基于粗集与遗传算法相结合的文本模糊聚类方法

王明春[***] 王正欧[*]

*（天津大学系统工程研究所 天津 300072）

**(天津工程师范学院数理与信息科学系 天津 300222)

摘 要：该文将粗集与遗传算法相结合的方法成功应用于文本模糊聚类。在聚类过程中，将权重参数的设定也通过编码由遗传算法确定，从而使得权重参数的设定具有科学性和可操作性，避免了在类似算法中确定权重时的主观性和不可靠性。最后的实例说明了算法的可行性。

关键词：粗集，遗传算法，文本挖掘，模糊聚类

中图分类号：TP18 **文献标识码**：A **文章编号**：1009-5896(2005)04-0548-04

Text Fuzzy Clustering Algorithm Based on Rough Set and Genetic Algorithm

Wang Ming-chun[***] Wang Zheng-ou[*]

[*](*Institute of Systems Engineering, Tianjin University, Tianjin* 300072, *China*)

[**](*Dept. of Math., Phys. and Info. Sci., Univ. of Tech. and Educa., Tianjin* 300222, *China*)

Abstract This paper presents a text fuzzy clustering algorithm which combines rough set and genetic algorithm fully. In the clustering process, the weight parameters are also described by genetic algorithm, thus it makes parameters more reasonable and operationable and avoids subjectivity and unreliability of describing weight parameters in the similar algorithms proposed by other researchers. The example demonstrates the feasibility of the algorithm.

Key words Rough set, Genetic algorithm, Text mining, Fuzzy clustering

1 引言

在信息过载的今天，绝大部分信息以文本的形式存储在计算机和网站上。这迫切需要人们开发出高效的文本挖掘方法，用以发现隐藏在文本内部的潜在的、新颖的、未知的知识。作为文本挖掘的主要内容之一，文本聚类是将文本集分组的全自动处理过程，使得类之间文本的不相似度最大，而类内部的文本之间的不相似度最小[1]。

典型的文本聚类算法：K-均值算法、球形 K-均值算法等都认为文本可以被唯一地归到一个类。但由于汉语文本的多样性和大量性，一个文本相对于一个类的成员关系有可能不能精确定义，一个文本有可能是一个以上类的候选成员。为了克服以上算法的缺点，文献[2]提出了模糊 C-原型（FCMdd）算法；文献[3]对此算法进行改进，提出了基于非欧几里德关系的竞争凝聚算法；但这两种算法的缺点是都需要预先给定隶属度。文献[4]和文献[5]将粗集与遗传算法相结合，分别对高速公路和网站访问者进行了聚类，其缺点是人工设定适应度函数中的权重参数，使得算法的执行缺乏科学性和可操作性。本文将粗集与遗传算法相结合的方法成功应用于文本模糊聚类；在聚类过程中，将权重参数的设定也通过编码由遗传算法确定，从而使得权重参数的设定具有科学性和可操作性，避免了前人在类似算法中确定权重时的主观性和不可靠性。最后的实例验证了算法的可行性。

2 粗集的有关概念

粗集（Rough Set, RS）理论是 80 年代初由 Pawlak 提出的一种处理模糊性和不确定性的数学工具，已经在很多领域取得了成功应用，下面介绍相关的几个定义和性质。

定义 1 称 $S=(U,C\cup D)$ 为信息系统，其中 U 为论域，是一非空有限集；C 为条件属性集合，D 为决策属性集合，$Q=C\cup D$ 称作属性集合。

定义 2 $A\subset Q$，$X\subset U$ 为感兴趣的对象集合，则 X 的 A 下近似 $\underline{A}(X)=\cup\{Y\in U/A:Y\subseteq X\}$，它是一定属于 X 的所有对象的集合。其中 U/A 表示对象集合 U 关于属性集合 A 的等价类集合。X 的 A 上近似 $\overline{A}(X)=\cup\{Y\in U/A:Y\cap X\neq\Phi\}$，它是可能属于 X 的所有对象的集合。

