基于粗集神经网络的分类方法①

Rough Set-Based Classification of Neural Network

徐妙君　(浙江海洋学院 数理与信息学院 浙江 舟山 316004)

摘　要： 数据挖掘是近年来发展快速的信息处理新技术，如何有效地从高维的、超大规模数据中提取隐藏的有用信息，是该领域的研究核心。针对海量数据的挖掘分类问题，将粗集和神经网络紧密结合建立一种新的高效数据挖掘模型，即利用粗糙集理论中的知识简化方法，去掉冗余的属性特征和样本，然后，利用性能优良的模糊 kohonen 聚类神经网络进行聚类分析，最后形成分类规则。该模型充分融合了粗集强大的规则提取能力和神经网络优良的分类能力。实验证明模型具有很好的分类效率，且有较高的精确性。

关键词： 粗集神经网络 数据挖掘 聚类分析 模糊 kohonen 聚类网络 分类

1　引言

数据挖掘是从数据库中提取隐含的、尚未发现的、潜在有用的信息的过程。面对高维的、超大规模的数据库，如何建立有效的、可扩展的数据挖掘算法是数据挖掘研究的方向之一。粗集理论是 20 世纪 80 年代由 Pawlak 提出的一种处理模糊性和不确性的数学工具[1]。在处理大数据量，消除冗余信息等方面，粗集理论有着良好的效果，因此广泛应用于数据挖掘的数据预处理、数据缩减、规则生成、数据依赖关系发现等方面。但是，由于粗集理论对错误描述的确定性机制过于简单，而且在约简的过程中缺乏交互验证功能、因此当数据中存在噪声时，其结果往往不稳定，精度不高。

神经网络由于分类精度高、鲁棒性强等优点，在机器学习、模式识别等领域得到了广泛的应用。然而，神经网络面对数据挖掘中的高维和超大规模问题，其学习速度缓慢、规则生成困难等缺陷表现更为明显，原有的神经网络算法在效率和可扩展方面都会出现问题。

由于粗集理论与神经网络具有很强的优势互补性，因此两种技术的有效结合是当前的一个研究热点。目前为止，经过许多专家学者的辛勤工作，粗集与神经网络结合这方面的成果也比较多。从目前掌握的外文资料来看，从 20 世纪 90 年代末到最近几年，粗集神经网络的集成技术主要有 3 种：粗集神经网络混合系统、基于粗边界神经元的粗边界神经网络、以信息颗粒计算为基础的粗—颗粒神经网络[2]。

本文将针对粗集神经网络混合系统的发展提出一种融合粗集理论和神经网络的数据挖掘新方法，适用于大型数据库的分类规则挖掘。其主要思想是首先由粗集理论对数据库进行初步约简；然后借助于神经网络在自学习过程中完成对数据的聚类和分类挖掘，并过滤数据中的噪声数据，最后由粗集理论对约简后的数据进行规则抽取，得到最终的挖掘知识。使得网络的学习速度大大加快，分类精度显著提高。

2　粗集神经网络结构

粗集理论基于属性依赖性、约简、核、规则提取、可区分函数和可区分矩阵等概念，实现对信息系统的预处理，去除冗余属性和冗余样本，压缩信息空间维数，精简知识系统。同时粗集对数据的分析获得的决策规则是对给定信息系统的客观描述，含有一定的先验知识来指导神经网络的结构设计、结构优化及参数的初始化，从而简化并优化网络结构，加快训练速度，同时可赋予网络语义解释性。本文主要研究从大型数

① 基金项目:浙江省教育厅科研项目(20070330);浙江海洋学院校级科研项目(21065005807)
收稿时间:2008-10-05

据库和海量数据中挖掘分类规则，结合粗集和神经网络的优点，提出了一种新的数据挖掘思路。

粗集作为神经网络的前端处理器，在保证分类能力不变的基础上通过属性约简压缩信息空间维数，实现特征选择，简化训练集。由精简的训练集构建神经网络和对神经网络进行训练，可保证网络既有很好的泛化能力又有较高的预测精度。这种耦合方式下的粗集神经网络的模式分类处理如图 1 所示。

