Journal of Risk Analysis and Crisis Response, Vol. 2, No. 2 (August 2012), 131-138

A Credit Rating Model for Enterprises Based on Projection Pursuit and *K*-Means Clustering Algorithm

Mu Zhang
School of Finance, Guizhou University of Finance and Economics
Guiyang 550004, Guizhou, China, E-mail: rim_007@163.com

Zongfang Zhou
School of Management and Economics, University of Electronic Science and Technology of China
Chengdu 610054, Sichuan, China

Received 7 February 2012
Accepted 14 March 2012

Abstract

This paper proposes a new credit rating model for enterprises based on Projection Pursuit and *K*-means clustering algorithm. Firstly, using Projection Pursuit, the comprehensive credit score of each sample is obtained, so as to reflect the structure or characteristics of original multi-dimensional data. Secondly, the distribution density of the comprehensive credit score series is estimated by the kernel density estimation method, and then the initial cluster centers in original high dimension space are determined according to the local maximum points of density function. Finally, starting from the initial cluster centers above, using *K*-means clustering algorithm, the final cluster centers are obtained, and then the credit grades are partitioned. Thus, the credit rating for enterprises is realized. Taking the high-tech listed companies in China as samples, it is proved that the model proposed by this paper is feasible and effective.

Keywords: enterprise credit rating; Projection Pursuit; kernel density estimation; initial cluster centers; *K*-means clustering algorithm

基于投影寻踪和 K-均值聚类的企业信用评级模型

张目 [1]　周宗放 [2]
1. 贵州财经大学/金融学院，贵阳 550004
2. 电子科技大学/经济与管理学院，成都 610054

摘要：提出一种基于投影寻踪和 *K*-均值聚类的企业信用评级模型。首先，运用投影寻踪对样本企业进行信用综合评分，以反映原高维数据的结构或特征；然后，利用核密度估计法对信用综合得分序列进行分布密度估计，并根据密度函数的局部极大值点确定原高维空间中的初始聚类中心；最后，从给出的初始聚类中心出发，运用 *K*-均值算法获得最终聚类中心，并划分企业信用等级，从而实现对样本企业的信用评级。以我国高技术产业上市公司为例，应用实例证明了该模型的可行性和有效性。

关键词：企业信用评级，投影寻踪，核密度估计，初始聚类中心，*K*-均值聚类算法

1. 引言

企业信用评级是运用科学的指标体系、定量分析和定性分析相结合的方法，通过对企业信用记录、经营水平、外部环境、财务状况、发展前景以及可能出现的各种风险等进行客观、科学、公正的分析研究之后，就其信用能力所做出的综合评价，并用特定的等级符号标定其信用等级[1]。信用评级有助于企业防范商业风险，为现代企业制度的建设提供良好条件；信用评级有利于资本市场的公平、公正和诚信；同时，信用评级也是商业银行确定贷款风险程度的依据和信贷资产风险管理的基础。

目前，信用评级最常用的方法是基于分类的方法。在 Altman（1968）[2]做出开创性工作之后，多元判别分析(MDA)[2-3]、Logistic 回归模型[4]、Probit 回归模型[5]等统计方法在信用评级中获得了广泛应

用。然而，这类统计方法存在着诸多局限，如：MDA 要求样本数据服从正态分布和等协方差，而现实中大量数据并不服从这些假定[6]；Logistic 回归模型不仅对中间区域的差别敏感性较强，而且当样本点完全分离时，模型参数的最大似然估计可能不存在[7]。20 世纪 90 年代以来，以聚类分析[8]和 *K*-近邻法[9]为代表的非参数统计方法被引入到信用风险分析中，其中，聚类分析具有不要求样本数据服从具体分布，并且，具有可对变量采用名义尺度和次序尺度等优点，适于信用风险分析中按照定量指标和定性指标对并不服从一定分布特性的数据信息分类的要求[10]。在众多的聚类算法中，*K*-均值（*K*-means）算法[11]是一种基于划分的聚类算法，因其理论上可靠、算法简单、收敛速度快、能有效处理大数据集而得到最为广泛的使用[12]。参考文献[13]和[14]对 *K*-均值算法在企业信用评级中的应用进行了有益的尝试，其基本思路是：首先采用 Z 评分法、因子分析法等对样本企业进行信用评分，然后，在系统自动指定初始聚类中心下，运用 *K*-均值算法对信用得分序列进行聚类。上述研究存在以下两个方面的问题：（1）将高维数据“降维”后进行聚类分析，易丢失数据信息；（2）由系统自动指定初始聚类中心，导致聚类结果缺乏可靠性。