2003-11-18 收到，2004-06-02 改回
国家自然科学基金（60275020）资助课题

设 $U=\{u_1,u_2,\cdots,u_n\}$ 为论域上的所有对象集合，它被属性 P 分成以下 m 类：$U/P=\{X_1,X_2,\cdots,X_m\}$，则所有 $X_i\in U/P$ 的上下近似满足下列基本性质[4]：

性质 1　$\Phi\subseteq\underline{A}(X_i)\subseteq\overline{A}(X_i)\subseteq U$。

性质 2　$\underline{A}(X_i)\cap\underline{A}(X_j)=\Phi,\ i\neq j$。

性质 3　$\underline{A}(X_i)\cap\overline{A}(X_j)=\Phi,\ i\neq j$。

性质 4　如果一个对象不属于任何一个下近似，则它必然属于两个或两个以上的上近似。

在上述性质中，性质 1 说明如果一个对象属于一个类 X_i 的下近似，则它必然属于 X_i 的上近似；性质 2 说明一个对象 $u_k\in U$ 至多为一个类 X_i 的下近似。

3　粗集与遗传算法相结合的文本聚类方法

遗传算法(Genetic Algorithm，GA)是一种集效率与效果于一身的优化搜索方法，它利用结构化的随机信息交换技术组合群体中各个结构中最好的生存因素，从而复制出最佳个体，并使之一代一代地进化，最终获得满意的优化结果。在设计遗传算法过程中，主要涉及到如何设计评价函数、遗传编码和遗传算子等几个问题[6]。

3.1　文本的表示

如何使文本易于被计算机处理，是文本挖掘所面临的最基本的前期工作，目前这方面的研究工作已经取得了一定的进展。矢量空间模型（Vector Space Model, VSM）是近年来应用较多且效果较好的方法之一。在该模型中，文本空间被看作是由一组正交词条向量所组成的向量空间，每个文档表示为其中的一个范化特征向量：

$$V(d)=(t_1,w_1(d);\cdots;t_i,w_i(d);\cdots;t_n,w_n(d)) \tag{1}$$

其中 t_i 为词条项，$w_i(d)$ 为 t_i 在文本 d 中的权值。$w_i(d)$ 一般被定义为 t_i 在 d 中出现频率 $tf_i(d)$ 的函数，即 $w_i(d)=\psi(tf_i(d))$。

在 VSM 中，TFIDF（Term Frequency Inverse Document Frequency）是一种常见的词条权重的确定方法，它的计算公式如下：

$$w_i(d)=tf_i(d)\times\log(N/n_i) \tag{2}$$

其中 N 为所有文本的数目，n_i 为含有词条 t_i 的文本数目。

3.2　评价函数

3.2.1 文本间相似度和不相似度的定义　在文本检索和文本挖掘中两个文本之间的相似度常用文本特征向量夹角的余弦来度量，即设文本对象 u_h 的特征向量表示为 $V(u_h)=(t_1,w_1(u_h);\cdots;t_i,w_i(u_h);\cdots;t_n,w_n(u_h))$，文本对象 u_k 的特征向量表示为 $V(u_k)=(t_1,w_1(u_k);\cdots;t_i,w_i(u_k);\cdots;t_n,w_n(u_k))$，则文本对象 u_h 和 u_k 之间的相似度为[7]

$$\cos(u_h,u_k)=\sum_{l=1}^{n}w_l(u_h)w_l(u_k)\Big/\sqrt{\sum_{l=1}^{n}w_l(u_h)^2\sum_{l=1}^{n}w_l(u_k)^2} \tag{3}$$

文本对象 u_h 和 u_k 之间的不相似度为

$$\mathrm{dist}(u_h,u_k)=1-\cos(u_h,u_k) \tag{4}$$

3.2.2 评价函数的定义　一个文本聚类模式的质量好坏可以由类内文本对象之间的不相似来确定，它的数学表达为

$$\Delta=\sum_{i=1}^{m}\sum_{u_h,u_k\in X_i}\mathrm{dist}(u_h,u_k) \tag{5}$$

其中 m 为最后形成的聚类的数目。

由于文本聚类最后所形成的类具有模糊的边界，两个对象 u_h 和 u_k 同属于一个类 X_i 有以下 3 种可能：(1) u_h 和 u_k 同时属于 $\underline{A}(X_i)$；(2) u_h 属于 $\underline{A}(X_i)$，而 u_k 属于 $\overline{A}(X_i)$；(3) u_h 和 u_k 同时属于 $\overline{A}(X_i)$。

