

# 改进的 K-means 算法在电信客户细分中的应用

耿筱媛, 张燕平, 闫屹

(安徽大学 计算智能与信号处理教育部重点实验室, 安徽 合肥 230039)

**摘要:**在 K-means 算法中, 选择不同的初始聚类中心会产生不同的聚类结果且有不同的准确率, 并且其迭代过程在时间上不是高效的。针对 K-means 算法的这两点不足做了一定程度上的改进, 理论分析表明, 改进后的算法具有较高的准确度和较低的时间复杂度。采用改进后 K-means 聚类算法对电信客户数据进行聚类分析, 得到具有不同特征的客户群组, 通过与统计分析的对比, 聚类结果分析更合理清晰, 更便于对不同群组采取不同的经营策略, 为管理者提供了合理的决策支持。

**关键词:**数据挖掘; 聚类算法; K-means 算法; 准则函数

**中图分类号:** TP301.6

**文献标识码:** A

**文章编号:** 1673-629X(2008)05-0163-05

## Application of Improved K-means Algorithm in Subdivision of Telecom Clients

GENG Xiao-yuan, ZHANG Yan-ping, YAN Yi

(Ministry of Education Key Lab. of Intelligent Computing and Signal Processing, Anhui University, Hefei 230039, China)

**Abstract:** In the K-means algorithm, selecting different initial centers that the algorithm begins with can produce different results and different accuracy, and the process of iterative is not high-efficiency. Investigates the standard K-means clustering algorithm and gives an improved algorithm by selecting better initial centers that the algorithm begins with, and improved its efficiency. The theoretic analysis shows that the improved K-means algorithm can get higher accuracy and better time-complexity. The application of this improved K-means algorithm in subdivision of telecom clients can improve the relationship between the organization and the customers and forecast the trend and behaviors to support people's decision. K-means algorithm is used to analyze sample data and discover varied characters of varied groups of customers. It can support enterprise to make an appropriate decision for each kind of customer.

**Key words:** data mining; clustering; K-means algorithm; Jc

## 1 数据挖掘与 K-mean 算法简介

### 1.1 数据挖掘过程

数据挖掘是一个处理过程。它是个多步骤过程, 包括挖掘数据、分析结果和采取行动, 被访问的数据可以存在于一个或多个操作型数据库中、一个数据仓库中或一个平面文件中。把数据库中的对象分类是数据挖掘的基本操作, 其准则是使属于同一类的个体间距离尽可能小, 而不同类个体间距离尽可能大, 它利用一种或多种计算机学习技术, 从数据库的数据中自动分

析并提取知识。为了找到效率高、通用性强的聚类方法人们从不同角度提出了近百种聚类方法, 典型的有 K-means 方法、K-medoids 方法、CLARANS 方法、BIRCH 方法等。

### 1.2 K-means 算法

K-means 算法是最常用的方法之一。K-means 算法产生局部最优解而不是全局最优解, 从不同的初始聚类中心出发会得到不同的聚类结果且准确率也不一样。通常 K-means 算法是从第  $n$  个数据对象中随机选择  $k$  个对象作为初始聚类中心, 这样就使得产生的聚类结果具有很大不确定性, 如何选择初始聚类中心点成了影响最后聚类结果的重要因素。

K-means 算法中, 一次迭代需要的总时间复杂度为  $O(nkd)$  ( $n$  指的是总的对象个数,  $k$  是指定的聚类数,  $d$  是数据对象的维数)。如果数据量比较大, 算法的时间开销也是相当可观的<sup>[1]</sup>, 处理大数据量时

收稿日期: 2007-08-28

基金项目: 国家自然科学基金项目 (60675031, 60475017); 973 计划 (国家重点基础研究) (2004CB318108); 安徽省教育厅重点自然科学基金项目 (2006kj015A); 安徽省教育厅自然科学基金项目 (2005kj053); 安徽大学 211 工程学术创新团队

作者简介: 耿筱媛 (1983-), 女, 安徽巢湖人, 硕士研究生, 研究方向为人工智能计算; 张燕平, 博士, 教授, 研究领域为人工神经网络、机器学习。

开支较大。该文对随机选择初始聚类中心的方式进行了改进,尽量使最初的初始聚类中心在空间分布上与数据实际的分布相一致,并且在算法迭代的过程中进行改进,减少每次迭代的计算次数。实验表明该文的改进算法较在准确率上有较大提高。