众所周知，*K*-均值算法对初始聚类中心较为敏感，对于给定的聚类数目 *K*，从不同的初始聚类中心出发，可能得到不同的聚类结果[15-16]。现有文献提出的优选初始聚类中心的方法主要有：密度评估法、距离优化法、基于遗传算法的方法和基于取样的方法等[17-22]。这些方法在一定程度上优化了初始聚类中心，减少了聚类的迭代次数。然而，上述方法均是在高维空间中进行计算，其算法复杂度较高，且某些方法存在输入参数难以确定的不足。

投影寻踪（Projection Pursuit，PP）[23-24]是一种直接由样本数据驱动的探索性数据分析方法，特别适用于分析和处理非线性、非正态的高维数据，其基本思想是把高维数据投影到低维子空间上，寻找出能反映原高维数据的结构或特征的投影，以达到研究分析高维数据的目的。有鉴于此，本文受参考文献[25]的启发，将投影寻踪与核密度估计结合运用于优选初始聚类中心，从而提出一种基于投影寻踪和 *K*-均值聚类的企业信用评级模型。本文的研究逻辑是：首先，运用投影寻踪对样本企业进行信用综合评分，以反映原高维数据的结构或特征；然后，利用核密度估计法对信用综合得分序列进行分布密度估计，并根据密度函数的局部极大值点确定原高维空间中的初始聚类中心；最后，从给出的初始聚类中心出发，运用 *K*-均值算法获得最终聚类中心，并划分企业信用等级，从而实现对样本企业的信用评级。

2. *K*-均值算法原理

K-均值算法的基本思想是通过迭代把数据对象划分到不同的簇中，以求目标函数最小化，从而使生成的簇尽可能的紧凑和独立。给定样本集和正整数 *K*，*K*-均值算法将样本集分割成 *K* 个簇，每个聚类中心是簇中样本的均值；将其余对象根据其与各个簇的中心的距离分配到最近的簇；然后，求出新形成的簇的中心。这个迭代重新定位过程不断重复，使得每个簇中所有样本与其中心的距离总和最小，直到目标函数最小化为止[11-12]。

K-均值聚类过程是通过反复移动簇中心以最小化簇集内的总度量（如：距离、相似度等）来完成的。设样本为 X_i（$i=1,2,\cdots,N$），给定一组初始聚类中心点 c_k（$k=1,2,\cdots,K$），初始聚类中心可以从样本集中随机选择，也可以根据实际需要来指定。*K*-均值聚类算法交替执行以下两步[11-12]：

（1）对每个样本 X_i，找出距离其最近的中心点（簇）

$$k=\arg\ \min_{k\in\{1,2,\cdots,K\}} d(c_k,X_i),\quad k=1,2,\cdots,K \tag{1}$$

（2）计算每个簇中样本的均值，该均值向量即成为该簇新的中心

$$c_k=\frac{1}{n_k}\sum_{j=1}^{n_k}X_j^{(k)},\quad k=1,2,\cdots,K \tag{2}$$

其中，n_k 为第 k 簇中的样本数。

重复以上两步，直到没有样本或很少的样本被分配到不同的簇中。

3. 企业信用评级模型的构建

对于多分类的企业信用评级问题，设有 m 个企业组成训练样本集 $A=\{A_i\,|\,i=1,\cdots,m\}$，企业信用评级指标集 $C=\{C_j\,|\,j=1,\cdots,n\}$，$x_{ij}$ 为训练样本 A_i 在

指标 C_j 下的指标值。基于投影寻踪和 *K*-均值聚类的企业信用评级模型构建步骤如下：

步骤 1：指标值的归一化处理。为消除各指标的量纲、统一各指标的变化范围和方向，须对指标值进行极值归一化处理。

对于成本型指标，令

$$y_{ij}=\frac{x_j^{\max}-x_{ij}}{x_j^{\max}-x_j^{\min}},\quad i=1,2,\cdots,m;\ j=1,2,\cdots,n \tag{3}$$

