二元数据子空间聚类算法的初始化研究 *

夏 英, 鲁 宁, 丰江帆
(重庆邮电大学 中韩合作空间信息系统研究所, 重庆 400065)

摘 要: 针对二元数据空间高维稀疏性的特点而提出的有限混合伯努利模型,能够快速寻找映射簇的模型框架;EM 算法是数学模型进行参数迭代的重要方法,其算法的优劣很大程度上取决于其初始参数。对于运用 EM 算法来实现有限混合伯努利模型聚类算法已有许多研究, EM 算法中参数的选取直接影响聚类算法的性能。引入 Binning 法和改变数据之间相似度测量方式、中心点的选取方式来进行初始化,从而大大减少聚类结果对初始参数的依赖,实验证明该算法是高效的、正确的。

关键词: 子空间聚类; 二元数据; 有限混合伯努利模型; EM 算法
中图分类号: TP18 **文献标志码**: A **文章编号**: 1001-3695(2009)01-0047-03

Research of initialization of subspace clustering algorithm in binary data

XIA Ying, LU Ning, FENG Jiang-fan
(*SIKO-GIS Research Center*, *Chongqing University of Posts & Telecommunications*, *Chongqing* 400065, *China*)

Abstract: Aiming at the characteristic of high-dimensionality and sparseness in binary data set, proposes the finite mixtures of Bernoulli distributions model for finding projected clusters fast. EM algorithm is the important method of iterative parameters, and the degree of good or bad with EM algorithm lies on initial parameters. As far as the finite mixtures of Bernoulli distributions model, there have been lots of researches about it. However, in EM algorithm, the initial parameters affect the clustering performance directly. Therefore, this paper introduced Binning method and computed parameters through changing the comparability measurement between dates and selection style about core-point, in order to reduce the dependence of the clustering for initial parameters. Experiment demonstrates the algorithm is efficient and accurate.

Key words: subspace clustering; binary data; the finite mixtures of Bernoulli distributions model; EM algorithm

子空间聚类是指在原始数据空间的子空间中寻找数据簇,使每个簇与相应的属性子集关联,而簇内数据点在这样的属性集合下高度相似。该概念在 1998 年由 Agrawal 等人[1,2]提出后,引起许多学者关注,并不断研究出新的成果。经典的算法有 CLIQUE[1]算法、改进了的 PROCLUS[3]/ORCLUS[4]等算法,这些算法虽然能有效地发现映射簇,但也存在许多不足。例如 CLIQUE 算法发现簇的形状总是超矩形的结合,而 PROCLUS/ORCLUS 算法的簇却总是外凸的;从 CLIQUE 到 PROCLUS/ORCLUS 都采用硬划分的方法把数据强制属于某一类;CLIQUE 和 PROCLUS/ORCLUS 都需要输入参数,且算法都对初始值非常敏感;CLUQYE 和 PROCLUS/ORCLUS 随着维数的增大,算法的运行时间均呈指数增长。基于模型的聚类方法[5]利用已知数学模型,通过逐渐逼近的方法使得给定数据集与数据模型之间达成最佳拟合。这种聚类算法属于软划分的方法,对簇的形状要求低,也可解决多密度的问题,虽然依赖初始参数,但通过改进可以降低这种依赖程度。

二元数据只有两种状态且大多是高维的,例如在进行购物篮分析时,数据属性代表超市商品,数据点代表购物事务。如果某商品在某事务中被购买,则在该商品对应的属性上取值为 1,否则取值为 0。超市商品数一般都有数千种,意味着每个数据点都有数千维的数据属性,但是每项事务所涉及的商品却很少,也许在数千维的属性中只有几个或几十个取值为 1,其他全部为 0;而且随着超市商品数增多,每项事务所涉及的商品却不会增多。由此可见,二元数据的特点是数据维度越大,数据的稀疏性越高。针对二元数据这种特点,A. Patrikainen[6]于 2002 年首次提出有限混合伯努利模型并且应用 EM 算法进行参数迭代的子空间聚类算法;此后不断有学者完善该模型,仍有对 EM 算法初始参数估计不足,致使迭代次数增多和聚类效果不理想的问题。本文针对 EM 算法的收敛程度取决于初始参数这一问题,在保持 EM 算法本身迭代的前提下,通过引入 Binning 法[7~9]、改变数据之间相似度测量方式和中心点的选取方式来细化 EM 初始值,从而改善 EM 的收敛,获得更好的聚类效果。

