基于关键点的不确定时间序列线性降维方法

汤其婕,朱小萍
(南京航空航天大学 计算机科学与技术学院,江苏 南京 211106)

摘 要:与确定时间序列相比,不确定时间序列在每个时间点上的取值不是一个确定的值,而是一个可能值的集合,这种不确定给时间数据的降维处理带来了巨大的挑战。加之时间序列固有的数据规模大、数据维度高的特点,对不确定时间序列进行预处理必不可少,现有的针对确定时间序列的降维方法已经不再适用。为解决此问题,建立适当的数据描述统计模型,将原始不确定时间序列归约为三条确定时间序列。同时,针对该模型,提出基于关键点的不确定时序数据线性降维算法。该算法综合考虑体现时序数据特征的极值点与转折点,在进行高效数据降维的同时避免了过度除噪的弊端。实验结果表明,该描述统计模型与基于关键点的线性降维算法的结合具有良好的降维效果,且对于不同领域的数据具有较好的普适性。

关键词:不确定时间序列;描述统计模型;关键点;线性降维
中图分类号:TP311 **文献标识码**:A **文章编号**:1673-629X(2018)08-0022-05
doi:10.3969/j.issn.1673-629X.2018.08.005

A Linear Dimensionality Reduction Method Based on Key Points for Uncertain Time Series

TANG Qi-jie,ZHU Xiao-ping
(School of Computer Science and Technology,Nanjing University of Aeronautics and Astronautics,
Nanjing 211106,China)

Abstract:Compared with traditional time series,the value of uncertain time series at each timestamp is a set of many possible values,which brings great challenges to linear dimensionality reduction for uncertain time series. Considering that uncertain time series data is large-scaled and multidimensional,it is necessary to preprocess raw data before proceeding to the next step. Traditional methods for uncertain time series dimensionality reduction are no longer applicable. To deal with the problem,we propose a descriptive statistical model which reduces the origin uncertain time series into three certain time series. In addition,a new time series data segmentation algorithm is proposed based on the model. The algorithm takes both extreme point and turning point into consideration,which makes efficient data dimensionality reduction while avoiding excessive noise cancellation. Experiment shows that the combination of linear dimensionality reduction method and statistical model has a great effect on dimensionality reduction. Furthermore,the method is also universal for data in different fields.

Key words:uncertain time series;descriptive statistical model;key points;linear dimensionality reduction

0 引 言

时间序列(time series)是一种典型的高维数据类型,也是数据挖掘[1-2]领域中主要研究的对象。一条时间序列是一组序列数据,它通常是在相等间隔的时间段内,依照给定的采样率,对某种潜在过程进行观测的结果。它广泛存在于工业、农业及商业等领域,与人们的生活息息相关。其典型数据包括航天飞船等重要仪器每一时刻的运行状态数据、医疗设备记录的病人每时每刻的心率变化等。研究如何有效地从这些复杂的海量时间序列中挖掘潜在的有用信息,具有重要的理论价值与现实意义[3-4]。

在实际生产生活中,时间序列的产生通常受不确定因素的影响,如数据采集设备的缺陷和环境影响,导致数据采集与实际数据有一定的偏差;或者出于隐私安全考虑,人为地将一定程度的偏差引入到数据中。这些偏差导致时间序列在每一个点上的取值对应一个

收稿日期:2017-09-10 **修回日期**:2018-01-15 **网络出版时间**:2018-04-28
基金项目:国家自然科学基金(61772269)
作者简介:汤其婕(1994-),女,硕士研究生,研究方向为数据与知识工程。

可能值的集合,无法给出其确定值。将这类型的数据定义为不确定时间序列。加之时间序列本身具有的海量、高维等特征,若直接对原始不确定数据进行数据挖掘等操作,效率很低。而解决这一问题最直接的方法就是对原始数据进行预处理操作。但是由于不确定时间序列与确定时间序列在取值上的不同之处,原有用于确定时间序列上的方法则不适用于不确定时间序列。因此,结合已有的针对确定数据的处理方法,找到适用于不确定时序数据的预处理方法,是目前针对不确定时间序列研究的重点,也是文中主要研究的问题。