对于效益型指标，令

$$y_{ij}=\frac{x_{ij}-x_j^{\min}}{x_j^{\max}-x_j^{\min}},\quad i=1,2,\cdots,m;\ j=1,2,\cdots,n \tag{4}$$

式（3）–（4）中，$x_j^{\max}$、$x_j^{\min}$ 分别为第 j 个指标的最大值和最小值。

对于固定型指标，即指标值越接近某一固定值越好的指标，有

$$y_{ij}=1-\frac{\left|x_{ij}-x_j^*\right|}{\max_i\left|x_{ij}-x_j^*\right|},\quad i=1,2,\cdots,m;\ j=1,2,\cdots,n \tag{5}$$

式（5）中，x_j^* 为第 j 个指标的最佳稳定值。

步骤 2：构造信用评分函数及投影指标函数。PP 方法就是把 n 维数据 $\{y_{ij}\,|\,j=1,\cdots,n\}$ 综合成以 $a=(a_1,a_2,\cdots,a_n)$ 为投影方向的一维投影值 Z_i：

$$Z_i=\sum_{j=1}^{n}a_j y_{ij},\quad i=1,2,\cdots,m \tag{6}$$

上式中，a 为单位长度向量。Z_i 近似刻画了样本企业的信用状况[7]，投影值越低，信用风险越高，则称式（6）为样本企业的信用评分函数，Z_i 为样本企业的信用综合得分。

PP 方法在综合 Z_i 时，要求 Z_i 的散布特征应为：局部投影点尽可能密集，最好凝聚成若干个点团，而在整体上投影点团之间尽可能散开。由此，投影指标函数可构造为[23]

$$Q(a)=S_Z D_Z \tag{7}$$

式中，S_Z 为 Z_i 的标准差，D_Z 为 Z_i 的局部密度，即

$$S_Z=\sqrt{\frac{1}{m-1}\sum_{i=1}^{m}(Z_i-\bar{Z})^2} \tag{8}$$

$$D_Z=\sum_{i=1}^{m}\sum_{j=1}^{m}(R-r_{ij})I(R-r_{ij}) \tag{9}$$

其中，$\bar{Z}=\frac{1}{m}\sum_{i=1}^{m}Z_i$ 为 Z_i 的均值；R 为求局部密度的窗口半径，它的选取既要使包含在窗口内的投影点的平均个数不太少，避免滑动平均偏差太大，又不能使它随着 m 的增大而增加太快，R 一般可取值为 $0.1S_Z$ [26-27]；点间距离 $r_{ij}=|Z_i-Z_j|$；$I(t)$ 为单位阶跃函数，当 $t<0$ 时其函数值为 0，当 $t\geq 0$ 时其函数值为 1。

步骤 3：优化投影指标函数。当样本集给定时，投影指标函数只随投影方向的变化而变化。不同的投影方向反映不同的数据结构特征，最佳投影方向就是最大可能暴露高维数据某类特征结构的投影方向 [26-27]。通过求解投影指标函数最大化问题可估计出最佳投影方向，即

$$\begin{aligned}&\max\quad Q(a)=S_Z D_Z\\&s.t.\ \sum_{j=1}^{n}a_j^2=1\end{aligned} \tag{10}$$

式（10）所设定的问题是一个以 $\{a_j\,|\,j=1,\cdots,n\}$ 为优化变量的复杂非线性优化问题，常规优化方法较难处理。模拟生物优胜劣汰规则与群体内部染色体信息交换机制的实码加速遗传算法（Real Coded Accelerating Genetic Algorithm， RAGA）是一种通用的全局优化方法，用它来求解该问题则十分简便而有效。RAGA 的具体算法参见参考文献[26]和[27]。

步骤 4：计算训练样本的信用综合得分，对信用综合得分序列进行分布密度估计。将步骤 3 估计出的最佳投影方向 a^* 代入式（6）后可得训练样本的信用综合得分 Z_i^*。初始聚类中心要求是一组能尽量反映数据分布特征的数据对象[28]；由投影寻踪原理可知，Z_i^* 的散布特征反映了原高维数据的某种结构或特征；因此，可以通过分析 Z_i^* 的散布特征来优选初始聚类中心。为此，本文利用核密度估计法对信用综合得分序列 Z_i^* 进行分布密度估计。其中，核密度估计定义如下[29-30]：