1 有限混合伯努利模型聚类算法

有限混合伯努利模型聚类算法的思想是:属性之间是相互独立的,每个属性均可看做一个独立事件 A_i,每个对象是独立事件实验样本。数据对象在某属性上是否取值为 1,取决于事件 A_i 发生的概率 p_i,数据对象在所有属性上取值的分布律可以用相应公式来表示。如果各对象在具体属性上取值的概率不同,则可以把概率相同的看做一类,形成数据集合中的聚类簇,同一簇的数据看做是由同一分布产生的数据点[10]。

1.1 有限混合伯努利模型

考虑 K 重 D 元混合伯努利分布,该分布的参数表示为 $P=(p_1,p_2,\cdots,p_k)$。其中:$p_k=(p_{k_1},p_{k_2},\cdots,p_{k_D})$ 表示第 k 个分布的参数;$\pi=(\pi_1,\pi_2,\cdots,\pi_k)$ 是权重向量,第 k 分量的概率可以由 π_k 给定,即 $P(k)=\pi_k$。给定模型参数的情况下,数据点 x_n

收稿日期: 2008-03-13; **修回日期**: 2008-05-28 **基金项目**: 国家"863"计划资助项目(2007AA12Z238)

作者简介:夏英(1972-),女,副教授,主要研究方向为数据库系统、空间数据挖掘;鲁宁(1984-),男,硕士研究生,主要研究方向为空间数据挖掘(sncwting915@126.com);丰江帆(1980-),男,讲师,主要研究方向为地理信息系统.

属于簇 k 的概率为 $P(x_n|k;p)=\prod_{d=1}^{D}p_k^{x_{nd}}(1-p_{k_d})^{(1-x_{nd})}$。考察特定的数据点 x_n,从而找出第 k 个分量产生它的概率,即后验概率,可以应用贝叶斯规则得到

$$P(k|x_n;\pi,p)=P(x_n|k;p)P(k)/P(x_n;\pi,p)=$$

$$P(x_n|k;p)\pi_k/[\sum_{k'=1}^{K}P(x_n|k';p)\pi_{k'}]$$

后验概率 $P(k|x_n;\pi,p)$ 有两个重要的性质。

性质 1 给定一数据点 x_n,求所有混合分量产生它的概率之和为 1,即

$$\sum_{k=1}^{K}P(k|x_n;\pi,p)=1 \tag{1}$$

性质 2 保持 k 不变,求该混合分量产生所有数据点的概率之和为

$$\sum_{n=1}^{N}P(k|x_n;\pi,p)=\pi_k N \tag{2}$$

1.2 EM 算法

假设存在一个完整数据集 $Y=(X,Z)$。其中:$X=\{x_1,\cdots,x_n\}$ 是不完整的数据集;Z_i 是引入的隐含变量,$Z_i\in\{1,2,\cdots,M\}$,M 是给定的有限整数。于是 $Y=\{(x_1,Z_1),\cdots,(x_n,Z_n)\}$,则完整数据的似然函数为

$$L(\theta|X,Z)=P(X,Z|\theta)=\prod_{i=1}^{n}P(x_i,z_i|\theta)\quad Z=(z_1,\cdots,z_n)$$

该似然函数的期望值 $E(L(\theta|X,Z))=\int_Z P(x_i,z_i|\theta)f(z)\,dz$ 采用 EM 算法的基本思想是针对不完整数据集 Y。对于混合伯努利分布模型,知道该模型的参数则可以根据该模型推出属于每个成分的各数据点的概率,然后修改每个成分的值,重复该过程直到收敛到结束条件。

