文章编号:1002 - 980X(2007)05 - 0051 - 06

聚类模型在客户关系管理中的应用以及对特征提取的探讨

谭元戎, 孙剑平

(南京理工大学 经济管理学院, 南京 210094)

摘要:随着数据挖掘技术的发展和信息的增长,企业和公司开始运用数据挖掘技术来分析和处理他们在商业活动中积累的关于客户的大量数据,以从中发现重要的规律,来指导公司的营销策略。客户聚类就是一个重要的问题。它根据客户的个人属性和行为属性,把相似的客户群聚集起来。公司可以根据不同类型的客户作出不同的营销策略。本文讨论了有关聚类模型的两个问题。第一,介绍了两种经典的聚类算法,以及他们的复杂度。并讨论它们在客户关系管理中的应用和有效性;第二,讨论了特征提取在聚类过程中的必要性,并给出了如何在聚类模型中进行特征提取的有效算法。

关键词:聚类;客户关系管理;数据挖掘;特征提取;非监督学习

中图分类号: F224.9 **文献标志码**:A

1 引言

随着数据挖掘技术的成熟,以及客户的个人信息和行为数据的积累,企业和公司开始重视运用数据挖掘技术来分析这些数据,希望从中得到有用的知识和规律,来指导公司的营销策略和发展计划。因此,数据挖掘技术在客户关系管理[1]领域中找到了很重要的应用。

例如,银行有一个数据库来保存客户数据。包括客户的个人信息,如性别,年龄,收入水平等,以及客户的行为数据,如客户在银行开的账户类型,每月的存储活动,信用卡客户的刷卡纪录等。移动公司可以掌握用户的个人信息,以及行为数据:每月通话的频率,时间长短,通话类型(本地,长途,漫游)等。综合用户的个人信息和行为数据,公司可以分析客户的消费特性。

基于对客户的分析,对客户群体进行聚类,是一个重要的问题。聚类就是对客户群体的一个划分——把相似的客户归为一类,把不相似的客户划分到不同的类中。比如移动客户中,如果每月话费在 1 000 元以上,有很多长途或者漫游的呼叫,这类客户通常是经常出差,公务繁忙的高额漫游客户群;如果每月话费在 100 元内,短信数量很大而通话比重较小的,这类客户往往是工作地点比较固定,并且比较经济型消费的低端客户群;如果用 GPRS 上网量大消费额高,这类客户就是 GPRS 大客户群。如果企业能够对客户进行有效的聚类,识别出客户子群体,就能够根据每种子群体,总结出他们有代表性的属性,并针对他们的消费行为,设计营销策略。对移动公司而言,可以针对不同客户群,设计不同的优惠移动套餐,以吸引相应的消费人群。

聚类(clustering)算法和分类(classification)算法的一个重要的区别是,聚类是一种非监督的学习(unsupervised learning),而分类是一种有监督的学习(supervised learning). 在分类问题中,数据除了属性之外,还有类属性,标示数据属于哪一类。分类算法就是根据这些训练数据来建立一个分类模型,模型的参数调整到使得模型可以最好的区分训练数据。而聚类问题只有数据的属性值,没有类属性。并且用户事先不知道有几个子类。聚类算法是通过数据属性的分布而找出相对密集的点,作为一个子类,相对分散的点作为其他的子类。

聚类模型为分析客户数据提供了一种有力的工具。在实际应用中,客户数据通常只有属性,没有类

收稿日期:2006 - 01 - 05

作者简介:谭元戎(1964 -),男,云南昆明人,南京理工大学经济管理学院,高级工程师,博士研究生,研究方向:人力资本管理;孙剑平,男,江苏南京人,南京理工大学经济管理学院,教授,博士生导师,从事人力资源与战略管理研究。

属性。或者类别通常由客户根据主题而定义,并人为地给每个数据点赋上类属性值。因此,给训练数据集人工的赋上类属性是很耗时,并且很主观的。而聚类算法就没有这样的要求,因此聚类算法很适合用于分析无类属性值的数据。

