一个高效的 KNN 分类算法

张著英　黄玉龙　王翰虎
(贵州大学计算机科学技术学院　贵阳 550025)

摘　要　KNN 算法是数据挖掘技术中比较常用的分类算法，由于其实现的简单性，在很多领域得到了广泛的应用。但是，当样本容量较大以及特征属性较多时，KNN 算法分类的效率就将大大降低。本文将粗糙集理论应用到 KNN 算法中，实现属性约简，提出了一种新的 KNN 分类方法，解决了 KNN 算法分类效率低的缺点，从而可使 KNN 算法能够得到更广泛的应用。

关键词　数据挖掘，KNN 分类，粗糙集，属性约简

A New KNN Classification Approach

ZHANG Zhu-Ying　HUANG Yu-Long　WANG Han-Hu
(College of Computer Science Technology, Guizhou University, Guiyang 550025)

Abstract　KNN algorithm has been widely used in many data mining areas due to its simplicity. When the samples become more and more large and characteristic attributes become more and more numerous, KNN algorithm becomes much lower. A new KNN algorithm based rough set theory is proposed in the paper, in order to improve the effectiveness.

Keywords　Data mining, KNN classification, Rough set, Attributes induction

1　引言

分类是数据挖掘领域中一种重要的技术，它是从一组已知的训练样本中发现分类模型，并且使用这个分类模型来预测待分类样本。目前比较常用的分类方法有：决策树、神经网络、KNN、SVM 和贝叶斯方法等，其中 KNN 算法以其实现的简单性及较高的分类准确性在中文文本自动分类等领域得到了广泛的应用。但是，该方法对属性较多的训练样本进行分类时，由于计算量大而使其效率大大降低，效果不很理想。为了在更广泛的应用领域使用 KNN 算法，如何提高该算法在处理属性较多的训练样本时的效率，受到了数据挖掘工作者极大的关注。

经过研究发现，如果在使用 KNN 算法之前对样本的属性进行约简，删除那些对分类结果影响较小的属性，则可以用 KNN 算法快速地得出待分类样本的类别，从而可以得到更好的效果。已经提出了一些属性约简的方法，但效果不十分理想，或者约简的属性不多，或者约简属性后影响分类的准确性。粗糙集理论在用于决策表的属性约简时，可在保持决策表中决策能力不变的前提下，删除其中不相关的冗余属性。我们采用粗糙集理论在 KNN 算法中进行属性约简，取得了较好的结果。

本文第 2 节描述基于粗糙集理论的决策信息表属性约简算法，接着给出改进的 KNN 算法和实验比较结果，最后是结束语。

2　属性约简

2.1　基本概念

本节描述算法需要的一些基本概念。

定义 1　信息系统是一个四元组 $S=\langle U, A, V, f\rangle$，其中 U 是信息系统样本的有限集合，被称为论域；A 为样本属性的集合；V 是属性 A 的值域集合；f 为信息函数，它是 $U\times A\to V$ 上的一个映射。

定义 2　假如 $A=C\cup D$，其中 C 为条件属性集合，D 为决策属性集合且 $C\cap D=\phi$，则我们把这样的信息系统称为决策系统，通常简记为 $S=\langle U, C\cup\{d\}\rangle$，$d$ 为单个的决策属性。

信息系统通常是以关系表的形式来表示的。关系表的行表示要研究的样本，列表示要研究样本的属性，样本的信息通过指定样本各属性的值来表示，这样我们也可以称此信息系统为决策表。

定义 3　假设 $S=\langle U, A\rangle$ 是一个信息系统，A 中所有的等价关系的交集是一个等价关系，称它为 A 上的不可区分关系，记为 $ind(A)$，其中 $(x, y)\in ind(A)$。

当且仅当 $\forall a\in A, f(x, a)=f(y, a)$ 对于 $a\in A$，$ind(A)=ind(A-\{a\})$，则 a 是 A 中不必要的。否则 a 是 A 中必要的。

定义 4　A 中所有必要的属性组成的集合，被称为 A 的核，记作 $CORE(A)$。如果每一个 $a\in A$ 都是必要的，则称 A 为独立的，否则称 A 为依赖的。设 $Q\subseteq A$，如果 Q 是独立的，且 $ind(Q)=ind(A)$，则称 Q 为 A 或信息系统 S 的一个约简。