用 $\pi(t)$、$p(t)$ 和 $p_k(t+1)$ 分别表示相应的参数值在 EM 算法中第 t 次的迭代值,构造基于混合伯努利分布的 EM 算法。第 t 次迭代后,通过下式计算 $\pi(t+1)$ 和 $p(t+1)$。

$$\pi_k(t+1)=(1/N)\sum_{n=1}^{N}P(k|x_n;\pi(t),p(t)) \tag{3}$$

$$P_k(t+1)=\sum_{n=1}^{N}x_nP(k|x_n;\pi(t),p(t))/(N\pi_k(t+1)) \tag{4}$$

由式(4)可以看出,第 k 个混合分量的参数 $p_k(t+1)$ 是全部数据点的加权平均数,而第 n 个数据点的权重则是第 k 个混合分量产生它的概率,然后通过迭代把 π 和 p 计算出来。

2 改进的初始化算法

该算法是以子空间聚类模型为基础,采用迭代的 EM 算法,从初始划分、参数估计开始,对划分方案进行迭代优化,最终形成聚类结果。

2.1 普通的初始化算法

在有限混合伯努利分布模型的 EM 算法中,需要对模型参数进行预估以及对属性进行初始化。常用的方法有 AGENES、K-means、随机中心等。AGENES 聚类,即首先将 n 个样本分成 n 类,使得每一类正好含有一个样本;然后将样本凝集成 $n-1$ 类、$n-2$ 类,直到所有的样本都凝集成所需要的 k 类为止。K-means 聚类属于聚类分析方法中应用最广泛的划分算法,目标就是要找到 k 个均值向量,k 就是聚类数目。基于给定的聚类目标函数,算法采用迭代更新的方法。每一次迭代过程都是向目标函数值减小的方向进行,最终的聚类结果使目标函数值取得极小值,以达到较优的聚类效果。因为二元数据取值只能是 0 或 1,所以平均数的取值也只是在[0,1]中,导致平均数在迭代过程中变化不大,这影响了聚类的效果。随机初始化方法即在数据集中随机抽取 n 个点作为聚类中心,n 为聚类数。然后,对数据集中的每个数据点(除了这 n 个点以外)计算其与这 n 个点的每个点的欧氏距离;根据距离最短原则,把每个点放入 n 类中的某一类。于是,对于每一类的数据,可以计算出其权重、期望、方差。但是由于二元数据取值有限,在用欧式距离进行测量从而判断两个数据点是否属于同一个簇可能会失效,因为大量的计算值会重复。基于以上原因,引入 Binning 法并改进其度量方式作为 EM 算法的初始化方法。

2.2 改进的 Binning 初始化算法

Binning 法可以想像成把数据空间在各维上划分为一个个的箱子,再把数据点投射到对应的箱子里去。这里,初始化 EM 的任务就是找到最优聚类中心,从而对参数以及属性集进行估值,想法是最好的聚类中心可能就是概率密度数最稠密的部分。于是,通过 Binning 法来寻找概率密度函数最稠密的部分。Bin 宽是指在 Binning 法中把整个数据空间分成若干个 Bin 并指定每一个 Bin 的宽度;然后把每个数据的每一维都投影到 Bin 中。因为 Bin 宽直接影响到数据的投影数量,所以选择一个最优的 Bin 宽就成为关键的问题。不过二元数据取值全部为{0,1},因此最优 Bin 宽无须考虑,这也是 Binning 法可以引入该算法的一个重要原因。另外,由于高维二元数据的特殊性,引入相异度和相似度两个概念对数据点进行度量以及在选取中心点时引入的相异度均值和相异度标准差的概念,对原有的 Binning 法进行改进。

a)相似值(simi)标准定义如下:

$$Y=\{X_1,X_2,\cdots,X_m\},X_t=(x_1,x_2,\cdots,x_n)$$

$$t\in\{1,2,\cdots,m\},x_p\in\{0,1\},p\in\{1,2,\cdots,n\}$$

$$\forall X_i,X_j\in Y,\text{simi}(X_i,X_j)=\sum_{k=1}^{n}x_{i_k}\odot x_{j_k},\text{且 simi}(X_i,X_j)\in R^+$$

b)相异值(repu)标准定义如下:

$$Y=\{X_1,X_2,\cdots,X_m\},X_t=(x_1,x_2,\cdots,x_n)$$

$$t\in\{1,2,\cdots,m\},x_p\in\{0,1\},p\in\{1,2,\cdots,n\}$$

$$\forall X_i,X_j\in Y,\text{repu}(X_i,X_j)=\sum_{k=1}^{n}x_{i_k}\otimes x_{j_k},\text{且 }x_{i_k}\cap x_{j_k}\neq 0,\text{repu}(X_i,X_j)\in R^+$$

c)相异度均值(repu_mean)标准定义如下:

$$C=(c_1,c_2,\cdots,c_n),c_i\in\{0,1\},Y=\{X_1,X_2,\cdots,X_m\}$$

$$X_t=(x_1,x_2,\cdots,x_n),t\in\{1,2,\cdots,m\},x_i\in\{0,1\}$$

$$t\in\{1,2,\cdots,n\},\text{repu\_mean}(C)=(\sum_{j=1}^{m}\text{repu}(C,X_j))/m$$

d)相异度标准差(repu_stddv)标准定义如下:

$$C=(c_1,c_2,\cdots,c_n),c_i\in\{0,1\},Y=\{X_1,X_2,\cdots,X_m\}$$

$$X_t=(x_1,x_2,\cdots,x_n),t\in\{1,2,\cdots,m\},x_i\in\{0,1\},i\in\{1,2,\cdots,n\}$$

$$\text{repu\_stddv}(C)=(\sum_{j=1}^{m}|\text{repu}(C,X_j)-\text{repu\_mean}(C)|)/m$$

2.3 实现步骤

通过 Binning 法来寻找概率密度函数最稠密的部分。首先根据一定的 Bin 宽将每一维上的整个数据空间分成若干个 Bin,然后把每个数据的每一维放到对应的 Bin 中,再计算每一个 Bin 中所含的数据点个数。含的点数多的则为概率密度相对大的区域,也就是聚类中心最可能存在的区域。在得到了每一个 Bin 中的数据点个数后,可以作进一步的优化:

a)给每一个数据点一个向量标记,表示其各维所在的 Bin 位置。

b)计算出 Bin 中数据点的均值,把小于均值的 Bin 去除考虑范围。

c)除去不满足条件的 Bin 后,再对所有的数据点进行筛选。

d)在筛选后的数据集中筛选中心点。

e)按照相似度重新排列数据点,相似度高的为一类。

2.4 算法伪代码

有限混合伯努利模型聚类算法

输入:num_clust //输入欲寻找簇的个数

输出: clusters //簇中心

dim: 数据维数;num_data: 数据;sele_core: 由步骤 d)筛选出的点的个数;

for i = 1 to num_data do

for j = 1 to dim do

把每一个数据点的每一维都投影到相应的 Bin 向量的属性中,并分别计算出映射到 Bin 向量中的数据个数之和

end for

end for

计算 bin 中数据点的均值。

for j = 1 to dim do

在 bin 向量中把小于均值的属性去除

end for

for i = 1 to num_data do

for j = 1 to dim do

根据 Bin 向量中属性的个数,从所有数据点中选出大于 Bin 向量中属性个数的点,数目为 sele_core

end for

end for

if num_clust > sele_core

算法结束,重新输入簇的个数

end if

if num_clust = = sele_core

把步骤 d)选出的点作为中心点

end if

if num_clust < sele_core

从已选出的聚类中心中进行第二次筛选,筛选的原则是彼此相异度最大的为簇中心

end if

for i = 1 to num_data do

for k = 1 to num_clust do

根据相似度标准和已知的簇中心重新计算,给相同类的点以相同的类标

end for

end for

3　算法分析与测试

实验使用的环境为 Genuine Intel(R) 2140@1.60 GHz 的 CPU,内存为 512 MB,操作系统为 Windows XP Professional,算法的编写使用 Visual C++ 6.0。