文中主要研究了基于关键点的线性降维算法与统计模型相结合的不确定时间序列线性降维问题。首先对不确定时间序列进行有效的描述统计建模,将不确定时间序列归约为三条确定的时间序列,进行空间维度的一次降维;然后分别对三条确定时间序列进行关键点的选取,进行时间维度的二次降维;最后,综合空间维度和时间维度的两次降维,得到整个不确定时序数据集的关键点,完成整个降维工作。

1　相关工作

1.1　线性表示的提出

针对确定时间序列的预处理,已有许多成熟的方法,如傅里叶变换[5](DFT)、离散小波变换[6](DWT)、奇异值分解[7](SVD)等等。以上几种方法在某些方面具有明显的优势,但是也存在不足。例如,傅里叶变换、离散小波变换存在一个共同缺点,即过度除噪。消除了局部极值点,造成重要信息遗失,数据表示误差较大;奇异值分解法依赖数据,该方法由于使用数据集产生新的基向量,数据项的任何改变都需要重新计算,时间复杂度大。基于以上几种算法的不足,时间序列线性表示[8-9]方法被提出,其主要目的是有效地保留数据的主要特征,对数据进行高效的降维处理。

1.2　传统时间序列的线性表示

传统的时间序列线性表示方法的主要思想就是用一系列前后相互连接的线段近似表示原始数据,其关键之处在于分段点的选用[10]。线性表示在时间序列的应用上有如下几个优点:理论简单,实现容易;压缩率可以调整。既保留数据的特点,又进行有效的降维。近年来,国内外许多学者对线性表示方法进行了深入研究,提出了许多优秀的线性表示方法。文献[11-12]各自独立提出了分段聚合近似(PAA)的时间序列线性表示方法,该方法的主要思想是首先将时间序列按照相同的时间跨度进行划分,然后以每个时间序列子段的平均值来近似表示相应子段。文献[8]介绍了一种基于重要点的线性表示方法(PLR-IP),即如果一个点在局部区间内与区间端点的比值超过设定的阈值 R,则认为它是重要点。通过调节阈值 R,可以获得不同精度的线性表示。文献[9]介绍了基于特征点的线性表示方法(PLR-FP),该方法的思想是首先提取时间序列的极值点,然后根据每个极值点保持的时间跨度去除噪声点。文献[13]介绍了一种基于斜率提取边缘点的时间序列线性表示方法(PLR-SEEP),该方法的主要思想是首先设定阈值,然后根据斜率变化选取分段点。

2　不确定时间序列描述统计模型建模

不确定时间序列与传统时间序列相比,主要的不同之处在于每个时间点的取值。前者是一个取值的集合,且每一个数值对应该数值在该时间戳发生的概率,后者是一个确定的数值。因此,针对传统时间数据对数值的处理,需要转换为对集合的处理。

定义 1(集中趋势):在描述统计学中,观察值在分布中心位置的聚集现象称为分布的集中趋势,一个分布中心特征的统计度量称为集中趋势的度量。数据集中,平均值是最常用、最具代表性的度量。其中,算数平均值,是整体数学期望的无偏估计。大数定律规定,随着重复次数接近无穷大,数值的算术平均值几乎肯定地收敛于期望值(expected value)。因此,文中选择观察值集合的期望值作为观察值的集中趋势度量,记为 $v_{t,\mathrm{ev}}$,表示不确定时序数据的集中趋势。例如,给定某一时间点的观察值及对应的概率{(5,0.3),(6,0.4),(7,0.2),(8,0.1)},则该点的集中趋势为 5×0.3+6×0.4+7×0.2+8×0.1=6.1。