但是,运用聚类算法也有一些难点。①有些算法需要用户预先指定子类的个数。可是用户往往面对着大量数据,很难准确地估计出这些数据可以自然的分割成几个子类。而且,聚类的结果也会因为用户指定子类的个数不同而不同。②数据里有很多属性。可是有些对于聚类并不相关,比如说,用户的家庭住址。怎样有效的选择相关的属性进行聚类,需要在聚类之前对数据进行分析,进行特征提取。③数据里有连续的属性,例如工资,年龄;也有离散的属性,例如性别,职位。对连续属性计算距离很自然,但是对于离散的属性,很难定义不同属性值之间的距离。④不同的连续属性,他们的取值有不同的范围。比如说,工资通常在(1 000,10 000)间取值,而年龄通常在(18,80)间取值。如果两个数据点,他们的工资差别是2 000,而年龄差别是5,如果把这两个距离简单的相加,那么在工资属性上的距离会大大超过年龄的差距,从而削弱年龄这个属性的影响力。因此,在聚类之前,要先对各属性进行归一化(normalization)。一种办法是把每个连续的属性值都归一化到一个统一的区间,比如[0,1]之间。

作者认为,在上面提到的对聚类的数据进行特征提取是一个很重要、值得研究的问题。在数据挖掘或者机器学习的领域中,绝大部分的对特征提取问题进行研究的工作都是在有监督学习的框架下,比如分类算法中,决策树C4.5[2]用Information Gain来进行特征提取。然而,聚类问题属于非监督学习。在这种情况下进行特征提取,难度会大大增加,由于缺乏类属性的信息,因此不能给每个单独的特征给出一个有效性的衡量。

尽管使用聚类算法有这些难度,它还是给用户分析公司的客户数据提供很好的模型。公司可以根据客户在属性上的差别,把他们分成不同的子类。然后分析每个子类的特点。本文将讨论两种重要的聚类算法,然后讨论如何将它们运用到客户关系管理的实例中。再给出在聚类问题中(非监督学习)进行特征提取的算法。本文将按照以下的结构组织。第二节讨论两种重要的聚类算法,他们的优缺点,以及他们的时间复杂度。这些算法将作为分析客户子群体的重要模型。第三节讨论将聚类算法运用到客户聚类问题上的步骤,和每一步应注意的问题。第四节,介绍聚类问题中的特征提取算法。第五节用一些实验来验证聚类算法和特征提取算法。第六节给出结论并结束全文。

2 聚类模型

聚类算法根据不同的方法,可以分成以下几种类型:基于划分的,例如K-Means[3],K-Medoids[4];基于层次的,例如BIRCH[5],CURE[6];基于密度的,例如DBSCAN[7],OPTICS[8]. 在这一节我们将详细介绍K-Means算法,并讨论它的一个改进版本,并分析它们的复杂度。

K-Means属于基于划分的聚类算法。其基本思路是,选择K个数据点作为子类的中心,然后根据所定义的距离的衡量,把其余的点都划分到不同的子类。K-Means是聚类模型中最有代表性的一个算法。它的思路是:①随机的挑选K个数据点作为初始的中心;②把每一个数据点归到离它最近的中心;③对每一个子类,计算它们的属性值的平均值,把算出的平均值作为新的中心;④重复步骤②,直到每一个数据点的归类都不再变化。K-Means的好处在于它的时间复杂度比较低,是O(tkn),t是算法循环的次数,k是子类的个数,n是数据点的个数。通常t,k<<n,因此K-Means可以看成是数据点个数的线性复杂度。但是K-Means算法的最终结果常常因为初始化中心的不同而不同。K-Means实际上是一个EM(Expectation-Maximization)[9]算法,所以它的结果是一个局部最优解,而不能保证全局最优解。另一个问题是每一步在计算每个子类的平均值。对于离散值,如何定义平均值是一个问题。而且K-Means算法对于容易受到数据中的噪音影响总体的聚类效果,比如说,有一个数据点,它的某一个属性值的值异常的大,用这个属性值计算出来的平均值会影响整个数据的分布,从而影响聚类的结果。

