\| * \|L_2\| * \|E\|)$。

证明：总算法的时间复杂度与三元概念中的外延、内涵和条件都有关系，设三元背景中的对象个数、属性个数和条件个数分别用 $\|L_1\|$，$\|L_2\|$ 和 $\|L_3\|$ 表示，三元概念的个数由 $\|E\|$ 表示。在过程 1 的描述中，主要时间取决于构建三维矩阵的过程。三维矩阵由文本数、词频个数以及类别数决定，而在满足三元关系的情况下，这三者可看成对象个数、属性个数和条件个数。因此过程 1 的时间复杂度为 $O(\|L_1\| * \|L_2\| * \|L_3\|)$。由构造三元概念的伪代码可知，过程 2 中的时间复杂度主要取决于三元生成算子的生成以及它与三元-集的个数。用 $\|M\|$ 表示三元生成算子的个数，$\|N\|$ 表示三元-集的个数。构建三元生成算子需要在单个类别的每个属性下对每个对象进行操作，可知时间复杂度为 $O(\|L_1\| * \|L_2\| * \|L_3\|)$，h 闭包算子的时间复杂度为 $O(\|M\| * \|L_2\|)$，扩充三元-集为三元概念的时间复杂度为 $O(\|N\| * \|L_3\|)$。由此，过程 2 的时间复杂度为 $O(\|L_1\| * \|L_2\| * \|L_3\|)$。过程 3 中的时间复杂度主要取决于对概念的最小隶属度赋值的三层嵌套，由伪码可知此算法最极端的情况为三元概念包含三元背景中的所有对象和属性，即 $O(\|L_1\| * \|L_2\| * \|E\|)$。考虑在三元概念中对象与属性的值一定分别小于三元背景中对象与属性的个数，因此构造 $R_{Con\text{-}T}$ 的时间复杂度小于 $O(\|L_1\| * \|L_2\| * \|E\|)$。由过程 4 的伪码可知，考虑极限情况即每个类别中均有全部三元概念时，时间复杂度为 $O(\|L_3\| * \|E\|)$，因此过程 4 的时间复杂度小于 $O(\|L_3\| * \|E\|)$。FTCA 模型总算法由上述 4 个模块构成，而过程 1—过程 3 可看成是整体对训练数据预处理的部分，过程 4 是最终分类部分，因此过程 1—过程 3 这一模块的时间复杂度可以表示为 $O(2(\|L_1\| * \|L_2\| * \|L_3\|) + \|L_1\| * \|L_2\| * \|E\|)$，过程 4 的时间复杂度为 $O(\|L_3\| * \|E\|)$，因此 FTCA 模型的时间复杂度为 $O(2(\|L_1\| * \|L_2\| * \|L_3\|) + \|L_1\| * \|L_2\| * \|E\| + \|L_3\| * \|E\|)$，即总算法的时间复杂度为 $O(\|L_1\| * \|L_2\| * \mathrm{Max}(\|L_3\|, \|E\|))$。考虑一般情况，在实际操作的数据中，类别即条件的个数一定小于训练数据集中所构造的三元概念的个数，$\|L_3\|$ 一定小于 $\|E\|$，因此该模型 FTCA 算法的时间复杂度一定小于 $O(\|L_1\| * \|L_2\| * \|E\|)$。

4 实验与分析

基于三元概念分析的文本分类算法是本文提出的一种新模型。为了验证该模型的有效性，分别用目前流行的深度学习中的 CNN 算法与经典机器学习算法 SVM(这里采用台湾大学林智仁等人开发的 libsvm 工具包)、KNN 算法以及本文提出的 FTCA 算法这 4 种方法在相同的中文数据集上进行实验对比；使用 Claudio Carpineto 等人提出的基于形式概念分析的 CL-SVM 分类模型[6] 作为英文数据集的对照组。对所有数据集采用准确率(Accuracy)的评价标准来衡量算法的性能[22]。为了避免仅由一次实验中训练集和测试集的随机选取带来的结果误差，对每组数据重复做 10 次分类，并对结果计算平均准确率(Average Accuracy)。

对实验中各参数的设置进行说明，其中 CNN 算法初始化模型参数为小的随机数，激活函数为 tanh，深度网络结构共有 3 层，特征维度不同时神经元个数也有所不同：在第一个数据集中，第一层卷积层有 36080 个神经元，第二层卷积层有 2880 个神经元，第三层回归层有 720 个神经元；在第二和第三个数据集中，第一层有 14440 个神经元，第二层有 1280 个神经元，第三层有 320 个神经元。SVM 算法选用比较简单且应用较广泛的 RBF 核函数；KNN 算法的 K 值取为 5。