定义 2(MME-Line 不确定时间序列):长度为 n 的 MM-Line 不确定时间序列由一条包含 n 个元素的序列构成,记为:

$$\begin{aligned}\mathrm{TSU}_{\mathrm{MME-Line}} = \{&([v_{1,\min},v_{1,\max}],v_{1,\mathrm{ev}}),([v_{1,\min},\\&v_{1,\max}],v_{1,\mathrm{ev}}),\cdots,([v_{n,\min},v_{n,\max}],\\&v_{n,\mathrm{ev}})\}\end{aligned}\tag{1}$$

其中,每个元素是一个二元组,由该时刻的观察值区间及其集中趋势组成;t_i 表示第 i 个时刻,序列中所有时刻的最大观察值所构成的序列叫作最大值序列,其相连之后所得曲线叫作最大值曲线,记为 Max-Line。同样,所有最小观察值构成最小值序列,相连后曲线记为 Min-Line;所有集中趋势构成集中趋势序列,记为 EV-Line。例如,不确定时间序列 $\mathrm{TSU}_{\mathrm{MME-Line}}$ = {([2.5,4.2],3.3),([3.6,7.9],5.5),([4.5,9.2],7.6),([3.8,7.2],6.2),([6.0,10.8],8.3),([5.1,9.6],9.6)},所有最大观察值(3.3,7.9,9.2,7.2,10.8,9.6)构成 Max-Line 序列,所有最小值(2.5,3.6,4.5,3.8,6.0,5.1)构成 Min-Line 序列,所有集中趋势(3.3,5.5,7.6,6.2,8.3,9.6)构成 EV-Line

序列。

定义 3(MME-Line 描述统计模型):不确定时间序列描述统计模型包括三条等长的确定时间序列:Max-Line、Min-Line 和 EV-Line。因此,不确定时间序列 MME-Line 的向量形式表示如下:

$$\boldsymbol{X}_{\text{MME-Line}} = \begin{bmatrix} X_{\text{Max-Line}} \\ X_{\text{Min-Line}} \\ X_{\text{EV-Line}} \end{bmatrix} \tag{2}$$

其中,$X_{\text{Max-Line}}$ 为 Max-Line 序列;$X_{\text{Min-Line}}$ 为 Min-Line 序列;$X_{\text{EV-Line}}$ 为集中趋势序列。

由此,对不确定时间序列的降维可以归约为对这三条确定时间序列的降维。

3 基于关键点的不确定时间序列线性降维

3.1 选取关键点

关键点是反映时间序列数据特征的重要点,也是对时序数据进行分段的点,体现了时间序列的轮廓和集中趋势。如图 1 所示,图中 1、2 处的关键点是典型的极值点,其可以通过极值法求出;但 3 ~ 5 处的转折点并不能通过极值法求出,而文献[8]证明,该类型的点在时间序列数据集中也是包含大量重要信息的数据点。文中关键点的选择包括极值点和转折点两部分。

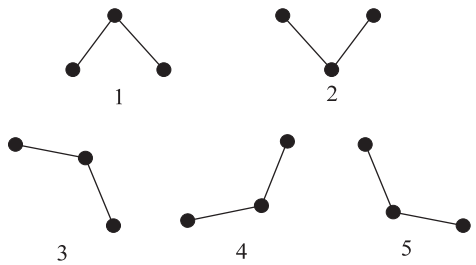


图 1 关键点举例

定义 4(转折点):在不超过序列的最大值和最小值下,由三点组成的简单时间序列,夹角越小,角处的顶点成为转折点的可能性就越大。将该点称为转折点。

定义 5(阈值参数):由三点组成的时间序列,A、O、B,其端点 O 是否为转折点,决定于三点数值是否满足条件 $|\text{Data}(A)+\text{Data}(B)-2*\text{Data}(O)| \geqslant C$。其中 C 即是自定义的阈值参数,其大小决定于所选数据的类型,Data(*)表示在时间点 * 处的数值。