K-Medoids是对K-Means的一个改进。算法的思路基本相同,但是在每一步不用子类中数据点的平均值作为子类的中心,而是选择子类中最中心的数据点作为这个子类的中心。K-Medoids的算法如下:①随机的选择K个数据点作为初始的中心;②选一个非中心的数据点A去代替一个中心B,计算这个代替会不会在带来更好的聚类效果。如果是,用A代替B作为中心;③重复步骤②,直到子类的中心不再变化。K-Medoids比K-Means能够更

好的处理数据中的噪音,因为用一个实际的数据点作为子类的中心,比用一个子类的平均值作中心,会更少的受数据中的噪音影响。但是 K-Medoids 时间复杂度比 K-Means 高,是 $O(k(n-k)^2)$,n 是数据点的个数,k 是子类的个数。因此对于小的数据集,K-Medoids 效果很不错,但是对于大的数据集,效率却不够高了。

3 对客户数据进行聚类分析

把聚类算法应用到客户数据上面,通常需要以下几个步骤:①特征提取;②归一化;③聚类;④分析聚类的结果。表 1 是关于移动客户的数据。本节将以这个移动数据作为例子,讨论进行聚类分析的每个步骤。

3.1 特征提取

特征属性提取是应用聚类算法,也是其他数据挖掘算法之前的一个重要的步骤。因为数据都是从现实世界中收集的,不像用于理论分析而人工生成的数据那样"纯净"。在收集数据时,常常会记录很多属性。但是,在对数据进行聚类的时候,并不是每一个属性都对聚类分析有用;每一个对聚类分析有用的属性,他们对聚类的影响也不是同等的。比如在表 1 中的移动数据,地址对移动客户消费群的划分是没有贡献的,因为移动客户不会因为他们的地理上的相近而产生相似的消费行为。像手机用户的年龄和性别,他们对聚类有一定的贡献,可是他们的影响不像用户的通话分钟数等消费属性那么直接。因此,在进行聚类分析前,需要把无关的属性去掉,并给一些影响小的属性确定一个较小的权重,以减轻它们在聚类过程中,对距离的贡献。

表 1 移动用户的数据

用户 ID	性别	年龄	职业	地址	短途分钟(分钟)	短途话费(元)	漫游分钟(分钟)	漫游话费(元)	GPRS 分钟	GPRS 话费	短信条数	短信话费
100	F	20	学生	…	100	40	0	0	0	0	260	26
200	M	35	销售	…	600	240	200	160	100	60	24	2.4
300	M	28	软工	…	250	100	30	24	500	300	180	18
400	M	50	教师	…	160	64	0	0	0	0	50	5
500	F	36	咨询	…	400	160	160	128	20	12	100	10
600	M	20	学生	…	80	32	0	0	200	120	1 000	100
…	…	…		…	…	…	…	…	…			…

另一个问题是,不同的属性之间可能会存在关联,甚至是完全相关。比如说,一个客户的月收入水平比较低,那么他的话费很可能就比较低;而一个高收入的客户,他的话费也比较可能偏高。因此月收入和话费这两个特征是有一定关联性的。再看一个例子,在表格一中,特征"短途分钟"和"短途话费"是完全相关的。从前者可以通过某种公式计算出后者(在我们的例子中,是"单价 * 短途分钟 = 短途话费")常常我们不希望在计算两个客户之间的距离时,计入一个属性的双重或者多重影响。所以我们需要进行特征提取,除去冗余和无关的属性,剩下的属性就是用于聚类分析的相关属性了。在第四节中,我们将会给出在非监督学习中,进行特征提取的具体算法,并且和监督学习中的特征提取算法进行比较。

3.2 离散属性值

聚类算法是基于数据点之间的距离,它的目标是把相似的数据点聚为一类,把不相似的数据点划分到不同的子类。对于连续的属性值,距离的衡量是很自然的。例如有两个数据点,P 和 Q。它们有 n 个属性。一个常用的距离公式是 Minkowski distance