根据以上 3 种可能，我们可以定义 3 种类型的类内不相似度 Δ_1，Δ_2 和 Δ_3 如下：

$$\Delta_1=\sum_{i=1}^{m}\sum_{u_h,u_k\in\underline{A}(X_i)}\mathrm{dist}(u_h,u_k) \tag{6}$$

$$\Delta_2=\sum_{i=1}^{m}\sum_{u_h\in\underline{A}(X_i),u_k\in\overline{A}(X_i);u_k\notin\underline{A}(X_i)}\mathrm{dist}(u_h,u_k) \tag{7}$$

$$\Delta_3=\sum_{i=1}^{m}\sum_{u_h,u_k\in\overline{A}(X_i);u_h,u_k\notin\underline{A}(X_i)}\mathrm{dist}(u_h,u_k) \tag{8}$$

聚类模式的总的类内不相似度应该是以上 3 种不相似度的加权和：

$$\Delta_{\mathrm{total}}=w_1\times\Delta_1+w_2\times\Delta_2+w_3\times\Delta_3$$

由于 Δ_1 对应着两个对象同时属于某一类这种情形，所以权重 w_1 应该具有最大的权值。相反，Δ_3对应 着两个对象都有可能属于某一类，因此权重 w_3 应该具有最小的权值。换句话说，3 个权重之间应该具有如下关系：$w_1>w_2>w_3$。

基于遗传算法的文本聚类过程就是使评价函数 Δ_{total} 取得近似最小值。

3.3　遗传编码

在此算法中，用来代表每个个体的二进制串由两部分组成，一部分用来表示评价函数中权重参数，另一部分用来表示所有文本对象分别对于每个类的成员关系。

3.3.1 权重参数的编码　在评价函数中，共有 3 个权重参数 w_1，w_2 和 w_3。让 w_1 取固定值 1.0，故可不进行编码表示；对于 w_2 和 w_3 可根据精度要求，分别用一定长度的二进制位串表示，同时根据前面的讨论，只有使 $w_2>w_3$ 成立的位串才算合格的编码。

3.3.2 文本对象和类之间的成员关系编码　设 $U=\{u_1,u_2,\cdots,u_n\}$ 为论域上的所有文本对象集合，将被分成以下 m 类：$U/P=\{X_1,X_2,\cdots,X_m\}$，则这一部分的编码由 n 个基因组成，其中每个对象一个基因，它是一个二进制的位串，用来描述这个对象所属的下近似和上近似。

根据前面所讲的粗集的性质，将用来表示基因的位串分

成两部分：下半部分和上半部分。下半部分和上半部分都分别有 m 位，即进行聚类的类的数目。基因的下半部分（上半部分）中的第 i 位用来说明对象是否属于类 X_i 的下近似（上近似）。

如果 $u_k \in \underline{A}(X_i)$，根据性质 1，有 $u_k \in \overline{A}(X_i)$。于是，基因中的下半部分和上半部分的第 i 位都应该置为 1，而根据性质 2 和 3，基因中的其他位都应该置为 0。

如果 u_k 不属于任何类的下近似，那么根据性质 4，它必然属于两个或两个以上类的上近似，则 u_k 所对应的基因的上半部分的相应位置为 1。

图 1 对类别数目 $m=3$ 时的所有有效基因和部分无效基因进行了举例说明。基因 g1 到 g7 是所有的有效基因。被 g1 所表示的对象属于 $\overline{A}(X_1)$ 和 $\overline{A}(X_2)$。被 g6 所表示的对象属于 $\underline{A}(X_2)$，那么根据性质 1，它也属于 $\overline{A}(X_2)$。除 g1 到 g7 外，其余的所有基因都是无效的。图 1 中也给出了 57 个无效基因中的 4 个。基因 ig1 是无效的，因为一个对象不可能属于 $\underline{A}(X_1)$ 而不属于 $\overline{A}(X_1)$。基因 ig2 是无效的，因为一个对象不可能同时属于 $\underline{A}(X_2)$ 和 $\overline{A}(X_3)$。基因 ig3 是无效的，因为一个对象不可能同时属于 $\underline{A}(X_1)$ 和 $\underline{A}(X_3)$。由于被基因 ig4 所表示的对象只属于 $\overline{A}(X_1)$，根据性质 4 这个基因是无效的。