实验环境：3.4GHz CPU，4GB 内存，Win10 操作系统，使用 Java 实验平台实现分类算法。

4.1 实验数据

为充分验证该模型的有效性，选用 3 种通用的中文数据集以及一种英文数据集进行对比实验。考虑到现构建三元概念算法的效率，选取小规模数据集。同时，在数据预处理阶段，中文数据集均采用中科院 ICTCLAS 的中文分词系统对文本进行分词处理。

4.1.1 复旦大学数据集

此文本分类语料库采用复旦大学计算机信息与技术系国际数据库中心自然语言处理小组整理的测试语料库。我们随机选用其中的 5 个类别，将文本集中 80%类别的文本作为训练数据，其余 20%作为测试数据。选取此数据集旨在说明提出的 FTCA 模型是有效的。选取此数据的特征维数为 4500。

4.1.2 旅游数据集

此数据集是旅游领域中文文本的分类语料，所有数据均来自互联网。选用其中 6 种类别(6-class)，分别是城市概况、地方文化、购物美食、交通指引、旅游服务和休闲娱乐。对测试数据集的选取，旨在观察在类别数目相同的情况下文本数目对实验结果的影响，以及在相对较小的相同数据集中 FTCA 模型与经典机器学习算法的性能。每次分别随机抽取每个类别中相同数目的文本数，共随机抽取 6 次，每次每个类别中抽取的文本数分别为 15 篇、20 篇、25 篇、30 篇、35 篇和 40 篇。每种类别中的训练样例与测试样例用交叉验证方法以 4:1的比例随机分配。测试数据集 6 种类别的属性如表 4 所列。

表 4 测试数据集 6-class 的属性

Dataset	Number (per Category)	Total number of Text	Feature dimension
*D*1	15	90	420
*D*2	20	120	550
*D*3	25	150	680
*D*4	30	180	760
*D*5	35	210	950
*D*6	40	240	1200

4.1.3 中文搜狗数据集

该数据集是搜狗实验室提供的文本挖掘数据集中的语料库，是大家公认的中文通用数据集，在分类和聚类中也得到广泛应用。分别随机选用其中的 8 种类别(财经、IT、健康、体

育、旅游、教育、招聘、文化）、6 种类别（财经、IT、体育、教育、招聘、文化）和 4 种类别（财经、IT、教育、招聘）进行实验。选取此数据集，旨在观察在相同文本数目的情况下类别数目对实验结果的影响，同时也可测试出相同类别中文本数目对分类结果的影响。同样，采用 k-fold 函数随机将该数据以4∶1的比例分为训练数据和测试数据。

4.1.4　20NewsGroup 英文数据集

该数据集是通用的、主要用于对英文文本分类训练及效果测试的语料，所有数据均来自互联网。选取与文献[6]中相同的数据集，即 8 个不同类别，每个类别分别随机抽取数目相同的文本进行分类处理。由于三元概念分析是在形式概念分析的基础上发展而来，选取此数据集的目的是将 FTCA 模型与基于形式概念分析的文本分类算法 CL-SVM[6] 作对比，观察两种模型的实验结果。

4.2　实验结果与分析

4.2.1　本文方法与基本模型的比较

在复旦大学数据集，不同方法的文本分类准确率结果如表 5 所列。

表 5　复旦大学数据集的分类准确率结果/%

Dataset	Method	Accuracy
Fudan University 5-class dataset	FTCA	84.45
	CNN	85.00
	SVM	80.00
	KNN	76.67

从表 5 可以看出，对于复旦大学数据集，FTCA 方法比传统 SVM 和 KNN 算法的分类准确率更高，可证明此方法是有效的。与此同时，KNN 算法的分类效果在此数据集中相对较差，而 CNN 算法的分类效果是相对最好的。在与 CNN 算法比较时，FTCA 模型的分类准确率与其相差近 0.5 个百分点，这是由于该数据集的特征维度较大，三元概念中的内涵个数以及其拥有的隶属度个数较多，没有对每个类别中相关度不高的三元概念进行适当删减，因此在构建训练集 $R_{Con\text{-}T}$ 时会生成过多冗余值，从而对测试数据的分类产生干扰。尽管本文提出的 FTCA 新模型在上述方法中没有达到最好的分类效果，但是与列举的最好方法相比，两者的分类能力是比较接近的。

