文章编号：1002-1566(2016)04-0641-08
DOI: 10.13860/j.cnki.sltj.20160109-002

智能信息处理的多指标面板数据聚类方法及其应用

林秀梅[1,3] 孙海波[1,2] 王丽敏[3]

(1. 吉林大学 数量经济研究中心，吉林 长春 130012；2. 吉林大学 商学院，吉林 长春 130012；
3. 吉林财经大学，吉林 长春 130117)

摘要：为提高具有先验知识样本的学习效率，本文在吸引子传播聚类模型基础上，引入半监督学习策略，并综合考虑样本动态信息变化，融合多指标面板数据，提出智能信息处理的多指标面板数据聚类模型。选取 30 家房地产业上市公司 2009-2013 年相关财务数据，利用此模型进行聚类和绩效评价分析，结果表明，智能信息处理的多指标面板数据聚类模型能更加有效地区分样本类别特征，可为上市公司绩效评价、金融管理与决策提供一个更加有效的方法和手段。

关键词：吸引子传播聚类模型；半监督学习；多指标面板数据；上市公司；绩效评价
中图分类号：O212 **文献标识码**：A

Intelligent Information Clustering Method Based on Multivariable Panel Data and Its Application

LIN Xiu-mei[1,3] SUN Hai-bo[1,2] WANG Li-min[3]

(1. Center for Quantitative Economics of Jilin University, Jilin Changchun 130012, China,
2. Business School of Jilin University, Jilin Changchun 130012, China,
3. Jilin University of Finance and Economics, Jilin Changchun 130117, China)

Abstract: In this paper, to affinity propagation clustering model, semi-supervised learning strategies is introduced on the basis of the original model in order to improve the learning efficiency of samples with prior knowledge. And then a clustering model by intelligent information processing is proposed in this paper, which fuses multivariable panel data and considers sample information dynamic change. This article analyzes 30 real estate industry listed companies multivariable panel data in 2009-2013, the results show that the model can provide a more effective method and means for financial management, performance evaluation and decision making.

Key words: affinity propagation clustering model, semi-supervised learning, multivariable panel data, listed companies, performance evaluation

0 引言

面对大量的上市公司财务数据，探索高效的、科学的、具有实用价值的分析方法，已成为金融数据挖掘领域研究的一个热点和难点。上市公司财务数据存在着数据量大、维度高、相关

收稿日期: 2014 年 9 月 2 日 **收到修改稿日期**: 2015 年 9 月 22 日
基金项目: 国家社会科学基金资助重点项目 (12AZD021)；国家自然科学基金资助项目 (61202306,61472049)；吉林省社科规划项目 (2012B143)；吉林省软科学项目 (20120620)。

性强等特点，传统研究常采用主成分分析 [1−2]、因子分析 [3] 以及数据包络分析 [4] 和层次分析 [5] 等方法，这些方法大多使用截面数据对上市公司绩效进行分析，较少考虑时间序列动态发展中所蕴含的信息。Bonzo D C 和 Hermosila A Y (2002)[6] 首次将多元统计方法引入到面板数据分析中，并利用自适应启发式模拟退火遗传算法改进了聚类方法；朱建平，陈民恳 (2007)[7] 从动态角度出，构建面板数据间相似性指标解决静态聚类问题，提高了面板数据聚类有效性；郑兵云 (2008)[8] 重新定义多指标面板数据距离函数和离差平方和函数，对我国各地区工业企业生产率数据进行聚类，验证了模型的有效性。任娟 (2012)[9] 在系统聚类基础上，加入面板数据增量指标和增量变化率指标，对多指标面板数据进行聚类。任娟 (2013)[10] 基于多元统计分析视角提出一个改进的因子分析和系统聚类方法，该方法依据 Fisher 有序聚类理论，构造了 Frobenius 范数形式的离差平方和函数，用于处理多指标面板数据有序聚类问题。王双英等 (2014)[11] 对我国 44 个行业一次能源消费面板数据降维处理后，利用自组织竞争神经网络进行聚类，并与传统聚类方法进行对比，证实提出的聚类模型优点显著。以上学者大多采用多元统计的方法构造距离函数和离差平方和函数进行系统聚类。系统聚类虽然类中心不断修正，但模式类别一旦指定后就不再改变，且系统聚类最终得到的是一个树状结构图，从图中不能确定类的最佳个数，存在灵活性差等弊端。本文尝试将扩大时间维度上动态信息的多指标面板数据与具有智能特性的吸引子传播模型相融合，有效的克服了这一缺陷。另外，考虑到上市公司数据的多样性、复杂性和较少的可预知性，以及从众多样本中发现特征明显的少量样本的容易性，本文还将半监督学习策略纳入到吸引子传播聚类模型体系内，提出一种智能信息处理的上市公司聚类绩效评价模型。