## 2 基本思想

### 2.1 K-means 算法思想

K-means 是数据挖掘技术中的一种基于划分的聚类算法,因其理论上可靠、算法简单、收敛速度快而被广泛用<sup>[2]</sup>。

K-means 算法的目标是根据输入参数  $k$ , 将数据集划分成  $k$  个簇。算法采用迭代更新方法: 在每一轮中, 依据  $k$  个聚类中心将其周围的点分别组成  $k$  个簇, 而每个簇的质心(即簇中所有点的平均值, 也是几何中心) 将被作为下一轮迭代的聚类中心。迭代使选取的聚类中心越来越接近真实的簇质心, 所以聚类效果越来越好。

聚类过程如图 1 所示。

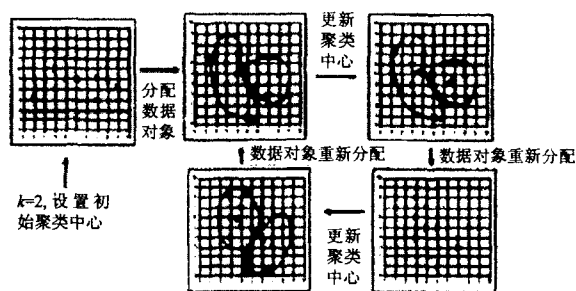


图 1 聚类过程

设将  $d$  维数据集  $X = \{x_i | x_i \in R^d, i = 1, 2, \dots, n\}$  聚集成  $k$  个簇  $w_1, w_2, \dots, w_k$ , 它们的质心依次为  $c_1, c_2, \dots, c_k$ , 其中  $c_i = \frac{1}{n_i} \sum_{x \in w_i} x$ ,  $n_i$  是簇  $w_i$  中数据点的个数。采用误差平方和准则函数作为目标函数来显式地判断算法是否结束, 见公式(1)。利用误差平方和准则函数能把真正属于同一类的样本聚合成一个类型的子集, 而把不同类的样本分开<sup>[3]</sup>。

$$J_c = \sum_{j=1}^k \sum_{i=1}^{n_j} |x_k^{(j)} - c_j(I)|^2 \quad (1)$$

当准则函数  $J_c$  收敛后, 算法结束。

K-means 算法步骤描述如下:

(1) 给定大小为  $n$  的数据集  $X$ , 令  $I = 1$ , 选取  $k$  个初始聚类中心  $c_j(I), j = 1, 2, 3, \dots, k$ 。

(2) 以  $c_j(I)$  为参照点对  $X$  进行划分, 计算每个样本数据对象与聚类中心的距离。若  $d(x_i, c_k(I)) = \min\{d(x_i, c_j(I)), i = 1, 2, \dots, n\}$ , 其中  $j = 1, 2, \dots,$

$k, i = 1, 2, \dots, n$ , 则将  $x_i$  划分到簇  $w_k$ 。

(3) 令  $I = I + 1$ , 根据公式  $c_j(I) = \frac{1}{n} \sum_{x \in w_j} x$  计算新的聚类中心和误差平方和准则函数值。

(4) 若  $|J_c(I+1) - J_c(I)| < \xi$  成立, 则算法结束。否则, 令  $I = I + 1$ , 返回(2) 执行。

### 2.2 算法改进

#### 2.2.1 聚类中心的选择

在 K-means 算法中, 选取  $k$  个点作为初始聚类中心点, 然后进行迭代操作, 初始点选取不同可能获得不同的聚类结果。进行数据划分目的是让一个聚类中的对象是相似的, 而不同聚类中的对象是不相似的<sup>[2]</sup>。用距离表示对象间的相似程度, 相似的对象间的距离比不相似的对象间的距离小。我们希望找到与数据在空间分布上相一致的初始聚类中心。如果能够找到  $k$  个初始中心, 它们分别代表了相似程度较大的数据集<sup>[4]</sup>, 那么就找到了与数据在空间分布上相一致的初始聚类中心。为了找到与数据在空间分布上相一致的、相似程度较大的数据集而采取下列步骤:

计算样本点两两之间的距离:

(1) 找出距离最近的两个点形成一个样本集  $U_1$ , 并将它们从总的样本集  $T$  中删除;

(2) 计算  $U_1$  中每一个样本与  $T$  中每一个样本的距离, 找出在  $T$  中与  $U_1$  中最近的点, 将它并入集合  $U_1$  并从  $T$  中删除;