因此，从第一种特定的测试数据集实例中可以得出，在该数据集下，本文提出的 FTCA 方法是有效的，且与列举的其他方法相比具有较高的分类准确率。

对于旅游数据集，4 种不同方法的分类准确率结果如表 6 所列。

表 6　旅游领域 6-class 数据集的实验结果/%

Text classification	*D1*	*D2*	*D3*	*D4*	*D5*	*D6*
CNN	66.67	65.83	70.00	67.57	76.20	79.17
SVM	61.11	68.51	68.21	69.62	70.90	71.83
KNN	55.56	62.50	62.04	61.21	63.17	64.58
FTCA	65.51	68.94	71.94	74.13	72.11	71.91

基于表 6 所列数据，随着文本数目的增加，4 种不同方法的分类准确率的变化趋势可由折线图反映出来，如图 2 所示。从表 6 的结果中选取 D1－D4 数据集进行分析，其分类结果如图 3 所示。

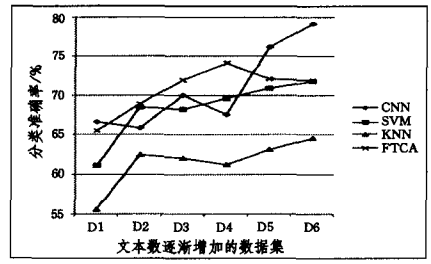

图 2　分类结果在不同文本数目下的变化趋势图

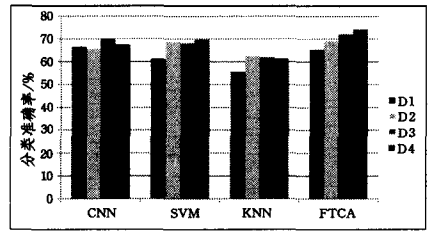

图 3　相同类别、不同文本数目下的分类准确率对比

结合图 2 和图 3 中显示的结果可以看出，在旅游数据的小数据集中，FTCA 的分类方法明显优于 KNN 算法，且与 SVM 算法的分类能力比较接近，证明该模型是有效的。直观上看，在这组数据中，对于类别数相同而文本数目不同的情况，KNN 算法的准确率在这 4 种方法中是最低的，同时 FTCA 算法在 D1－D6 数据集中的准确率均优于 SVM 和 KNN 算法，说明 FTCA 算法将数据集构建成带有隶属度的三元概念形式有利于文本的分类。首先，将数据概念化为模糊三元概念，构建起概念与特征以及概念与类别之间的直接关系，以利于测试文本与训练数据的对比分类；其次，三元概念的构建本身就是一种聚类的过程，FTCA 正是利用了三元概念分析中对其结构属性处理的特点，将同一类别的数据聚集在一起，突出了每种类别的特点，随着文本数目和特征维度适当的增加，类别信息更加突出，继而对于测试数据的分类效果也有所提升。

由表 4 可知，在保持类别不变的情况下，随着文本数目的增加，特征维数也在不断上升，结合图 2 显示的内容可知 SVM 算法和 KNN 算法的准确率均呈上升趋势，说明在一定范围内训练数据的增加会提高算法的准确率；同时 CNN 算法的分类波动性较大，分类准确率不稳定。由图 3 可知，FTCA 算法随着数据维度的增大准确率也在稳步上升，提高了近 9 个百分点；在 D5 和 D6 数据集中百分比仍然高于 SVM 和 KNN 算法，但低于 CNN 算法，说明 FTCA 对于特征维度及三元概念的选择有待进一步探索。

从第二种旅游数据集得出的实验结果以及变化趋势图可以分析出，FTCA 算法在随着文本数适量增加的同时，分类准确率也在逐渐提高，说明在选择特定范围的数据维度的情况下，FTCA 有较高的分类准确率。

表 7 列出了 4 种不同分类方法对于搜狗文本语料库的数据集的准确率对比结果。由表 7 可以看出，在对该数据集的文本进行四分类、六分类或八分类时，不论是相较于经典 SVM 和 KNN 方法还是 CNN 方法，FTCA 的方法均有较高的分类效果，同理可以证明该文提出的 FTCA 模型是有效的。