1 多指标面板数据

多指标面板数据的结构具有三个维度：时间维度、样本维度、指标维度。为了更清晰展现多指标面板数据形式，可将其转换成一个二维表，如表 1 所示。

表 1 多指标面板数据二维表

时间	1	⋯	t	⋯	T
样本	$M_1\cdots M_j\cdots M_p$	⋯	$M_1\cdots M_j\cdots M_p$	⋯	$M_1\cdots M_j\cdots M_p$
1	$X_{11}(1)\cdots X_{1j}(1)\cdots X_{1p}(1)$	⋯	$X_{11}(t)\cdots X_{1j}(t)\cdots X_{1p}(t)$	⋯	$X_{11}(T)\cdots X_{1j}(T)\cdots X_{1p}(T)$
2	$X_{21}(1)\cdots X_{2j}(1)\cdots X_{2p}(1)$	⋯	$X_{21}(t)\cdots X_{2j}(t)\cdots X_{2p}(t)$	⋯	$X_{21}(T)\cdots X_{2j}(T)\cdots X_{2p}(T)$
3	$X_{31}(1)\cdots X_{3j}(1)\cdots X_{3p}(1)$	⋯	$X_{31}(t)\cdots X_{3j}(t)\cdots X_{3p}(t)$	⋯	$X_{31}(T)\cdots X_{3j}(T)\cdots X_{3p}(T)$
⋯	⋯	⋯	⋯	⋯	⋯
i	$X_{i1}(1)\cdots X_{ij}(1)\cdots X_{ip}(1)$	⋯	$X_{i1}(t)\cdots X_{ij}(t)\cdots X_{ip}(t)$	⋯	$X_{i1}(T)\cdots X_{ij}(T)\cdots X_{ip}(T)$
⋯	⋯	⋯	⋯	⋯	⋯
N	$X_{N1}(1)\cdots X_{Nj}(1)\cdots X_{Np}(1)$	⋯	$X_{N1}(t)\cdots X_{Nj}(t)\cdots X_{Np}(t)$	⋯	$X_{N1}(T)\cdots X_{Nj}(T)\cdots X_{Np}(T)$

其中，样本总量为 N，每个指标用 M_j 表示，共有 p 个指标数，时间长度为 T。$X_{ij}(t)$ 表示第 i 个样本在 t 时间的第 j 个指标数值。多指标面板数据可以获得样本在不同时间点上的水平指标值，为了全面反映样本指标的动态变化特征，本文在原有指标基础上，引入每个指标的变化增量和变化速度。这样便增加了样本时间维度上的动态信息，有利于分析样本的各种特性。

2 吸引子传播聚类与半监督学习

吸引子传播聚类模型是 2007 年由 Frey 和 Dueck[12] 在《Science》上提出的。与其他聚类模型相比，该模型具有快速、高效、聚类效果稳定且不必预先给定聚类数目的特点，能很好

