一种非均匀分布数据的非线性标准化方法

梁 路 黎 剑 霍颖翔 滕少华
(广东工业大学计算机学院 广州 510006)

摘 要 传统的数据标准化处理通常采用的是线性的变换方法,其在处理非均匀分布的数据集时,容易因局部区间内数据点间距过小导致后续的数据挖掘(尤其是基于距离的挖掘)结果不够精确。因此,为非均匀分布数据提出一种基于数据拟合的非线性变换标准化方法,该方法能够在不改变数据整体分布规律的前提下,依据统计找出对应的非线性变换函数,根据函数对各数据点的取值进行非线性放缩,将数据稠密的区间进行扩大的同时将数据稀疏的区间进行压缩,让挖掘的结果更加精确。实验采用BP(Back Propagation)神经网络、支持向量机(Support Vector Machine,SVM)、最近邻分类(K-Nearest Neighbor,KNN)3种经典分类算法结合不同的数据集进行了挖掘,结果表明,分类的错误率有不同程度的下降,同时 F_1 度量有所提高。

关键词 非均匀分布,非线性标准化,数据预处理
中图法分类号 TP301.6 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2016.4.054

Nonlinear Normalization for Non-uniformly Distributed Data

LIANG Lu LI Jian HUO Ying-xiang TENG Shao-hua
(School of Computer Science and Technology, Guangdong University of Technology, Guangzhou 510006, China)

Abstract Traditional normalization method for continuous attributes is usually a linear transformation. When using linear normalization to deal with some non-uniform datasets, it's easy to cause the subsequent data mining (particularly some mining methods based on distance) results are inaccurate enough for the interval of each data point in the local space is too small. This paper suggested a nonlinear normalization based on data fitting, and we could find out the corresponding nonlinear transformation function in the premise of not changing the distribution rules of data. According to the function, we could nonlinearly zoom the data interval, expand the interval of dense data and shrink the interval of sparse data at the same time. It can make the data mining more accurate. We used the neural network, SVM and KNN combining with different data set to test. The results show that the error rate decreases and the F_1 measure increases at the same time.

Keywords Non-uniform distribution, Nonlinear normalization, Data preprocessing

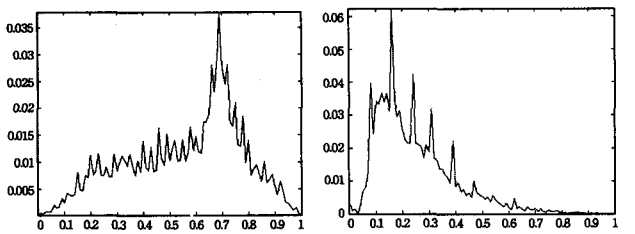
1 引言

当前,随着数据挖掘的各类模型逐渐趋于稳定与成熟,而现实世界中的原始数据集大多具有不准确、不完整、不一致等特性,如何合理、高质量地进行数据预处理以得到更优的样本输入,已经成为提高数据挖掘模型的效率与准确度的关键[1,2]。在一次完整的数据挖掘过程中,数据预处理所耗费的工作量有时可以占到总工作量的60%以上,而此时进行挖掘的工作量仅有10%左右[3]。在数据预处理技术中,数据标准化是一种常用的对数据进行无量纲化处理的方法,目的在于赋予所有属性相同的权重,对基于距离度量的数据挖掘模型有重要作用[4]。尤其是对于神经网络、支持向量机等模型,数据集中的属性存在不同的取值范围往往会导致模型的不稳定[5]。而对数据进行标准化处理后,不仅能加快模型的训练速度,还能提高模型的精确度[6]。因此,对不同取值范围的属性进行标准化是非常必要和重要的[7]。

关于数据标准化的研究,目前存在大量的依据数据的特性来处理的方法。按照映射方式的不同,可以划分为两类:线性标准化和非线性标准化[8]。过去在数据预处理阶段主要采用的是一些线性的标准化方法,这些方法简单有效,共同特点是通过某种线性变换将数据映射到一个较小的区间[0,1]或者[-1,1]之间,如Min-Max标准化、Z-Score标准化法与小数定标法[9]等。然而,实际的生产生活当中存在着帕累托法则(又称二八定律),它揭示了数据分布不平衡这一现象广泛存在于各个领域中,几乎所有未经处理的原始数据都存在着分布不均匀的现象。图1为两种常见的数据非均匀分布。

到稿日期:2015-03-26 返修日期:2015-07-24 本文受国家863计划重大项目(2013AA01A212),国家自然科学基金资助项目(61272067,61104156,61402118),广东省自然科学基金(9451009001002777)资助。

