不确定噪声下海量文本数据的模糊挖掘算法研究

潘大胜

（百色学院 信息工程学院，广西 百色 533000）

摘　要：针对传统的数据挖掘方法一直存在挖掘精度低、运行时间长的问题，提出基于小波变换与关联规则的不确定噪声下海量文本数据的模糊数据挖掘算法，首先利用小波变换对不确定噪声下海量文本数据的模糊数据进行预处理，将模糊海量文本数据时间序列转换至频谱空间中，获得频谱空间内距离最小、类间聚类最大的变换基系数，并将其作为海量文本模糊数据特征系数，利用数据特征系数计算出其从属于各类别的隶属度，确定模糊文本数据集的关联规则，依据多维海量数据集之间的相关程度进行区间划分，由此实现对不确定噪声下海量文本数据的有效挖掘.实验结果证明，所提算法能有效提高海量文本数据挖掘精度，且挖掘效率较高.

关键词：不确定噪声；海量文本数据；模糊数据挖掘算法；特征系数；关联规则

Research on Fuzzy Mining Algorithm for Massive Text Data Under Uncertain Noise

PAN Da-sheng

(School of Information Engineering, Baise University,Baise 533000,China)

Abstract：According to the traditional data mining methods have been mining of low precision, long running time, the wavelet transform and the uncertain fuzzy association rules data text data mining algorithm based on noise, firstly using wavelet transform for uncertain fuzzy data of massive text data noise preprocessing, fuzzy time massive text data sequence conversion to spectrum space, get the distance transform based clustering coefficient, the maximum of the minimum inter class spectrum space, and as a massive text data feature data using fuzzy coefficient, calculate the feature coefficients from membership belonging to the respective categories, determine the fuzzy association rules text data sets, interval division basis the multidimensional degree between the massive data sets, thus the uncertainty of effective mining of massive text data noise. The experimental results show that the proposed algorithm can effectively improve the accuracy of massive text data mining, and the mining efficiency is high.

Key words：uncertain noise; massive text data; the fuzzy data mining algorithm; characteristics of the coefficient of; association rules

作者简介：

潘大胜　男（壮族），（1975-），硕士，副教授.研究方向为数据挖掘技术.

E-mail：bspandsh@163.com.