当 k = 1 $d(P,Q) = \sqrt[k]{(P_1-Q_1)^k+(P_2-Q_2)^k+\cdots+(P_n-Q_n)^k}$

$d(P,Q)$ 就是曼哈顿(Manhattan)距离,当 k = 2,$d(P,Q)$ 就是欧氏(Euclidean)距离。

然而,对于离散的属性值,距离的定义却不是那么自然。离散的属性的一种特殊情况是二元变量,只能取 0 或 1 两个值。对于二元变量计算距离,可以用以下的办法。假设两个数据点 P 和 Q。它们有 n 个二元变量。假设其中有 a 个二元变量 P 和 Q 都取 1,有 b 个变量 P 取 1 而 Q 取 0,有 c 个变量 P 取 0 而 Q 取 1,有 d 个变量 P 和 Q 都取 0。那么 P 和 Q 之间的距离可以用以下公式计算。

$$d(P,Q) = \frac{b+c}{a+b+c+d}$$

如果离散的属性可以取多个值,比如说,职业可以是学生,教师,医生,公务员,经理等等,对这种离

散的属性计算距离的时候,可以先把它们进行转化,对每一个可能的值都创造一个二元变量。对于职业这个变量,创造 5 个二元变量,分别对应于它的五个可能的取值。如果一个客户,他的职业是教师,那么他在教师这个二元变量上取 1,在学生,医生,公务员等其他四个二元变量上都取 0。在进行转化之后,就可以套用二元变量计算距离的公式进行计算。这种转化的例子如表 2 所示。

表 2 离散变量转化成二元变量

用户 ID	职业	转化后的二元变量				
		学生	教师	医生	公务员	经理
100	学生	1	0	0	0	0
200	医生	0	0	1	0	0
300	经理	0	0	0	0	1
…	…	…	…	…	…	…

3.3 归一化

归一化是数据预处理的另一个重要步骤。在聚类的不同属性中,每一个属性有不同的取值范围。比如在表 1 中的移动数据,年龄通常在[18, 80],然而,通话分钟数的取值通常在几百。如果把这两个距离简单的相加,那么在通话分钟属性上的距离会大大超过年龄上的距离,从而削弱年龄差距的影响。一种办法是把每个属性的取值范围都归一化到[0, 1]的范围内。也可以给不同的属性根据他们重要性的大小,赋上不同的权值。

3.4 聚类分析

对数据预处理好之后,就可以运用聚类算法进行聚类分析了。可以根据问题的需要,选择相应的算法。

运用聚类算法之后,需要管理人员对聚类的结果进行分析,并作出反馈。比如说,从聚类的结果来看,是否合理,是否符合他们的预期结果。比如,用户在运用 K-Means 分析移动客户的时候,指定 k = 3。但是从聚类的结果来看,分成 3 个子类,并不能很好的区分客户群体。因此用户需要调整聚类的参数,重新对数据进行聚类分析,直到聚类的结果合理为止。

4 非监督学习中的特征提取算法

在这一节中,我们介绍在非监督学习中的特征提取算法。在监督学习中,特征提取问题被研究的很多。主要有两类代表性的算法。第一类叫 wrapper approach[10],就是把各种特征组合成不同的特征子集,用一个分类算法来测试在不同特征子集上的分类正确率,以此来选择最好的特征子集。特征子集通常有以下两种方法产生:①向前选择,从单个特征开始,然后每次往当前的特征子集里面增加一个好的特征,再测试正确率;②向后消除,现选中所有的特征,然后每次剔除一个差的特征,在剩下的特征子集中,测试正确率。第二类算法是对每个特征计算一个分数,衡量它的重要性。对所有特征按分数排序,剔除那些分数低的特征。Fisher Score 和 Information Gain 就属于这一类。在进行这一类分数计算的时候,通常需要类属性值来帮助衡量一个特征的重要性。

