不确定域环境下基于DKC值改进的K-means聚类算法

任培花[1] 王丽珍[2]
(山西大同大学数学与计算机科学学院 大同 037009)[1] (山西大同大学教育科学与技术学院 大同 037009)[2]

摘 要 提出一种不确定域环境下基于DKC值改进的K-means聚类算法,即U2d-Kmeans。该算法首先考虑到数据对象的不确定性因素,引入不确定域对数据对象进行描述;其次吸取2d-Kmeans的优点,对数据集进行预处理(剔除孤立点),并且采用累积距离的方法确定初始聚类中心,从而避免了随机选取聚类初始点造成聚类不稳定的缺陷;最后经过算法有效性对比实验证明得出,U2d-Kmeans算法比前两种算法更客观、有效。

关键词 不确定域,DKC值,2d-距离,聚类算法
中图法分类号 TP311.13 **文献标识码** A

Improved K-means Clustering Algorithm Based on DKC in Uncertain Region Environment

REN Pei-hua[1] WANG Li-zhen[2]
(School of Mathematics and Computer Science, Shanxi Datong University, Datong 037009, China)[1]
(School of Education Science and Technology, Shanxi Datong University, Datong 037009, China)[2]

Abstract This paper presented an improved K-means clustering algorithm based on DKC in uncertain region environment, namely U2d-Kmeans. Firstly, the algorithm takes uncertainty factors into account of the data object description, then uses new pretreatment method(removing isolated point) of data set and the cumulative distance method of determining the initial clustering center that is mentioned in the 2d-Kmeans algorithm. These methods avoid the defect of clustering instability caused by the random selection of clustering initial point. Finally, comparison experiment of the algorithm proves that the improved U2d-Kmeans is more objective and effective than the other two algorithms.

Keywords Uncertain region, DKC, 2d-distance, Clustering algorithm

聚类就是按照事物间的相似性进行区分和分类的过程。目前常用的主要聚类算法有基于划分的聚类算法、基于层次的聚类算法、基于密度的聚类算法、基于网格的聚类算法和基于神经网络的聚类算法[1,2]。基于划分的聚类算法是将数据集划分成多个簇,每个簇中至少包含一个数据对象,根据划分方式的不同,每个数据对象可以属于多个簇(模糊划分)或仅属于一个簇(确定性划分)。

K-means属于一种基于确定性划分的聚类算法,在数据集较小的情况下具有较好的鲁棒性,再加上其简单、实用、易于扩展等而得到了各行各业的普遍青睐。但该方法也存在3个明显的缺陷:1)只能用确定性的数值描述数据对象,因为测量误差、属性值遗失等原因很容易造成数据对象描述不准确,进而影响聚类效果;2)预先指定初始聚类中心点,显然在孤立点未被剔除的情况下,这种选取方式会影响聚类的效率和效果;3)凭经验随机剔除孤立点,因为初始聚类中心点是预先指定好的,所以会影响后期孤立点的发现,漏掉重要的孤立点,孤立点的存在和误删会影响聚类效果。因此,如何准确地描述数据对象、合理地选取聚类中心点和剔除孤立点是聚类问题中讨论的热点。

目前的文献都只是针对上述缺陷中的某项进行改进,如基于信息熵的精确属性赋权K-means聚类算法。原福永等人[3]采用熵值法计算数据对象各属性的权值,但还是没考虑到数据对象各属性值存在的各种各样的不确定性,如值误差、取值在一个范围内、属性值遗失等。姚丽娟等人[4]提出的一种基于粒子群的聚类算法利用粒子群的全局搜索能力解决K-means算法的局部收敛问题,提高了聚类质量和降低了时间复杂度,但是她们只对初始化聚类中心的选取给出了改进算法,并没有考虑孤立点的处理方法。储岳中等人[5]利用DBSCAN算法对初始数据集进行预处理,以去除部分孤立点,减少样本空间的规模,但初始化聚类中心的选取还是采用随机选取方式。

上述所有文献都是针对K-means聚类算法的3个缺陷分别采用方法加以改进。鉴于此,本文尝试从以下两个方面对基于2d-距离的K-means算法[6,7](在文中简称2d-Kmeans算法)加以改进:①对数据对象定义一个不确定域和不确定向量[8]来表达数据的不确定程度;②利用DKC值对初始数据集进行预处理,即剔除部分孤立点并确定初始聚类中心。这样不仅会适当减少样本空间的规模,而且很好地避免了因孤立

