

基于模糊聚类的神经网络在数据挖掘分类中的应用研究

王 刚, 黄丽华, 张成洪

复旦大学管理学院, 上海 200433

[摘要] 常见的决策树分类算法、贝叶斯分类算法、神经网络分类算法为数据挖掘分类算法研究提供了重要基础。但面对海量数据时, 在时间效率、鲁棒性和精确性上都显示出了不足。为此, 本文将模糊聚类的思想引入到神经网络分类算法中, 首先通过模糊聚类子模型, 将样本数据聚为几个数据子集, 然后再采用不同的神经网络对各个数据子集同时进行训练学习。由于经过了模糊聚类子模型的预处理, 每个神经网络训练学习样本的复杂性大大减少, 使神经网络的学习效率大大提高。最后通过 UCI 下的实际数据库, 对提出的分类算法进行了检验, 结果显示了基于模糊聚类的神经网络在数据挖掘分类中应用的有效性。

[关键词] 数据挖掘; 分类; 神经网络; 模糊聚类

[中图分类号] TP18

[文献标识码] A

[文章编号] 1000-7857(2007)15-0058-04

Application of Neural Network Based on Fuzzy Clustering in the Classification Algorithm of Data Mining

WANG Gang, HUANG Lihua, ZHANG Chenghong

School of Management, Fudan University, Shanghai 200433, China

Abstract: The common classification algorithms, such as Decision Tree, Bayesian classification and Artificial Neural Network, provide an important foundation for the classification algorithm of data mining. But due to huge amounts of data, these algorithms face some challenges in time efficiency, robustness and accuracy. This paper proposes an artificial neural network model based on fuzzy clustering. First, it clusters the training data into different sub-data using the Fuzzy Clustering Model. Subsequently, it uses various artificial neural networks to train the sub-data. After that, the number and complexity of training data is reduced and the efficiency of the artificial neural network is enhanced greatly. In the end, the UCI's databases are used to prove the usefulness of the new classification algorithm. The results show the validity of application of the Neural Network based on fuzzy clustering in the classification of data mining.

Key Words: data Mining; classification; neural network; fuzzy clustering

CLC Number: TP18

Document Code: A

Article ID: 1000-7857(2007)15-0058-04

0 引言

数据库中收集了大量的数据, 分类分析已成为数据挖掘研究领域中的一个非常活跃的研究课题。现在关于分类的算法已有很多, 如基于决策树的分类、基于统计的贝叶斯分类、基于神经网络的分类, 还有近几年兴起的基于关联规则的分类、基于数据库的分类、K-最临

近分类、基于案例推理的分类、基于遗传算法的分类、基于粗糙集的分类和基于模糊集的分类等^[1]。

虽然现在关于分类的算法有很多, 并且有的还很成熟, 但是面对海量数据, 上述算法在时间效率、鲁棒性、精确性上还存在一些问题。Lin 等^[2]用实验的方法对各种分类算法进行比较后, 尚未发现一种方法对所有

收稿日期: 2007-03-12

基金项目: 国家自然科学基金项目 (70571016, 70471011)

作者简介: 王 刚, 上海市国顺路 670 号复旦大学管理学院, 博士研究生, 研究方向为商务智能、混合智能系统、知识管理等;

E-mail: wgedison@gmail.com

张成洪 (通讯作者), 上海市国顺路 670 号复旦大学管理学院, 副教授, 研究方向为信息管理与信息系统;

E-mail: chzhang@fudan.edu.cn

的数据都优于其他方法。并且从数据挖掘在国际上的发展来看,数据挖掘的研究重点现已从提出概念和发现方法转向系统应用和方法创新上,研究中注重多种发现策略和技术集成,以及多种学科之间的相互渗透。一些学者通过将各种方法有机结合起来,扬长避短,以获得各种模型整合功效,从而解决目前数据挖掘分类的新问题^[3]。