表 7 基本方法对于搜狗数据集的实验结果/%

Category	Text Number	Method of Text Classification			
		CNN	SVM	KNN	FTCA
4-class	180	69.45	71.46	65.07	83.02
	200	77.50	76.56	65.50	79.44
	240	72.92	74.02	66.54	79.00
	280	68.43	75.50	70.81	78.90
	300	65.00	75.54	73.16	81.97
6-class	180	69.45	67.59	60.26	72.98
	240	70.94	69.46	60.95	78.24
	300	69.67	69.06	58.33	71.97
	360	70.94	68.72	60.63	71.03
8-class	200	50.00	58.01	53.47	60.00
	240	60.00	60.08	56.25	63.00
	280	62.25	63.69	55.36	62.36
	360	69.45	63.46	56.78	65.38
	480	63.50	60.15	53.13	60.20

为了清晰表达出文本数目相同的情况下类别数目对分类效果的影响，给出文本数目为 240 篇时 4 种分类方法在四类、六类和八类的分类柱状图，如图 4 所示。

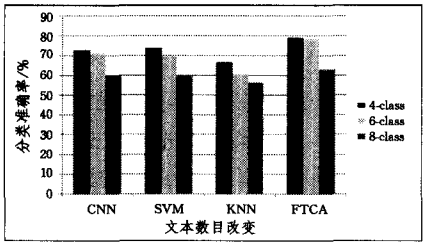

图 4 不同类别数目下的分类准确率值对比

从图 4 可以清晰看出，在文本数目相同的前提下，类别数目对实验分类效果的影响比较大。图 4 显示在文本数目为 240 篇时，4 种分类方法中四类、六类和八类的分类准确率均大幅降低。由于类别数目增多而文本数目不变，那么每个类别中的文本数目会减少，对于算法来说训练数据集的数目就会相应减少，干扰信息的比重增大，因此算法的分类效果就会有所下降。尽管类别中文本数目减少，但是对于 FTCA 模型，仍将每一类的信息聚类构建三元概念，以尽可能突出类别特征，为测试数据提供有效的训练信息，从而对其准确分类，所以 FTCA 算法的分类效果仍然优于其他 3 种算法。此数据集可以说明，在文本数目相同的情况下，当类别数目适当增多时，选择 FTCA 算法也具有较高的分类准确率。

4.2.2 本文方法与同类工作的比较

表 8 给出了对于 20NewsGroup 英文数据集，基于形式概念分析的分类算法 CL-SVM 与基于三元概念分析的分类算法 FTCA 的分类准确率对比结果。

表 8 FTCA 与 CL-SVM 模型的分类准确率对比/%

	40docs	80docs	120docs	160docs
linear CL-SVM	71.79	65.00	69.09	76.42
gaussian CL-SVM	71.79	66.25	70.00	77.14
FTCA	70.84	68.75	70.83	77.20

通过表 8 可以看出，不论是相对于线性核 CL-SVM 的模型还是高斯核 CL-SVM 的模型，FTCA 都能够达到较好的分类效果。从 4 组数据的分类结果可知，随着文本数目的逐渐递增，FTCA 的分类准确率在一定程度上也有所增加。

由此可以得出结论，在此数据集中，基于三元概念分析的分类算法的分类算法比基于形式概念分析的分类准确率高，证明与现有的同类工作相比，FTCA 模型的分类效果更好；也说明了三元概念中增加条件集合，将一类信息聚集处理的形式相较于二元形式更具有优势。

同时在此公共数据集下，对比目前工作，改进的经典模型中具有较好分类性能的是由 Li Qiang 提出的基于类别的模糊关联度算法(CFCD)[18]，其分类准确率最高达到 80.10%。可以看出，基于概念的分类结果没有 CFCD 最好的效果好，基于概念的分类算法在提取概念的过程中限制了对数据集的选择，考虑到算法的复杂度，选取了较小数据集使得训练集数目较少，影响了分类的最终结果。

4.2.3 不同数据集在 FTCA 模型下的运行时间

参考文献[23]中对算法的耗时分析，对 FTCA 模型进行运行时间的分析。分别选取类别数一定但文本数目有变化的数据集 2，以及文本数目一定而类别数不同的数据集 3，其中数据集 3 选用 4-class 和 8-class 的数据进行时间对比分析。