定义 1：设 $K(\square)$ 为 R^1 上一个给定的概率密度函数，$h_m>0$ 是一个与 m 有关的常数，满足 $m\to\infty$，$h_m\to 0$，则称

$$f_m(z^*)=\frac{1}{mh_m}\sum_{i=1}^{m}K(\frac{z^*-Z_i^*}{h_m}) \qquad (11)$$

为 $f(z^*)$ 的一个核密度估计，其中 $K(\square)$ 为一已知核函数，满足

$$\sup_{-\infty<u<+\infty}\left|K(u)\right|<+\infty,\quad K(u)=K(-u) \qquad (12)$$

$$\int_{-\infty}^{+\infty}K(u)du<+\infty \qquad (13)$$

$$\lim_{|u|\to\infty}\left|uK(u)\right|=0 \qquad (14)$$

h_m 称为窗宽或光滑参数。

步骤 5：确定初始聚类中心，运用 *K*-均值算法划分信用等级。由步骤 4 得出密度函数 $f_m(z^*)$ 及相应的核密度估计曲线。在已知数据分布的条件下，一个优良的初始聚类中心应满足[31]：（1）选择的初始聚类中心点各属于不同的类，即任意两个初始聚类中心点不能属于同一类；（2）选择的初始聚类中心点应能够作为该类代表，即应该尽量靠近类中心。据此，可直观搜索出密度函数 $f_m(z^*)$ 的局部极大值点，并选取与局部极大值点最临近的样本投影点在原高维空间中所对应的点为初始聚类中心点。

在运用 *K*-均值算法对企业进行信用评级时，首先根据信用评级的实际需要设定 K 个信用等级，则应有 K 个聚类数目与之对应，从而需选取 K 个样本点组成初始聚类中心。假设密度函数 $f_m(z^*)$ 有 N 个局部极大值点，当 $K=N$ 时，初始聚类中心随即确定；当 $K<N$ 时，从我国商业银行“区别对待，择优扶持”的信贷原则出发，在 N 个局部极大值点中选取数值较大的前 K 个点来确定初始聚类中心；当 $K>N$ 时，则需通过增加训练样本数量来使得 $K\leqslant N$。

在确定初始聚类中心后，运用 *K*-均值算法对训练样本进行聚类分析，从而得到 K 个最终聚类中心点。由式（6）计算 K 个最终聚类中心点的信用综合得分，然后，根据信用综合得分的大小，建立聚类类别与信用等级的一一对应关系，从而划分出 K 个信用等级，并实现对训练样本的信用评级。

步骤 6：对新样本进行信用评级。对于一个新的测试样本，首先，运用式（3）–（5）对测试样本的信用评级指标值进行标准化处理，特别地，当测试样本的第 j 个指标值在训练样本指标值区间 $\left[x_j^{\min},x_j^{\max}\right]$（$j=1,2,\cdots,n$）内时，即为归一化处理。然后，分别计算测试样本与步骤 5 得出的 K 个最终聚类中心点的欧式距离，找出距离其最近的中心点，该中心点对应的信用等级即为测试样本所属的信用等级。特别地，当测试样本与 2 个或 2 个以上最终聚类中心点的欧式距离相等时，则可通过计算联系向量距离[32]来加以区分。

4. 应用实例

4.1. *指标体系与样本数据*

本文参照国家财政部统计评价司的企业绩效评价指标体系和中国工商银行企业资信评估指标体系，遵循指标选取的系统性、科学性、客观性、可比性及可操作性等原则，从偿债能力、营运能力和盈利能力等三个方面构建企业信用评级指标体系。该指标体系包括以下 12 个指标：流动比率、速动比率、资产负债率、利息保障倍数、存货周转率、应收账款周转率、总资产周转率、固定资产周转率、总资产报酬率、净资产报酬率、销售净利率、股本报酬率等。

选取沪、深股市中的高技术产业上市公司作为实验样本，样本区间选定为 2005-2007 年，数据来源于国泰安数据库。剔除异常数据样本后，最终获得 112 家样本企业，其中，有 74 家为“非 *ST* ”企业，这类企业称之为“正常企业”；其余 38 家为

“*ST* 或*ST”企业，这类企业称之为“违约企业”。将实验样本集划分为训练样本集和测试样本集。随机抽取 50 家“正常企业”和 25 家“违约企业”作为训练样本，剩余的 24 家“正常企业”和 13 家“违约企业”作为测试样本。使用 Matlab7.1 工具包、Eviews6.0 和 SPSS16.0 软件进行实验分析。