数据的压缩程度可以通过调节阈值参数来设置。阈值参数设置越大,则数据压缩程度越大;反之,数据压缩程度越小,数据分段越精细。

3.2 选取关键点算法

由 3.1 可知,关键点的选择包含极值点和转折点两部分。通过对描述统计模型中的三条确定时间序列分别进行关键点的选取,可以得到三个关键点的集合,分别是 Max-Line 关键点集合、Min-Line 关键点集合和 EV-Line 关键点集合。算法的具体过程如下。

算法:MME-KP

输入:不确定时间序列 TSU = $\{(t_1,V_1,P_1),(t_2,V_2,P_2),\cdots,(t_n,V_n,P_n)\}$,阈值参数 C_1、C_2、C_3,时阈参数 T

输出:三条确定时间序列关键点集合

```
1  double [ ]MaxNum; double [ ]MinNum; double [ ]EVNum;
2  for( i =0; i <1; i ++){
3  sort( V_i );
4  v_{i,ey} = p_{i,1} * v_{i,1} + p_{i,2} * v_{i,2} + ⋯ + p_{i,n} * v_{i,n} ;
5  MaxNum[ i ] = v_{i,max} ; MinNum[ i ] = v_{i,min} ; EVNum[ i ] = v_{i,ey} ;}
6  ArrayList MaxKP; ArrayList MinKP; ArrayList EVKP;
7  for( i =0; i <MaxNum. length; i ++){
8  if(MaxNum[ i ]是极值点){
9  MaxKP. add( i );
10 }else if(MaxKP[ i ]是转折点 && ! MaxKP. contains( i ){
11 MaxKP. add( i );
12 }
13 }//MinKP 和 MaxPV 的计算与 MaxPV 类似
14 MaxKP = getKPArrayList(MaxKP, T); //对关键点进行进一步筛选,使其满足时间跨度 T
15 MinKP = getKPArrayList(MinKP, T);
16 EVKP = getKPArrayList(EVKP, T);
```

3.3 算法分析

MME-KP 算法首先通过提出的描述统计模型将不确定时间序列归约为 3 条确定时间序列,即最大值序列、最小值序列和集中趋势序列,有效提取了数据的三个主要特征。其次,对三条归约得到的确定时间序列分别进行两遍扫描选取关键点。其中关键点包含了极值点和转折点两部分,基本保留了数据的全部特征。同时,该时间序列线性降维算法综合考虑了时间跨度的选择,可以根据不同数据的特点动态设置阈值参数 C 和时阈参数 T,改变数据选取的精细程度和数据的压缩度。综上,该算法既能较好地保留时序数据的特征关键点,同时又能有效降维。算法的复杂度为 $O(n)$,n 为时间序列的长度[14]。

4 实 验

4.1 实验目的和实验环境

为了验证文中提出的基于关键点的线性降维算法与描述统计模型相结合的实际效果,选取了 10 个不同领域的时间序列数据集进行实验。实验以压缩率和拟合误差作为评价优劣的指标,将 MME-KP 算法与 PAA 算法以及 PLR-FP 算法进行对比。

定义 6(拟合误差):给定一组时间序列 $X = \{x_1, x_2, \cdots, x_n\}$,通过对时间序列 X 进行分段取点得分段点集合 X_{PPS},再利用线性插值法对 X_{PPS} 进行填充得到衍生的拟合时间序列,记为 $X^C = < x_1^c, x_2^c, \cdots, x_n^c >$,则拟合序列与原始序列的拟合误差为:

$$E = \sqrt{\sum_{i=1}^{n} (x_i - x_i^c)^2} \tag{3}$$

拟合误差是用来衡量拟合时间序列与原始时间序列差异性的一个重要指标。在同等压缩率的情况下,拟合误差越小,表示拟合效果越好,拟合数据越接近真实数据。

4.2　实验数据处理

确定时间序列是一组按时间先后顺序排列的精确数据,这些数据通过加入扰动成为不确定化数据,由此成为不确定时间序列。给定某一时刻 i,不确定时间序列在该时刻的值可表示为:

$$V_i = d_i + e_i \tag{4}$$

其中,d_i 为确定时间序列 i 时刻的精确值;e_i 为该时刻误差,通常服从某种概率分布。

文中实验数据取自于 www. cs. ucr. edu/~eamonn/tutorials. html 公布的用于数据挖掘的通用实验数据集(KP-Data),如表 1 所示。将每个数据集的训练子集和测试子集重新配置整合,获得实验数据集。同时,通过不确定化模型将不确定性引入到确定序列中,人为加入扰动形成误差 e_i,并使 e_i 服从某种分布,从而将序列转化成不确定时间序列。

