使用 BP 网络改进 K-means 聚类效果

王银辉　熊忠阳
（重庆大学计算机学院　重庆 400040）

摘　要　K-means 算法中的 k 值的确定和初始聚类中心的选择严重影响聚类效果。针对这一问题，本文提出使用 BP 神经网络改进 K-means 聚类效果的方法。通过对聚类结果进行反复训练，调整聚类数，K-means 的聚类效果得到改善。采用人工数据和实际商业数据的实验证明该方法能有效地改善传统的聚类效果。

关键词　K-Means，BP，聚类，神经网络

Using BP Neural Network to Improve K-mean'Result

WANG Yin-Hui　XIONG Zhong-Yang
(School of Computer, Chongqing University, Chongqing 400040)

Abstract　The value of K and the selection of initial centers heavily affect the result of K-means algorithm. Aiming at this problem, this paper puts forward a method that uses BP neural network to improve K-means' result. Training clustering result by BP network, the effect of K-means can be improved. Experiments using artificial data and actual business data testify the validity of this method. It can improve the traditional K-means effect well.

Keywords　K-Means, BP, Clustering, Neural network

1　引言

K-means(MacQueen, 1967)是一种经典动态聚类算法。该算法具有简单、高效的特点，能够有效处理大数据集。但算法的聚类效果受聚类数 K 的影响。对未知聚类数的数据集，K 通常由人工确定。这种随机选择的方法常常严重影响聚类效果，同时聚类结果也受到初始聚类中心的影响。为了得到更好的结果，必须修正 K 值和初始聚类中心。一种常见的修正方法是多次聚类，改进每次聚类的效果。这种思想和反馈机制很容易联系在一起，因此想到用 BP 网络改善 K-means 的效果。本文探索了如何使用 BP 神经网络改善 K-means 聚类效果的方法。

BP 神经网络是一种重要的人工神经网络模型，由于其高度非线性映射能力[2]而被广泛应用于各个领域。在研究 BP 算法的过程中，经过实验得出一种使用 BP 网络改进 K-means 的方法。在 K-means 中，当 k 和初始聚类中心确定后，在样本集和最后形成的中心之间存在一种映射关系，即：

$$\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \xrightarrow{f} \begin{bmatrix} c_1 \\ c_2 \\ \vdots \\ c_K \end{bmatrix}$$

BP 网络可以用来反映这种变换关系。把样本数据和各样本点到中心的距离输入 BP 网络，在训练过程中，把 BP 网络隐层节点的调整结果传递回 K-means。通过这种方法，K-means 不断地修正直到得到满意的解。

2　K-means 聚类

主要思想：首先初始化聚类中心 C，然后将各样本按"最近邻"原则分类。描述如下：输入一样本集 X 和整数 K（聚类数），输出 X 的一个划分：

$P_k=\{C_1,\cdots,C_k\}$。任意选择 K 个初始中心，根据距离将 $X=\{X_1,\cdots,X_n\}$ 归入到 K 个簇中。重新计算每个簇的中心直到其不再变化为止。

算法描述

①从给定的样本集中随机选择 K 个初始中心

$M^{(0)}(j)=[m_1^{(0)}(j),m_2^{(0)}(j)\cdots,m_n^{(0)}(j)],j=1,2,\cdots,k$，令 $t=0$

②划分样本集 $X(i)=[x_1(i),x_2(i),\cdots,x_n(i)],i=1,2,\cdots$，根据各点到中心的距离将样本划分到一个簇中：

$X(i)\in C^{(t)}(j)$ if $D(X(i),M^{(t)}(j)\leqslant D(X(i),M^{(t)}(l),l=1,2,\cdots,k$

③更新簇的中心，得到：

$M^{(t+1)}(j)=[m_1^{(t+1)}(j),m_2^{(t+1)}(j),\cdots,m_n^{(t+1)}(j)],j=1,2,\cdots,k$

$M^{(t+1)}(j)=\frac{1}{\Omega_j}\sum_{X(i)\in C(j)}X(i)$，$\Omega_j$ 是簇 C_j 中样本的个数