4.2. *信用评分及分布密度估计*

按照第 3 节步骤 1，对训练样本指标值进行归一化处理。运用 RAGA 求解式（10）所设定的最优化问题，得出最大投影指标函数值：$Q_{\max}(a)$ =0.9539，最佳投影方向：a^* =（0.0171，0.0140，0.3387，0.1517，0.4217，0.3066，0.3451，0.2032，0.2904，0.2252，0.3880，0.3793）。将 a^* 代入式（6），计算出训练样本的信用综合得分 Z_i^*。

根据定义 1 对信用综合得分序列 Z_i^* 进行分布密度估计。首先，采用 Silverman（1986）提出的经验法则[33]计算初始光滑参数，即：假定 $f(z^*)$ 为正态密度函数 $N(0,\sigma^2)$，选取正态核函数，则根据经验法可得最佳渐进光滑参数为：

$$\hat{h}_{AMISE} \approx 1.06\hat{\sigma}m^{-\frac{1}{5}} \qquad (15)$$

其中 $\hat{\sigma}$ 为信用综合得分序列 Z_i^* 的标准差估计值。

将 $m=75$，$\hat{\sigma}=0.1166$ 代入式（15）得出 $\hat{h}_{AMISE} \approx 0.0521$。

其次，选取正态（Gaussian）核函数：

$$K_G(u)=(\sqrt{2\pi})^{-1}\exp(-\frac{u^2}{2}),\quad u\in\{-\infty,+\infty\} \qquad (16)$$

设置格点数为 200，利用 Eviews6.0 软件实现核密度估计，由于拟合曲线不光滑，本文还采用尝试法[30]对光滑参数进行适当调整，当光滑参数为 0.0180 时得到较为满意的结果。信用综合得分序列 Z_i^* 的核密度估计曲线见图 1。

4.3. *初始聚类中心的确定*

从图 1 可以看出，密度函数 $f_m(z^*)$ 共有 7 个局部极大值点，通过对 Eviews6.0 软件输出的数据矩阵的直观搜索，得出这 7 个局部极大值点分别为：0.7459，0.8152，1.0193，1.1156，1.2350，1.3082，1.4700。本文根据我国商业银行贷款五级分类的实际需要，设定 5 个信用等级，则应有 5 个聚类数目与之对应。由第 3 节步骤 5，在上述 7 个局部极大值点中选取数值较大的前 5 个点来确定原高维空间中的初始聚类中心。

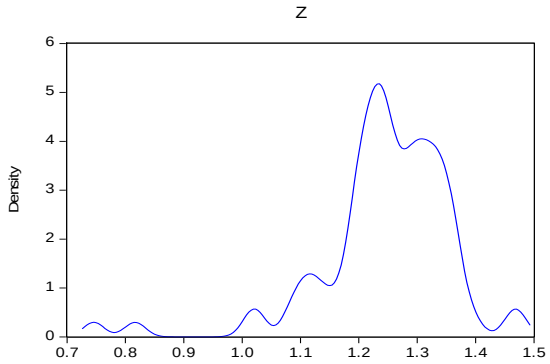

Fig. 1. Kernel Density Estimation Curve

图 1 核密度估计曲线

4.4. *K-均值聚类分析与信用等级的划分*

导入 4.3 节得出的初始聚类中心进行 *K*-均值聚类分析。设置聚类数目为 5，最大迭代次数为 20，收敛准则为 0.01，经过 5 次迭代后，达到聚类结果的要求，聚类分析结束，得出最终聚类中心。由式（6）计算出 5 个最终聚类中心点的信用综合得分为：1 类-1.1080、2 类-1.2293、3 类-1.2486、4 类-1.2700、5 类-1.3501。根据信用综合得分的大小，建立聚类类别与信用等级的一一对应关系，即有：5 类-Ⅰ、4 类-Ⅱ、3 类-Ⅲ、2 类-Ⅳ、1 类-Ⅴ，其中，数字序号Ⅰ-Ⅴ分别代表 5 个信用风险从低到高的信用等级。