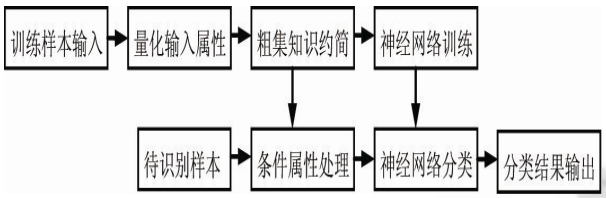

图 1　用于模式分类的粗集神经网络结构

粗集作为前端处理器外，也可作后期处理，或协同神经网络进行综合评判，或优化网络结构。

2.1 神经网络

本文采用先聚类再分类的方法。聚类分析是数据挖掘中一个很活跃的研究领域。聚类分析的算法和方法很多，一种好的聚类算法既要考虑模糊划分同一类中的紧凑程度又要考虑类与类之间的离散程度。一方面，如果仅仅考虑同一类中的紧凑程度，那么可以把每个样本单独划分为一类，这样同一类的紧凑程度最好；另一方面，如果仅仅考虑类与类之间的分离程度，那么就要把数据样本划分为类与类之间的距离之和为最大。本文采用的模糊 Kohonen 聚类网络(Fuzzy Kohonen Clustering Network，简称 FKCN)。是用隶属度来确定每个数据点属于某个聚类的程度的一种聚类算法。模糊 Kohonen 聚类网络跟一般的 Kohonen 网络的不同点在于它的学习规则是把模糊 C 均值模型(简称 FCM)和一般的 Kohonen 网络学习规则和更改策略结合起来，模糊 C 均值算法是对某个函数最小化的过程，因此模糊 Kohonen 聚类网络的学习过程也是对某一目标函数的优化过程，而不是一个启发式强制收敛过程，因此能够保证最优值。比上述一般的 Kohonen 网络和模糊 C 均值算法无论在分类精度上还是使用范围上都有一定优越性。

2.2 模糊 Kohonen 聚类网络

Tsao 将 kohonen 聚类网络(简称 KCN)与模糊 C-均值算法结合[3]，在 KCN 的学习机制中引入模糊 C-均值算法，提出了模糊 Kohonen 聚类神经网络。FKCN 为一个两层神经网络，第一层为输入层，含有 p 个神经元，p 为输入数据维数；第二层为竞争输出模糊神经元，其个数等于类数 c，当第个神经元输出为 $u_{ik} \in [0,1]$ 时，表示输入样本 $X_k, k \in \{1,2,...,n\}$ 属于第 i 类的隶属度为 u_{ik}。具体算法：

步骤 1：给定采样空间 $X = \{x_1, x_2, \cdots, x_n\}$，定义距离为 $\| \|_A$，聚类数为 $c, 2 \le c \le n$，误差阈值为 $\varepsilon > 0$。

步骤 2：初始化网络权值 $W = \{W_1, W_2, ..., W_c\}$，给定模糊参数 $m_0 > 1$，最大迭代次数 $t_{\max}$，初始迭代次数 t=0。

步骤 3：根据式(1)迭代更新模糊隶属度函数 $\{u_{ik}\}$，并根据式(2)计算学习率 $\{a_{ik}\}$。

$$u_{ik} = \frac{1}{\sum_{j=1}^{c} \left(\frac{\|X_k - W_i\|}{\|X_k - W_j\|} \right)^{\frac{2}{(m_t - 1)}}} \tag{1}$$

$$a_{ik}(t) = (u_{ik}(t))^{m_t} \tag{2}$$

其中 $m_t = m_0 - t\delta m,\ \delta m = (m_0 - 1)/t_{\max}$。

步骤 4：根据式(3)调整网络权值 $W_i(t)$。

$$W_i(t) = W_i(t-1) + \frac{\sum_{k=1}^{n} a_{ik}(X_k - W_i(t))}{\sum_{s=1}^{n} a_{is}(t)} \tag{3}$$

步骤 5：计算能量函数 $E(t) = \|W(t) - W(t-1)\|^2 = \sum_{i=1}^{c} \|W_i(t) - W_i(t-1)\|^2$，如果 $t > t_{\max}$ 或 $E(t) < \varepsilon$，停止迭代，否则进入步骤 3。

3 粗集神经网络模型的建立

通过上面的分析建立一个数据分类模型，用于大数据集的模式分类。

3.1 建模过程

第一步　数据预处理。从数据仓库中提取数据作为样本输入，考虑样本的数目不能太少，随机选择样本，这样样本更具有代表性。但是数据可能是连续的或是离散的，也可能有部分数据已经缺失。就要利用粗糙集的优点进行数据分析（Rough Sets Data Analysis, RSDA），RSDA 之前首先应将连续的数据量化处理[4]。常用的离散量化方法有“等间隔均匀量化”、“等频度量化”、“最大熵法”等。