表 1　KP-Data 数据集

序列名称	序列长度	序列名称	序列长度
MALLAT	1 024	UWave	945
InlineSkate	1 882	Haptics	1 092
HandOutlines	2 709	Fish	463
CinC_ECG_torso	1 639	Symbols	398
StarLightCurves	1 024	Worms	900

4.3　实验方法

实验主要分为两部分:

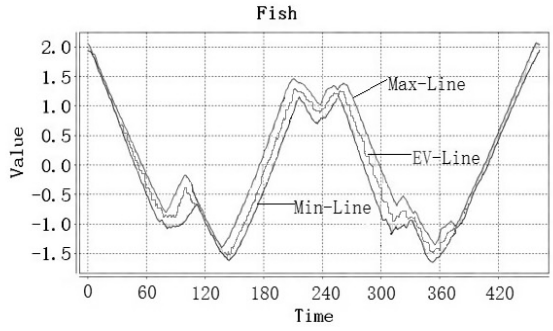
(1)将原始数据进行统计模型建模,每种类型的数据将分别建模得到三条确定时间序列数据:最大值序列、最小值序列和集中趋势序列;

(2)在步骤 1 建立的模型基础上,分别用提出的 MME-KP 算法、Yi 和 Keogh 提出的分段聚集近似(PAA)线性表示方法,以及文献[9]提出的 PLR-FP 线性表示方法对数据分别进行处理,得到整个不确定时序数据集的关键点集合。最终以同种压缩率的标准下拟合误差的大小作为评估算法质量的指标。拟合误差越小,算法性能越好。

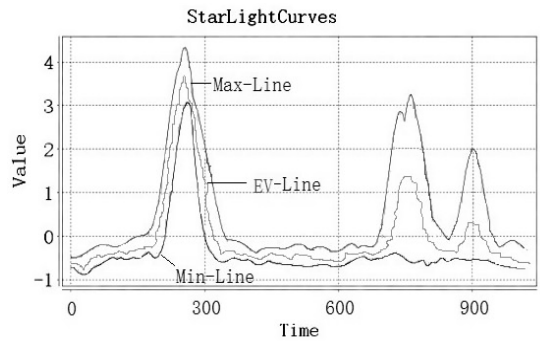
4.4　实验结果与分析

(1)第一部分。

在提出的统计模型下,部分数据的建模表示如图 2 所示。



(a) Fish 原始数据图



(b) StarLightCurves 原始数据图

图 2　部分原始数据建模表示

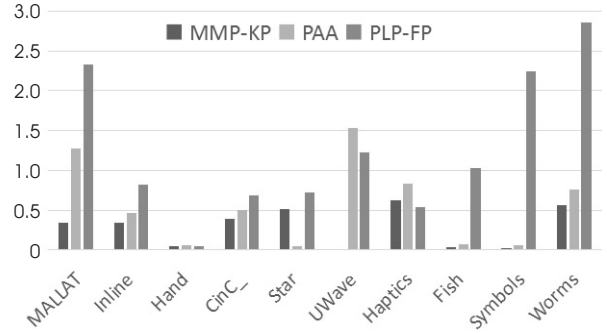
(2)第二部分。

文中提出的 MME-KP 需要输入 3 个阈值参数:C_1、C_2 和 C_3,分别对应三条确定时间序列,即最大值时间序列、最小值时间序列及集中趋势时间序列的关键点阈值参数。参数设置越大,数据的压缩率越小,拟合误差越大;参数设置越小,数据压缩率越大,数据分段越精细,拟合误差越小。实验选取的数据来自 10 个不同领域,数据的差异性较大。为了对比实验的公平性,将根据不同数据类型调整参数,保证在 10 组数据的压缩率都在 50% 的基准下,进行三种算法的拟合误差对比。在同一压缩率情况下,拟合误差越小,算法性能越好。