由 3.4 节中对算法时间复杂度的分析可知，该模型的运行时间分为两个大模块，第一个模块是对数据的整体预处理，包括对文本进行 TF-IDF 处理、三元背景的形成、三元概念的构建以及概念与特征关系的搭建；第二个模块则是对测试文本的分类。分别对这两大模块进行时间分析，数据集 2 和选用的数据集 3 的耗时结果如表 9 和表 10 所列。

表 9 FTCA 对数据集 2 的耗时结果/s

数据集 2	预处理时间	分类时间
D1	8.051	0.452
D2	9.823	0.806
D3	49.679	1.445
D4	101.350	3.090
D5	1460.238	5.351
D6	1108.404	4.247

表 10 FTCA 对数据集 3 的耗时结果/s

数据集 3	预处理时间	分类时间
4-class 文本数 200	124.270	4.006
8-class 文本数 200	104.113	3.261
4-class 文本数 240	165.432	5.130
8-class 文本数 240	165.304	4.973
4-class 文本数 280	297.584	10.27
8-class 文本数 280	240.733	8.596

从表 9 和表 10 可以看出，FTCA 模型在预处理和分类中所消耗的时间均随着文本数目的增加而增多。从表 9 中可以明显看出分类时间呈缓增趋势而预处理时间呈猛增趋势，这是由于在预处理阶段随着文本数目的增加，三元概念的对象个数、属性个数及三元概念的个数均会增加，三元概念构建速度较慢，相应时间消耗会很大。尽管此模型预处理时间较长，但其对数目相同且对应特征一致的数据只进行一次预处理，便可以对不同测试数据进行多次分类测量。表 10 中，不论在预处理时间还是在分类时间上，不同类别数对于相同文本数目的数据来说耗时相差不多，因为类别数目的变化远小于三元概念的个数，因此在此数据集中变化的类别数目对 FTCA 模型耗时的影响不大。

综合以上 3 种测试数据集的实验对比可以得出，与经典算法、现流行的卷积神经网络算法以及 CL-SVM 模型相比，本文提出的 FTCA 算法在特定的数据集中均取得了较好的

结果，证明 FTCA 模型是有效的。对同一条件下的数据作优先处理，突出了同类别的信息特点，体现了运用三元概念的优势，在选取一定数据规模的情况下具有较好的分类效果。

结束语　本文提出了一种基于三元概念分析并与模糊理论相结合的文本分类模型，主要研究了如何构建文本数据经预处理后的多值三元背景以及带有隶属度的三元概念间的联系，通过使用模糊值和模糊关系提出了新颖的模糊三元概念 FTCA 和概念间相似函数，并利用三元关系将文本数据抽象为形式化的三元概念，从而有效地进行分类。FTCA 模型的提出既是对三元概念分析的扩展，为传统机器学习分类算法提供了新的思路，也为三元概念的应用奠定了基础，在现实生活中具有广阔的应用前景。

虽然三元概念分析有很大的发展空间，但 FTCA 模型是对三元概念分析在应用中的初步探索，仍有不足之处，如数据规模有待进一步增大，特征维度的增大对分类效果的影响也有待进一步研究。下一步工作是优化该模型，将模型运用在大规模数据下，针对不同类型的数据提出高效且实用的分类方法，同时如何将数据信息更好地抽象为三元概念的形式也是未来的研究方向。

参 考 文 献

[1] LEHMANN F, WILLE R. A triadic approach to formal concept analysis[C]// International Conference on Conceptual Structures: Applications, Implementation and Theory (LNCS954). Heidelberg: Springer-Verlag, 1995: 32-43.

[2] GANTER B, WILLE R. Formal concept analysis: mathematical foundations[M]. Berlin: Springer-Verlag, 1999: 66-68.

[3] BELOHLAVEK R, GLODEANU C, VYCHODIL V. Optimal factorization of three-way binary data using triadic concepts[J]. Order-A Journal on the Theory of Ordered Sets and Its Applications, 2013, 30(2): 437-454.

[4] TANG Y Q, FAN M, LI J H. Cognitive system model and approach to transformation of information granules under triadic formal concept analysis[J]. Journal of Shangdong University (Natural Science), 2014, 49(8): 102-106. (in Chinese)
汤亚强，范敏，李金海. 三元形式概念分析下的认知系统模型及信息粒转化方法[J]. 山东大学学报(理学版)，2014，49(8)：102-106.