第二步　属性的约简。属性约简是粗糙集应用于数据挖掘的核心概念之一。通过约简的计算，粗糙集可以用于特征约简或特征提取，属性关联分析。粗糙集是计算密集的，已经证明求取所有约简和最小约简

的问题都是 NP — hard 的[5]。计算属性约简类似于机器学习中的最小属性子集选择问题。高效的约简算法是粗糙集理论应用于数据挖掘与知识发现领域的基础，寻求快速的约简算法仍是主要研究课题之一。目前，已提出了许多属性约简算法。如文献[6]提出了区分矩阵算法，这种算法直观，易于理解，能够计算出核与所有约简。但是逻辑公式化简的过程计算量很大，降低了属性约简算法的效率。文献[7]提出了基于属性重要性的算法，该算法使用核作为计算约简的出发点，以属性的重要性作为启发规则，计算一个最好的或者用户指定的最小约简。文献[8]提出了的基于遗传算法的约简方法，遗传算法是一种非常有效的搜索和优化技术，有着隐含并行性、鲁棒性和全局搜索等特点，已经被广泛应用到众多领域，遗传算法用于求解决策表的相对属性约简是快速有效的。文献[9]提出的复合系统的约简，即怎样利用现有的子系统的约简求复合系统的约简。其主要思想是将布尔函数的化简问题转化成集合空间中的边界搜索问题，而在已知子系统的约简的情况下，复合系统的搜索空间将得到简化。文献[10]提出了扩展法则的约简，其中定义了一种强等价，若它们在区分函数的所有项中同时出现或不出现，则称两个属性称为局部强等价，此时它们就可以仅用一个属性代替。实验表明该算法比基本算法快数十到数百倍，因而能处理更大的数据集。文献[11]提出了动态约简，动态约简能够有效的增强约简的抗噪音能力，动态约简的计算过程较为简明，主要是对决策表进行采样，然后对采样后的决策表计算所有约简。在所有的子表中保持不变或近似保持不变的约简就是动态约简。文献[12]提出了基于数据库操作的约简。文献[13]提出了并行算法。各种属性约简既有优点，又有缺点，在具体应用中根据实际情况选取一种约简简单，效率高的算法。

第三步　不完备系统的数据处理。实际应用中，由于一些噪声因素(例如人为记录或仪器故障等) 的影响可能造成系统的数据丢失，或者因为某种原因(例如实验条件限制) 系统的一些数据根本得不到， 这些都导致了对象集合U 的描述的不完备，从而造成了不完备信息系统的出现。这些遗失的数据影响了随后的数据分析，必须采取一种可行的数据填补方法， 在保证使填补后的数据产生的分类规则具有尽可能高的支持度且在规则集中的前提下， 对数据进行填补。目前填补遗失值的方法有均值法、最大频法[14]、不完备数据分析方法(ROUSTIDA)[15]等。

第四步　根据简化的决策规则，确定分类的预期目标数。

第五步　把经过约简后的数据输入到神经网络中训练，本文采用 FKCN 进行聚类。

第六步　从数据仓库中提取测试数据进行分类测试。

第七步　对数据的分类结果进行分析，如果是协调系统，转下一步。如果是不协调系统，利用粗集知识的广义决策函数处理，形成广义决策表。根据广义决策计算每条规则的可信度。

第八步　提取最后的决策规则，数据挖掘完毕。

3.2 基于粗集神经网络的数据分类仿真研究

采用的数据来自机器学习网站 http://www.ics.uci.edu/~mlearn/databases/ 中 credit-screening,主题是关于银行用户信用卡使用的信誉问题。

该文件中的数据共有 15 个属性值，共 690 组数据。由于考虑到持卡人的机密，所有的属性名和属性值已经被处理成无意义的各种数值和符号，并且其中有一部分数据是缺失的，缺失的记录见表 3 中。此数据集被研究人员多次使用，用于机器学习。这里采用此数据集进行仿真，是因为具有代表性。属性值中既有连续数据，又有离散数据，连续数据中既有值很大的，又有很小的，特别是一部分缺失数据，正是符合我们研究的主题。

表 1　属性值信息表

A1:　a, b	A2:　continuous.
A3:　continuous.	A4:　l, t, u, y ,
A5:　g, gg, p	A6:　aa, c,cc, d,e,ff, i, j, k, m, q, r, w, x.
A7:bb, dd, ff, h, j, n, o. v, z.	A8:　continuous.
A9:　t, f.	A10: t, f.
A11:　continuous.	A12: t, f.
A13:　g, p, s.	A14: continuous.
A15:　continuous.	

