数据挖掘 聚类 机器学习 数据库 (12)

计算机科学2000Vol. 27№. 4

42-45

数据挖掘中的聚类方法

Clustering Method in Data Mining

王 实 高 文 TP311.13

(中国科学院计算技术研究所 北京 100080)

Abstract In this paper we introduce clustering method of Data Mining. Clustering has been studied very deeply. In the field of Data Mining, clustering is facing the new situation. We summarize the major clustering methods and introduce four kinds of clustering method that have been used broadly in Data Mining. Finally we draw a conclusion that the partitional clustering method based on distance in data mining is a typical two phase iteration process: 1) appoint cluster; 2) update the center of cluster.

Keywords Data mining, Clustering

从空间 X 中给定一个有限的取样点集(或从数据库中取得有限例子的集合),$\{x^i\}_{i=1}^n$,聚类的目标是将数据聚集成类,使得类间的相似性尽量小,而类内的相似性尽量大。

分类问题(监督)和聚类问题根本的不同是:分类问题中,我们知道训练例的分类属性值,而在聚类问题中,就需要我们在训练例中找到这个分类属性值。

1 数据挖掘领域中的聚类研究

把数据库中的对象集合分割成一组聚类是数据挖掘的基本操作[1],可以用于分类(无监督的)[2],聚合和分割[3],剖析[2],数据缩减,预测。聚类方法基于一些定义好的标准,统计聚类方法基于相似性测量[4~6],而概念聚类方法基于对象具有的概念[7,8]。

数据库中的聚类对象是例子,每个例子由不同的属性构成,这些属性主要分为两类:数值属性(Numeric Attributes,可以比较大小)和符号属性(Categorical Attributes,不能比较大小)。在数据挖掘领域中,由于要处理非常大而复杂的数据集,所以对传统的聚类方法提出两个需要尽量满足的要求:①能同时处理数值属性和符号属性。②算法的效率要满足大数据集的大数量、高复杂性、增量的要求,在现存的聚类方法中,如果能同时处理数值属性和符号属性,那么一般来说,效率很低;而对那些效率高的算法而言,它们大都只能处理数值属性。

现存的聚类算法一般分为分割和分层两种。分割聚类算法通过优化一个评价函数把数据集分割为 k 个部分。分层聚类是由不同层次的分割聚类组成,层次之间的分割具有嵌套的关系。

在数据挖掘中,新近提出的一阶逻辑决策树可以用作分层聚类[9],一个层次上的所有节点定义了一个例子的分布,并且一个节点的所有子节点定义了相应于那个节点的所有例子的分布。

通过使用 Gower 的相似性系数[10]和其它一些不相似测量方法[11],标准的层次聚类方法[4,12]能够同时处理数值属性和符号属性,但不足之处是二次方的计算代价。

机器学习领域中的概念聚类算法通过符号属性来进行聚类,并得出聚类的概念描述。因为这后一种特性有助于解释聚类结果,所以非常适合于数据挖掘。与统计聚类方法不同的是,这些算法基于一种对带有同一或相似概念的对象的搜索,因而其效率依赖于搜索策略。如果数据集具有许多概念和非常多的例子,那么这种基于概念的搜索方法就不合适。

神经网络中的 SOM 方法[13]通过反复的学习来聚类数据。矢量量化 VQ 方法中的 LBG 方法[14]只能对数值属性进行聚类。这两种方法的效率都比较低。

PAM[6,15]方法可以聚类数值属性和符号属性,但效率不高。一种将 PAM 和一个采样过程结合起来的改进方法 CLARA[6]提高了其效率。

很多努力被用于提高现有算法的执行效率以处理非常大的数据集,如在 CLARANS[15]中仔细设计搜索方法(采用随机搜索,新的改进包括使用 R*-树[16,17],在 DBSCAN[18] 中组织索引(采用 R*-树),在 BIRCH[19,20]中组织数据结构(采用类似 R*-树的一种变化:CF 树),以及对 BIRCH 的进一步改进以增进其伸缩性的 BUBBLE 和 BUBBLE-FM[20](这两种算法只需对数据库扫描一次,并产生很高的聚类质量)。这