使得 M 满足 C_j 中所有点到新的聚类中心距离和最小，即：

$$\sum_{X(i)\in C^t(j)} D(X(i),M^{t+1}(j))\rightarrow\min$$

④如果算法收敛则退出

$M^{(t+1)}(j)=M^t(j)$，对所有 j

否则，令 $t=t+1$，转向②

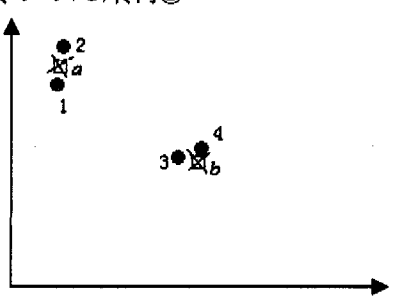

图 1

从上面所介绍的步骤得知，在聚类过程中，每次循环都要计算各样本点到簇中心的距离。图 2 描述了图 1 的拓扑结构。

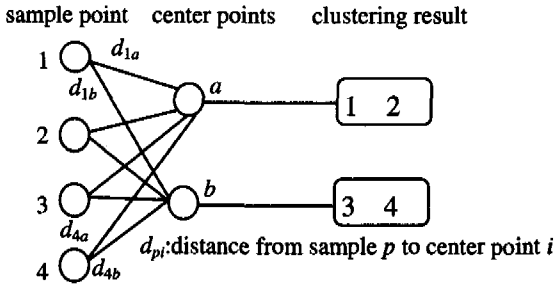

图 2

3 BP 神经网络

3.1 BP 神经网络模型

BP 神经网络是一种利用后向传播来训练的前馈网络。BP 是一种有导师的学习算法，采用非线性多层前馈的网络结构[2]，使网络的均方差最小。

BP 网络中，输入层神经元接收要被网络处理的数据，输出层产生全局计算结果。隐层表示问题的复杂度，通常采用一个隐层。所有输入层的神经元与隐层连接，隐层又与输出层连接。每个连接都用一个权值表示。结构如图 3 所示。

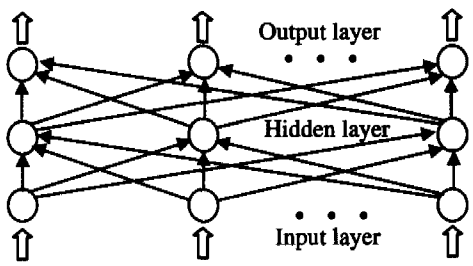

图 3

3.2 BP 网络隐层节点的分析

在 BP 网络的拓扑结构中，输入节点和输出节点数量与要解决的问题有关。隐层节点数量的选择没有固定方法。

设 O_{pi} 是样本 p 在训练中经隐节点 i 的输出，O_{pj} 是样本 p 经隐节点 j 的输出，n 是样本总数。

$$O_i=\frac{1}{n}\sum_{p=1}^{n}O_{pi} \quad O_j=\frac{1}{n}\sum_{p=1}^{n}O_{pj}$$

令 $\alpha_p=O_{pi}-O_i \quad \beta_p=O_{pj}-O_j$

对隐节点 i、j，它们相应的输出序列分别为 O_{pi} 和 O_{pj}，这两个序列的相关度 ρ_{ij} 定义为

$$\rho_{ij}=\frac{\sum_{p=1}^{n}\alpha_p\beta_p}{\sqrt{\sum_{p=1}^{n}\alpha_p^2}\cdot\sqrt{\sum_{p=1}^{n}\beta_p^2}}$$

显然 $|\rho_{ij}|\leqslant 1$，ρ_{ij} 表示了隐节点 i 和 j 的相关度。如果 ρ_{ij} 太大，则表明 i 和 j 相似，因此这两个节点可以被合并成一个节点。