梁 路(1980-),女,博士,副教授,CCF会员,主要研究方向为数据挖掘、协同计算,E-mail:lianglu@gdut.edu.cn;黎 剑(1989-),男,硕士生,主要研究方向为数据挖掘、机器学习,E-mail:leejian1989@foxmail.com;霍颖翔(1989-),男,主要研究方向为机器学习、声纹处理;滕少华(1962-),男,博士,教授,主要研究方向为大数据、数据挖掘与协同计算、网络安全,E-mail:shteng@gdut.edu.cn。



(a)常见数据非均匀分布 1　　(b)常见数据非均匀分布 2

图 1

当数据分布较为集中即稠密区间数据点之间的相对距离过小时，数据挖掘算法(尤其是对距离较为敏感的算法)无法将数据点精确区分开。此时因识别精度的问题，在稠密区不易划出准确的分类面；同时，分类面的细小误差能轻易带来大量的错误识别。对于这一问题，线性的映射并不会改变数据的分布状况，即如果在原样本空间下数据间距过小，经线性标准化之后，数据点相对间距依旧过小；并且标准化处理后由于值域的变换，数据点在数据稠密空间的间距的实际值会更小，这对于基于距离度量的挖掘模型无疑是非常不利的。因此，一种可以在不改变数据相对分布规律的前提下让非均匀分布的数据趋向于均匀分布的非线性变换方法，将能改善大多数基于距离度量的挖掘模型的效果。这种标准化方法不仅具有传统标准化方法所具有的无量纲化处理、加快模型训练速度等作用，还能更多地起到提高模型精确度的作用。在大多数的实际应用中，它能够替代传统的线性标准化方法。

传统的标准化方法中也存在一些非线性的标准化方法，它们大多是专家根据领域专业知识设计的函数，主要有 Logistic 标准化和模糊标准化[7]，其函数公式如下。

Logistic 标准化：

$$F(x)=\frac{1}{1+e^{-x}}$$

模糊量化标准化：

$$F(x)=\frac{1}{2}+\frac{1}{2}\sin\left(\frac{\pi}{\max(x)-\min(x)}\times\frac{x-(\max(x)-\min(x))}{2}\right)$$

这些方法的思路是通过特定的非线性的函数变换将数据映射到新的空间下，适用于非线性数据[10]，可以在一定程度上实现对数据的非线性放缩。但其变换函数依赖于专家经验而设计，并且其目的在于对指定区间的数据进行放缩，而不是根据数据样本实际的分布做出相应的调整，因此这一类非线性变换函数只能处理特定分布(尤其是有特定领域背景)的一类数据；然而在处理实际放缩需求与函数的放缩区间不一致的数据时，往往容易引起整体数据分布规律的变化进而导致数据挖掘的效果下降。

针对以上问题，本文提出了一种基于数据拟合的非线性标准化方法，其特点是能够根据数据样本的分布动态地确定区间放缩的程度，解决了线性标准化无法改变数据稠密分布的问题，同时又比传统非线性标准化方法的适用范围更广。首先，基于数据的频率直方图构建一个函数映射，对非均匀分布的原数据进行非线性变换，将数据稠密的区间适当拉伸扩大，增大稠密区间的数据点间的距离，压缩稀疏区间的数据点间的距离。以降低挖掘算法在数据稀疏区的识别精度为代价，换取数据稠密区识别精度的提高，从而从整体上让基于距离度量的挖掘算法的挖掘结果更加精确。这是一种无监督的、对无标签数据与有标签数据都适用的方法。进一步地，对有标签数据提出一种有监督的非线性标准化方法，通过选取 Gini 指数对数据区间的不纯度进行统计，进而对数据不纯的区间适当拉伸，提高分类算法的识别效果。最后，为了证明该方法有助于改进数据挖掘(尤其是基于距离度量的模型)的结果，同时选取了 BP(Back Propagation)神经网络、支持向量机(Support Vector Machine，SVM)、最近邻分类(K-Nearest Neighbor，KNN)等 3 种基本的、有代表性的、同时又对标准化有一定要求的挖掘算法作为实验模型。在此基础上，将本文方法与主流的线性标准化方法和非线性标准化方法进行了对比实验，结果表明，本文方法能够有效提高非均匀分布的数据集的识别精确度。

2　基于数据拟合的非线性标准化

本文提出的基于数据拟合的非线性标准化是在数据的每一个维度上分别进行标准化，能够根据数据样本分布来确定区间的放缩程度，因此适用于大多数的非均匀分布的数据。其主要思想是通过分布统计指标对数据进行按列统计得到分布曲线，并对该曲线进行积分，得到单调递增的非线性变换函数，数据通过该函数映射进行非线性标准化。