第一步 首先不考虑缺失的数据，对可分类数据进行处理，形成类别数据，由粗集进行属性约简。对属性组：$A1, A4, A5, A6, A7, A9, A10, A12, A13$ 求区分矩阵。

表 2　区分矩阵

A1	A4	A5	A6	A7	A9	A10	A12	A13
0	0	0	1	0	0	0	0	0
0	0	0	0	0	0	0	1	0
0	0	0	0	0	0	1	0	0
1	0	0	0	0	1	0	0	0
0	0	0	0	1	0	0	0	0
0	1	1	0	0	1	0	0	0

注：每行的 1 表示该属性的析取，行与行之间的合取就是区分函数。得到区分函数为：

$$\Delta = A6 \times A7 \times A10 \times A12 \times (A1 + A9)(A4 + A5 + A9)$$

从中可以得到它的核属性(关键属性)为：$A6, A7, A10, A12$。

第二步 去除约简属性，得到后续处理的属性集为：$A2, A3, A6, A7, A8, A10, A12, A14, A15$

对所有关键属性的数据进行缺失补齐，共有 28 项缺失数据需要补齐。利用基于粗集的量化容差的不完备数据处理方法，得到结果见表 3，下划线的数据就是补齐的数据。

表 3　缺失数据的补漏

记录号	A2	A3	A6	A7	A8	A10	A12	A14	A15
85	27.3	3.5	4	8	3	1	2	300	0
88	30.3	0.38	4	8	0.88	1	2	928	0
.....									
603	42.3	1.75	6	3	0	1	2	150	1
624	25.6	0	6	3	0	1	1	0	0
73	34.8	4	4	1	12.5	1	2	292	0
.....									
280	24.6	13.5	6	3	0	1	1	80	0
408	40.3	8.13	9	8	0.17	2	1	160	18
628	22	7.835	7	1	0.165	1	2	480	0

第三步 对所有 690 项先进行归一化处理，进行 FKCN 聚类分析，采用 600 条数据训练，90 项(其中包括缺失的 28 条记录)作为测试数据。根据简化的决策表把用户的信誉聚类为二个类别，网络稳定最后可以得到输出权值矩阵为：

W=[30.243 3.036 7.104 6.273 1.857 1.219 1.590 284.764 9.175 29.203 4.982 6.634 6.215 1.833 1.336 1.412 100.676 8.318]

把测试数据输入训练好的网络，90 项数据其中 62 条数据，没有误分，28 条缺失数据，误分 2 条。

表 4　属性约简后的决策表

序号	A2	A3	A6	A7	A8	A10	A12	A14	A15	决策属性
1	30.83	0	13	8	1.25	2	1	202	0	1
2	58.67	4.46	11	4	3.04	2	1	43	560	1
3	24.5	0.5	11	4	1.5	1	1	280	824	1
.....										
648	24.08	9	1	8	0.25	1	2	0	0	0
687	22.67	0.75	2	8	2	2	2	200	394	0
688	25.25	13.5	6	3	2	2	2	200	1	0
689	17.92	0.205	1	8	0.04	1	1	280	750	0
690	35	3.375	2	4	8.29	1	2	0	0	0

第四步 把数据的分类结果交给银行分析师处理，并把模糊隶属度函数交给银行分析师，由于篇幅的关系，取其中几个模糊隶属度进行观测。

600 条记录被分成 2 类，一类 218 条，一类 382 条。

Columns 586 through 594

U=[0.0246 0.9741 0.0051 0.4789 0.0091 0.0077 0.0178 0.9800 0.0088 0.9754 0.0259 0.9949 0.5211 0.9909 0.9923 0.9822 0.0200 0.9912]

从模糊隶属度的值可以看到，还有一部分虽然把某个数据项分为某一个类，但实际它属于这一类的值可能只是比其它类高一点(0.5211 与 0.4789)，所以这些数据都属于边界点。在实际情况下，可能这个人的信誉度不是按照分类的结果，需要在决策的时候进行进一步的处理，这也是基于粗集的好处，可以采用不协调系统处理方法进行处理。特别地，在对数据进行聚类地时候，能尽量多几个聚类中心，让数据的确定性更高，就象好跟优秀就都属于好的一类。