到稿日期:2012-06-11 返修日期:2012-09-08 本文受2011年山西省科技基础条件平台建设"大同地区科学数据共享服务平台"项目(2011091002-0102)资助。

任培花(1980—),女,硕士,讲师,主要研究方向为聚类算法、自主计算,E-mail:rphren@163.com;王丽珍(1970—),女,硕士,教授,主要研究方向为教育技术、数据共享平台建设。

点和初始聚类中心点的相互影响造成的聚类效果不佳。

1 关键概念定义

1.1 不确定域和不确定向量[8]的定义

在对大量数据进行聚类分析前，聚类样本数据可能存在各种不确定性的情况，不确定性通常由采样、测量、人为因素引起。为了获取聚类的最佳效果，对于不确定性数据对象的表达不再用精确的数据来表示，而是采用一个数据集不确定向量来辅助描述。

定义 1(不确定域) 给定 p 维空间的数据集 $X=\{x_1,x_2,\cdots,x_n\}$，其中 x_i 代表第 i 个数据对象，用 $x_i=(x_{i1},x_{i2},\cdots,x_{ip})^T\in R^p$ 表示，设 x_i 带有一个容许误差范围$[-UR_i,UR_i]$(其中$UR_i=(ur_{i1},\cdots,ur_{ip})^T$，$ur_{ij}\geqslant 0$)，这里的$UR_i$ 称为x_i 的不确定域。

定义 2(不确定向量) 给数据集 X 定义一个不确定向量集 $UV=\{uv_1,\cdots,uv_i,\cdots,uv_n\}$，其中 $uv_i=(uv_{i1},\cdots,uv_{ip})^T\in R^p$，$uv_i$ 是数据对象 x_i 的不确定向量，分量 uv_{ij} 满足如下约束条件：$|uv_{ij}|\leqslant ur_{ij}$ (ur_{ij} 是不确定域 UR_i 的分量)。

本文给出的改进算法对不确定数据对象的处理均采用不确定向量辅助表达的方式。不确定向量 uv_i 是定义在不确定域 UR_i 上的不确定向量，因此不确定数据对象 x_i 用 x_i+uv_i 表示。另一方面，考虑到不确定向量分量 uv_{ij} 和不确定域分量 ur_{ij} 之间的约束关系 $|uv_{ij}|\leqslant ur_{ij}$，图 1 是不确定二维数据对象 x_i 在不确定域 UR_i 上的表示。

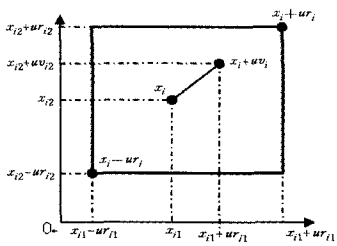

图 1 不确定数据对象 x_i 在不确定域 UR_i 上的表示形式

1.2 半径邻居(数)、DKC[7]值的定义

利用 DKC 值(数据对象邻居数的比值)去除孤立点，并且进行初始聚类中心点的选取，很大程度上可以避免因孤立点和初始聚类中心点相互影响造成的聚类效果不佳的缺陷。以下给出了 d-距离、半径邻居、半径邻居数和 DKC 值的定义。

定义 3(d-距离) 设数据集中的数据对象为 x_i，距离 x_i 最近的 d 个距离中最大的距离为 $d_{\max}$，称为 x_i 的 d-距离。

定义 4(半径邻居) 数据对象 x_i 的 d-距离半径邻居为以 x_i 为圆心、以 $d_{\max}$ 为半径的圆内的所有数据对象。

定义 5(半径邻居数) 设数据集中的数据对象为 x_i，以 d-距离(即 $d_{\max}$)为半径的圆内的所有数据对象的个数称为 x_i 的 d-距离半径邻居数，用 d_{x_i} 表示：

$$d_{x_i}=|x_j\in X|\quad dist(x_i,x_j)\leqslant d_{\max},i\neq j$$

同理，以 $2d$-距离(即 $2d_{\max}$)为半径的圆内的所有数据对象的个数称为 x_i 的 $2d$-距离半径邻居数，用 $2d_{x_i}$ 表示：