Table 1. The Result of Enterprise Credit Rating Based on the New Model

表 1　企业信用评级结果（本文模型）

聚类类别	5	4	3	2	1
信用等级	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ
			训练样本（75）		
企业总数	5	21	25	12	12
违约企业比例	0.0000	0.0476	0.1600	0.7500	0.9167
			测试样本（37）		
企业总数	2	11	13	6	5
违约企业比例	0.0000	0.0909	0.1538	0.8333	1.0000

注：表中括号内为样本个数。

4.5. *信用评级结果与对比分析*

根据聚类类别与信用等级的对应关系及聚类成员，即可实现对训练样本的信用评级。对于新的测试样本，则按照第 3 节步骤 6 评定其信用等级。训练样本和测试样本的信用评级结果见表 1。

由表 1 可知，训练样本的信用评级结果表现为：从第Ⅰ级到第Ⅴ级，随着信用等级的降低，违约企业比例（可近似看成违约率）呈单调递增趋势。即，企业信用等级越低，其违约概率越大，信用风险越高，这与信用风险管理理论相吻合。测试样本的信用评级结果呈现出与训练样本相似的特征，表明本文模型具有良好的泛化能力，能够满足实际应用的需要。

为便于比较，本文还采用由系统自动指定初始聚类中心的 *K*-均值算法（以下简称为传统模型）对样本企业进行信用评级。设置聚类数目为 5，最大迭代次数为 20，收敛准则为 0.01，经过 8 次迭代后，达到聚类结果的要求，聚类分析结束，得出最终聚类中心。由式（6）计算出 5 个最终聚类中心点的信用综合得分为：1 类−0.8896、2 类−1.1551、3 类−1.3143、4 类−1.2287、5 类−1.3153。训练样本和测试样本的信用评级结果见表 2。

由表 2 可知，虽然训练样本的信用评级结果与信用风险管理理论相符，但测试样本的信用评级结果未呈现出与训练样本相似的特征，说明传统模型的泛化能力较差，不能满足实际应用的需要。另外，传统模型的聚类分析迭代次数为 8 次，高于本文模型的 5 次，表明本文模型优选的初始聚类中心减少了 *K*-均值算法的迭代次数，加快了算法的收敛速度，提高了算法的运算效率。

此外，为进一步考察本文模型的聚类效果，本文还比较了上述两个模型的最小目标函数值：$J=\sum_{k=1}^{5}J_k$，其中，J_k 表示第 k 类中聚类成员与其中

Table 2. The Result of Enterprise Credit Rating Based on the Traditional Model

表 2　企业信用评级结果（传统模型）

聚类类别	5	3	4	2	1
信用等级	Ⅰ	Ⅱ	Ⅲ	Ⅳ	Ⅴ
			训练样本（75）		
企业总数	8	27	18	20	2
违约企业比例	0.0000	0.0370	0.2778	0.8500	1.0000
			测试样本（37）		
企业总数	5	12	9	11	0
违约企业比例	0.2000	0.0833	0.3333	0.7273	——

注：表中括号内为样本个数。

心的距离总和。计算结果显示，本文模型的 J 值为 18.57，小于传统模型的 19.08，表明本文模型的聚类效果优于传统模型。

5. 结束语

本文将投影寻踪与核密度估计结合运用于优选初始聚类中心，从而提出一种基于投影寻踪和 *K*-均值聚类的企业信用评级模型。该模型具有以下特点：（1）运用投影寻踪对样本企业进行信用综合评分，以反映原高维数据的结构或特征；利用核密度估计法对信用综合得分序列进行分布密度估计，并根据密度函数的局部极大值点来确定原高维空间中的初始聚类中心，具有合理性和可操作性；（2）把高维数据投影到低维子空间上，在低维子空间进行初始聚类中心的优选，计算相对简单，且不需要任何输入参数，具有直观性和便捷性；（3）从给出的初始聚类中心出发，在原高维空间中运用 *K*-均值算法进行聚类分析，最大限度的保留了原始数据的信息，并提高了聚类结果的可靠性。本文的研究为拓展 *K*-均值算法在企业信用评级中的应用提供了新的方法和思路。*K*-均值算法是基于梯度下降的算法，不可避免地常常陷入局部极优[16]，因此，将基于遗传算法、免疫规划或粒子群优化的 *K*-均值算法引入到企业信用评级中有待于进一步研究。

致谢

本文获得国家自然科学基金面上项目（70971015）、教育部人文社会科学研究规划基金项目（11YJA630196）和贵州财经学院金融学院科研项目（2009-04）的资助，在此表示衷心的感谢。

参考文献

[1] Li Shimei, The Theoretical Thinking of Credit Capacity Evaluation of the Industrial Enterprises in China, *J. Jilin University Journal Social Sciences Edition.* **48**(4) (2008) 107–112.
李士梅.我国工业企业信用能力评价的理论思考[J].吉林大学社会科学学报, 2008,48(4):107–112.