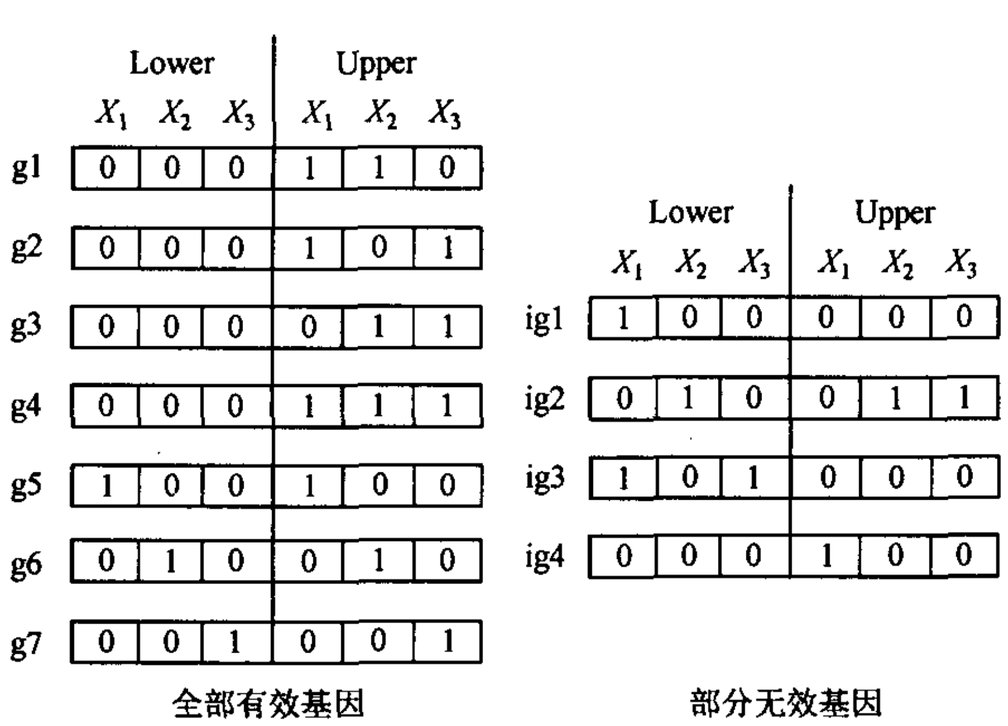

图 1 类别数目 $m=3$ 时的基因举例

3.4 遗传算子的设计

3.4.1 复制 对父代种群中的个体，按照各自的评价函数值在整个种群的个体评价函数值的总和中所占的比例，采用轮盘选择方法进行选择。

3.4.2 交叉 在交叉操作中，以一定的概率 P_c 选择个体参与交换。对于参与交叉的两个父辈个体，交叉点的选择分两种情况讨论：

对于权重参数编码部分，由于编码较短，采用单点交叉，交叉以每个二进制位为单位进行。由于交叉有可能引起 $w_2 > w_3$ 不成立，在这种情况下重新选择交叉点，直到使交叉后满足约束为止。

对于基因编码部分，由于编码较长，采用双点交叉，交叉以每个基因为单位进行。

3.4.3 变异 在变异操作中，以一定的概率 P_m 选择个体进行变异。对于参与变异的父辈个体，权重参数编码部分仍以每个二进制位为单位进行，变异取补运算；对于基因编码部分，仍以每个基因为单位进行，在合格基因组中随机选择一个基因替换原来的基因。

3.5 运算终止条件

事先规定一个最小迭代次数，当迭代次数超过该值后，开始检查每代群体中最优个体评价函数值的变化情况。一旦最优个体评价函数值不再增长或增长的非常缓慢时，即可终止计算。

4 实验研究

我们从 2003 年 5 月到 10 月的人民日报上共抽取了 253 篇文章，其中经济类文章 97 篇政治类文章 84 篇，教育类文章 72 篇。进行分词后，得到文本向量的维数为 2437 维。