(3) 直到  $U_1$  中的样本个数到达一定阈值;

(4) 再从  $T$  中找到样本两两间距离最近的两个点构成  $U_2$ ;

(5) 重复上面的过程直到形成  $k$  个点集, 最后对  $k$  个点集分别进行算术平均形成  $k$  个初始聚类中心。

假设有一个 2 维数据集包含有 10 个样本, 它们的分布如图 2 所示。

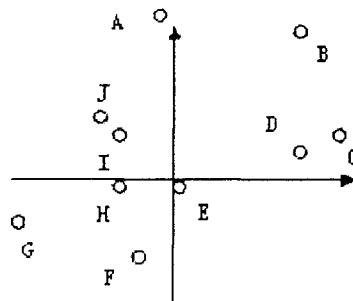


图 2 二维数据集样本分布图

阈值  $\alpha$  的取值因实验数据不同而有所不同,  $\alpha$  的取值过小则可能使几个初始聚类中心点在同一区域得到,  $\alpha$  的取值过大则可能使初始聚类中心点偏离密集区域, 从实验的情况来看  $\alpha$  取 0.75 时效果是比较好的。

的<sup>[4]</sup>。

假设要把它们划分为两类,按照上面的思想寻找初始聚类中心。 $J, I$  之间的距离最近,那么选择  $J, I$  构成一个样本集  $U_1$  并将它们从总的集合  $T$  中删除,  $U$  中与  $U_1$  相邻最近的点是  $H$ , 这样便将  $H$  加入  $U_1$  集合并将它从  $T$  中删除。如果规定每个样本集中样本最大个数为 4 则  $U_1$  中将会再添加样本  $E$ , 然后在  $T$  中再找出相互之间距离最近的两个点  $C, D$  构成  $U_2$ , 并将它们从  $T$  中删除。 $T$  中与  $U_2$  相邻最近的点是  $B$ , 这样便将  $B$  加入  $U_2$  并将它从  $T$  中删除, 同样  $A$  也会并入  $U_2$ 。最后将这两个样本集分别进行算术平均形成两个初始聚类中心。这样得到的初始聚类中心与实际样本的分布更加相符, 从而可以得到更好的划分效果。在聚类中欧氏距离有着非常直观的意义, 文中样本点之间的距离采用欧氏距离样本<sup>[5]</sup>。

$X = (x_1, x_2, x_3, \dots, x_n)$  和样本  $Y = (y_1, y_2, \dots, y_n)$  之间的距离按下式计算:

$$d(X, Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2}$$

一个样本点与一个样本集的距离定义为这个样本点与这个样本集中所有样本点当中最近的距离, 则一个样本点  $X$  和一个样本集合  $V$  之间的距离定义如下:

$$d(X, V) = \min(d(X, Y), Y \in V)$$

### 2.2.2 减少迭代次数

在 K-means 算法中, 每次迭代是把每一个数据对象分到离它最近的聚类中心所在类, 这个过程的时间复杂度为  $O(nkd)$ 。 $n$  指的是总的对象个数,  $k$  是指定的聚类数,  $d$  是数据对象的维数。每次迭代后产生新的分类, 对于每个新的分类都要重新计算其聚类中心, 这个过程的时间复杂度为  $O(nd)$ 。因此这个算法一次迭代需要的总时间复杂度为  $O(nkd)$ 。当  $n$  较大时, 算法的时间开销也较大<sup>[1]</sup>, 所以算法在处理大的数据集时时间开支较大。文中运用三角形两边之和大于第三边的定律<sup>[6]</sup>, 减少 K-means 算法中每次迭代的计算次数。

在 K-means 算法的第一个循环阶段, 每次迭代中要计算每一个样本数据到各个聚类中心的距离, 依次比较得到与之距离最小的一个聚类中心, 并被分配到这个类中。在 K-means 算法中采用的是欧几里得距离, 因此可以考虑借用几何三角形中三边关系定理: 两边之和大于第三边, 从而简化比较过程, 减少运行时间开支。令  $x_i \in X$ ,  $d(c_m, c_n)$  为二个聚类中心的距离,  $d(c_m, c_n)$ 、 $d(x_i, c_m)$  与  $d(x_i, c_n)$  三边构成了一个如图 3 所示的三角形, 则有:

$$d(c_m, c_n) \leq d(x_i, c_m) + d(x_i, c_n)$$

$$\text{即: } d(c_m, c_n) - d(x_i, c_m) \leq d(x_i, c_n)$$

如果  $d(c_m, c_n) \geq 2d(x_i, c_m)$ , 则有:  $d(x_i, c_m) \leq d(x_i, c_n)$ , 即  $x_i$  到中心  $c_n$  的距离比到  $c_m$  的距离大。因此在  $d(c_m, c_n) \geq 2d(x_i, c_m)$  的前提下, 就不必计算  $d(x_i, c_n)$  了。