本文正是按照这样的思路,同时考虑到神经网络对噪声数据具有高承受能力、分类精度高、可用训练过的神经网络提取分类规则的优点,以及具有学习训练时间长、不易收敛的缺点,将模糊聚类引入到神经网络的学习过程中,通过模糊聚类将初始样本根据其不同特征进行“归类”,然后再对各组“相似”的样本采用不同的神经网络进行训练。由于各组样本在数量和复杂性上较原始样本都有所减少,因此可大大提高此时神经网络的训练效率,从而达到减少训练时间、提高收敛性的目的。

1 基于模糊聚类的神经网络分类算法

1.1 基于模糊聚类的神经网络模型总体框架

为了达到提高网络的训练效率、减少训练时间,而不损失信息量、保持使用精度的目的,本文首先采用模糊聚类的方法,将训练数据 D 聚为 k 类: $D_1, D_2, \dots, D_k$ 。由于每一类的样本之间存在着一定的相似性,并且样本个数也较原始样本个数少,当再使用神经网络训练时,单个神经网络的训练时间必然会减少。接下来再对聚类后的数据 $D_1, D_2, \dots, D_k$ 采用不同的神经网络 $ANN_1, ANN_2, \dots, ANN_k$ 同时进行训练,根据并行性原理,这样就可以使整个网络效率得到提高。最后为了能对各个神经网络训练的不同结果进行综合,采用第一步中得到的模糊权值 $\omega_1, \omega_2, \dots, \omega_k$ 将各个网络的训练结果 $O_1, O_2, \dots, O_k$ 综合起来,从而完成整个网络的训练。基于模糊聚类的人工神经网络模型可分为 FCM 子模型和 FNN 子模型,其总体结构如图 1 所示,图中虚线表示模型的训练过程,实线表示模型的工作阶段。

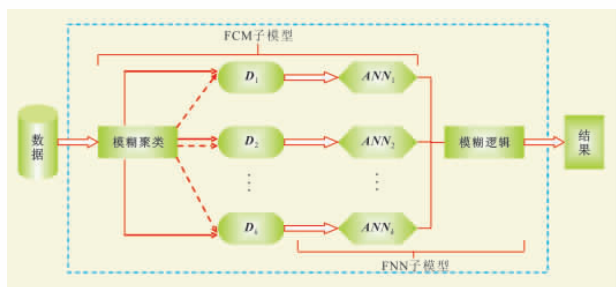


图 1 基于模糊聚类的神经网络分类算法

Fig.1 Classification algorithm of neural network based on fuzzy clustering

1.2 FCM 子模型

FCM 子模型完成数据的模糊聚类任务,这个过程主要有两个任务:首先要确定聚为几类,然后再确定采用什么方法来聚类。

1.2.1 聚类中心个数的确定

关于这个问题,不同的学者从不同的角度作了探讨。有的学者根据研究发现^[4]:样本的可分性越好,其对应的最优加权指数 m^* 就越大;反之,样本的可分性越差, m^* 越低。从这一结论,可以设计一种聚类有效的判别方法——比较法来解决此问题,即设聚类个数 C 从 2 开始增加到某一常数值,分别计算 m^* ,然后根据所得 m^* 来选取最适宜的 C 值。只有得到最优的 C^* 划分时,对应的 m^* 才最大。

虽然上述方法能解决这个问题,但需要很多的搜索时间,并有可能陷入局部最小。本文采用 Chiu 于 1994 年提出的 Subtractive Clustering 算法。该算法是对 Yanger 的 Mountain Clustering 的扩展,它假定每一个数据都是一个潜在的聚类中心,基于数据样本的点的密度,计算每一个可能定义为聚类中心的样本点的可能性量度来进行模糊聚类。具体过程详见文献[5]。