的解决大规模数据处理问题。在 Frey 和 Dueck 提出吸引子传播聚类模型一年后，Brusco 和 Frey[13−14] 在《Science》杂志中再次发表了讨论该模型的两篇论文，进一步证明该模型用于大规模数据聚类时明显优于其它聚类模型。因此，吸引子传播聚类模型已经成为数据挖掘领域的一种重要工具。

2.1 吸引子传播聚类模型

给定数据样本点集合 $D=\{x_1,x_2,\cdots,x_N\}$，$x_i=\{x_{i1},x_{i2},\cdots,x_{id}\}$ $(i=1,2,\cdots,N)$ 为集合中一个数据样本点。其中，N 为样本数目，d 为样本属性维度。该模型是在 N 个样本点的相似度矩阵上进行聚类。首先，计算样本 x_i 与样本 x_k 之间相似度为 $s(i,k)$，形成一个 $N\times N$ 的相似度矩阵作为模型的输入。其中，$s(i,k)=-||x_i-x_k||^2$ $(i,k=1,2,\cdots,N;\ i\neq k)$ 用其来表示样本 x_k 适合做样本 x_i 类代表点的程度。若 $s(i,k)\to 0$ 表示样本 x_k 与样本 x_i 之间的相似度最大；反之，若 $s(i,k)\to -\infty$ 则表示样本 x_k 与样本 x_i 之间不存在相似之处，即二者不在同一个类中。此外，吸引子传播聚类模型还为每个样本设定一个偏向参数 $s(k,k)$，反映样本 x_k 成为类代表点的可能性大小，偏向参数越大说明样本 x_k 成为类代表点的可能性越大。模型初始假定每个样本被选作类代表点的可能性是相同的，即设定所有 $s(k,k)$ 为相同值 P，为 S 中所有非对角线元素均值。吸引子传播聚类模型基本思想是将所有样本作为潜在的聚类中心，样本之间通过归属度 $a(i,k)$ 和吸引度 $r(i,k)$ 两种消息不断传递，目的是寻找到最优类代表点集合，使得所有样本点与最近的类代表点之间相似度之和最大。其中，$a(i,k)$ 是样本 x_k 发送给样本 x_i 的信息量，表示样本 x_i 所积累的证据，反映样本 x_i 选则样本 x_k 为类代表点的适合程度；$r(i,k)$ 是样本 x_i 发送给样本 x_k 的信息量，表示样本 x_k 所积累的证据，反映样本 x_k 为样本 x_i 类代表点的代表程度。图 1 给出了归属度与吸引度传递的示意图。

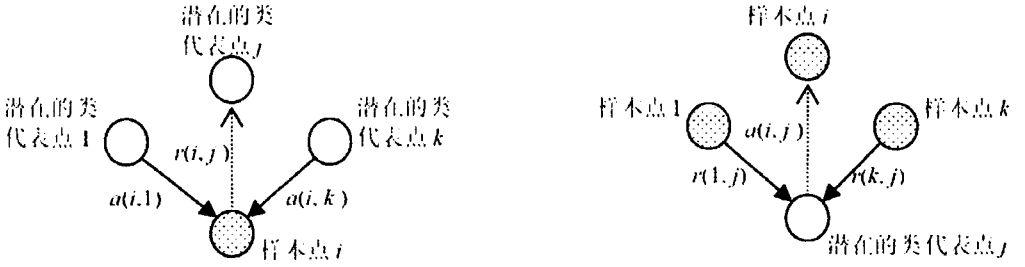

图 1 归属度与吸引度传递的示意图

吸引子传播聚类模型的迭代过程就是 $a(i,k)$ 和 $r(i,k)$ 迭代更新的过程，初始阶段设定 $a(i,k)=0$，$r(i,k)=0$。两种信息量具体迭代过程如下：

$$a(i,k)=\begin{cases}\min\left\{0,r(k,k)+\sum\limits_{i's.t.i'\notin\{i,k\}}\max\{0,r(i',k)\}\right\}, & i\neq k,\\ \sum\limits_{i's.t.i'\neq k}\max\{0,r(i',k)\}, & i=k,\end{cases}\tag{1}$$