本方法主要包括以下几个步骤：

- 线性标准化
- 分布统计
- 高斯模糊
- 规范化与积分
- 函数拟合

同时，按照分布统计指标有无使用类标签信息，这种方法可以划分为无监督的非线性标准化与有监督的非线性标准化两类。这两种方法在步骤上一致，仅在分布统计指标的选择上存在差别。

2.1　线性标准化

不考虑缺失值的影响，首先，统计原数据的区间范围，按照线性标准化方法对原数据进行标准化。考虑对数据先采用线性标准化处理是因为当原始数据的值域取值过大时，会对整个方法过程的计算造成一定困难，并且可能导致最后拟合函数的过程难以收敛。因此，可以先采用线性的标准化方法(例如 Min-Max 标准化)将数据的值域压缩到[0，1]区间，以便于后续计算。

2.2　分布统计

想要得到一种能够根据数据分布的稠密程度确定区间放缩程度的方法，就必须先对数据的分布进行一个粗略的统计，确定稠密的区间以及稠密的程度。在分布统计的过程中，首先要选择适当的采样点数 K。对于 K 的设定，采样点数越多，分布曲线越细致，采样精度越高，但也会造成计算量过大的问题；而采样点数过少，则会引起对数据分布统计的失真，进而影响最终的标准化效果。因此，在实际的应用中，应该根据样本的数量来设定采样点数。经多次实验，综合考虑计算量在可接受的范围内，同时又能够大致地反映数据分布曲线，将 K 设定在样本数量的 $\frac{1}{100}$ 到 $\frac{1}{10}$ 得到的效果较好，按照采样点数 K，将区间划分成 K 个子区间，依据统计指标统计数据点落在这些子区间的频数，并计算指标数值所占总的百分比。对于指标的选择，有频数和 Gini 指数两种。

2.2.1 无监督非线性标准化

无监督的非线性标准化方法是一种基本的非线性变换数据标准化方法，它以数据点的频数作为统计指标，而不需要考虑数据点的类标签。它在数据集类标签缺失或者无标签数据集聚类预处理时有重要的作用。

定义分布曲线 $f(x)$ 如下：

$$f(x_D)=\frac{x_D}{m},D=1,2,\cdots,K$$

其中，x_D 为在区间 D 内的数据点数，m 为总的数据点数。由此得到一个数据的分布图，图 2 所示为 $K=1000$ 的数据分布曲线。

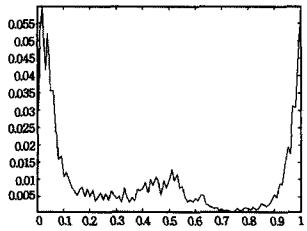


图 2 分布曲线

2.2.2 有监督非线性标准化

在有标签数据集以及分类问题的处理上，可能面临如图 3 所示的问题。假设有 A、B 两类数据，A 类数据主要分布在数据稠密区间，B 类数据分布在数据稀疏区间，其分类界线可能存在于数据稀疏的区间，这时如果采用常规的无监督非线性标准化压缩数据稀疏的区间，则不利于分类器的构建。进一步，针对此类问题，本文提出了有监督的非线性标准化方法。

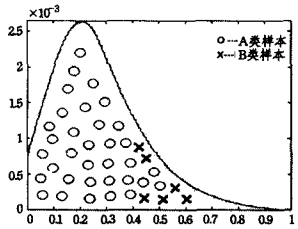


图 3 特殊分布曲线

有监督非线性标准化方法在对数据进行非线性变换时，考虑数据的标签信息，引入信息论中的 Gini 指数[4]。Gini 指数是一种定义区间不纯度的指标，其定义如下：

$$Gini(D)=1-\sum_{j=1}^{n}p_j^2$$

其中，p_j 为类 j 在区间 D 内的频率。以 Gini 指数作为分布曲线 f 的统计指标，在分类问题上情况会有很大的不同。仍然以图 2 中的数据为例，对其进行 Gini 指数的统计，当 $K=1000$ 时得到如图 4 所示的结果。