1.2.2 模糊 C 均值聚类

FCM 子模型除了要对数据进行聚类外,还要产生一定的隶属度,为 FNN 子模型打下“基础”。我们知道在聚类分析中,按照划分的结果不同可分为硬聚类和软聚类。其中硬聚类是不同聚类间的界限明显的聚类,而软聚类(基本上指模糊聚类)的每个输入样本可能以不同的隶属度或概率属于一个或多个聚类,隶属函数或概率是输入样本和聚类中心的关系表述的 0~1 的值,该输入与所有分类的关系值总和为 1。根据 FCM 子模型的特点和要求,本文选取软聚类中的模糊聚类 FCM 算法作为 FCM 子模型的聚类方法。限于篇幅,FCM 的具体步骤详见文献[5]。

1.3 FNN 子模型

FNN 子模型完成对模糊聚类子模型的训练,不同的聚类采用不同的神经网络训练,并通过模糊聚类子模型产生的隶属度将各神经网络的结果结合起来,从而达到学习训练的目的。

考虑到经典 BP 算法的低效性,本文的 FNN 子模型采用了一种基于 L-M (Levenberg-Marquardt) 算法的神经网络^[5]。

L-M 学习算法用在多层前向神经网络中,对应于权值的调整率为

$$\Delta W = (J^T J + \mu I)^{-1} J^T r \quad (1)$$

其中, J 为网络学习误差对权值微分的 Jacob 矩阵, r 为网络输出的误差向量, μ 为一个标量。

L-M 学习算法具体过程如下所示。

1) 算法: LM 算法。

- 2) 输入:训练样本 samples。
- 3) 输出:一个训练的对样本分类的神经网络。
- 4) 方法:
 - ① 初始化网络,初始权值为 W_0 ;
 - ② $k=0$;
 - ③ 输入样本 X_k , 计算 G_k ;
 - ④ 当 $\|G_k\| \geq \varepsilon$ 为真时,继续下行,否则程序结束;
 - ⑤ 分解 $H_k + \mu_k I$, 若不正定,置 $\mu_k = 4\mu_k$, 并重复这一步直到 $H_k + \mu_k I$ 正定;
 - ⑥ 解方程 $(H_k + \mu_k I)s = -G_k$, 得 s_k ;
 - ⑦ 求 $E(H_k + s_k)$, $Q_k(s_k)$, 和 r_k , 这里 $Q_k(s_k) = E(W_k) + G_k^T s + 1/2 \cdot s^T H_k s$;
 - ⑧ 若 $r_k < 0.25$, 置 $\mu_{k+1} = 4\mu_k$; 若 $r_k > 0.75$, 置 $\mu_{k+1} = \mu_k/2$; 否则置 $\mu_{k+1} = \mu_k$;
 - ⑨ 若 $r_k \leq 0$, 置换 $W_{k+1} = W_k$; 否则置 $W_{k+1} = W_k + s_k$;
 - ⑩ 令 $k=k+1$, 跳至步骤④。

其中, E 为网络学习的误差函数; $G_k = \nabla E$, H_k 表示网络学习误差函数对权值微分的 Hesse 矩阵; X_k 表示输入的样本, I 为单位矩阵。

通过使用不同的神经网络训练后,可以得到各网络的输出向量 $O_1, O_2, \dots, O_k$, 再使用模糊逻辑将各神经网络的训练结果结合起来: $O = \omega_1 \cdot O_1, \omega_2 \cdot O_2, \dots, \omega_k \cdot O_k$, 从而得到最终的训练结果。

2 实验及结果分析

为了验证基于模糊聚类的神经网络在数据挖掘分类中的有效性,本文使用 UCI 机器学习数据库^[6]中的 6 个数据集作为实验数据集,各数据集的基本信息如表 1 所示。实验环境为 600 MHz Celeron II CPU, 128 MB 内存, 操作系统为 Microsoft Windows 2000, 编程软件为 Matlab 6.5。