$$r(i,k)=s(i,k)-\max_{k's.t.k'\neq k}\{a(i,k')+s(i,k')\}.\tag{2}$$

为了避免模型在迭代过程中发生振荡，在信息更新过程中引入阻尼因子 $\lambda\in[0,1)$，分别对当前 $a(i,k)$ 和 $r(i,k)$ 的值与上一步迭代结果加权求和得到更新的 $a(i,k)$ 和 $r(i,k)$，本文选取 λ 为默认值 0.5。

$$r(i,k)^{(t+1)}=\lambda r(i,k)^{(t)}+(1-\lambda)r(i,k)^{(t-1)},\tag{3}$$

$$a(i,k)^{(t+1)} = \lambda a(i,k)^{(t)} + (1-\lambda)a(i,k)^{(t-1)}. \tag{4}$$

计算所有样本点的归属度 $a(i,k)$ 和吸引度 $r(i,k)$ 之和，当聚类中心稳定或者达到最大迭代次数时，根据公式 (5) 获得最优类代表点以及样本点与其类代表点的从属关系。

$$\underset{k}{\operatorname{argmax}}\big(a(i,k)+r(i,k)\big). \tag{5}$$

模型中需要设定的参数只有最大迭代次数、λ 和偏向参数。偏向参数通过公式 (2) 进入模型，对吸引度 $r(i,k)$ 产生影响，进而影响归属度 $a(i,k)$。偏向参数的大小会影响聚类的个数，降低 P 值会减少类的数目，增大 P 值会增加类的数目。

2.2 半监督学习

半监督学习是一种介于无监督和监督学习之间的方法，既能依靠学习规则，挖掘样本自身蕴含的信息，同时也可以通过少量已知样本指导大量未知样本的学习，避免有监督学习时标记大量样本所带来的困惑。具体可以描述为：给定一个未知分布的样本集合 $X = L \cup U$，$L = \{(x_1,y_1),(x_2,y_2),\cdots,(x_{NL},y_{NL})\}$ 为已标签的样本集合，$U = \{x'_1,x'_2,\cdots,x'_{NU}\}$ 为未标签样本集合，希望找到一个学习器 $f := X \to Y$ 能够准确对样本 x 预测其标签 y，其中 NL 与 NU 分别表示集合 L 和集合 U 中样本数目 [15]。

半监督聚类是半监督学习的一个重要研究方向，利用少量的具有先验信息数据辅助聚类，以提高聚类模型的精度，具有较好的聚类性能。先验信息数据一般分为两类：第一类是某些样本已带有确定类标签；第二类是成对点约束信息。Wagstaff[16] 提出两种类型的成对点约束，即 Must-linked，$M = \{(x_i,x_j)\}$ 集合和 Cannot-linked，$C = \{(x_i,x_j)\}$ 集合。$M = \{(x_i,x_j)\}$ 集合表示样本点 x_i 和 x_j 必须属于同一类，$C = \{(x_i,x_j)\}$ 集合限定样本点 x_i 和 x_j 不在同一类。

3 智能信息处理的多指标面板数据聚类模型

3.1 相似度矩阵计算

在考虑多指标面板数据时间序列动态发展特征后，即增加指标变化增量、变化速度。具有时间动态信息的样本 X_i 第 j 个指标表示为：

$$\begin{aligned} X_{ij} = \ & \big(X_{ij}(1)\cdots X_{ij}(t)\cdots X_{ij}(T), Growth_{ij}(2)\cdots Growth_{ij}(t)\cdots Growth_{ij}(T), \\ & Growth\_rate_{ij}(2)\cdots Growth\_rate_{ij}(t)\cdots Growth\_rate_{ij}(T)\big), \end{aligned} \tag{6}$$