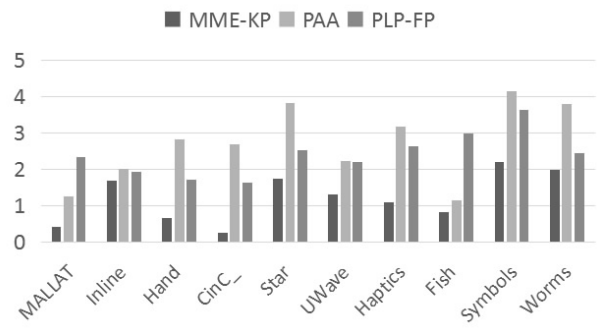
部分实验结果如图 3 所示。

从图 3 中可以看出,在 10 个通用时间序列数据集中,MME-KP 算法在其中 7 个数据集上 Max-Line 拟合序列的误差最小,在另外三个数据集中,误差与最好的算法相当。同时,由图(b)可以看出,在 EV-Line 上的拟合误差,MME-KP 算法明显优于其他两种算法,在 10 个数据集中都保持优势。实验结果说明,采用了极值点与转折点相结合的选取关键点算法,在数据降维上具有明显优势。因为其在保留了极值点这一数据的明显特征外,还适当选取了体现数据特征的转折点,

减小了一些重要点被粗暴舍弃的概率,提高了数据选取的精细度,从而降低了拟合误差。



(a)三种算法在 Max-Line 上拟合误差的比较

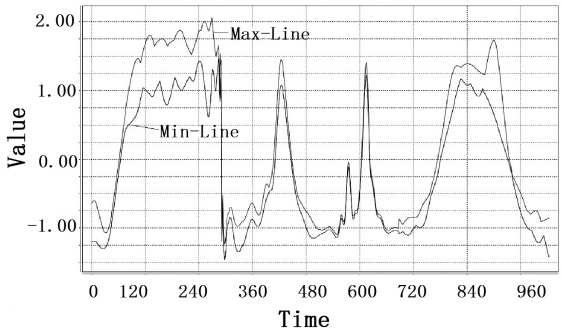


(b)三种算法在 ME-Line 上拟合误差的比较

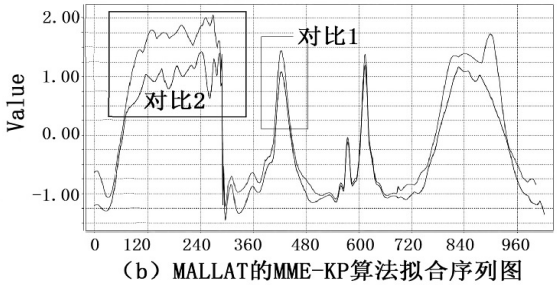
图 3 实验结果

同时实验结果显示:对于短时间内波动频率比较平缓的时间序列,MME-KP 算法的实验效果与其他两种算法相比,优势并不是很大,如图 4 中的对比 1 所示,MME-KP 算法和 PAA 算法都能较好地拟合 MALLAT 的原始数据。但是对于短时间内波动频率剧烈的时间序列,如图 4 中的对比 2 所示,数据在短时间内波动较大,MME-KP 算法的拟合效果就远胜于 PAA 算法。主要原因是,在短时间内,时间序列波动较大,会产生大量极值点,对于以找极值点和转折点为主要关键点的 MME-KP 算法来说会捕获大量关键点,从而保证了数据主要特征的不流失。而对于 PAA 算法,它主要思想是以一段时间内的数据均值来代替这段序列,所以,在短时间内时间序列波动较大的情况下,在这段时间内的数据均值波动并不会很大,相对地,它的拟合序列与原序列相比,就会丢失很多重要点信息,造成拟合误差较大。