对各实验数据集采用基于模糊聚类的神经网络训练时,FCM 子模型中聚类中心个数的确定采用了 Matlab 6.5 下的 subtractive 模块完成,各参数取默认值,模糊 C 均值聚类则可由 Matlab 6.5 下的 FCM 模块完成。对 FNN 子模型中神经网络结构的确定,本文采用 3 层

表 1 数据集概况
Table 1 Outline of data sets

数据集	BREAST	CAR	CMC	IRIS	YEAST	ZOO
属性数	9	6	9	4	8	16
分类数	2	4	3	3	10	7
实际样本数	683	1 728	1 473	150	1 484	101

结构的网络,隐含层节点个数根据经验公式 $n_1 = \sqrt{m+n} + \beta$ 确定^[5], 其中 m 为输出层节点数,即为各数据集的属性数, n 为输入层节点数,即为各数据集的分类数, β 为 1~10 之间的常数, FNN 子模型中神经网络的误差值统一取 0.1 (接下来的对比实验中,为了有可比性,各神经网络的误差值也统一取 0.1)。关于训练集和测试集的获得本文采用保持法,即随机将原始数据集 2/3 的数据用作训练集,剩下的 1/3 用作测试集。限于篇幅,具体的训练过程这里略去。

将本文提出的分类算法与经典的 BP 算法、LM 算法进行比较,其中各算法的参数取值同基于模糊聚类的神经网络模型一致,30 次重复实验的平均结果如表 2 所示。其中 BP, LM 表示分别采用 BP, LM 算法的神经网络, FCM-FNN 表示基于模糊聚类的神经网络;表中“⊕”表示实验过程中网络都收敛,“○”表示实验过程中网络存在发散的情况,“-”表示在此精度下,网络无法收敛;对每一种算法,第 1 列数据为训练所耗时间,单位为 s, 第 2 列数据为分类的精确。

通过对比实验可以看出,随着样本数量的增加,原有的 BP、LM 算法都会出现不能收敛的情况,但基于模糊聚类的神经网络则在保证了一定的分类精度的条件下基本上都能收敛,这就使得算法在时间效率、鲁棒性和精确性上得到了提高。

3 结论

20 世纪 80 年代末,随着数据库、互联网技术的迅速发展,以及管理信息系统、网络数据中心的推广应用,数据的存取、查询、描述统计等技术日臻完善,但高

表 2 实验结果
Table 2 Experiment Results

数据集	BP			LM			FCM-FNN		
BREAST	848.9	70.66%	○	395.8	73.32%	○	234.9	73.33%	⊕
CAR	—	—	○	1 001.2	67.28%	○	383.7	65.33%	⊕
CMC	—	—	○	727.1	63.57%	○	237.5	68.53%	⊕
IRIS	694.2	76.67%	⊕	268.4	80.00%	⊕	15.6	80.00%	⊕
YEAST	—	—	○	784.5	66.24%	○	388.2	68.62%	○
ZOO	631.1	81.26%	⊕	217.4	83.66%	⊕	26.1	82.63%	⊕

层次的决策分析、知识发现等实用技术还很不成熟,导致了“信息爆炸”但“知识贫乏”的现象,数据挖掘技术因此适时而生。这当中的分类问题,更是由于其使用的广泛性,得到了越来越多的关注。

虽然不同领域对分类问题都已作过研究,但面对如今海量的数据,各传统算法都显得“力不从心”。为此,本文通过将模糊聚类的思想引入到传统的神经网络模型中,通过对原始数据集的聚类简化,从而达到提高训练效率、提高系统鲁棒性的目的,并通过实验验证了这一想法。

但同时我们也应该看到,本文提出的基于模糊聚类的神经网络还有很多值得继续研究的问题,如聚类的个数对系统效率和分类精度的影响等。相信随着对基于模糊聚类的神经网络研究的不断深入,这些问题一定可以得到一个满意的答案,并且数据挖掘分类算法也一定会不断向前发展。

参考文献 (References)

- [1] LAST M, MAIMON O. A compact and accurate model for classification [J]. IEEE Transactions on Knowledge and

Data Engineering, 2004, 16(2): 203-215.