其中，$X_{ij}(t)$ 表示第 i 个样本 t 时刻的第 j 个指标数值，$Growth_{ij}(t)$ 表示第 i 个样本 t 时刻第 j 个指标变化增量，$Growth\_rate_{ij}(t)$ 表示第 i 个样本 t 时刻指第 j 个指标变化速度。

$$Growth_{ij}(t+1) = X_{ij}(t+1) - X_{ij}(t), \tag{7}$$

$$Growth\_rate_{ij}(t+1) = \frac{X_{ij}(t+1) - X_{ij}(t)}{X_{ij}(t)}. \tag{8}$$

因此，样本 α 和样本 β 的相似度选择欧氏距离公式计算，可表示为：

$$S(\alpha,\beta) = -\sqrt{(X_\alpha - X_\beta)(X_\alpha - X_\beta)^T}. \tag{9}$$

获得相似度矩阵后，根据先验信息对相似度矩阵进行调整：①如果样本 x_i 和 x_j 属于 Must-linked 集合，则 $S(x_i,x_j) = S(x_j,x_i) = 0$；②如果样本 x_i 和 x_k 不属于 Must-linked 集合，

而样本 x_i 和 x_j 属于 Must-linked 集合，样本 x_j 和 x_k 也属于 Must-linked 集合，则 $S(x_i,x_k)=S(x_k,x_i)=0$；③如果样本 x_i 和 x_j 属于 Cannot-linked 集合，则 $S(x_i,x_j)=S(x_j,x_i)=-\infty$。为了获得较为理想的分类数目，本文采用公式 (10) 对偏向参数 P 进行计算。

$$s(i,i)=\frac{\varphi\sum_{i,j=1;i\neq j}^{N}s(i,j)}{N(N-1)}, \tag{10}$$

其中，φ 为调节权，N 为样本数。

3.2 聚类流程

智能信息处理的多指标面板数据聚类模型具体流程如下：

1) 根据公式 (9) 计算样本相似度矩阵 S，根据公式 (10) 计算偏向参数 P，并根据成对点约束原则对相似度进行调整；

2) 设定最大迭代次数并初始化 $a(i,k)$ 和 $r(i,k)$，利用公式 (1) 至公式 (4) 获得每一次迭代后的 $A=\big(a(i,k)\big)_{N\times N}$ 和 $R=\big(r(i,k)\big)_{N\times N}$，根据 $R(k,k)+A(k,k)$ 值来判断是否为聚类中心，当 $R(k,k)+A(k,k)>0$ 时，认为是一个聚类中心；

3) 当达到最大迭代次数或聚类中心连续多少次不发生改变时，输出聚类结果。否则，更新吸引度矩阵和归属度矩阵，重新确定类代表点。

4 实证应用

4.1 聚类评价指标

(1) Silhouette 指标 [17]

假定一个样本容量为 N 的数据集被分为 k 类 C_i $(i=1,2,\cdots,k)$，其中，$a(m)$ 表示类 C_j 中的样本 m 与 C_j 内其他样本的平均距离，$d(m,C_i)$ 表示 C_j 中的样本 m 到另一个类 C_i 中所有样本的平均距离，$b(m)=\min\{d(m,C_i)\}$，其中 $i=1,2,\cdots,k$ 且 $i\neq j$，则样本 m 的 Silhouette 指标计算公式为 $Sil=\big(b(m)-a(m)\big)/\max\{a(m),b(m)\}$。一个数据集中所有样本的平均 Sil 值不仅能够反映聚类结构的类内紧凑性并且也能很好的反映类间可分性，Sil 值越大说明聚类质量越好。

(2) DB 指标 [18]