定义 5　令 $S=\langle U, C\cup d, V, f\rangle$ 是一个决策系统，其中 $U=\{x_1, x_2, x_3, \cdots, x_n\}$ 是决策系统的论域，C 为条件属性集合，d 为决策属性，$f(x, a)$ 是记录 x 在属性 a 上的值，则可辨识矩阵

$$M_{ij}=\begin{cases}\{a \mid 当\ a\in C\ 且 f(x_i, a)\neq f(x_j, a)\} & d(x_i)\neq d(x_j)\\ \phi & 其他\end{cases}$$

当两个样本的决策属性的取值相同时，它们所对应的可辨识矩阵元素的取值为 ϕ；当两个样本的决策属性不同且可以通过一些条件属性的取值不同加以区分时，它们所对应的

可辨识矩阵元素的取值为这两个样本取值不同的条件属性集合，也就是可以区分这两个样本的条件属性集合。在可辨识矩阵 M_{ij} 中，如果存在一个元素，它的值是单个属性元素的集合，则表明该属性是区分这个元素所对应的两个样本所必须的唯一的属性，这样的属性我们称为矩阵 M_{ij} 的核属性。矩阵 M_{ij} 中包含的核属性组成的集合，我们称之为决策信息系统的核属性集。从定义 5 我们可以发现，可辨识矩阵是一个对称矩阵，因此我们只需要考虑矩阵的下三角就可以了。

2.2 决策信息表属性约简算法

决策信息表属性约简算法的主要思想为：计算决策信息表的可辨识矩阵 M_{ij}；将 M_{ij} 中的核属性取出，并将 M_{ij} 中含核属性的元素置为 ϕ；对于 M_{ij} 中所有取值不为空的元素，分别构建相应的析取逻辑表达式 p_{ij}。对表达式进行合取和析取变换后得到表达式 P'，最后将可辨识别矩阵 M_{ij} 中的核属性合取后加入析取表达式 P' 中的每一个合取项中，得出决策信息表 S 的约简结果。在实际应用中，决策表 S 的约简可能会得到多个结果，我们可以选择一种包含对决策表分类有重要影响的非核属性的约简结果。

算法 1 决策信息表属性约简算法

输入：训练样本集$(X_1, X_2, \cdots, X_n)$，其中 $a_1, a_2, \cdots, a_n$ 为训练样本集的条件属性，d 为训练样本集的决策属性。

输出：训练样本集$(X_1, X_2, \cdots, X_n)$的一个属性约简，其中 $a_1, a_2, \cdots, a_m$($m \leqslant n$)为约简后训练样本集的条件属性，d 为约简后的决策属性。

步骤 1 for $i=1$ to n do
　　按定义 5 分别求出 M_{ij} 中的元素 C_{ij}；
　　end for;

步骤 2 for $i=1$ to n do
　　for $j=1$ to i do
　　// M_{ij} 为对称矩阵，在此只考虑下矩阵的下三角；
　　if $|C_{ij}|=1$ then 将 C_{ij} 加入核属性集合 H 中；
　　end for;
　　end for;

步骤 3 将 M_{ij} 中含核属性的元素 C_{ij} 置为 ϕ；
　　for $i=1$ to n do
　　for $j=1$ to i do
　　if $|C_{ij}|\neq 0$ the 将 C_{ij} 中含的属性进行析取，得到析取表达式 P_{ij}；
　　end for;
　　end for;

步骤 4 for $i=1$ to $|p_{ij}|$ do
　　$P=P\wedge P_{ij}$
　　end for;

步骤 5 将表达式 P 转换为析取形式 P'；

步骤 6 将核属性集合 H 中的元素进行合取，得到合取表达式 M；

步骤 7 将 M 与表达式 P' 中的每个合取项进行合取；合取得到的结果即为训练样本集$(X_1, X_2, \cdots, X_n)$的属性约简结果。

2.3 属性约简例

表 1 是一个已知的决策表。

表 1 决策表

u	a	b	c	d
1	2	2	0	1
2	1	2	0	0
3	1	2	0	1
4	0	0	0	0
5	1	0	1	0
6	2	0	1	1

按定义 5，我们可以得出表 1 的可辨识矩阵 M_{ij} 如下：

$$\begin{bmatrix} a & & & & \\ a & & & & \\ a,b & a,b & a,b & & \\ a,b,c & b,c & b,c & & \\ & a,b,c & a,b,c & a,c & a \end{bmatrix}$$

通过核属性的定义，我们可以得到 M_{ij} 矩阵的核属性为 $\{a\}$；在 M_{ij} 中只有一种属性组合 bc 不包含核属性，使用约简算法构造析取表达式 $P=b\vee c$；由于 P 只有一种组合，所以不需要再进一步进行形式化变换。这样，原来决策表的条件属性可以简化为$\{a, b\}$或者是$\{a, c\}$。约简后的决策表如表 2 和表 3 所示。