些努力显著地提高了原有的算法对大数据集的执行效率，不足之处是它们都不能处理符号属性。CLARANS是分割聚类方法，BIRCH是分层聚类方法，这两种方法只能处理具有凸形或球形边界的聚类；而分层聚类方法DBSCAN可以处理具有任意边界外形的聚类。

CURE[21]方法是一种基于随机采样和分割的分层聚类方法，能够发现具有任意边界外形的聚类。

ROCK[22]方法是一种分层聚类方法，通过采用links来测量一对具有符号属性的数据点之间的相似性以实现符号属性聚类。

K-means聚类方法[23,4]和K-median[24]方法在处理大数据集方面非常有效，非常适合应用于数据挖掘，但其只能处理数值属性。而另一种方法Class-Entropy[25]方法与K-means方法相比，效率要低。

一种对K-means改进的方法由Ralambondrainy[26]提出，其主要思想是把符号属性转换成二进制属性（用0和1表示：一个符号在或不在），并在算法中把这些二进制属性当成数值来处理。这种方法需要一个转换过程来处理符号属性，如果符号很多的话，转换后的空间会很大。而另一种方法K-prototypes[27~29]结合了K-means方法和根据K-means方法改进的能够处理符号属性的K-modes方法[27,28]，算法的执行效率较好。

数据挖掘中得到广泛应用的基于距离的分割聚类算法典型地采用两阶段反复循环过程：

1）指定聚类，即指定 x^i 到某一个聚类，使得它与这个聚类中心的距离比它与其它聚类中心的距离要近。

2）修改聚类中心。

算法的结束条件是不再有例子被重新分配。

这一类算法中，有的算法在对每一个例子的每一次指定后就修改一次聚类中心（如SOM方法），有的算法当对所有的例子都指定完后才修改一次聚类中心（如LBG方法）。所以对这一类方法来说，存在两个基本问题，即：(1)如何计算距离；(2)如何修改聚类中心。在计算距离时，对数值属性主要的方法是采用明考夫斯基距离中的欧氏距离，而对符号属性则可以采用海明距离。以下将讨论在数据挖掘领域中得到广泛应用的一些方法。

2 数据挖掘领域中常用的聚类方法

2.1 神经网络方法

神经网络方法中用于聚类的方法主要是SOM（Self-Organizing Feature Map）神经网络，它由输入层和竞争层组成。输入层由 N 个输入神经元组成，竞争层由 $m\times m=M$ 个输出神经元组成，且形成一个二维平面阵列。输入层各神经元与竞争层各神经元之间实现全互连接。该网络根据其学习规则，通过对输入模式的反复学习，捕捉住各个输入模式中所含的模式特征，并对其进行自组织，在竞争层将聚类结果表现出来，进行自动聚类。竞争层的任何一个神经元都可以代表聚类结果。如图1所示，

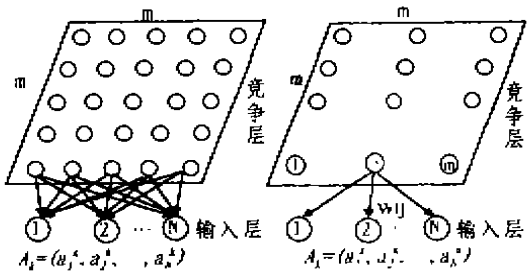

图1 SOM网络基本结构

图2 输入神经元与竞争层神经元j的连接情况

为清楚起见，将图1所示结构中各输入神经元与竞争层神经元j的连接情况抽出，如图2所示。设网络的输入模式为 $A_k=(a_1^k,a_2^k,\cdots,a_N^k)$，$k=1,2,\cdots,p$；竞争层神经元向量为 $B_j(b_{j1},b_{j2},\cdots,b_{jn})$，$j=1,2,\cdots,m$；其中 A_k 为连续值，B_j 为数字量。网络的连接权为$\{w_{ij}\}$，$i=1,2,\cdots,N$；$j=1,2,\cdots,M$。初始为[0,1]区间内的随机值。

值得说明的是：SOM方法是一种两阶段（指定聚类、聚类中心的修改）基于欧式距离的反复循环过程。显然这种方法只能针对于数值属性。

该网络寻找与输入模式 A_k 最接近的连接权向量 $W_g=(w_{g1},w_{g2},\cdots,w_{gN})$，将该连接权向量 W_g 进一步朝与输入模式 A_k 接近的方向调整，而且还调整邻域内的各个连接权向量 W_j，$j\in N_g(t)$。随着学习次数的增加，邻域逐渐缩小，最终得到聚类结果。