[5] WEI L, WAN Q, QIAN T, et al. An overview of triadic concept analysis[J]. Journal of Northwest University (Natural Science Edition), 2014, 44(5): 689-699. (in Chinese)
魏玲，万青，钱婷，等. 三元概念分析综述[J]. 西北大学学报(自然科学版)，2014，44(5)：689-699.

[6] CARPINETO C, MICHINI C, NICOLUSSI R. A Concept Lattice-Based Kernel for SVM Text Classification[C]// Formal Concept Analysis, International Conference (ICFCA 2009). Darmstadt, Germany, 2009: 237-250.

[7] BELOHLAVEK R, BAETS B D, VYCHODIL J O V. Inducing Decision Trees via Concept Lattices[J]. International Journal of General Systems, 2009, 38(4): 455-467.

[8] KANG X, LI D, WANG S. A multi-instance ensemble learning model based on concept lattice[J]. Knowledge-Based Systems, 2011, 24(8): 1203-1213.

[9] LI S T, TSAI F C. A fuzzy conceptualization model for text mining with application in opinion polarity classification[J]. Knowledge-Based Systems, 2013, 39(2): 23-33.

[10] LI S T, TSAI F C. Noise control in document classification based on fuzzy formal concept analysis[C]// IEEE International Conference on Fuzzy Systems (FUZZ). 2011: 2583-2588.

[11] POELMANS J, IGNATOV D I, KUZNETSOV S O, et al. Formal concept analysis in knowledge processing: A survey on applications[J]. Expert Systems with Applications, 2013, 40(16): 6538-6560.

[12] LIU G J, WANG W Y. Research on the application of concept lattice in intelligent learning[C]// Cross Strait Quad-Regional Radio Science and Wireless Technology Conference (CSQRWC), 2011. IEEE, 2011: 1499-1501.

[13] PRISS U. Formal concept analysis in information science[J]. Annual Review of Information Science & Technology, 2006, 40(1): 521-543.

[14] BELOHLAVEK R, GLODEANU C, VYCHODIL V. Optimal Factorization of Three-Way Binary Data Using Triadic Concepts[J]. Order-A Journal on the Theory of Ordered Sets & Its Applications, 2013, 30(2): 437-454.

[15] IGNATOV D I, GNATYSHAK D V, KUZNETSOV S O, et al. Triadic Formal Concept Analysis and triclustering: searching for optimal patterns[J]. Machine Learning, 2015, 101(1): 271-302.

[16] TADRAT J, BOONJING V, PATTARAINTAKORN P. Building classification rules for case-based classifier using fuzzy sets and formal concept analysis[C]// International Conference on Soft Computing As Transdisciplinary Science and Technology. ACM, Cergy-Pontoise, France, 2008: 13-18.

[17] FORMICA A. Concept similarity in Formal Concept Analysis: An information content approach[J]. Knowledge-Based Systems, 2008, 21(1): 80-87.

[18] LI Q, HE L, LIN X. Dimension reduction based on categorical fuzzy correlation degree for document categorization[C]// IEEE International Conference on Granular Computing. 2013: 186-190.

[19] LIU X J. Study on the Construction Algorithm of Concept Trilattices and Its Application [D]. Xi'an: Xidian University, 2013. (in Chinese)
刘晓今. 概念三元格构造算法及应用研究[D]. 西安：西安电子科技大学，2013.

[20] ZHANG Z, DU J, WANG L. Formal concept analysis approach for data extraction from a limited deep web database[J]. Journal of Intelligent Information Systems, 2013, 41(2): 211-234.

[21] TRABELSI C, JELASSI N, YAHIA S B. Scalable mining of frequent tri-concepts from folksonomies[M]// Advances in Knowledge Discovery and Data Mining. Springer Berlin Heidelberg, 2012: 231-242.

[22] FENG G H. Review of Performance Evaluation of Text Classification[J]. Journal of Intelligence, 2011(8): 66-70. (in Chinese)
奉国和. 文本分类性能评价研究[J]. 情报杂志，2011(8)：66-70.

[23] CHAI Y M, ZHANG Z, WANG L M. An Algorithm for Mining Global Closed Frequent Itemsets Based on Distributed Frequent Concept Direct Product [J]. Chinese Journal of Computers, 2012, 35(5): 990-1001. (in Chinese)
柴玉梅，张卓，王黎明. 基于频繁概念直乘分布的全局闭频繁项集挖掘算法[J]. 计算机学报，2012，35(5)：990-1001.