在文献[11]中提出了一个新的特征提取算法 Laplacian Score (LS). 这个算法的思想是,假设在一个 n 维空间,两个数据点距离很近。如果我们选择一个 k (k < n) 维的子空间,在这个子空间中,这两个数据点仍然很近。那么我们认为这个 k 维的子空间能够很好的保持数据点在原来 n 维空间的临近关系。这种性质叫做 Locality Preserving. 在 LS 算法中,对每一个特征都计算出它的 Laplacian Score,来反映它的 Locality Preserving power。为了描述数据点附近的几何结构,我们建一个 nearest neighbor 图。LS 要提取出那些能够保持这个图的结构的特征。

我们认为,LS 很适用于解决聚类问题中的特征选择问题。因为 ①LS 不需要知道类属性,这正是聚类问题所不能提供的信息。②LS 寻找的是可以保持数据点临近结构(距离)的特征,这与聚类问题的目标是一致的。

我们在下面给出 LS 算法的详细描述。让 L_r 表示第 r 个特征的 Laplacian Score. 让 f_{ri} 表示第 i 个数据点的第 r 个特征值, i = 1, …, m. 该算法的理论分析可以参见文献[11]。

在第五节,我们将在实验中检验 LS 算法在聚类算法中的有效性。

5 实验分析

在这一节中,我们在三个数据集上使用聚类算法和特征提取,进行实验分析。

5.1 实验设计

我们使用的是 Matlab 提供的 K-Means 算法。我们把在原始数据的所有特征上进行聚类的结果作为基准 (baseline)。此外,再测试经过 LS 算法进行特征提取之后,在提取出来的 k 维子空间上进行聚类的效果。

5.2 实验精度

我们使用的数据都是带有类属性(class label)的。在聚类和特征提取过程中,这些类属性不被利用。在聚类之后,我们利用类属性作为一个标准,来衡量我们聚类的精度。让 r_i 和 s_i 分别表示数据点 i 的聚类号(cluster label)和类属性(class label)。我们的实验精度定义为:

$$AC = \frac{\sum_{i}^{n}\delta(s_i, map(r_i))}{n}$$

这里的 n 是所有的数据点的个数,$\delta(x,y)$ 是 delta 函数,如果 x = y,那么函数取值为 1,否则为 0。map 是一个映射函数,把每一个聚类号 r 映射到它对应的类属性上。

5.3 数据

我们使用了 UCI Machine Learning Repository 的三个数据集。

数据一 Optical Recognition of Handwritten Digits。这个数据集是对手写的数字进行识别的数据。一共分为 10 类,对应数字 0,…,9。每个数字有 64 个特征,每个特征取值为 0 到 16 之间的整数。我们用 K-Means 算法在原始数据(64 个特征)上做聚类,另外用 LS 算法提取 k 维子空间(k = 10,15,…,60)。我们用前面定义的精度来衡量在不同子空间和原始空间上的聚类效果。图 1 给出了在数据一上的聚类结果。

1. 建立一个有 m 个节点的 nearest neighbor graph G。节点 i 对应于数据点 X_i,节点 i 和 j 之间有一条边连接,如果 X_i 是 X_j 的 k nearest neighbors 之一,或者反之。

2. 如果节点 i 和 j 是有边相连的,给出 $S_{ij} = e^{-\frac{\|x_i - x_j\|^2}{t}}$,如果它们无边相连,$S_{ij} = 0$。S 是图 G 的权重矩阵,以描述数据点的临近结构。