表 2 约简后的决策表 2

u	a	b	d
1	2	2	1
2	1	2	0
3	1	2	1
4	0	0	0
5	1	0	0
6	2	0	0

表 3 约简后的决策表 3

u	a	c	d
1	2	0	1
2	1	0	0
3	1	0	1
4	0	0	0
5	1	1	0
6	2	1	0

3 改进的 KNN 分类算法

3.1 KNN 分类算法

KNN 分类算法的主要思想是：先计算待分类样本与已知类别的训练样本之间的距离或相似度，找到距离或相似度与待分类样本数据最近的 K 个邻居；再根据这些邻居所属的类别来判断待分类样本数据的类别。如果待分类样本数据的 K 个邻居都属于一个类别，那么待分类样本也属于这个类别。否则，对每一个候选类别进行评分，按照某种规则来确定待分类样本数据的类别。

我们采用欧氏距离来确定样本的相似性。欧氏距离的计算公式为

$$d(x, y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2}$$

其中 $X=(x_1, x_2, \cdots, x_n)$和 $Y=(y_1, y_2, \cdots, y_n)$代表两个样本数据，$n$ 为样本特征属性的个数。

3.2 改进的 KNN 分类算法

我们首先用决策信息表属性约简算法 1 对训练样本数据集进行属性约简。通过约简，可以删除冗余的属性，从而可以提高 KNN 算法的分类效率。然后，将待分类样本用 KNN 算法进行分类，算法 2 是该算法的详细描述。

算法 2 改进的 KNN 分类算法

输入：训练数据集 $D=\{(X_i, Y_i), 1\leqslant i\leqslant n\}$，其中 X_i 是第 i 个样本的条件属性，Y_i 是第 i 个样本的决策属性，即第 i 个样本的类别属性，新样本为 X，距离函数为 d。

输出：待分类样本 X 的决策属性(类别属性)Y。

步骤 1 for $i=1$ to n do
　　用算法 1 约简样本的属性；
　　end for

步骤 2 在约简后的训练样本中删除相同的样本；

步骤 3 按训练样本的约简结果对待分类样本数据 X 的属性进行约简；
　　for $i=1$ to m do
　　// m 是对样本数据进行属性约简后的特征// 属性的个数 $m\leqslant n$
　　计算 X 和 X_i 之间的距离 $d(X_i, X)$；
　　end for

步骤 4 对上面计算得到的距离进行排序，得到 $d(X, X_{i1})\leqslant d(X, X_{i2})\leqslant\cdots\leqslant d(X, X_{in})$；

步骤 5 选择前 K 个样本：$S=\{(X_{i1}, Y_{i1}), (X_{i2}, Y_{i2}), \cdots, (X_{ik}, Y_{ik})\}$；

步骤 6 统计 S 中每一个训练所属类别出现的次数，从而将出现次数最多的类别确定为待分类样本数据 X 的类别 Y。

3.3 实验比较

我们用某地区 1985 年到 1998 年以来有关林业方面的统计数据作为数据集来验证算法 2 的有效性。为了简单起见，我们只选择了其中的部分属性：人数、平均工资、投资额、原木产量、等外材产量、小规格树产量和总销量。{人数，平均工资，投资额，原木产量，等外材产量，小规格树产量}为条件属性集合，总销量为决策属性。我们还对表中的数据进行了离散化处理：采用相对比较的方法，将数据转化为前一年的增幅表，然后通过离散编码的方式，再将其转化为 Boolean 表。相对于前一年，数据增加的记为 1，数据减少的记为 0。这样，我们可以较方便地用 KNN 算法来分析总销量的变化与其他因素之间的关系。同时，为了提高分类结果的可靠性，在这里我们采用交叉验证的方法对数据集进行分类。

为便于理解，在这里我们只给出几个简单的实验结果。按传统的 KNN 算法进行分类的一些结果如下：当 K 值为 3、训练数据为 9 个(1985 年到 1993 年数据)、测试数据为 5 个(1994 年到 1998 年数据)时，我们得到的分类准确度为 60%。当 K 值为 3、训练数据为 9 个(1990 年到 1998 年数据)、测试数据为 5 个(1985 年到 1989 年数据)时，我们得到的分类准确度为 40%。