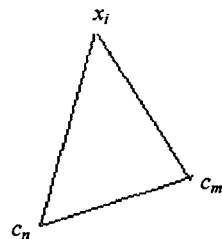


图3 三角形

## 3 算法和实验

### 3.1 算法

经过上述 2 点改进后的 K-means\_1 算法描述如下:

(1) 计算任意两个样本间的距离  $d(X, Y)$ , 找到集合  $T$  中距离最近的两个点形成集合  $U_m (1 \leq m \leq k)$ , 从集合  $T$  中删除这两个点。

(2) 在  $T$  中找到距离集合  $U_m$  最近的点将其加入集合  $U_m$ , 并从集合  $T$  中删除该点。

(3) 重复第 2 步直到集合中的样本点个数大于等于  $\alpha n/k (0 < \alpha \leq 1)$ 。

(4) 如果  $m < k$ ,  $m++$ , 再从集合  $T$  中找到距离最近的两个点形成新的集合  $U_m$ , 并从集合  $T$  中删除这两个点, 返回第 2 步执行。

(5) 将最终形成的  $k$  个集合中的样本点分别进行算术平均从而形成  $k$  个初始聚类中心  $c_j(I), j = 1, 2, 3, \dots, k$ 。

(6) 从这  $k$  个初始聚类中心出发计算每两个聚类中心间的距离  $d(c_i(I), c_j(I))$ , 其中  $i = 1, 2, \dots, k; j = 1, 2, \dots, k$ 。

(7) 设  $x_i$  当前所在类为  $U_m$ , 计算  $x_i$  与  $U_m$  类中心的距离  $d(x_i, c_m(I))$ , 若  $d(c_m(I), c_j(I)) \geq 2d(x_i, c_m(I))$  不成立, 则计算  $d(x_i, c_j(I))$ ; 若  $d(x_i, c_j(I)) < d(x_i, c_m(I))$ , 则暂时将  $x_i$  分配到  $A_j$ , 返回(7)循环运行, 最终将  $x_i$  划分到簇  $U_m$  中。其中  $j = 1, 2, \dots, k; i = 1, 2, \dots, n; m = 1, 2, \dots, n$ 。

(8) 令  $I = I + 1$ , 根据公式  $c_i = \frac{1}{n} \sum_{x \in w_i} x$  计算新的聚类中心和误差平方和准则函数的值。

(9) 若  $|J_c(I+1) - J_c(I)| < \xi$  成立, 则算法结

束。否则,令  $I = I + 1$ , 返回(6) 执行。

对算法 K-means 和 K-means<sub>1</sub> 做一个比较。在第二个循环阶段重新计算聚类中心时, 这二个算法的时间复杂度是相同的。但是在第一个循环阶段指定聚类簇时, 改进的 K-means<sub>1</sub> 算法显然减少了计算量。先考虑一个样本点的情况。在 K-means 算法中, 计算样本点到各中心点的距离的次数是  $k$  次, 而 K-means<sub>1</sub> 算法中, 在最好情况下计算样本点到各中心点的距离的次数是 1 次, 最坏情况下计算样本点到各中心点的距离的次数是  $k$  次。假设  $\alpha$  为第一循环阶段一次迭代时一个样本点的平均计算次数, 则有  $\alpha < k$ 。K-means<sub>1</sub> 算法一次迭代需要的总的时间复杂度为  $O(nkd)$ , K-means<sub>1</sub> 算法一次迭代需要总的时间复杂度为  $O(nad)$ 。如果数据集中的样本点较多, 即  $n$  比较大时, 算法的优越性就显示出来了。

### 3.2 实验结果分析

文中实验中采用某市电信公司 2007 年 1 月份一万多位客户的消费数据, 其中包括小灵通和座机。对原始数据进行预处理后产生 10 498 条有效记录。经过处理后的数据共有市话费、长途费、短信费、月租费、优惠费、总话费、长途占比、其他费这 8 维特征, 文中正是根据这 8 维特征对数据进行分析比较。