SOM网络的最大局限性是，当学习模式较少时，网络的聚类效果取决于输入模式的先后顺序，而且网络连接权向量的初始状态对网络的收敛性能有很大影响。

2.2 矢量量化方法

矢量量化方法VQ中LBG方法用来进行聚类，通常的做法是将所有要识别矢量的集合分成若干子集，各子集中的矢量具有相似特征，因而能用一个具有代表性的矢量来表示。该具有代表性的矢量称为码字。全部码字的集合称为码本。

为了使这种方法的迭代运算不至于无限循环下去，设置了 δ 和 L 两个阈值参数。δ 的值设置得远小于1，当 $\delta^{(m)}<\delta$ 时，表明再进行迭代运算畸变的减小是极有限的，这时可以停止运算。L 是限制最大迭代次数的

参数，以防止 δ 设置得较低时迭代次数过多。

这种方法也是一种两阶段（指定聚类、聚类中心的修改）基于欧式距离的反复循环过程。显然这种方法只能针对于数值属性。此算法的关键是"指定聚类"和"聚类中心的修改"两项。前者完成的工作是以第$(m-1)$步形成的 M 个码字 $Y_i^{(m-1)}$ 为基准，将全部 X 的集合按照最近邻准则划分为 M 个子集 $S_i^{(m)}$，$i=1,\cdots,M$，每一个子集可以看成是一个小区，即"聚类区"。对于 $Y_i^{(m-1)}$ 而言，它所给出的总畸变 $D^{(m)}$ 是最小的，后者完成的工作是按照前者完成的划分求出新的码字 $Y_i^{(m)}$。当采用欧式距离来计算畸变时 $Y_i^{(m)}$ 应是 $S_i^{(m)}$ 中所有质量的质心。由于 $Y_i^{(m-1)}$ 不一定是 $S_i^{(m)}$ 的质心，用 $Y_i^{(m)}$ 代替 $Y_i^{(m-1)}$ 必然能使总畸变下降。这样每完成一次迭代计算，总畸变必然有所下降。因此该算法是一种使总畸变单调下降的算法。

系统的总畸变是它的 M 个码字决定的状态空间点的函数。在大多数实际情况中，该函数并非凸函数，既有全局最小点，又有多个局部最小点。所以算法一般取得的是一个局部最优解。

2.3 K-means 方法

K-means 算法是一种分割的而非分层的聚类方法，在数据挖掘领域中得到了最广泛应用。给定一个例子的集合 X，其每一个属性均为数值属性，和一个整数 $k(\leqslant n)$，K-means 算法将 X 分割为 k 个聚类并使得在每个聚类中所有值与该聚类中心距离的总和最小，每个聚类的聚类中心是每个聚类的均值。该过程可以被描述为如下数学问题：

$$\text{最小化：}P(W,Q)=\sum_{l=1}^{k}\sum_{i=1}^{n}w_{i,l}d(X_i,Q_l) \tag{1}$$

$$\text{满足：}\sum_{l=1}^{k}w_{i,l}=1,w_{i,l}\geqslant 0,i=1,\cdots,n,l=1,\cdots,k \tag{2}$$

其中，W 是一个 n×k 分割矩阵，用以表示每个例子在哪个聚类中，通常它的每一行的和为1，$Q=\{Q_1,Q_2,\cdots,Q_k\}$是聚类结果的集合。$d(\cdot,\cdot)$是两个对象的欧氏距离的平方。

问题 P 可以通过反复求解如下两个子问题 P1和 P2而得到解决：

1）问题 P1：固定 $Q=\bar{Q}$，解决简化后的问题 $P(W,\bar{Q})$，即指定聚类。

2）问题 P2：固定 $W=\bar{W}$，解决简化后的问题 $P(\bar{W},Q)$，即聚类中心的修改。

对问题 P1采用如下办法解决：

$$w_{i,l}=1 \quad \text{如果 } d(X_i,Q_l)\leqslant d(X_i,Q_t),\text{对}1\leqslant t\leqslant k$$
$$w_{i,t}=0 \quad \text{对 } t\neq 1 \tag{3}$$