假定类内部离散度为 S_i，两类之间分离度为 D_{ij}，DB 指标对类间相似度 R_{ij} 满足下列条件：① $R_{ij}\geq 0$；② $R_{ij}=R_{ji}$；③若 $S_i=0$ 且 $S_j=0$，则 $R_{ij}=0$；④若 $S_i>S_j$ 且 $D_{ij}=D_{ik}$，则 $R_{ij}=R_{ik}$；⑤若 $S_i<S_j$ 且 $D_{ij}<D_{ik}$，则 $R_{ij}>R_{ik}$。DB 指标定义为

$$DB=\sum_{i=1}^{N_C}\frac{R_i}{N_C}, \tag{11}$$

其中，$R_{ij}=(S_i-S_j)/D_{ij}$，$S_i=\sum_{x\in C_i}D(x,v_i)/N_i$，$D_{ij}=D(v_i,v_j)$，$R_i=\max_{j=1,\cdots,N_C}(R_{ij})$，$v_i,v_j$ 表示类 C_i 和 C_j 的质心，N_i 表示类 C_i 中样本个数，$D(\cdot,\cdot)$ 表示距离函数，N_C 表示类的数目。DB 指标数值越小，表示聚类结果越好。

4.2 数据选取

本文选取 30 家房地产业上市公司 2009-2013 年报财务数据，应用智能信息处理的多指标面板数据聚类模型，基于获利能力、运营能力、偿债能力和发展能力四个方面对上市公司的绩

效状况进行评价。其中，获利能力由净资产收益率 (ROE)、主营业务利润率 (MOM)、每股收益 (EPS) 三个指标测度，运营能力由存货周转率 (ITR)、应收账款周转率 (RTR)、总资产周转率 (TAT) 三个指标测度，偿债能力由流动比率 (CR)、速动比率 (ATR)、资产负债率 (DAR) 三个指标测度，发展能力由净利润增长率 (NPGR)、主营业务收入增长率 (MBRG)、总资产增长率 (TAGR) 三个指标测度。相应数据来自 RESSET 金融研究数据库。首先将原始数据标准化，转换为无量纲的数值，标准化公式为：$X_{ij} = (X_{ij} - \overline{X}_j)/S_j$，其中，$\overline{X}_j = \sum_{i=1}^{n} X_{ij}/N$，$S_j = \sqrt{\sum_{i=1}^{n}(X_{ij} - \overline{X}_j)^2/N}$，$N$ 表示样本总数，X_{ij} 中的 i 是指上市公司的数量，j 指加入时间序列动态发展特征的多指标面板数据的列数，X_{ij} 表示标准化后的数据；$\overline{X}_j$ 表示第 j 列的均值，S_j 表示第 j 列的方差。

4.3 聚类结果

本文采取了 Silhouette 指标和 DB 指标评价聚类的有效性。为能更好的说明本文聚类方法的优越性，我们选用 AP 模型、SOM 神经网络算法和系统聚类方法与本文模型进行对比。对于 AP 模型和 SOM 神经网络模型聚类数目均为自动产生，因此无需指定。而系统聚类数目需要人为确定，为能够最大限度减少误差，设定系统聚类与本文模型获得相同类数，具体结果如表 2 所示。

表 2 聚类评价指标对比表

评价指标	本文模型	AP 模型	SOM 神经模型	系统聚类
Silhouette	0.791	0.502	0.366	0.607
DB	0.211	0.417	0.435	0.383

通过对比可以看出，智能信息处理的多指标面板数据聚类模型得到的聚类结果 Silhouette 指标值最高，DB 指标值最低，优于其他三种聚类方法，表 3 列出了本文聚类算法获得的聚类结果。

表 3 智能信息处理的多指标面板数据聚类模型聚类结果表

类别	上市公司代码			
第一类	C000002	C000024	C002146	C002305
	C600048	C600266	C600743	
第二类	C000608	C002077	C600239	
第三类	C000040	C000502	C000505	C600053
第四类	C000537	C000558	C000965	C002016
	C600173	C600223	C600240	C600246
	C600322	C600325	C600376	C600383
	C600657	C600665	C600745	C600185