[2] Altman E I, Financial ratios, discriminant analysis and the prediction of corporate bankruptcy, *J. Journal of Finance.* **23**(4) (1968) 589–609.

[3] Altman E I, Haldeman R G and Narayanan P, Zeta analysis: a new model to identify bankruptcy risk of corporations, *J. Journal of Banking and Finance.* **1**(1) (1977) 29–54.

[4] Ohlson J, Financial rations and the probabilistic prediction of bankruptcy, *J. Journal of Accounting Research.* **18**(1) (1980) 109–130.

[5] Gentry J A, Whitford D T and Newbold P, Predicting Industrial Bond Ratings with a Probit Model and Funds Flow Components, *J. The Financial Review.* **23**(3) (1988) 269–286.

[6] Eisenbeis Robert A, Pitfalls in the Application of Discriminant Analysis in Business, Finance, and Economics, *J. Journal of Finance.* **32**(3) (1977) 875–900.

[7] Wang Chunfeng and Li Wenhua, Credit Risk Assessment in Commercial Banks: Projection Pursuit Discriminant Model, *J. Journal of Industrial Engineering and Engineering Management.* **14**(2) (2000) 43–46.
王春峰,李汶华.商业银行信用风险评估:投影寻踪判别分析模型[J].管理工程学报, 2000,14(2):43–46.

[8] Lundy M (eds.), *Cluster analysis in credit scoring. Credit Scoring and Credit Control*, 1nd edn. (Oxford University Press, New York, 1993, 78–90).

[9] Henley W E and Hand D J, A k-nearest-neighbor classifier for assessing consumer credit risk, *J. Statistician.* **45**(1) (1996) 77–95.

[10] Zhang Wei and Li Yushuang, Credit Risk Analysis in Commercial Bank: An Overview, *J. Journal of Management Sciences in China.* **1**(3) (1998) 20–27.
张维,李玉霜.商业银行信用风险分析综述[J].管理科学学报,1998,1(3):20–27.

[11] MacQueen J, Some methods for classification and analysis of multivariate observations, in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, (University of California Press, Berkeley, Calif., 1967, 1), pp. 281–297.

[12] Liang Xun (eds.), *Data Mining Algorithms and Applications*, 1nd edn. (Peking University Press, Peking, 2006, 16–22, 193).
梁循,数据挖掘算法与应用[M].北京:北京大学出版社,2006.16–22,193.

[13] Zuo Ziye and Zhu Yangyong, Credit Scoring and Rating Based on Clustering Technology of Data Mining, *J. Computer Applications and Software.* **21**(4) (2004) 1–3.
左子叶,朱扬勇.基于数据挖掘聚类技术的信用评分评级[J].计算机应用与软件,2004,21(4):1–3.

[14] Zhang Guiqing and Liu Shulin, Empirical Study of Credit Risk Evaluation in China's Commercial Banks, *J.* Journal of Hebei University of Economics and Trade. **26**(4) (2005) 41–45.
张贵清,刘树林.我国商业银行信用风险评级实证分析[J].河北经贸大学学报,2005,26(4):41–45.

[15] Pena J M, Lozano J A and Larranaga P, An Empirical Comparison of Four Initialization Methods for the K-Means Algorithm, *J. Pattern Recognition Letters.* **20**(10) (1999) 1027–1040.

[16] Liu Jingming, Han Lichuan and Hou Liwen, Cluster Analysis Based on Particle Swarm Optimization Algorithm, *J. Systems Engineering-Theory and Practice.* **25**(6) (2005) 54–58.

刘靖明,韩丽川,侯立文.基于粒子群的 K 均值聚类算法[J].系统工程理论与实践,2005,25(6):54–58.