定义隐节点 i 的发散度 S_i 为：

$$S_i=\frac{1}{n}\sum_{p=1}^{n}O_{pi}^2-O_i^2$$

如果 S_i 太小，则表明所有样本对应于节点 i 的输出变化小，即节点 i 在训练过程中没有起明显作用。

根据相关度和发散度的定义，下面给出调整隐层节点的规则：

合并规则：如果 $|\rho_{ij}|\geqslant\varepsilon_1$ 且 $s_i,s_j\geqslant\varepsilon_2$，合并 i 和 j。（ε_1 和 ε_2 是给定的阈值，本文中，$0.85\leqslant\varepsilon_1\leqslant 0.90$，$0.005\leqslant\varepsilon_2\leqslant 0.01$）。

删除规则：如果 $S_i<\varepsilon_2$，则删除节点 i。

4 BP-K-Means

4.1 设计思想

把 K-means 算法的中心点看作隐节点，归一化处理后的样本点到中心的欧式距离看作连接输入层节点和隐层节点的权值。ρ_{ij} 反映了节点 i 和 j 的相关度。如果 $|\rho_{ij}|\geqslant\varepsilon_1$，表明 i 和 j 所代表的簇之间的差异不明显，因此合并簇 i 和簇 j，即两个中心点被合并为一个。如果发散度太小，表明当前的中心点不能很好地划分各样本，因此试着找出更好的聚类中心。

4.2 算法实现

①输入样本集 P，初始化 BP 网络的参数：系统误差设为 0.001，学习率为 0.15，惰性因子为 0.075，训练次数为 20000，初始化两个集合 N 和 M 均为空，循环次数 $t=0$。

②确定初始簇的数量 K'（K' 应为一个较大的值，例如：$n/5$，n 为样本个数），随机选择初始的聚类中心 $C'=\{C'_1,C'_2,\cdots,C'_{K'}\}$。

③聚类。产生 K' 个簇和各簇的中心 $C=\{C_1,C_2,\cdots,C_{k'}\}$，保存各样本点到 C 的距离，则得到距离矩阵 d_{pi}，（$p=1,\cdots,n,i=1,\cdots,K'$）

设当前聚类结果为 $R[t]$，$t=t+1$。如果 $R[t-1]$的效果好于 $R[t]$，算法结束并输出 $R[t]$，衡量聚类效果的标准是所有簇之间的距离差异概率。

④构造 BP 网络，结构为 $n-K'-1$，网络的期望输出是 K-Means 的划分结果。根据集合 C 中的节点的序列给每个隐层节点编号，$O=\{O_1,O_2,\cdots,O_{K'}\}$。经过多次实验比较，选择双曲线函数 $f(x)=\frac{1-e^{-x}}{1+e^{-x}}$ 作为隐层节点的激励函数效果较好。初始化隐节点的阈值。

⑤ 初始化网络的权矩阵 w：

$$w_{pi}=\left\{\begin{array}{l}\dfrac{d_p-\bar{d}_i}{\max\{d_{pi}\}-\min\{d_{pi}\}},d_p\neq\bar{d}\\ \omega(\omega=0.05),d_p=\bar{d}\end{array}\right\}(p=1,\cdots,n,i=1,\cdots,K')$$

$\bar{d}_i=\frac{1}{n}\sum_{p=1}^{n}d_{pi}$，$d$ 是样本点到簇中心的欧式距离，n 是样本数量。

⑥计算各隐节点的输出值。

⑦如果隐节点 i 和 j 符合删除规则，则删除 i，把 i 的编号存入集合 M 中。如果节点 i 和 j 符合合并规则，则将 i 和 j 合并为一个节点，把 j 的编号存入集合 N 中。

⑧重复正向计算、后向传播、修改权值直到系统误差足够小，如果一次循环中训练次数超过规定的次数则退出算法并输出 $R[t]$。

⑨如果 N 非空，则修改初始聚类中心：$C'=C'-N$，$K'=K'-|N|$，如果 M 不为空，则随机选择其他中心点 Q，修改初始聚类中心：$C'=\{C'-M\}\cup Q$，$|Q|=|M|$，K'不变，如果 N 和 M 都为空，则退出并输出 $R[t]$，否则转向步骤③。