$$2d_{x_i}=|x_j\in X|\quad dist(x_i,x_j)\leqslant 2d_{\max},i\neq j$$

定义 6(DKC 值) d_{x_i} 是对象 x_i 的 d-距离半径邻居数，$2d_{x_i}$ 是对象 x_i 的 $2d$-距离半径邻居数，DKC_{x_i} 值则为 $2d_{x_i}$ 和 d_{x_i} 的比值，即：

$$DKC_{x_i}=\frac{2d_{x_i}}{d_{x_i}}$$

DKC 值越高，说明数据对象 x_i 周围是一个稀疏区域，此时可以将数据对象 x_i 认定为一个孤立点；相反，DKC 值越低，说明对象 x_i 周围是一个密集区域，即 x_i 不可能是孤立点。

2 不确定域环境下基于 DKC 值改进的 K-means 聚类算法

2.1 算法基本思想

为了保证数据对象的真实性，引入不确定域和不确定向量的概念，用不确定向量对数据对象进行辅助描述。根据这种描述方式，本文提出一种改进的 K-means 聚类算法，即不确定域环境下基于 DKC 值改进的 K-means 聚类算法(在文中简称 U2d-Kmeans 算法)。

2.2 算法过程

步骤 1 用不确定向量对数据对象进行描述。

首先把连续型、符号型属性归化为离散型、数值型属性，使其转化为利于新 K-means 算法的实现，本文设数据对象用 p 维属性描述。

设不确定数据对象集为 $X=\{x_1,x_2,\cdots,x_n\}$，其中 x_i 代表第 i 个数据对象，用 $x_i=(x_{i1},x_{i2},\cdots,x_{ip})^T\in R^p$ 表示，对应 X 的不确定向量集为 $UV=\{uv_1,\cdots,uv_i,\cdots,uv_n\}$，其中 $uv_i=(uv_{i1},\cdots,uv_{ip})^T\in R^p$，是数据对象 x_i 的不确定向量，分量 uv_{ij} 满足如下约束条件：$|uv_{ij}|\leqslant ur_{ij}$ (ur_{ij} 是不确定域 UR_i 的分量，$UR_i=(ur_{i1},\cdots,ur_{ip})^T$，$ur_{ij}\geqslant 0$)。因此不确定数据对象 x_i 用 x_i+uv_i 表示，即：

$$x_i+uv_i=\{x_{i1}+uv_{i1},x_{i2}+uv_{i2},\cdots,x_{ip}+uv_{ip}\}$$

步骤 2 计算所有数据对象的 DKC 值，利用 DKC 值对数据对象集进行预处理(剔除孤立点、选取初始聚类中心点)。

步骤 3 计算每个对象与初始聚类中心的最短距离，把对象加入最近的簇中。

步骤 4 更新聚类中心点，计算准则目标函数 $J=\sum_{i=1}^{K}\sum_{x\in s_i}d(x,s_i)$(其中 s_i 是聚类中心点，$d(x,s_i)$表示各个簇中数据对象 x 到簇中心的欧式距离)。

步骤 5 重复步骤 3、步骤 4，直到聚类中心点不再变化，最小化准则函数，使其收敛稳定 $|J^{t+1}-J^t|<0.00001$(t 为迭代次数)。

以下是具体过程：

输入：数据集 X=$\{x_1,x_2,\cdots,x_n\}$，不确定向量集 UV=$\{uv_1,\cdots,uv_i,\cdots,uv_n\}$，最近距离参数 d，孤立点的个数为 a，簇数目参数 K，t 为迭代次数。

输出：K 个簇 S=$\{S_1,S_2,\cdots,S_K\}$

初始化：i=1，j=1，S={}

Begin

Step 1 计算所有不确定数据对象的 DKC 值

Repeat

①利用欧式距离公式计算 x_i 的 d_{max} 值：

$$d(x_i,x_j)=\sum_{k=1}^{n}(x_{ik}+uv_{ik}-x_{jk}-uv_{jk})^2\ (1\leqslant j\leqslant n)$$