对问题 P2采用如下办法解决：

$$q_{l,j}=\frac{\sum_{i=1}^{n}w_{i,l}x_{i,j}}{\sum_{i=1}^{n}w_{i,l}} \quad \text{对}1\leqslant l\leqslant k,1\leqslant j\leqslant m \tag{4}$$

所以解决问题 P 的基本算法同样是一个两阶段的反复循环的过程。

因为 $P(\cdot,\cdot)$是非凸的并且由算法产生的序列 $P(\cdot,\cdot)$是严格的降序，所以经过有限步循环后，算法将收敛于一个局部最小点。算法的时间复杂度为 $O(Tkn)$，其中 T 为循环次数，n 为输入数据集的对象的个数。

K-means 算法具有如下重要的特点：1）能有效地处理大数据集，2）它经常中止于一个局部最优解。3）由于欧式距离的局限性，它仅能处理数值属性。4）聚类结果具有凸的外形。5）算法的执行结果和例子的顺序有关。

2.4 K-prototypes 方法

先介绍根据 K-means 方法改进的处理符号属性的 K-modes 方法。

1）在不相似性测量中，采用海明距离取代欧氏距离；其中 X,Y 为具有 m 个符号属性的例子，x_j,y_j 是对应的属性。

$$d_1(X,Y)=\sum_{j=1}^{m}\delta(x_j,y_j)$$
$$\delta(x_j,y_j)=\begin{cases}0(x_j=y_j)\\1(x_j\neq y_j)\end{cases} \tag{5}$$

2）一个集合的 mode：X 为具有 $A_1,A_2,\cdots,A_m$ 符号属性的例子的集合，$X=\{X_1,X_2,\cdots,X_n\}$，其中 $X_i,i=1,\cdots,n$，为 X 中的一个例子。X 的一个 mode 被定义为一个向量 $Q=[q_1,q_2,\cdots,q_m]$满足最小化下式：

$$D(X,Q)=\sum_{i=1}^{n}d_1(X_i,Q) \tag{6}$$

Q 不必是 X 的一个例子。

3）寻找一个集合的一个 mode：在 X 中，具有属性 A_j 的第 k 个符号属性值 $c_{k,j}$ 的例子数为 $n_{c_{k,j}}$，且符号属性 $c_{k,j}$ 在 X 中的相对频率为：

$$f_r(A_j=c_{k,j}|X)=\frac{n_{c_{k,j}}}{n} \tag{7}$$

定理1[29] 函数 $D(X,Q)$是最小化的当且仅当对所有的 $j=1,\cdots,m$，和 $q_j\neq c_{k,j}$，$f_r(A_j=q_j|X)\geqslant f_r(A_j=c_{k,j}|X)$。

该定理给出如何在一个给定 X 中寻求 Q 的方法。该定理也隐含指出一个数据集 X 的 mode 不是唯一的。例如，集合{[a,b],[a,c],[c,b],[b,c]}的 mode 为[a,b]或[a,c]。

4）K-modes 算法：当采用式(5)作为不相似性测量方法后，代价函数式(1)变为：

$$P(W,Q)=\sum_{l=1}^{k}\sum_{i=1}^{n}\sum_{j=1}^{m}w_{i,l}\delta(x_{i,j},q_{l,j}) \tag{8}$$

其中 $w_{i,j}\in W$，并且 $Q_i=[q_{i,1},q_{i,2},\cdots,q_{i,m}]\in Q$。

对 K-means 算法做如下改进：对 P1问题用(5)式来解决。对 P2问题，采用聚类的 modes 代替 means 并根据定理1选 modes，该算法的收敛情况同 K-means 方法。

其次介绍将 K-means 和 K-modes 结合的 K-prototypes 方法。

1)例子 X,Y，其属性为 $A_1,A_2,\cdots,A_p,A_{p+1},\cdots,A_m$，其中 $A_1,A_2,\cdots,A_p$ 属性为数值属性，$A_{p+1},\cdots,A_m$ 为符号属性，则测量它们两个的下相似性如下式：

$$d_2(X,Y)=\sum_{j=1}^{p}(x_j-y_j)^2+\gamma\sum_{j=p+1}^{m}\delta(x_j,y_j) \qquad (9)$$