在未进行聚类前, 可以看出数据分布比较杂乱, 从统计学角度来看, 总话费消费在 100 元以下的客户有 6 603 位, 占了总人数的一半以上, 这类用户基本上的话费消费以市话费为主, 其他消费非常少; 总话费在 100~300 元之间的客户有 2710 位, 占总人数的 25.8%, 这类客户的消费特征并不十分明显, 主要集中在市话和长途消费上, 其他各项消费均有涉及, 且有部分用户享有很大程度的优惠; 300~500 元之间的的客户数只有 167 人, 约占总人数的百分之一, 可以看出这类用户的长途消费占比大多较高, 并享有优惠; 500~

1000 元的客户数有 1 014 位, 约为总人数的十分之一, 这类客户的市话消费在总话费中占比相当高, 其次是长途消费, 享有的优惠也很高, 可能是一些企业用户; 1000 元以上的客户只有 4 位, 长途话费构成他们消费的主要部分。

用文中的改进后的 K-means 算法对这些数据进行聚类, 并分为如下 10 类客户, 如表 1 所示。

#### 数据结果分析:

其中第八类是这样的一类群体, 有 4 679 个样本, 约占样本总数的 40%, 这一聚类人群的基本特征如下: 长途占比=0, 长话费=0, 其他费=0, 优惠费=0。这类人群是电信的主要客户, 由于他们的收入不高, 每月的话费也比较低, 主要是进行市话通话和短信, 没有长途通话。

其中第五类和第六类用户的长途话费明显高于其他用户, 这类客户的月消费额较高, 长途电话打的比较多, 占总话费的一半以上。虽然总人数不多, 但是给企业带来了可观的利润。

另外, 第二类 and 第十类用户的市话消费明显比长途话费要高, 并且这类客户的帐单优惠占比也很大, 因为他们使用了相应的优惠套餐。

第三类客户的短信费较高, 且帐单优惠较少。

鉴于以上对客户的分类, 可以调整一下营销策略:

对于第八类客户人数最多, 且只打市话, 对他们应该降低市话费用。

第二类 and 第十类用户鼓励继续使用原优惠套餐。

第五、六类用户长途占比高, 对其实行长途优惠, 如满 1000 分钟七折优惠, 并鼓励使用优惠套餐。

第三类用户短信费占比较高, 应鼓励其使用短信包月等优惠套餐。

以此吸引他们继续保持消费。

经过比较可以看出, 对数据进行聚类后, 数据分类

表 1 实验结果

|    | 市话费   | 长途费   | 其他费  | 短信费   | 月租费  | 优惠费   | 总话费    | 客户数   | 长途占比% |
|----|-------|-------|------|-------|------|-------|--------|-------|-------|
| 1  | 20.1  | 11.3  | 3.6  | 12.8  | 6    | 0     | 53.8   | 1 924 | 21.03 |
| 2  | 412.3 | 39.5  | 38.6 | 40.5  | 29.5 | -264  | 296.4  | 1 046 | 13.32 |
| 3  | 46.1  | 32.8  | 81.6 | 139.5 | 21.4 | -7.4  | 324.6  | 142   | 10.11 |
| 4  | 203.8 | 108.4 | 24.7 | 42.5  | 27.4 | -31.3 | 426.5  | 15    | 25.4  |
| 5  | 190.2 | 817.5 | 28.8 | 26.5  | 15   | -5.4  | 1072.6 | 4     | 76.22 |
| 6  | 451   | 317.6 | 18   | 17.4  | 0    | -471  | 333    | 10    | 95.1  |
| 7  | 33.6  | 23.7  | 11.9 | 24.6  | 8.8  | -1.6  | 114.2  | 735   | 20.77 |
| 8  | 44    | 0     | 0    | 0.3   | 5.8  | 0     | 50.1   | 4 679 | 0     |
| 9  | 88.7  | 53    | 18.2 | 34.2  | 15.7 | -7.7  | 211.8  | 929   | 25.01 |
| 10 | 672.3 | 145.1 | 62.5 | 34.7  | 95.2 | -457  | 564.9  | 1 014 | 25.69 |

更清晰,把具有相似特征的客户归为一类,不同特征的客户分为不同的类,可以清楚地看出各类用户的主要消费方向,比仅仅从统计学角度的分类更合理,便于有关部门对不同的客户制定更加合理有效的营销策略,提高企业效益。