3. 对于每一个特征 r,定义

$f_r = [f_{r1}, f_{r2}, \cdots, f_{rm}]^T$

$D = diag(S1), 1 = [1, \cdots, 1]^T, L = D - S$。

那么定义

$$\tilde{f}_r = f_r - \frac{f_r^T D 1}{1^T D 1} 1$$

4. 对每一个特征 r,计算它的 Laplacian Score:

$$L_r = \frac{\tilde{f}_r^T L \tilde{f}_r}{\tilde{f}_r^T D \tilde{f}_r}$$

图 1 Digit Data 的实验结果

在图 2 中,我们可以看出,当提取的特征数目比较小的时候(例如 10,15),在上面聚类的效果比较差。因为这个子空间不足以充分的描述数据点在原始空间(64 维)上面的距离关系。当特征提取的维数增长到 25 的时候,取得了最好的聚类效果。此后当特征数渐渐增加的时候,聚类的效果反而下降了。曲线的最后一个点是使用了所有 64 个特征的准确率。这个曲线表示,在这个数据集上,有些特征是相互关联相互影响,甚至是冗余的。它们的存在不仅不能带来有用的信息,反而会干扰聚类的效果。因此,无监督学习中的特征提取算法在这类数据上,用来消除冗余的特征是很有用的。

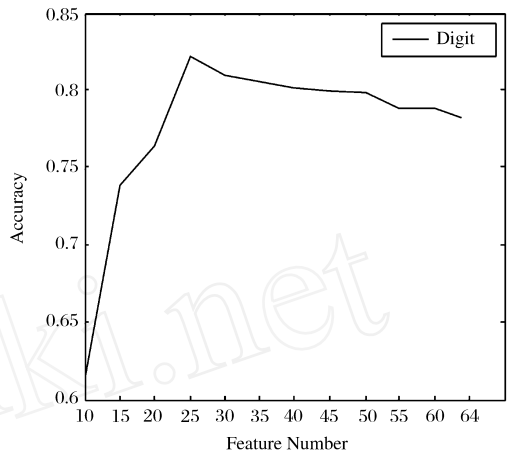

图 2 LS 算法的描述

数据二 Wine Recognition Database。这个数据是对酒进行识别的数据集。一共分为 3 类,每个酒的样本数据有 13 个特征。我们用 K-Means 算法在原始数据(13 个特征)上做聚类,另外用 LS 算法提取 k 维子空间(k = 1,2,…,12)。我们用前面定义的精度来衡量在不同子空间和原始空间上的聚类效果。

图 3 给出了在数据二上的聚类结果。在图 3 中,我们可以看出,当提取的特征数目为 1 或 2 时,聚类的效果比较差。说明只用这两维特征是不足以描述原始数据的空间,也难以达到理想的聚类效果的。当提取的特征数据增加,聚类的效果有了明显的提高,并且稳定下来。这说明,在原始数据上,不同特征之间的关联性比较小。所有特征一起使用的时候,它们之间的相互影响和干扰并不会很大。

数据三 Wisconsin Breast Cancer Databases (WDBC)。这个数据是对每个样本进行判断是否患有 Breast Cancer 的数据集,分为两类(良性和恶性)。每个样本数据有 30 个特征。我们用 K-Means 算法在原始数据上做聚类,另外用 LS 算法提取 k 维子空间(k = 1,2,…,30)。我们用前面定义的精度来衡量在不同的子空间和原始空间上的聚类效果。