4.4 结果分析

为更明晰地反映每一类上市公司财务指标整体经营状况，本文给出每类上市公司各指标均值表，如表 4 所示。

从获利能力角度分析，第一类上市公司的三个指标平均值均处于最高水平，其净资产收益率和主营业务利润率分别超过行业绩效评价优秀值 3.3 个百分点和 1.2 个百分点，2009 年到 2013 年间净资产收益率年均增加 0.77 个百分点，增速达到 5%，每股收益年均增量为 0.15

元，增速达到 16%；第三类上市公司的获利指标平均值都处于最后一位，且净资产收益率和每股收益表现为负增长，净资产收益率年均下降 1.38 个百分点，增速为 −78%，每股收益年均下降 0.02 元，增速为 -77%，获利能力明显不如其他三类上市公司。可以看出，以深万科 A 为代表的第一类 7 家上市公司获利能力处于行业领先地位，可为投资者带来稳定收益，而第三类上市公司获利能力处于行业下游，有待提升。

表 4 各类上市公司指标均值表

类别	获利能力			运营能力			偿债能力			发展能力		
	ROE (%)	MOM (%)	EPS (元)	ITR (次)	RTR (次)	TAT (次)	CR (%)	ATR (%)	DAR (%)	NPGR (%)	MBRG (%)	TAGR (%)
第一类	19.854	30.926	0.956	0.264	6.556	0.290	1.891	0.619	71.618	33.200	39.719	36.821
第二类	8.729	30.605	0.310	0.829	2.651	0.293	1.797	1.044	69.409	131.185	87.790	54.797
第三类	2.541	25.411	0.078	0.656	3.357	0.305	1.634	0.522	57.982	−5.721	64.803	5.294
第四类	14.679	26.599	0.419	0.273	5.139	0.282	1.954	0.502	68.138	26.885	21.815	29.786

从运营能力角度分析，四类总资产周转率五年的平均水平相差不大，存货周转率第二类平均值最高为 0.829 次，但从第二类三家上市公司的五年时间序列数据发现呈明显的下降趋势，阳光新业地产股份有限公司 (C000608) 的存货周转率由 2009 年的 0.126 次降至 2013 年的 0.027 次，下降 78.58%，江苏大港股份有限公司 (C002077) 由 2009 年 2.051 次降至 2013 年的 1.712 次，下降 16.53%，云南城投置业股份有限公司 (C600239) 由 2009 年 1.034 次降至 2013 年的 0.181 次，下降 82.50%。第一类上市公司存货周转率明显低于其他几类，这是由于该类企业在扩大业务规模同时确保一定量货源，导致存货规模增长速度大于其销售增长速度，表现出低存货周转率，其应收账款周转率最高，年均增量为 0.21 次，增速为 17%，其他三类均出现不同程度的负增长。第一类公司在获利能力逐年增加的同时，总资产周转率保持较好的增长势头，表明该类企业经营稳健且运营能力强。

从偿债能力角度分析，第三类上市公司流动比率在五年中平均增量为 0.09%，增速为 67%，而第一、二类出现负增长，第四类虽有增长，但与第三类相比增长缓慢。第三类速动比率表现出正增长，其余三类均为负增长，且资产负债率平均值低于行业平均水平 58.14%，平均增速最低。这表明第三类上市公司资本结构较合理，财务风险较低，偿债能力较强。

从发展能力角度分析，第二类上市公司获利能力、运营能力、偿债能力并不优秀，但第二类上市公司发展能力指标都是最高值，表现出良好的发展能力。该类公司拥有一定的潜在发展能力，未来将会有一定的发展空间。

除了从获利能力、运营能力、偿债能力、发展能力进行分析，进一步对每一类公司进行综合评价，给出了每一类上市公司的经营特征，可以更清楚了解每类上市公司经营获利状况，如表 5 所示。