#### 4 结束语

随着数据库和网络的飞速发展,聚类任务所涉及的数据规模越来越大,K-means 方法是聚类方法中常用的一种,当有计算资源和计算时间约束的情况时,它存在一定的局限性,而当数据规模很大时,这个矛盾更加突出<sup>[7]</sup>。文中提出的改进算法是一种适用于大规模数据处理的方法,它可以比较有目的地选取初始聚类中心,减小聚类结果对初值的依赖性,提高聚类的稳定性,使改进后的算法在准确度和耗费时间上都有所改善。文中采用聚类分析中的 K-means 算法,对电信行业的客户进行聚类,主要是根据客户的消费特征来进行的。以此有针对性地提供服务,提高营销政策的针对性和有效性。这样才能在激烈的市场竞争中获得主动地位,提高电信行业企业的效益和竞争力。

文中采用改进的 K-means 聚类算法对电信客户

数据进行聚类分析,得到具有不同特征的客户群组,对不同群组采取不同的经营策略,帮助管理者提供了合理的决策支持。该算法的改进思想也可以为其他领域客户细分析提供参考。

#### 参考文献:

- [1] Moore A W. The anchors hierarchy: Using the triangle inequality to survive high dimensional data [C]//In: Proc. UAI - 2000: The Sixteenth Conference on Uncertainty in Artificial Intelligence. New York: Springer, 2000.
- [2] Han J W, Kamber M. Data Mining: Concepts and Techniques [M]. San Francisco: Morgan Kaufmann, 2000.
- [3] Bult J R, Wansbeek T. Optimal selection for direct mail [J]. Marketing Science, 1997, 14(4): 321 - 324.
- [4] 袁方, 孟增辉, 于戈. 对 k-means 聚类算法的改进 [J]. 计算机工程与应用, 2004(36): 44 - 48.
- [5] Forgy E. Cluster analysis of multivariate data: Efficiency vs. interpretability of classifications [J]. Biometrics, 1965, 21(3): 768 - 771.
- [6] 陈光宇, 胡丽英, 苏勇. 聚类分析在电信行业客户关系管理中的应用 [J]. 微计算机信息, 2006, 22(11): 57 - 60.
- [7] 易瑛, 路璐, 曹东. 改进的 k-means 算法在客户细分中的应用研究 [J]. 微型机与应用, 2005(12): 34 - 37.

(上接第 155 页)

- [4] 周利军, 周源华. 数字图像水印的扩频实现 [J]. 红外与激光工程, 2000(5): 27 - 31.
- [5] 陈海永, 刘泊, 邢佳. 基于 DCT 变换和 m 序列的二值水印嵌入算法 [J]. 哈尔滨理工大学学报, 2004(5): 76 - 79.

(上接第 162 页)

指标为 498080, 同时得到的最好的调度为 (5, 3, 4, 2, 1)。

表 1 成本参数表

| 主产品号 | 超产惩罚系数 | 欠产惩罚系数 |
|------|--------|--------|
| 1    | 1200   | 5000   |
| 2    | 1000   | 4000   |
| 3    | 1000   | 4000   |
| 4    | 1450   | 3000   |
| 5    | 1456   | 6000   |

#### 6 结束语

研究了汽车装配车间生产计划与调度的同时优化问题,着重讨论了生产计划和调度同时优化模型的建立和求解方法,同时给出了系统体系架构的四层模型。

文中所讨论的算法已在汽车装配车间生产计划与调度系统中使用,并给南京某汽车总装厂带来了良好

的经济效益。而且该系统是一个组件化的开放式系统,具有较好的灵活性和适应性,能够适应企业的未来发展。

#### 参考文献:

- [1] 高筱芸, 严洪森, 路致远. 基于模型重构的生产计划优化系统设计 [J]. 计算机技术与发展, 2006, 16(3): 167 - 169.
- [2] 严洪森, 夏琦峰, 朱旻如, 等. 汽车装配车间生产计划与调度的集成优化方法 [J]. 自动化学报, 2002, 28(6): 911 - 919.
- [3] 缪红萍. 免疫遗传算法及应用研究 [D]. 北京: 北京化工大学, 2005.
- [4] 钟远晖. NET 平台下企业生产管理软件系统的研究和开发 [D]. 南京: 东南大学, 2004.
- [5] 盛蕾, 方华. 基于 ASP.NET 的四层 WEB 应用模型设计与实现 [J]. 计算机与数字工程, 2006, 34(7): 147 - 150.