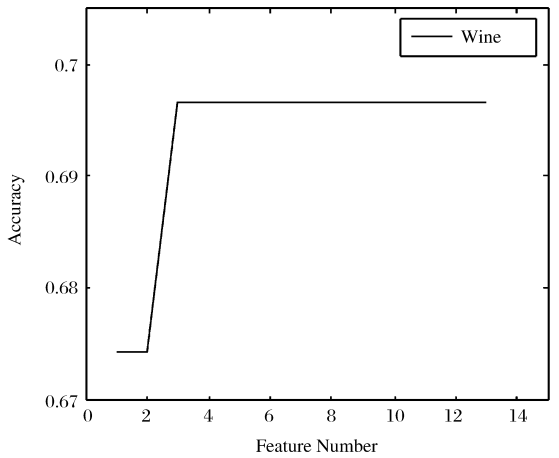

图 3 Wine Data 的实验结果

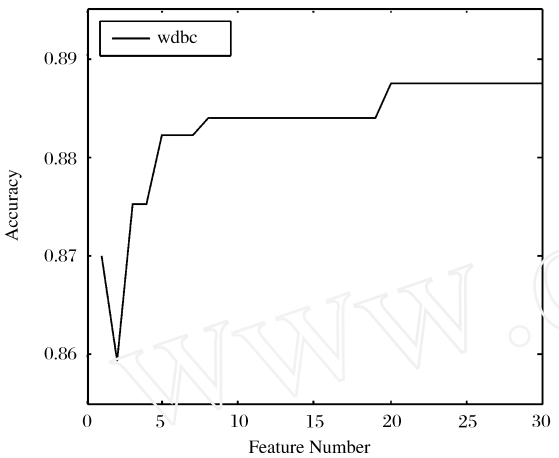

图 4 WDBC Data 的实验结果

图 4 给出了在数据三上的聚类结果。在图 4 中,我们可以看到当选取的特征数目比较少的时候,聚类的效果不好。当提取的特征数目比较多的时候,聚类效果逐渐提高。当提取的特征数达到 20 的时候,聚类结果稳定下来。这说明,在这个数据上,属性之间的冗余性比较小,因此相互干扰比较少。

在实践聚类问题的时候,如果用户有每个数据对应的类属性,那么可以根据这个信息来调整聚类的参数,包括聚成几类,是否需要进行特征提取,需要提取几个特征,等等,以获得最好的聚类效果。如果用户的数据不包含类属性,用户可以尝试自己手工标识一些数据,在此基础上进行聚类,然后推而广之到大的数据集上。或者可以尝试不同的 K 值,然后分析得出的结果,能否合理的反映客户的特征,是否符合用户心中的期望值。经过多次的尝试之后,可以选择一个比较合理的分析结果。在聚类问题中,参数如何设定没有一个确定的答案,应该视不同的应用/数据而定。用户也应该结合自己的行业知识,对聚类的效果进行分析,进而反馈,如有需要,再改进聚类的效果。

6 结论

在本文中,我们讨论了聚类算法及其在客户关系管理中的应用。我们分析了两种有代表性的聚类算法,并讨论他们的优缺点和时间复杂度。本文还介绍了对客户数据进行聚类分析的具体步骤,并介绍了在无监督学习下进行特征提取的算法。我们通过实验及其结果的分析,验证了聚类算法和特征提取算法的运用。通过试验,可以看出聚类算法对实际数据分析带来的效果。

参考文献.

[1] STEPHEN J SMITH, ALEX BERSON, KURT THEARLING. Building Data Mining Applications for CRM[M]. McGraw - Hill, 1999.

[2] J ROSS QUINLAN. C4.5: Programs for Machine Learning [M]. Morgan Kaufmann, 1993.

[3] MACQUEEN J B. Some methods for classification and analysis of multivariate observations[C]. In Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley, CA: University of California Press, 1967: 281 - 297.

[4] L KAUFMAN, P J ROUSSEEUW. Finding groups in data: an introduction to cluster analysis[M]. John Wiley & Son, 1990.

[5] T ZHANG, R RAMAKRISHNAN, M LIVNY. BIRCH: an efficient data clustering method for very large databases [C]. In Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD'96), 1996.

[6] S GUHA, R RASTOGI, K SHIM. Cure: An efficient clustering algorithm for large databases[C]. In Proceedings of the ACM SIGMOD International Conference on Management of Data (SIGMOD'98), 1998.

[7] M ESTER, H - P KRIEGEL, J SANDER, X XU. A density - based algorithm for discovering clusters in large spatial databases[C]. In Proceedings of the 1996 International Conference on Knowledge Discovery and Data Mining (KDD'96), 1996.

[8] M ANKERST, M BREUNIG, H - P KRIEGEL, J SANDER. Optics: Ordering points to identify the clustering structure[C]. Proc. ACM SIGMOD'99 Int. Conf. on Management of Data, Philadelphia PA, 1999.

[9] J HAN, M KAMBER. Data Mining: Concepts and Techniques[M]. Morgan Kaufmann, 2001.