为了比较算法的优劣，我们以下面的标准评价聚类结果：一个文本在聚类后仍属于原来所在类的下近似或上近似则认为该文本的聚类结果是正确的。

利用文献[5]中人工给定的权重参数 w_1 =1.0， w_2 =0.5, w_3 =0.25；在遗传算法中交换概率设为 0.35，变异概率设为 0.1，在迭代 500 次后，最终的聚类结果如表 1 所示。通过观察，最后的结果并不理想，因为（1）只有 187 篇聚类正确，文本聚类正确率只有 73.9%。（2）每个文本都精确属于某一类，这与实际情况中一个文本可能同时属于多个类不符。

表 1 利用文献[5]算法得到的聚类结果

结果对比		文本聚类结果						
		只属于经济类	只属于政治类	只属于教育类	同时属于经济类和政治类	同时属于经济类和教育类	同时属于政治类和教育类	同时属于经济类、政治类和教育类
原分类结果	经济类（97）	82	11	4	0	0	0	0
	政治类（84）	15	63	6	0	0	0	0
	教育类（72）	14	19	41	0	0	0	0

表 2　利用本文算法得到的聚类结果

结果对比		文本聚类结果						
		只属于经济类	只属于政治类	只属于教育类	同时属于经济类和政治类	同时属于经济类和教育类	同时属于政治类和教育类	同时属于经济类、政治类和教育类
原分类结果	经济类（97）	41	4	2	33	13	2	2
	政治类（84）	3	37	1	28	2	11	2
	教育类（72）	4	0	24	3	26	14	1

利用本文给出的模糊聚类方法，保持遗传算法中各参数设置不变，限制权重参数 w_2 在 0.1 和 0.9 之间取值，w_3 在 0.05 和 0.85 之间取值，在算法的执行过程中，让权重参数进行自适应优化，在迭代 500 次后，最终 3 个权重参数的取值为 w_1 =1.0，w_2 =0.3135，w_3 =0.1062，聚类结果如表 2 所示。在 253 篇文章中，只有 21 篇文章的聚类结果是错误的，正确率明显提高，为 91.7%。并且有一些文章同时属于多个类，经过阅读这些文章，发现聚类结果与实际情况基本吻合。

5　结论

本文将粗集和遗传算法相结合的方法成功应用于文本模糊聚类；在聚类过程中，将权重参数的设定也通过编码由遗传算法确定，从而使得权重参数的设定具有科学性和可操作性。对于如何确定聚类的数目，是我们下一步研究的内容。

参 考 文 献

[1] 王伟强，高文. Internet 上的文本数据挖掘[J]. 计算机科学, 2000, 27(4): 32 – 37.

[2] Krishnapuram R, Joshi A,Yi L. A fuzzy relative of the *k*-Medoids algorithm with application to web document and snippet clustering[A]. Proc. IEEE Intl. Conf. Fuzzy Systems-FUZZ IEEE 1999[C], Korea. 1999-08, Vol.3: 1281 – 1286.

[3] 李家福，张亚菲，陆建江. 模糊聚类算法在汉语文本聚类中的应用[J]. 计算机工程, 2002, 28(4): 15 – 16.

[4] Pawan Lingras. Unsupervised rough set classification using GAs[J]. *Journal of Intelligent Information Systems*, 2001, 16(3): 215 – 228.

[5] Pawan Lingras. Rough set clustering for web mining. Proc of the 2002 IEEE Conf. on Fuzzy Systems, USA, 2002, Vol.2: 1039 – 1044.

[6] 郭嗣琮. 信息科学中的软计算方法[M]. 沈阳：东北大学出版社, 2001 年 11 月: 263 – 279.

[7] 史忠植. 知识发现[M]. 北京：清华大学出版社, 2002 年 1 月: 359 – 360.

王明春：　男，1971 年生，讲师，博士生，主要研究方向：数据挖掘、文本挖掘.

王正欧：　男，1938 年生，教授，博士生导师，主要研究方向：神经网络、系统辨识、系统优化等.