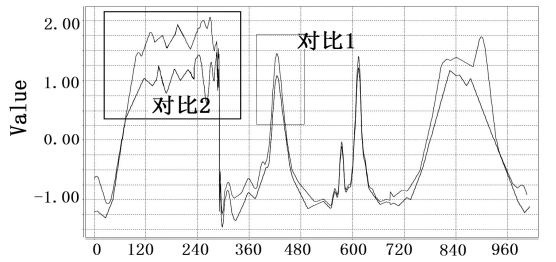
综上,MME-KP 降维方法在与提出的描述统计模型的结合上,有以下几点优势:在降维效果上,可以通过参数的改变,进行不同压缩率的降维。参数设置越高,数据压缩程度越大。在现实应用中可以根据实际需求选择所需的降维效果;在与其他线性降维方法 PAA 以及 PLR-FP 的比较过程中,MME-KP 方法体现出了更为良好的拟合效果,误差更小,与原始数据保持高度的



(a) MALLAT原始序列图



(b) MALLAT的MME-KP算法拟合序列图



(c) MALLAT的PAA算法拟合序列图

图 4 实验对比

5 结束语

针对不确定时序数据区别于确定时序数据的主要特征,首先提出不确定时间序列向确定时间序列归约的描述统计模型,同时综合考虑已有的确定时间序列的线性分段思想,提出综合考虑极值点与转折点的关键点选取算法。综合以上两点,将不确定时序数据进行空间和时间上的有效降维。实验结果表明,该算法在数据降维效果上有良好的性能,且能高度还原原始数据。下一步的研究工作是在所提出模型的基础上找到更合适的原始数据拟合方案。

参考文献:

[1] 毛国君,段立娟. 数据挖掘原理与算法[M]. 第 3 版. 北京:清华大学出版社,2016.

[2] 潘 定,沈钧毅. 时态数据挖掘的相似性发现技术[J]. 软件学报,2007,18(2):246-258.

[3] 王光宏,蒋 平. 数据挖掘综述[J]. 同济大学学报:自然科学版,2004,32(2):246-252.

[4] 钟 晓,马少平,张 钹,等. 数据挖掘综述[J]. 模式识别与人工智能,2001,14(1):48-55.

(下转第 31 页)

和传统 TFIDF 算法的分类效果要好。召回率、正确率、*F* 值、精确率都有一定的提高,这也证明了改进后的 TFIDF 算法更有效。

3　结束语

TFIDF 特征权值计算方法在垃圾邮件过滤中被大量使用,但是仍然存在一些不足之处。文中对其没有考虑到特征词在类中分布情况、没有考虑到特征词在另一类的出现频率、低估了频繁出现的特征词的权值并高估了出现频率低的特征词的权值这三方面进行改进。以朴素贝叶斯算法作为分类器,分别计算测试邮件是正常邮件和垃圾邮件的概率。最终以召回率、正确率、*F* 值、精确率作为评估指标,将改进后的 TFIDF 和 CHI 统计特征提取、传统的 TFIDF 进行比较。通过实验数据进行分析,说明改进后的 TFIDF 算法是有效的。

参考文献:

[1] 薛正元. 基于改进贝叶斯决策的邮件过滤[J]. 计算机工程与应用,2013,49(7):98-101.

[2] 杜　选. 基于加权补集的朴素贝叶斯文本分类算法研究[J]. 计算机应用与软件,2014,31(9):253-255.

[3] 卢宏涛,张秦川. 深度卷积神经网络在计算机视觉中的应用研究综述[J]. 数据采集与处理,2016,31(1):1-17.

[4] 奉国和,吴敬学. KNN 分类算法改进研究进展[J]. 图书情报工作,2012,56(21):97-100.

[5] 石铁峰. 支持向量机在电子邮件分类中的应用研究[J]. 计算机仿真,2011,28(8):156-158.