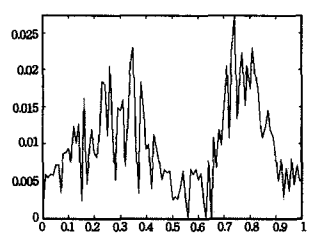


图 4 Gini 指数统计分布图

可以看出，数据虽然聚集在靠近 0 和 1 的两端，但是恰巧在数据稀疏的区间有较大的不纯度，也就是说，在面对分类问题时，采用无监督非线性标准化的方法对数据稀疏的区间进行压缩有时可能是不合理的；相反，我们必须对不纯度较高的区间进行拉伸，而不是仅仅考虑数据聚集的区间。

2.3 高斯模糊

在粗略的分布统计后，从分布曲线可以看出，曲线在稠密区的某些子区间上存在间断，这是由样本数量有限而采样精度过高造成的，此时部分子区间上没有数据点并不代表实际的数据不会分布在这些区间上，而不光滑连续的曲线不仅给函数拟合带来困难，还会造成区间放缩的距离过小，效果不明显等后果。对于这种情况，可以适当忽略子区间的落差，只关注这些小区间组成的局部区间的整体放缩需求。因此，必须做一次插值操作，使曲线光滑连续，更贴近实际。为了让曲线光滑连续，本文的具体做法是对曲线进行高斯模糊。高斯模糊的参数选择必须使曲线光滑连续同时又能保证与频数图尽量拟合，μ 与 sigma 的设定由实验确定，其将对整个方法的最终效果起到决定性作用，高斯模糊光滑后得到的曲线 f 如图 5 所示。

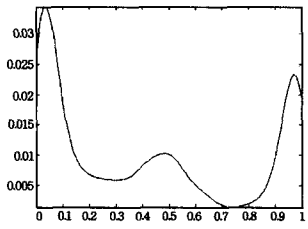
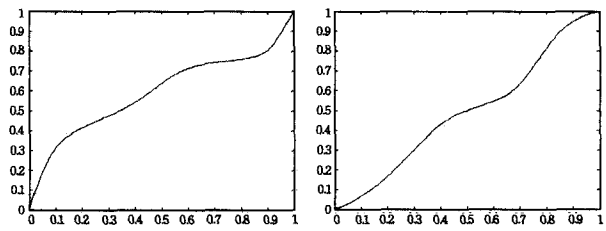


图 5 高斯模糊后的分布曲线

2.4 规范化与积分

将分布曲线高斯模糊之后，初步得到了一条光滑连续并且能大致反映数据分布规律的曲线 f。由于希望得到的非线性映射的函数的值域是[0,1]，因此必须进一步对曲线 f 进行规范化处理，即对曲线 f 每一个取值点除以所有取值点的累加和。然后，对曲 f 线进行积分，最后得到一个定义域与值域同为[0,1]的非线性变换曲线 F。曲线 f 反映了数据在各个区间的稠密程度，而曲线 F 则根据区间的稠密程度反映了数据在区间内拉伸的幅度，曲线 F 即最终得到的非线性变换曲线。根据曲线 x 值与 y 值的映射关系，可以把原数据从 x 轴空间映射到 y 轴空间，从而实现根据数据稠密区间以及稠密程度确定的非线性变换的数据标准化。图 6(a)与图 6(b)所示分别为由无监督与有监督方法得到的非线性变换曲线。



(a)无监督非线性变换曲线　　(b)有监督非线性变换曲线

图 6

2.5 函数拟合

由于前面的步骤在实际操作中是以 PCM(Pulse Code Modulation)方式近似处理的，得到的曲线 F 实际上带有锯齿，并不是一条真正光滑的曲线。在实际的应用中，我们希望

数据是通过具体的函数表达式进行标准化的。因此，最终采用最小二乘法产生与 F 相似的高阶多项式函数来得到一条合理、光滑的形变函数。

值得注意的是，在采用最小二乘法对数据进行拟合时，必须检查其与实际数据的均方误差。如果误差过大，必须增加拟合多项式的阶数来减小误差，考虑到算法的效率，不建议采用 10 阶以上多项式进行拟合。并且对于这样一个函数，如果 10 阶多项式拟合依然无法将误差控制在 0.01 以内，则需考虑是否是发生了“龙格现象”[11]。并非所有函数都能用插值多项式进行良好的逼近。在出现“龙格现象”的情况下，可考虑采用分段线性插值、Spline 样条插值或贝赛尔曲线等替代方法。另外，线性分段插值效率最高，且误差范围可控，所以在数据量庞大或者对计算效率有要求时，可优先采用此法。

2.6 复杂度分析

本文所提出的基于数据拟合的非线性标准化方法在计算开销上较常规的线性预处理方法更高，统计分布的时间复杂度为 $O(N)$，高斯模糊与最小二乘法的计算开销取决于分布曲线的采样精度，而与实际样本数据的量无关。并且以上步骤所耗费的计算开销仅在第一次计算产生，后续的数据可按照分析得到的函数进行变换。