计算 X 中每个数据对象 x_i 与所有其他对象的欧式距离 $d(x_i,x_j)$，选出 d 个最小的距离，将其中最大的距离取出，记为 d_{max}。

②求出 x_i 的 d_{x_i} 和 $2d_{x_i}$ 值。

③求出 x_i 的 DKC_{x_i} 值：$DKC_{x_i}=\frac{2d_{x_i}}{d_{x_i}}$。

i=i+1

Until i>n

将 n 个 DKC 值用简单选择排序法按降序排序。

Step 2 剔除孤立点，得到新的数据集 X′

Repeat

Delete DKC_j 对应的数据对象

j=j+1

Until j>a

$X'=\{x_1', x_2', \cdots, x_{n-a}'\}$

Step 3 设置 X′的初始聚类中心

计算每个簇的平均数据对象个数 $L=[\frac{n-a}{K}]$，第一个聚类中心点为 DKC_n 对应的数据对象，第二个为 DKC_{n-L} 对应的数据对象，第三个为 DKC_{n-2L} 对应的，以此类推，一共可以得到 K 个初始聚类中心点。

Step 4 将 X′形成 K 个簇

计算不确定数据对象 $x_i'(i=1,2,3,\cdots,n-a)$ 与所有聚类中心 v_j $(j=1,2,\cdots,K)$ 间的欧式距离 $d(x_i', v_j)$：

$$d(x_i', v_j)=\| x_i'+uv_i'-v_j \|=\sqrt{\sum_{l=1}^{p}(x_{il}'+u_{il}'-v_{jl})^2}$$

$(1\leqslant j\leqslant K)$

当 $d(x_i', v_j)$ 值达到最小时，将 x_i 分配到该聚类中心代表的簇中。形成 K 个簇：

$S=\{S_1, S_2, \cdots, S_K\}$

Step 5 更新每个簇的聚类中心

$$v_j'=\frac{1}{N_j}\sum_{x_i\in S_j}(x_i+uv_i) \quad (j=1,2,\cdots,K)$$

N_j 是第 j 个簇 S_j 中数据对象的个数。

Step 6 计算准则函数

$$J=\sum_{i=1}^{K}\sum_{x\in s_i}d(x,s_i)$$

s_i 是聚类中心点，$d(x,s_i)$ 表示各个簇中数据对象 x 到簇中心的欧式距离。

Step 7

Repeat

Step 4;

Step 5;

Step 6;

Until K 个簇中心不再发生变化并且 $|J^{t+1}-J^t|<0.00001$

End

3 实验结果及性能分析

实验的运行环境：CPU：Intel(R) Core(TM)2 Duo，CPU T7100@1.80GHz；内存：1GB；操作系统：Microsoft Windows XP Professional；仿真软件工具：Matlab2007，其中可以直接调用聚类算法工具箱的 K-means 算法函数等。

数据集说明：为了验证本文算法的有效性和正确性，本文从 UCI(美国加州大学湾儿分校即 University of California Irvine)机器学习数据库[10]上选取 Wisconsin Prognostic Breast Cancer (WPBC)数据集作为实验数据。说明如表 1 所列。

表 1 选用数据集说明

数据集名称	样本数	属性数	类数	缺失值数
Wisconsin Prognostic Breast Cancer (WPBC)	198	34	2	4

聚类性能分析方法：我们在该实验中比较本文算法、文献[7]中的算法(基于 2d-距离改进的 K-means 聚类算法)和传统的 K-means 算法在 Iris、Wine、Breast-cancer 数据集上的聚类性能。实验采用聚类精度[11]来衡量 3 种算法的有效性，计算方法如下。

(1)每类的精度 p_i：

$$p_i=\frac{a_i}{a_i+b_i} \quad (1\leqslant i\leqslant C)$$

说明：设数据集 X 分为 C 类，a_i 表示正确分配到 c_i 类的数据对象数，b_i 表示错误分配到 c_i 中的数据对象数，p_i 为每个类的精度，数值越高，代表该类的精确度越高。