3.3 算法分析

采用本文的方法一方面大大简化了数据集，数据集中的 10 个属性值最后约简只得到 4 个核属性，简化了神经网络的训练集，第二方面又对缺失数据按粗集理论的方法进行补齐。如果单用神经网络就无法处理，这部分数据遗漏的后果可能影响以后的决策。图 2 显示了去除缺失数据后的样本，直接采用模糊自组织特征神经网络聚类的效果和经过本文的方法在误差

变化上的比较(取 m=2,tmax=1000)。从速度上看，本文的方法在 8 次迭代后就收敛到精度(1e-5)，而原始数据经过 11 次才能收敛。可以看出本文的方法在效率上有所提高，经过测试，数据量越多效果越明显。在准确率上，从表 5 中也可以看出粗集的应用提高了分类精度。

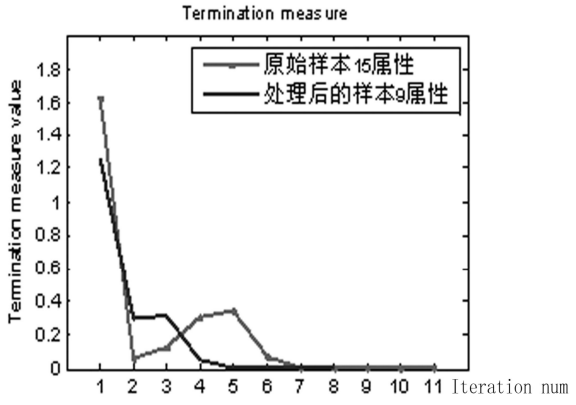

图 2　误差比较

表 5　分类结果的比较

	FKCN	本文的方法
平均误分数	23.8	20.6
平均误分率	3.97	3.43

4　结论

针对粗集理论和神经网络的各自优点，本文提出一种数据挖掘新方法，即利用粗糙集理论中的知识简化方法，去掉冗余的属性特征和样本，然后，利用性能优良的模糊 kohonen 聚类网络进行聚类分析，最后形成分类规则。实验表明，该方法快速有效，得到的分类规则简单准确，可理解程度高。

参考文献

1 Pawlak Z. Rough Sets: Theoretical Aspects of Reasoning about Data. Kluwer Academic Publishers, Boston, 1991.

2 张东波,王耀南,易灵芝.粗集神经网络及其在智能信息处理领域的应用.控制与决策, 2005,20(5):121-126.

3 Tsao EC, Bezdek JC, Pal NR. Fuzzy Kohonen Clustering Network. Pattern Recognition, 1994,27(5): 757－764.

4 苗夺谦.粗糙集理论中连续属性的离散化方法.自动化学报, 2001,27(3):296－302.

5 康立军,吴丽丽.基于可变精度的不完备信息系统粗集扩展模型及属性约简.甘肃农业大学学报, 2008, 43(2):152－155.

6 Skowron A, Rayszer C. The discernibility matrices and functions in information systems. Intelligent Decision Support: Handbook of Applications and Advances of the Rough Sets Theory, 1992:331－362.

7 Hu X H, Cercone N. Learning in relational database: a rough sets approach. Computational Intelligence, 1995, 11:323－358.

8 代建华,李元香.粗集中属性约简的一种启发式遗传算法.西安交通大学学报, 2002,36(12):1286－1290.

9 Kryszkuewicz M, Rybinski H. Finding reducts in composed information systems.Proceedings of RSKD'93. London: Springer-Verlag,1993,261－273.

10 Starzyk J, Nelson D E, Sturtz K. A mathematical foundation forimproved reduct generation in information systems. Knowledge and Information Systems, 2000:131－146.

11 Bazan J, Skowron A, Synak P. Dynamic reducts as a tool for extracting laws from decisions tables. Methodologies for Intelligent Systems.1994.Berlin: Springer-Verlag:346－355.

12 Hu XH, Lin TY, Han JC. A new rough sets model based on database systems. Wang G, et al.(eds): RSFDGrC 2003, LNAI 2639, 2003:114－121.

13 Robert S. Parallel Computation of Reducts. Polkowski L and Skowron A (eds.): RSCTC'98, 1998:450－458.

14 Kryszk iew icz M. Rough set approach to incomplete information systems. Information Sciences, 1998,112 (1- 4):39－49.

15 孟军,刘永超,莫海波.基于粗糙集理论的不完备数据填补方法.计算机工程与应用,2008,44(6):175－177.