3 实验与分析

3.1 数据集及预处理

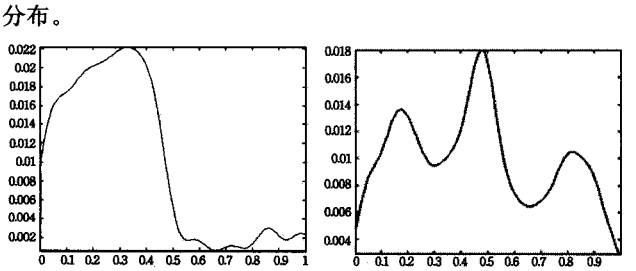
本文选取来源于 UCI 的公开数据集 Abalone、Breast Cancer(BC)、Spambase、Wall Follow Robot(WFR)作为实验数据集。数据集详细信息如表 1 所列。

表 1 数据集详细信息

	样本数	维度	类别
Abalone	4177	8	3
BC	699	10	2
Spambase	4601	57	2
WFR	5456	30	4

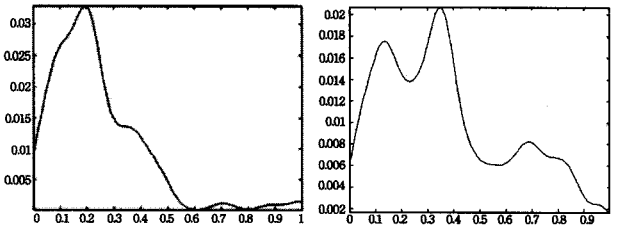
数据集不存在缺失值，但都存在轻微的样本不平衡现象。对数据按维度进行统计分析，发现均存在分布不均匀现象。由于线性的标准化方法并不会改变分布，因此选取线性标准化方法中常用的 Min-Max 标准化作为比对。以 WFR 数据集为例，图 7(a)和图 8(a)为采用 Min-Max 标准化方法对其中一个维度进行标准化之后的频数统计与 Gini 统计。

分别采用本文提出的非线性标准化方法对数据集进行预处理，图 7(b)为无监督非线性标准化方法处理之后的统计分布，图 8(b)为有监督非线性标准化方法处理之后的 Gini 统计分布。



(a)处理前频数统计分布　(b)处理后频数统计分布

图 7



(a)处理前 Gini 统计分布　(b)处理后 Gini 统计分布

图 8

3.2 实验

选取线性标准化方法中的 Min-Max 标准化、非线性标准化方法中的 Logistic 标准化和模糊标准化(Fuzzy)分别对数据集进行标准化预处理，采用 BP 神经网络、SVM、KNN 3 种不同的分类算法对数据集进行训练，实验重复多次，每次实验均随机抽取 63.2%的数据作为训练集，余下的数据作为测试集。所训练的分类器的性能根据数据挖掘模型的不同而有所不同。主要的考察指标为分类器的错误率，由于本文所采用的测试数据集存在一定程度的样本不平衡现象，因此引入在样本不平衡情况下能反映分类器性能的 F_1 度量作为评价指标。

3.2.1 BP 神经网络

对于 BP 神经网络分类器的性能，除了考虑分类器总体的错误率以及对于各类样本的 F_1 度量外，同时还必须考虑训练时分类器能否快速收敛。因此，经本文所提出的非线性标准化方法对数据进行预处理后，综合考察 BP 神经网络分类器的总体错误率、F_1 度量、均方差等指标。由于 BP 神经网络本身的随机性质可能给实验结果带来误差，我们取多次实验数据的平均值进行比较，以期得到一个较为客观的比较结果。BP 神经网络的错误率实验结果如表 2 所列。

表 2 BP 神经网络错误率分析

	Abalone	BC	Spambase	WFR
Min-Max	34.04	2.83	6.99	4.51
Logistic	35.74	3.94	6.41	4.38
Fuzzy	35.03	3.21	23.89	7.18
无监督	**33.51**	2.95	6.58	**1.80**
有监督	33.84	**2.78**	**6.00**	3.94

实验结果表明，对于 BP 神经网络，相比于传统的线性和非线性标准化方法，本文方法能够将误分类的样本数量最大降低约 50%。多次实验，取各类样本的 F_1 度量的平均值总和来考察分类器在样本不平衡情况下的性能，如表 3 所列。