(2)整个数据集 X 的精确度 $micro\text{-}p$：

$$micro\text{-}p=\frac{\sum_{i=1}^{C}a_i}{n} \quad (n=\sum_{i=1}^{C}(a_i+b_i))$$

说明：n 为数据集对象总数，$micro\text{-}p$ 为整个数据集的聚类精确度。

实验结果分析：传统的 K-means 算法既没考虑孤立点的剔除，也没制定初始点的选取规则，因此只能采用运算若干次(这里取 $T=10$ 次)后所求的平均精确度来判断算法的有效性，由表 2 可知该值只有 0.5459。而 2d-Kmeans[7]算法同时考虑孤立点的剔除和初始点的选取方式，能够提高聚类的精度。从聚类实验结果可以看出，聚类精度平均值为 0.6398，比传统的 K-means 算法的精确度稍高，但该算法的精度受 a 和 d 不同取值的影响很严重。从表 3 可知，$a=5$、$d=5$ 时的精确度最高为 0.7032，$a=20$、$d=20$ 时仅为 0.4709，也是很不稳定的。而本文的 U2d-Kmeans 算法与以上两种算法相比，有两方面优势：一方面，还原了数据聚类的本来面目，做聚类实验的时候首先考虑缺失值、误差等不确定因素，实验放在不确定域上进行，这样就直接避免了缩小样本规模的危险，提高了聚类的有效性；另一方面，本文算法引用了 2d-Kmeans[7]算法中对孤立点和初始点的处理方式，这样可以提高聚类精度。通过选取不确定域 $ur_{kj}\in[0,4]$ $(k=7,29,86,197, j=34)$ 的不同值，验证 $|J^{(t+1)}-J^{(t)}|<0.00001$ (t 为迭代次数)时对聚类精确度的影响，结果如表 4 所列，聚类精度平均值为 0.7993，当不确定域取 2 的时候精度最高。总体来说，这种算法较前两种算法性能更好。图 2 是 3 种算法的对比图。

表 2 K-means 算法 10 次的聚类精确度

算法	聚类精度
第 1 次	0.5000
第 2 次	0.5825
第 3 次	0.5309
第 4 次	0.5825
第 5 次	0.5000
第 6 次	0.5000
第 7 次	0.5567
第 8 次	0.5876
第 9 次	0.5309
第 10 次	0.5876
平均值	0.5459

表 3 2d-Kmeans 算法在 a 和 d 不同取值下的聚类精度对比

a 和 d 取值	Breast Cancer Wisconsin (Diagnostic)
a=2,d=20	0.6873
a=5,d=20	0.6873
a=10,d=3	0.6450
a=20,d=5	0.6450
a=3,d=3	0.5921
a=5,d=5	0.7032
a=10,d=10	0.6873
a=20,d=20	0.4709
聚类精度平均值	0.6398

表 4　聚类算法结果的精确度

ur_{kj} 取值	结果的精度
0	0.7855
1	0.6855
2	0.8959
3	0.7337
4	0.8959
聚类精度平均值	0.7993

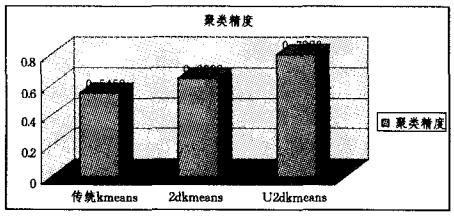

图 2　数据集 Breast Cancer Wisconsin (Diagnostic)的对比图

为了更加直观地对比各个算法的效果，将各个算法在 Breast Cancer Wisconsin (Diagnostic)数据集上的聚类效果用直方图的形式描述出来，如图 2 所示。显然 U2d-Kmeans 算法对 Breast Cancer Wisconsin (Diagnostic)数据集的聚类效果既客观又合理，效果优于其他两种。

结束语　传统聚类算法大多没有考虑数据对象的不确定因素，只是简单地消除数据的不确定成分，这种数据预处理方式会影响聚类效果。为了得到真实的聚类结果，本文考虑到数据对象的不确定成分，提出一种不确定域环境下基于 DKC 值改进的 K-means 聚类算法。另一方面，该算法还借鉴了 2d-Kmeans 算法中对孤立点和初始点的处理方法，分 3 步进行：(1)计算每个数据对象的 DKC 值；(2)根据 DKC 值对原始样数据集剔除孤立点；(3)对 DKC 值排序，根据累积距离的方法确定初始聚类中心。实验结果表明，该算法比传统的 K-means、2d-Kmeans 聚类算法有更好的聚类效果。但是，值得注意的是，聚类算法中引入数据对象的不确定性因素会给算法带来复杂性问题，这是今后研究的重点。

参 考 文 献

[1] Han Jia-wei, Kamber M. Data Mining: Concepts and Techniques [M]. Morgan Kaufmann Publishers, 2001