[6] TONG S. Support vector machine active learning with applications to text classification[J]. Journal of Machine Learning Research,2008,2(1):45-66.

[7] 张保富,施化吉,马素琴. 基于 TFIDF 文本特征加权方法的改进研究[J]. 计算机应用与软件,2011,28(2):17-20.

[8] JOACHIMS T. A probabilistic analysis of the rocchio algorithm with TFIDF for text categorization[C]//Proceedings of the fourteenth international conference on machine learning. [s. l.]:Morgan Kaufmann Publishers Inc. ,1996.

[9] ZHANG Wen, YOSHIDA T, TANG Xijin. TFIDF, LSI and multi-word in information retrieval and text categorization [C]//IEEE international conference on systems, man and cybernetics. Singapore:IEEE,2009:108-113.

[10] 李学明,李海瑞,薛　亮,等. 基于信息增益与信息熵的 TFIDF 算法[J]. 计算机工程,2012,38(8):37-40.

[11] 任永功,杨荣杰,尹明飞,等. 基于信息增益的文本特征选择方法[J]. 计算机科学,2012,39(11):127-130.

[12] 刘庆和,梁正友. 一种基于信息增益的特征优化选择方法[J]. 计算机工程与应用,2011,47(12):130-132.

[13] 徐峻岭,周毓明,陈　林,等. 基于互信息的无监督特征选择[J]. 计算机研究与发展,2012,49(2):372-382.

[14] 刘海峰,苏　展,刘守生. 一种基于词频信息的改进 CHI 文本特征选择[J]. 计算机工程与应用,2013,49(22):110-114.

[15] 程克非,张　聪. 基于特征加权的朴素贝叶斯分类器[J]. 计算机仿真,2006,23(10):92-94.

[16] 王行甫,杜　婷. 基于属性选择的改进加权朴素贝叶斯分类算法[J]. 计算机系统应用,2015,24(8):149-154.

(上接第 26 页)

[5] AGRAWAL R, FALOUTSOS C, SWAMI A. Efficient similarity search in sequence databases[C]//International conference on foundations of data organization and algorithms. [s. l.]:Springer-Verlag,1993:69-84.

[6] STRUZIK Z R, SIEBES A. Wavelet transform in similarity paradigm[C]//Pacific-Asia conference on research and development in knowledge discovery and data mining. [s. l.]: Springer-Verlag,1998:295-309.

[7] KORN F, JAGADISH H V, FALOUTSOS C. Efficiently supporting ad hoc queries in large datasets of time sequences [J]. ACM SIGMOD Record,1997,26(2):289-300.

[8] PRATT K B, FINK E. Search for patterns in compressed time series[J]. International Journal of Image and Graphics, 2001,2(1):89-106.

[9] XIAO Hui, FENG Xiaofei, HU Yunfu. A new segmented time warping distance for data mining in time series database [C]//Proceedings of 2004 international conference on machine learning and cybernetics. Shanghai:IEEE,2004:1277-1281.

[10] 刘世元,江　浩. 面向相似性搜索的时间序列表示方法述评[J]. 计算机工程与应用,2004,40(27):53-59.

[11] KEOGH E J, PAZZANI M J. A simple dimensionality reduction technique for fast similarity search in large time series databases[C]//Pacific-Asia conference on knowledge discovery and data mining. [s. l.]:Springer,2000:122-133.

[12] YI B K, FALOUTSOS C. Fast time sequence indexing for arbitrary Lp norms[C]//Proceedings of the 26th international conference on very large data bases. [s. l.]:Morgan Kaufmann Publishers Inc. ,2000:385-394.

[13] 詹艳艳,徐荣聪,陈晓云. 基于斜率提取边缘点的时间序列分段线性表示方法[J]. 计算机科学,2006,33(11):139-142.

[14] 刘永志,皮德常,陈传明. 基于关键点的不同长度时间序列相似性度量[J]. 计算机工程与应用,2014,50(20):1-4.