前一部分基于欧氏距离，后一部分基于海明距离。

2)则(1)式改进为：

$$P(W,Q)=\sum_{i=1}^{k}\left[\sum_{j=1}^{n}w_{i,j}\sum_{j=1}^{p}(x_{i,j}-q_{i,j})^2+\gamma\sum_{j=1}^{n}w_{i,j}\sum_{j=p+1}^{m}\delta(x_{i,j},q_{i,j})\right] \qquad (10)$$

对公式两边分别求解就可以最终寻求一个局部最优解 Q^*,W^*。在这里，γ 为平衡 K-means 和 K-modes 两边的一个参数。

3 聚类的质量

评估聚类的质量分为：描述质量，纯粹用于描述(在数据集中识别聚类)；预测质量，用于预测。

1)描述质量：测量聚类的描述质量很困难，没有一个意见一致的标准。一个比较通用的描述质量标准是 partition utility[30]，定义为：

$$PU(\{C_1,\cdots,C_N\})=\sum_k CU(C_k)/N \qquad (11)$$

即，每个聚类的 category utility 的一个平均[31]，其中 category utility 定义如下：

$$CU(C_k)=P(C_k)\sum_r\sum_j(P(A_r=V_{rj}|C_k)^2-P(A_r=V_{rj})^2) \qquad (12)$$

这里 $C_k,k=1,\cdots,N$，为一个聚类。A_r 是属性，其属性域由 V_{rj} 组成。

2)预测质量：如果一个聚类用于预测，那么就要遵循如下的质量标准：

①绝对标准，一个聚类 C 在实例集 E 上，最大化下式：

$$Q(C)=-E(d(x,p(f(x,C)))^2) \qquad (13)$$

其中，f 是一个聚类指定函数，p 是求聚类中心的函数，d 是距离，E 是期望。最大化 Q 意味着在聚类过程中，最小化聚类内部的方差。

②相对标准，相对标准即是相对差：

$$RE=\frac{\sum_{i=1}^{n}d(e_i,p(f(e_i,C)))^2}{\sum_{i=1}^{n}d(e_i,p(Tr))^2} \qquad (14)$$

其中，e_i 是在测试集中的例子，$p(f(e_i,C))$是相应的预测，$p(Tr)$是训练集的中心，最大化 Q 意味着最小化 RE。

4 总结

数据挖掘中的聚类方法主要存在如下的问题：

1)符号属性的问题：大部分聚类方法因为是基于欧式距离的，所以只能处理数值属性，新出现的一些聚类方法，例如 K-prototypes 方法不但可以处理数值属性也可以处理符号属性，这样就可以大大扩展聚类方法的应用范围。

2)算法的效率：通过对现有的聚类算法进行改进，使之具有伸缩性，具有增量聚类的能力，在处理大数据集时，对数据库只需扫描一次[30,24,12]。这些是当前数据挖掘领域研究的一个重要问题。

3)初值的选择：初值的选择对聚类算法的最终结果有很大的影响。在数据挖掘领域中可以实现的方法是采用多组不同的初值并进行多次迭代，最后选其中的最佳者作为运算结果，但不能保证一定能够达到全局最小值。

4)输入顺序：许多算法对数据的输入顺序非常敏感，如 CLARANS[16]。

5)最优解的问题：聚类问题本质是一个优化问题，这就是通过一种迭代运算使得系统的目标函数达到一个极小值。但是这个目标函数在状态空间中是一个非凸函数，它有很多极小值，其中只有一个是全局最小值，而其它都是局部最小值。优化的目标是达到前者而非后者中的任意一个，因而后者可能较之前者大得多，所以待解决的问题是一个非凸优化问题。这里的问题出在算法使目标函数在迭代中只降不升，因此一旦落入某个局部最小点就再也拔不出来。解决这个问题的办法在于是算法具有随机性，这就是让每一次迭代运算中目标函数上升的概率不等于零相应的算法称为"随机松弛算法"，简记为 SR(Stochastic Relaxation)算法。能够从理论上严格证明，在满足一定的条件时，SR 算法能够保证目标函数收敛到全局最小点的概率为 1[33]。但为了求得最优解往往需要耗费大量的时间。在数据挖掘领域中，大数据集使得求最优解问题变得不太可能，所以一般情况下，采用初值的选择来求得一个近似最优解。

(参考文献共44篇，略)