5 实验结果

5.1 训练过程中的聚类结果和簇中心变化情况(人工产生的数据)

A

图 4 初始状态 $K'=5$

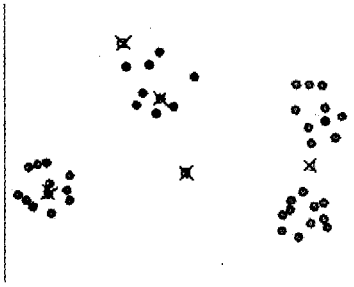

图 5 第一次聚类后的结果 $K'=5$

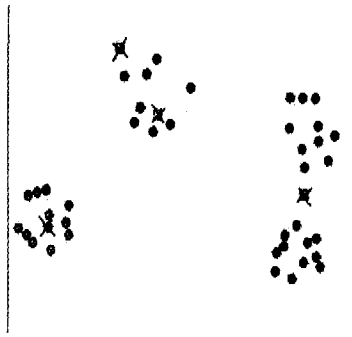

图 6 第一次训练后的结果 $K'=4$

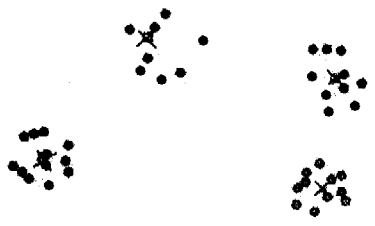

图 7 最后结果 $K'=4$

5.2 算法效果(样本来自重庆医药公司 2004 年 11 月的销售数据,样本采用 3 维属性:价格、区域、药品种类)

表 1

样本数	初始簇数	最终簇数	总时间(毫秒)
50	6	3	31
100	10	4	74
500	15	9	518
1,000	20	15	1237

注:初始簇的数量会影响整个计算时间,但是该值不能被设为太小,否则不会得到最优解。

结束语 本文给出了使用 BP 网络的隐层激励函数反映 K-means 算法簇的性质的方法。通过这种方法来改善聚类效果。实验数据验证了新方法的效果。BP 网络的特性使采用这种方法适合于解决复杂问题,特别是多维数据。

需要注意的是:聚类结果很大程度上依赖于合适的隐层节点激励函数。实验中也曾经采用 S 函数作为激活函数,但是聚类效果失真。下一步的工作将是找出从样本数据和中心点到激励函数之间的变换的数学映射关系,一旦实现这一目标,该方法将能够适用于更多场合。

参 考 文 献

1 Kanungo T, Mount D M, Netanyahu N S, et al. An Efficient k-Means Clustering Algorithm: Analysis and Implementation [J]. IEEE Transaction On Pattern Analysis And Machine Intelligence, 2002, 24(7)

2 Yuan Zengren. The artificial neural network and its application [M]. Tsinghua University Publishing House, 1999. 66~101

3 Cyberko G. Approximations by superpositions of a sigmoidal function [J]. Math Control Singnal System, 1989. 34~89

4 Jain A K, Dubes R C. Algorithms for Clustering Data [M]. Prentice-Hall, 1988

5 Kandel E R, Schwarts J. Principles of neural science [M]. Elsevier, 1985. 1~60

6 Lippmann R P. An Introduction to computing with neural nets [J]. IEEE ASSP Magazine, April 1987, 12~35

7 Fu Jinguang, Xu Gang, Wang Yuguo. Clustering Based on Genetic Algorithm Computer Engineering. 2004, 30(4)

8 Huang Z. Extensions to the k-means algorithm for clustering large data sets with categorical values. Data Mining and Knowledge Discovery, 1998, 2

9 Han Jiawei, Kamber M. Data Mining: Concepts and Techniques [M] Beijing: Publishing House of Mechanical Industry, 2001, 8

10 Hecht-Nielsen R. Theory of the Back-Propagation neural network. In: Proc. IEEE Intl. Conf. Neural Network, 1989

11 Wu Jinlan, Zhu Wenxing. An Iterated Local Search Algorithm for K-Means Clustering Computer engineering and application. 2004, 22