[10] RON KOHAVI, GEORGE H JOHN. Wrappers for feature subset selection[J]. Artificial Intelligence, 1997 (1 - 2): 273 - 324.

[11] XIAOFEI HE, DENG CAI, PARTHA NIYOGI. Laplacian Score for Feature Selection[C]. Advances in Neural Information Processing Systems18(NIPS), 2005.

(下转第 83 页)

[5]张鹤丹,王惺,付峰,等.中国城市能源指标体系初探[J].中国能源,2006(5).

[6]梁凯丽,冯俊小.谈能源利用评价办法[J].中国冶金,2005(10).

Demonstrable Research on China Urban Energy - conserving Evaluation

WU Guo-hua, YAN Shu-ping

(The Research Center for Resources Economy and Strategy, Shandong University of Finance, Jinan 250014, China)

Abstract: Firstly, based on analyzing the limitations of the current energy - conserving evaluation indexes, this article defines the urban energy - conserving scope. Secondly, the article has established urban energy - conserving evaluation model and evaluation index system. It contains 27 indexes in 3 layers with the basis of the industry, building, urban transportation and live consumption. Thirdly, according to the certain principles, it divides the urban energy - conserving level to four grades. Based on calculating industry energy consumption index A1 and synthesis index of building, transportation and live consumption B, it has set up the standing of energy - conserving level, that is. In the end, it takes some urban as an example to carry on a demonstrable research. The result proved the urban energy - conserving level was the Ⅲ class in 2005, belonging to advanced energy - conserving level. The industrial energy - saving result is better, but the job of energy - conserving about building, transportation and life consumption needs to be strengthened. The analysis result reflects the actual condition of the urban energy consumption basically.

Key words: urban; energy - conserving; index system; evaluation model

(上接第 8 页)

The Research of Determinants of Female Entrepreneurial Intentions

QIAN Yong-hong, WANG Zhong-ming

(Management School of Zhejiang University, Hangzhou 310058, China)

Abstract: With the advent of Knowledge Economics epoch, women possess the truly equate opportunities to compete with man in business world and their entrepreneurship, oppressed for centuries, is unprecedentedly released. The tendency to start private business for women flourished since 1990s, and nowadays business operated by women has been recognized to have significant contribution to the economic miracle of PRC. In this article, the authors are motivated to explore the idiosyncratic factors impacting women entrepreneurial intentions. Based on priori research fruits and structural interview, this paper proposed that gender identification and family commitment are two paramount factors for women to engender entrepreneurial intentions. A sampling pool of 351 women entrepreneurs were used to test our hypotheses. Empirical results indicated that gender identification and family commitment moderate the relationship between women entrepreneurial intention and its relevant antecedents. These findings provide insights in how to stimulate intrinsic entrepreneurship for women and facilitate our understanding of why women as a whole group are less likely to choose selecting self - employment (entrepreneur) as professional career.

Key words: entrepreneurial intentions; gender identification; family commitment.

(上接第 56 页)

The Practice of Clustering in CRM and Discussion on Characteristic Distill

TAN Yuan-rong, SUN Jian-ping

(School of Economics & Management, Nanjing University of Science and Technology, Nanjing 210049, China)

Abstract: With the development of Data Mining and the growth of information, many enterprises and cooperates begin to use the Data Mining technique to analyze and process the large amount of data which is accumulated in the business activities in order to find out important rules for marketing strategy. Customer clustering is a crucial issue. According to customer clustering, similar customer groups cluster on the basis of their personal and behavioral attributes. Then, the enterprises can work out different marketing strategies and apply them to each customer group respectively. The thesis focuses on two questions about clustering formation. First, several typical clustering algorithms and their complexities are introduced. Meanwhile, the application and efficiency of the clustering algorithm are discussed in it, and then the effectual arithmetic on how to carry through characteristic distill in the clustering formation is given out.

Key words: clustering; customer relationship management (CRM); data mining (DM); characteristic distill; unsupervised learning (UL).