- [2] LIN T S, LOH W Y, SHIH Y S. A comparison of prediction accuracy, complexity and training time of thirty-three old and new classification algorithms [J]. Machine Learning, 2000, 39: 203-228.

- [3] CASTILLO O, MELIN P. Hybrid intelligent systems for time series prediction using neural networks, fuzzy logic, and fractal theory [J]. IEEE Transactions on Neural Networks, 2002, 13(6): 1395-1408.

- [4] 黎俊锋, 朱锋峰. 基于样本密度的 FCM 改进算法 [J]. 科学技术与工程, 2007, 7(4): 636-638.

LI Junfeng, ZHU Fengfeng. Improved FCM algorithm based on the density of samples [J]. Science Technology and Engineering, 2007, 7(4): 636-638.

- [5] 王刚. 基于混合智能系统的数据挖掘分类算法研究 [D]. 长沙: 中南大学, 2004.

WANG Gang. Study of classification in data mining based on HIS[D]. Changsha: Central South University, 2004.

- [6] UCI Machine Learning Repository[DB/OL]. [2007-03-10]. <http://www.ics.uci.edu/~mllearn/MLRepository.html>.

(责任编辑 王士泉)

·学术动态·

中国科协

“新观点新学说学术沙龙”欢迎申报

中国科协从 2006 年起组织举办“新观点新学说学术沙龙”系列活动。其目的是为认真贯彻全国科技大会和中国科学技术协会第七次全国代表大会精神,充分发挥学术交流作为原始创新源头之一的作用,倡导自由探究,活跃学术空气,激发科技工作者的创造力,促进创新人才成长。

学术沙龙每期邀请科学家领衔主持,由科学家定义题、定议程、定方式、定时间、定地点、定人员,力求主题新颖、人员专业、形式发散、学科交叉和成果科学,努力为萌芽时期尚未获得主流认可的学术观点、理论以及灵感提供交流平台,为科技工作者营造多学科交叉并自由探讨学术问题的良好环境。每期沙龙采用现场全程记录的方法,会后整理出版文集。活动会务费用(包括出版文集)由中国科协提供。

各学会可组织科学家申报学术沙龙方案。凡主题新颖,领衔科学家有创新思想和观点,能够邀请一批不同领域、具有共同兴趣的专家围绕主题进行有建设性争鸣讨论的方案均可提出申请。

联系人: 赵崇海

电子信箱: xsjl@cast.org.cn

电话: 010-68788562, 68580342

·学术动态·

“青年科学家论坛”欢迎申报

——每个活动资助 5 万元

中国科协青年科学家论坛创办于 1995 年,旨在推动青年科技工作者了解世界科技发展动态,拓宽视野,增长知识和才干,营造良好的学术讨论环境和学术交流氛围,搭建高层次的学术交流平台,培养优秀的科技人才和各个学科的学术和技术带头人。论坛坚持服务国家科技目标、鼓励自由探索、突出自主创新的原则,重点围绕基础学科、新兴学科、边缘学科、交叉学科等领域的前沿、热点问题,以及与国民经济建设密切相关的重大科技问题进行研讨,促进创新人才的成长和原始创新能力的提升。凡具有良好的学风和科学道德,从事科研工作并已取得较为突出的成绩,年龄在 40 岁以下,具有高级专业技术职称和博士学位的人员均可作为执行主席申报该项目。在不违反参会原则的情况下,论坛代表由执行主席确定。该论坛充分体现青年人的特点,采用丰富灵活多样的举办形式,在倡导学术民主和学术争鸣的基础上提倡举办形式的创新。青年科学家论坛项目每季度审批一次,每年拟组织 20 期以上,不定期举行,每个活动资助 5 万元。

详细信息请浏览中国科协网页(<http://www.cast.org.cn>)“学会与学术”栏目。