表 5 各类上市公司绩效综合评价表

类别	获利能力	运营能力	偿债能力	发展能力	特征
第一类	优秀	优秀	中等	良好	运营获利型
第二类	中等	一般	一般	优秀	发展型
第三类	一般	中等	优秀	一般	偿债型
第四类	良好	良好	良好	中等	一般型

5 结论

本文在吸引子传播聚类模型基础上，引入半监督学习策略，充分利用少量先验信息引导聚类，并融合多指标面板数据提取上市公司时间维度信息，改善了原模型聚类性能。针对上市公司绩效评价问题，本文的评价模型给出了全新视角，通过对评价结果进行分析，政府和金融监管机构能够及时、准确发现问题，为制定相应经济政策和监管措施提供有效的决策依据。同时，投资者也可以多角度了解上市公司，进行科学合理的投资组合并防范和规避投资风险。

[参考文献]

[1] Hao H, Wang Z, Xu H. The evaluation of listed banks' competitiveness based on principal component analysis [A]. International Conference on Management Science and Engineering [C]. IEEE, 2010, 5: 110–114.

[2] Liu H F, Wang J. Integrating independent component analysis and principal component analysis with neural network to predict Chinese stock market [J]. Mathematical Problems in Engineering, DOI: 10.1155/2011/382659, 2011.

[3] Huang Y D, Du L B. Evaluation of performance of listed companies using factor analysis-take listed companies in Luzhong region as an example [A]. International Conference on Regional Management Science and Engineering [C]. Melbourne: IEEE, 2010: 711–715.

[4] Gao C, Fan Z, Zhang J. New energy listed companies competitiveness evaluation based on modified data envelopment analysis model [A]. International Conference on Intelligent Computing and Information Science [C]. Springer, 2011, 135: 613–618.

[5] Chen J. The analysis of the investment value of the real estate listed companies based on the AHP [A]. International Conference on Engineering and Business Management [C]. Scientific Research Publishing, 2011: 2058–2062.

[6] Bonzo D C, Hermosilla A Y. Clustering panel data via perturbed adaptive simulated annealing and genetic algorithms [J]. Advances in Complex Systems, 2002, (4): 339–360.

[7] 朱建平, 陈民恳. 面板数据的聚类分析及其应用 [J]. 统计研究, 2007, 24(4): 11–14.

[8] 郑兵云. 多指标面板数据的聚类分析及其应用 [J]. 数理统计与管理, 2008, 27(2): 265–270.

[9] 任娟. 多指标面板数据聚类方法及其应用 [J]. 统计与决策, 2012, (4): 92–95.

[10] 任娟. 多指标面板数据融合聚类分析 [J]. 数理统计与管理, 2013, (32)1: 57–67.

[11] 王双英, 王群伟, 曹泽. 多指标面板数据聚类方法及应用 — 以行业一次能源消费面板数据为例 [J]. 数理统计与管理, 2014, 01: 42–49.

[12] Frey B J, Dueck D. Clustering by passing messages between data points [J]. Science, 2007, 315(5814): 972–976.

[13] Brusco M J, Kohn H F. Comment on clustering by passing messages between data points [J]. Science, 2008, 319(5864): 726c.

[14] Frey B J, Dueck D. Response to comment on Clustering by passing messages between data points [J]. Science, 2008, 319(5864): 726d.

[15] 梁吉业, 高嘉伟, 常瑜. 半监督学习研究进展 [J]. 山西大学学报 (自然科学版), 2009, 4: 528–534.

[16] Wagstaff K, Cardie C, Rogers S, et al. Constrained K-means clustering with background knowledge [A]. 18th International Conference on Machine Learning [C]. San Francisco: Morgan Kaufmann Publishers Inc., 2001: 577–584.

[17] Rouseeuw P J. Silhouettes:a graphical aid to the interpretation and validation of cluster analysis [J]. Journal of Computational and Applied Mathematics, 1987, 20(1): 53–65.

[18] Davies D L, Bouldin D W. A cluster separation measure [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1979, (2): 224–227.