表 3 BP 神经网络 F_1 度量分析表

	Abalone	BC	Spambase	WFR
Min-Max	1.9616	1.9379	1.8536	3.7980
Logistic	1.9024	1.9141	1.8654	3.8093
Fuzzy	1.9233	1.9294	1.7014	3.6720
无监督	**1.9710**	1.9355	1.8616	**3.9148**
有监督	1.9683	**1.9393**	**1.8738**	3.8093

实验数据显示，在本文方法标准化后的数据集上构建的 BP 神经网络分类器在 F_1 度量上有 0.01 到 0.1 不同程度的提高，也就是说 BP 神经网络分类器对各类样本的召回率和精度均有不同程度的提高，说明了本文方法能够让 BP 神经网络的挖掘结果更加精确；而能否快速收敛到精确的结果也是 BP 神经网络的一个重要方面。我们在 WFR 数据集上进行了实验，图 9 显示了 BP 神经网络某次训练时在线性标准

化处理的数据集和非线性标准化处理的数据集上的收敛过程。结果表明，BP 神经网络在非线性标准化处理过的数据集上能够更快地收敛，同时也能达到更低的均方误差。

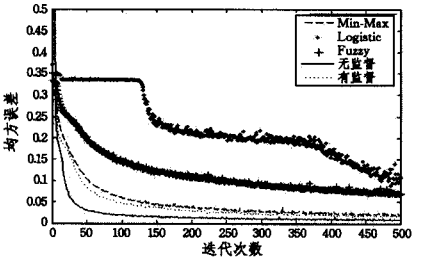


图 9　BP 神经网络收敛过程

3.2.2　SVM

在 SVM 算法中，对数据进行标准化有助于参数的选取和加快训练速度。因此，我们采用线性与非线性等 3 种不同的标准化对数据进行处理并进行实验。本文采用了 SVM 算法中较为常用的 C-SVM 分类算法进行训练，根据 5-折交叉验证选取合适的参数，选取在不同数据集分类的错误率上表现最好的线性核作为核函数。经多次实验，采用 SVM 算法的分类错误率与 F_1 度量如表 4 和表 5 所列。实验结果表明，经本文方法处理后，SVM 在错误率以及 F_1 度量上均有不同程度的改善。

表 4　SVM 分类错误率分析表

	Abalone	BC	Spambase	WFR
Min-Max	36.68	3.02	10.46	11.70
Logistic	41.56	3.25	**6.79**	10.91
Fuzzy	40.57	3.31	37.68	23.70
无监督	**35.41**	**2.80**	6.91	**3.39**
有监督	35.84	2.90	7.70	8.81

表 5　SVM 的 F_1 度量分析表

	Abalone	BC	Spambase	WFR
Min-Max	1.8821	1.9336	1.7754	3.4513
Logistic	1.6491	1.9292	**1.8567**	3.5182
Fuzzy	1.7494	1.9269	0.8675	2.9427
无监督	**1.8934**	1.9388	1.8537	**3.8403**
有监督	1.8946	**1.9364**	1.8376	3.5748

核函数的选择对于提高 SVM 的分类准确率有重要的作用，选取 WFR 数据集对在不同的核函数选择下非线性标准化方法的表现进行了实验。实验结果如图 10 所示，在线性核、多项式核、RBF 核和 sigmoid 核上，本文方法的表现从整体上都较其他标准化方法更好。

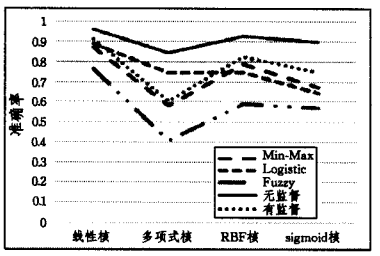


图 10　不同核函数下的准确率

3.2.3　KNN

对于 KNN 算法，选取不同的 K 时，本文方法在 WFR 上的表现如图 11 所示。依然对 KNN 算法在错误率以及 F_1 度量方面进行考察，实验结果如表 6 与表 7 所列，经本文方法处理后，KNN 算法亦能在各项指标上有一定的提高。

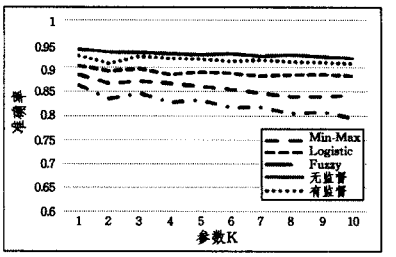


图 11　不同参数 K 下的准确率

表 6　KNN 分类错误率分析表

	Abalone	BC	Spambase	WFR
Min-Max	38.54	**2.98**	10.86	13.96
Logistic	**38.08**	3.47	**7.31**	8.11
Fuzzy	39.59	3.48	15.28	16.63
无监督	38.51	3.24	8.58	7.36
有监督	38.67	3.20	9.07	10.87