[2] 李光宇. 基于改进的 CLARANS 算法在数据挖掘中的研究[J]. 中南林业科技大学学报，2010，3：142-145

[3] 原福永，张晓彩，罗思标. 基于信息熵的精确属性赋权 K-means 聚类算法[J]. 计算机应用，2011，31(6)：1675-1677

[4] 姚丽娟，罗可，孟颖. 一种基于粒子群的聚类算法[J]. 计算机工程与应用，2012，13

[5] 储岳中，徐波. 动态最近邻聚类算法的优化研究[J]. 计算机工程与设计，2011，32(5)：1687-1690

[6] 杨臻. 基于 2k-距离的孤立点算法研究[J]. 福建电脑，2009，2：77-78

[7] 陈福集，蒋芳. 基于 2d-距离改进的 K-means 聚类算法研究[J]. 太原理工大学学报，2012，43(2)：114-118

[8] 刘位龙. 面向不确定性数据的聚类算法研究[D]. 济南：山东师范大学，2011

[9] Pfoser D, Jensen C S. Capturing the Uncertainty of Moving-ObjectRepresentations[C]// Proceedings of the 6th International Symposium on Advances inSpatial Databases. 1999：111-132

[10] UCI Machine Learning Repository [DB/OL]. http://archive.ics.uci.edu/ml/, 1992-07-16

[11] Ahmad A, Dey L. A K-mean clustering algorithm for mixed numeric and categorical data[J]. Data and Knowledge Engineering, 2007, 63：503-527

[12] 王茜，张鲲鹏. 隐私保护数据挖掘算法 MASK 的改进[J]. 重庆理工大学学报：自然科学版，2012，26(6)：63-66

(上接第 168 页)

分量和剩余分量，然后用 GEP 算法对上述的各个分量进行演化训练，最后将产生的各个预测模型进行重构，得到最终预测模型。根据对比实验可知，EMD&GEP 模型在预测精度上效果更为理想，稳定性能上比 GEP 模型更为优秀。

参 考 文 献

[1] Lyu M R. Handbook of software reliability engineering[M]. New York: McGraw Hill, 1996

[2] 赵亮，王建民，孙家广. 统计测试的软件可靠性保障能力研究[J]. 软件学报，2008，19(6)：1379-1385

[3] Yang B, Li X, Xie M, et al. A generic data-driven software reliability model with model mining technique[J]. Reliability Engineering and System Safety, 2010, 95：671-678

[4] Raja U, Hale D P, Hale J E. Modeling software evolution defects: a time series approach[J]. J. Softw. Maint. Evol.: Res. Pract., 2009, 21：49-71

[5] 贾治宇，康锐. 软件可靠性预测的 ARIMA 方法研究[J]. 计算机工程与应用，2008，44(35)：17-19

[6] Su Y S, Huang C Y. Neural-network-based approaches for software reliability estimation using dynamic weighted combinational models[J]. The Journal of Systems and Software, 2007, 80：606-615

[7] Moura M C, Zio E, Lins I D, et al. Failure and reliability prediction by support vector machines regression of time series data [J]. Reliability Engineering and System Safety, 2011, 96：1527-1534

[8] Lo J H. A study of applying ARIMA and SVM model to software reliability prediction[C]// 2011 International Conference on Uncertainty Reasoning and Knowledge Engineering. 2011：141-144

[9] 李海峰，陆民燕，王智新. 基于灰色系统理论的软件可靠性综合评价框架[J]. 北京航空航天大学学报，2008，34(11)：1261-1265

[10] 李海峰，陆民燕，曾敏，等. 基因表达式编程在软件可靠性建模中的应用[J]. 计算机科学与探索，2011，5(6)：534-546

[11] Huang N E, Shen Z, Long S R. Anew view of nonlinear waves: the hilbert spectrum[J]. Annual Review of Fluid Mechanics, 1999, 31：417-457

[12] 玄兆燕，杨公训. 经验模态分解法在大气时间序列预测中的应用[J]. 自动化学报，2008，34(1)：97-101

[13] Dong Y, Wang J Z, Jiang H, et al. Short-term electricity price forecast based on the improved hybrid model[J]. Energy Conversion and Management, 2011, 52：2987-2995

[14] 马飒飒，陈自力，赵守伟. 基于聚类的软件失效数据预处理[J]. 计算机工程与应用，2006，11：106-109

[15] Bandara P K, Wikramanayake G N, Goonethillake J S. Software reliability estimation based on cubic splines[C]// Proceedings of the World Congress on Engineering. 2009：12-15