[17] Kaufman L and Rousseeuw P J (eds.), *Finding Groups in Data: An Introduction to Cluster Analysis*, 1nd edn. (John Wiley and Sons, New York, 1990, 64–75).

[18] Li Chunsheng and Wang Yaonan, New initialization method for cluster center, *J. Control Theory and Applications.* **27**(10) (2010) 1435–1440.
李春生,王耀南.聚类中心初始化的新方法[J].控制理论与应用,2010,27(10):1435–1440.

[19] Katsavounidis I, Jay Kuo C.-C. and Zhang Zhen, A New Initialization Technique for Generalized Lloyd Iteration, *J. IEEE Signal Processing Letters.* **1**(10) (1994) 144–146.

[20] Xiong zhongyang, Chen ruotian and Zhang Yufang, Effective method for cluster centers' initialization in K-means clustering, *J. Application Research of Computers.* **28**(11) (2011) 4188–4190.
熊忠阳,陈若田,张玉芳.一种有效的 K-means 聚类中心初始化方法[J].计算机应用研究,2011,28(11):4188–4190.

[21] Phanendra Babu G and Narasimha Murty M, A near-optimal initial seed value selection in K-means algorithm using a genetic algorithm, *J. Pattern Recognition Letters.* **14**(10) (1993) 763–769.

[22] Bradley P S and Fayyad Usama M, Rifining Initial Pionts for K-Means Clustering, in *Proceedings of the Fifteenth International Conference on Machine Learning*, (Morgan Kaufmann Publishers, San Francisco, CA., 1998), pp. 91–99.

[23] Friedman J H and Turkey J W, A projection pursuit algorithm for exploratory data analysis, *J. IEEE Transactions on computer.* **23**(9) (1974) 881–890.

[24] Huber P J, Projection pursuit (with discussions), *J. The Annals of Statistics.* **13**(2) (1985) 435–475.

[25] Gan Wenyan and Li Deyi, Hierarchical Clustering based on Kernel Density Estimation, *J. Journal of System Simulation.* **16**(2) (2004) 302–305, 309.
淦文燕,李德毅.基于核密度估计的层次聚类算法[J].系统仿真学报,2004,16(2):302–305,309.

[26] Fu Qiang and Zhao Xiaoyong (eds.), *The Principle and Application of Projection Pursuit Model*, 1nd edn. (Science Press, Peking, 2006, 1–119).
付强,赵小勇.投影寻踪模型原理及其应用[M].北京:科学出版社,2006.1–119.

[27] Jin Juliang and Ding Jing (eds.), *Water Resources Systems Engineering*, 1nd edn. (Sichuan Science and Technology Press, Chengdu, 2002, 37–179).
金菊良,丁晶.水资源系统工程[M].成都:四川科学技术出版社,2002.37–179.

[28] Lai Yuxia and Liu Jianping, Optimization Study on Initial Center of K-means Algorithm, *J. Computer Engineering and Applications.* **44**(10) (2008) 147–149.
赖玉霞,刘建平.K-means 算法的初始聚类中心的优化[J].计算机工程与应用,2008,44(10):147–149.

[29] Parzen E, On Estimation of a Probability Density Function and the Mode, *J. The Annals of Mathematical Statistics.* **33**(3) (1962) 1065–1076.

[30] Li Zhuyu, Lu Wanbo and Gong Jinguo (eds.), *The Non-parametric Estimation Techniques in Economic, Financial Econometrics*, 1nd edn. (Science Press, Peking, 2007, 7–58).
李竹渝,鲁万波,龚金国.经济、金融计量学中的非参数估计技术[M].北京:科学出版社,2007.7–58.

[31] Liu Liping and Meng Zhiqing, An Initial Centrepoints Selection Method for k-means Clustering, *J. Computer Engineering and Applications.* **40**(8) (2004) 179–180.
刘立平,孟志青.一种选取初始聚类中心的方法[J].计算机工程与应用,2004,40(8):179–180.

[32] Zhang Mu and Zhou Zongfang, An Improved TOPSIS Method Based on Connection Degree, *J. Systems Engineering.* **26**(8) (2008) 102–107.
张目,周宗放.一种基于联系度的改进 TOPSIS 法[J].系统工程,2008,26(8):102–107.

[33] Silverman B W (eds.), *Density Estimation for Statistics and Data Analysis*, 1nd edn. (Chapman and Hall, London, 1986, 43–60).