表 7　KNN 的 F_1 度量分析表

	Abalone	BC	Spambase	WFR
Min-Max	1.8314	1.9245	1.7703	3.4303
Logistic	**1.8448**	1.9246	**1.8463**	3.6375
Fuzzy	1.8004	1.9233	1.6756	3.2848
无监督	1.8313	1.9292	1.8173	**3.6800**
有监督	1.8254	**1.9294**	1.8093	3.5347

3.3　分析

从实验结果可以看出，本文方法整体上对提高挖掘结果的精确度有所帮助，但对于不同数据集、不同模型的挖掘结果的提升幅度并不一致，这是由数据集自身的分布决定的。实验结果中非线性标准化方法整体上较线性标准化方法更好，而在 Spambase 数据集上 Logistic 标准化方法的表现优于其他方法，这也恰好说明了传统的非线性标准化方法在其特定放缩区间与数据样本分布恰好匹配的时候能够取得不错的效果，但 Logistic 方法并没有在所有数据集上都表现得很好，而且即使在 Spambase 数据集上与本文方法相比也并没有体现出较大的差距。在这一点上，本文方法较传统的非线性标准化方法的适用范围更广。现实中也存在部分数据集，即使在数据稠密区间，数据点间距也足够大，尤其一些数据集的属性实际取值为整数(例如 BC)，其采用本文算法所带来的提升空间有限。而对于一些分布极端倾斜的数据集(99%的属性取值集中在 1%的值域)，即使采用本文算法对间距进行拉伸之后，也不能达到识别的精度，挖掘的结果也不会有太大的变化。

结束语　本文所提出的非线性标准化方法是一种基于统计的预处理方法，适用于无缺失值的连续值属性，能够在不改变数据分布规律的条件下适当提高数据稠密区的数据点间距，并且能依据数据本身分布的状况对变换函数做出调整。因此该方法在数据分布不均匀的数据集上有较好的效果，在均匀分布的数据集上退化为传统的线性标准化方法。虽然本方法的计算量相对于传统标准化方法所需要的计算开销有所提升，但本方法仅在首次标准化预处理时需要计算，新的数据只需要按照计算结果进行一次函数变换即可。然而，该方法同时也面临一些问题，对数据集逐个属性进行处理切断了不同属性之间的联系，容易对相关度高的属性造成影响，从而造成某些样本数据在某一维度上的间距扩大的同时在另一维度上的间距缩小。但是，我们认为样本在全空间下稠密的区域在单一维度下也一定稠密，所以本文提出的方法仅对全空间

下数据稀疏的区域可能造成一定的不良影响，但并不影响数据挖掘模型整体挖掘精确度的提升，实验结果也说明了这一点。而如何消除该方法可能带来的些许不良影响，则可作为非线性标准化方法未来的一个研究方向。

参 考 文 献

[1] Kamiran F, Calders T. Data preprocessing techniques for classification without discrimination[J]. Knowledge and Information Systems, 2012, 33(1): 1-33

[2] Guo Xi-yue, He Ting-ting. Survey about Research on Information Extraction[J]. Computer Science, 2015, 42(2): 14-17 (in Chinese)
郭喜跃，何婷婷. 信息抽取研究综述[J]. 计算机科学，2015，42(2)：14-17

[3] Wang R Y, Storey V C, Firth C P. A framework for analysis of data quality research[J]. IEEE Transactions on Knowledge and Data Engineering, 1995, 7(4): 623-640

[4] Jiawei H, Kamber M. Data mining: concepts and techniques [M]. San Francisco, CA, Itd: Morgan Kaufmann, 2001

[5] Weigend A S. Time series prediction: forecasting the future and understanding the past[R]. Santa Fe Institute Studies in the Sciences of Complexity, 1994

[6] Mendelsohn L. Preprocessing data for neural networks [OL]. https://www.tradertech.com/mendelsohn/library/neural-networks/preprocessing-data

[7] Yu L, Wang S, Lai K K. An integrated data preparation scheme for neural network data analysis[J]. IEEE Transactions on Knowledge and Data Engineering, 2006, 18(2): 217-230

[8] Liping Y, Yuntao P, Yishan W. Research on data normalization methods in multi-attribute evaluation[C]// International Conference on Computational Intelligence and Software Engineering, 2009 (CiSE 2009). IEEE, 2009: 1-5