为了验证本文改进的 Binning 算法初始化有限混合伯努利模型聚类算法的精确性和可操作性,本文采用三个数据集进行了实验。在同一数据集上运行 K-means 算法和 AGENES 算法与本文算法进行比较。

a)第一个数据集是由数据生成器生成的 3 000 个模拟数据,但每 1 000 个数据点的维数各不相同,有 2% 的噪点分布。表 1 与图 1 就是对用改进的 Binning 初始化算法、K-means 与 AGENES 算法运行数据集时所耗费的时间以及数据点出错数的比较说明。实验结果表明,在数据个数不变、维数增大的情况下,改进的 Binning 法进行初始化的精确度要高于 K-means,与 AGENES 基本持平,且改进的 Binning 初始化算法处理速度是最快的。

表1　三算法运行时间的比较　　单位:ms

数据个数(维)	算法		
	Binning	K-means	AGENES
1 000(100)	62	78	203 139
1 000(150)	93	93	432 810
1 000(200)	93	125	717 340

表2　三算法运行时间的比较　　单位:ms

数据个数(维)	算法		
	Binning	K-means	AGENES
1 200(120)	78	109	409 938
1 800(120)	109	140	1 455 343
2 000(120)	109	156	2 160 922

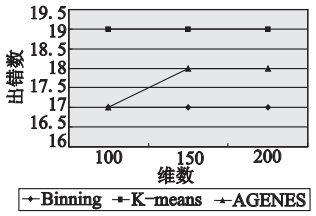

图1　三算法出错数的比较

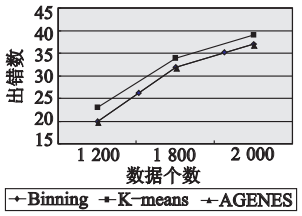

图2　三算法出错数的比较

c)第三个数据集来源于 http://www-users.cs.umn.edu/~han/data/上的 la1 的文本数据库,其文本数量为 3 204,属性为 31 472。原文档集合被划分为 financial、foreign、national、metro、sports 和 entertainment 六类。运用本文算法在同一 PC 机上对该数据进行聚类操作,运行时间为 62 ms,运行 K-means 与 AGENES 算法的时间分别是 165 ms 和 31 285 ms,精确度依次为 0.981 9、0.911 4、0.981 0。

4　结束语

本文针对 EM 算法在收敛上的缺陷,在保持算法迭代的简单性前提下,通过改变 EM 的初始化方法来优化 EM 算法,从而使有限混合伯努利模型聚类算法得到更好的聚类效果。一方面引入 Binning 法来初始化 EM;另一方面针对二元数据的数据特点,改变了数据之间相似度测量方式和中心点的选取方式,并比较了传统的初始化方法。实验表明,该方法有着较好的实际聚类效果和较快的处理速度。下一步的工作是促进有限混合伯努利模型聚类算法,如何减少迭代次数并使参数自适应。

参考文献:

[1] CHENG C, FU A W, ZHANG Yi. Entropy-based subspace clustering for mining numerical data[C]//Proc of the 5th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York:ACM Press, 1999:84-93.

[2] AGRAWAL R, GEHRKE J, GUNOPULOS D, *et al*. Automatic subspace clustering of high dimensional data for data mining applications [J]. ACM SIGMOD Record, 1998,27(2):94-105.

[3] AGGARWAL C C, WOLF J L, YLIP S, *et al*. Fast algorithms for projected clustering[J]. ACM SIGMOD. 1999,28(2):61-72.

[4] AGGARWAL C C, YU P S. Finding generalized projected clusters in high dimensional space[J]. ACM SIGMOD. 2000,29(2):70-81.

[5] FRALEY C. Algorithms for model-based Gaussian hierarchical clustering[J]. SIAM Journal on Scientific Computing, 1999,20(1):270-281.

[6] PATRIKAINEN A. Projected clustering of high-dimensional binary data[D]. Helsinki: Helsinki University of Technology, 2002.

[7] 岳佳,王士同. 高斯混合模型聚类中 EM 算法及初始化的研究[J]. 微计算机信息,2006, 11(3):244-246.

[8] BIERNACKI C. Initializing EM using the properties of its trajectories in Gaussian mixtures[J]. Statistics and Computing, 2004, 14(3):267-279.

[9] SCOTT D W. On optimal and data-based histograms[J]. Biometrika,1979,66(3):605-610.

[10] 和亚丽. 基于高维空间的聚类技术研究[D]. 太原:中北大学, 2005.