[9] Pyle D. Data preparation for data mining [M]. Morgan Kaufmann, 1999

[10] Uragun B, Rajan R. Developing an appropriate data normalization method[C]// 2011 10th International Conference on Machine Learning and Applications and Workshops (ICMLA). IEEE, 2011, 2: 195-199

[11] Zhang Yu-nong, Li Ming-ming, Chen Jin-hao, et al. Solving the problem of Runge phenomenon by coefficients and order determination method[J]. Computer Engineering and Applications, 2013, 49(3): 44-49 (in Chinese)
张雨浓，李名鸣，陈锦浩，等. 龙格现象难题被解之系数与阶次双确定方法[J]. 计算机工程与应用，2013，49(3)：44-49

(上接第 259 页)

詹卫东. 面向中文信息处理的现代汉语短语结构规则研究[M]. 北京：清华大学出版社，2000：106-115

[13] Zhou Ming, Huang Chang-ning. Approach to the Chinese Dependency Formalism[J]. Journal of Chinese Information Processing, 1994, 3: 35-52 (in Chinese)
周明，黄昌宁. 面向语料库标注的汉语依存体系的探讨[J]. 中文信息学报，1994，3：35-52

[14] Androutsopoulos I, Koutsias J, Chandrinos K V, et al. An evaluation of Naive Bayesian anti-spam filtering[C]// Proc of the Workshop on Machine Learning in the New Information Age Joint 11th European Conference on Machine Learning. Barcelona, Spain, 2000: 9-17

[15] Agrawal R, Srikant R. Fast Algorithms for Mining Association Rules in Large Databases[C]// Proceedings of 20th International Conference on Very Large Data Bases (VLDB 1994). Santiagode Chile, Morgan Kaufmann, 1994: 487-499

[16] Park J S, Chen M S, Yu P S. An Effective Hash-Based Algorithm for Mining Association Rules[C]// Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD'95). San Jose, 1995: 175-186

[17] Savasere A, Omiecinski E, Navathe S. An Efficient Algorithm for Mining Association Rules in Large Databases[C]// 21st VLDB Conf. Zurich, Switzerland, 1995: 432-444

[18] Zhang Yu-qi, Zhou Qiang. Automatic Identification of Chinese Base Phrases [J]. Journal of Chinese Information Processing, 2002, 16(6): 1-8 (in Chinese)
张昱琪，周强. 汉语基本短语的自动识别[J]. 中文信息学报，2002，16(6)：1-8

[19] Croft W. Syntactic Categories and Grammatical Relations: The Cognitive Organization of Information [M]. Chicago and London: The University of Chicago Press, 1991: 66-78

[20] Zhao Jun, Huang Chang-ning. A Probabilistic Chinese BaseNP Recognition Model Combined with Syntactic Composition Templates[J]. Journal of Computer Research and Development, 1999, 36(11): 1384-1390 (in Chinese)
赵军，黄昌宁. 结合句法组成模板识别汉语基本名词短语的概率模型[J]. 计算机研究与发展，1999，36(11)：1384-1390

[21] Li Mu, Gao Jian-feng, Huang Chang-ning, et al. Unsupervised Training for Overlapping Ambiguity Resolution in Chinese Word Segmentation[J]. Proceedings of the Second SIGHAN Workshop on Chinese Language Processing (SIGHAN'03). 2003: 1-7

[22] Zhao Lei-lei. Feature extraction method based on the pattern of words and basic phrases[D]. Baoding: Hebei University, 2009 (in Chinese)
赵蕾蕾. 基于词和基本短语模式的特征提取方法[D]. 保定：河北大学，2009

[23] Langley P, Wayne L, Thompson K. An analysis of Bayesian classifiers[C]// Proc of the 10th National Conf on Artificial Intelligence. San Jose, California, 1992: 223-227

[24] Dominggos P, Pazzani M. On the Optimality of the Simple Bayesian Classifier under Zero-One Loss[J]. Machine Learning, 1997, 29: 103-130

[25] Wang Guo-yin, Zheng Zheng, Zhang Yi. RIDAS-A Rough Set Based Intelligent Data Analysis System [C]// Proc of First IEEE International Conference on Machine Learning and Cybernitics (ICMLC2002). Beijing, 2002: 646-649

[26] Gu Yi-jun, Fan Xiao-zhong, Wang Jian-hua, et al. Automatic Selection of Chinese Stoplist [J]. Transactions of Beijing Institute of Technology, 2005, 25(4): 337-340 (in Chinese)
顾益军，樊孝忠，王建华，等. 中文停用词表的自动选取[J]. 北京理工大学学报，2005，25(4)：337-340