采用算法 2 对与上面相同的数据集进行分类的结果如下：先采用粗糙集理论对布尔表进行属性约简，得到的属性约简结果为：{人数、投资额、原木产量、等外材产量}。删除不相容以及冗余的属性值{平均工资，小规格树产量}后，再用 KNN 算法对其进行分类。当 K 的值为 3、训练数据为 9 个(1985 年到 1993 年数据)、测试数据为 5 个(1994 年到 1998 年数据)时，我们得到的分类准确度为 80%。当 K 值为 3、训练数据为 9 个(1990 年到 1998 年数据)、测试数据为 5 个(1985 年到 1989 年数据)时，我们得到的分类准确度为 60%。

通过上述实验比较，我们可以看出，当样本数据的特征属性较多以及样本的容量较大的时，用算法 2 得到的结果比用传统 KNN 分类算法要好。

结束语 KNN 算法是目前数据挖掘领域一种比较常见的分类算法，由于其实现的简单性，在许多领域有着广泛的应用。由于 KNN 算法不需要构建分类模型，所有的有关分类的计算都是在对新样本数据分类的时候进行的，因此当样本数据的特征属性的数量较多、样本的容量较大时，分类的时间代价很大，分类的效果不是很好，这会对实际应用产生很大的影响。本文提出的算法 2 是在对新的样本数据进行分类之前，先对它们用算法 1 进行属性约简，删除那些对样本的决策影响很小或者是根本没有影响的冗余属性，从而使 KNN 分类能够比较顺利地进行，提高了分类的效率，扩大了 KNN 算法的应用范围，同时保证了分类的准确性。但是，在现实中，样本数据集中可能会包括一些噪声样本数据，同时也可能会包括一些属性缺失的样本数据，这将会对 KNN 分类造成很大的影响。如何用粗糙集理论对这些噪声样本数据以及属性缺失的样本数据进行处理，从而可以使 KNN 分类能够顺利地进行，将是我们下一步要研究的问题。

参 考 文 献

1 Pawlak Z D. Rough set theory and its application to data analysis [J]. Cybernetics and Systems, 1998,29(9):611～6685

2 Pawlak Z D. Rough set theory and its applications[J]. Journal of Telecommunications and Information Technology, 2002(3)

3 陈安，陈宁，周龙骧，等. 数据挖掘技术及应用. 北京：科学出版社，2006

4 胡学刚，郭光亚. 一种基于粗糙集的朴素贝叶斯分类算法. 合肥工业大学学报，2006(2)

5 张冬玲. 基于粗糙集理论的属性约简算法的实现. 计算机应用，2006(2)

(上接第 133 页)

体网络，并采用非线性集成所有个体网络。此外，本文选择均方差(RMSE)作为比较标准，具体比较结果如表 1 所示。

表 1 预测结果比较

集成方法	RMSE	名次	2005
NNEBag	0.0087	5	0.11180
SNNEBag	0.0071	3	0.10947
NNEIPCABag	0.0081	4	0.11235
SNNEIPCABag	0.0053	2	0.11249
NSNNEIPCABag	0.0052	1	0.11245

从试验结果可以发现：(1)尽管只有 17 个训练样本，所有集成方法的测试结果仍都非常接近于实际值；(2)集成预测效果明显好于任何一个个体网络；(3)总体上，本文提出的 NSNNEIPCABag 方法优于其它所有集成方法；(4)江门市 2005 年 GDP 的增长率稍高于 11%；(5) NSNNEIPCABag 方法适用于经济预测。

结束语 本文试图提供一种新的基于特征提取的选择性神经网络集成方法——NSNNEIPCABag，该方法综合集成了 Bagging 算法、IPCA 特征提取方法、选择性集成技术以及非线性集成技术，充分发挥它们的协同优势，在一定程度上解决了目前神经网络集成研究中存在的主要问题。虽然仅用了 17 个学习样本，本文提出的 NSNNEIPCABag 方法的泛化能力总体上比其它集成方法更好。本文的研究为小样本经济预测提供了一种新的有效途径。

参 考 文 献

1 Hansen L K, Salamon P. Neural Network Ensembles [J]. IEEE Trans Pattern Analysis and Machine Intelligence, 1990, 12(10): 993～1001

2 Krogh A, Vedelsby J. Neural Network Ensembles:Cross Validation, and Active Learning [J]. Advances in Neural Information Processing System, 1995(7):231～238

3 凌锦江，陈兆乾，周至华. 基于特征选择的神经网络集成方法[J]. 复旦学报(自然科学版)[J], 2004, 43(5):685～688

4 Zhou Z H, Wu J X, Tang W. Ensemble Neural Networks:Many Could Be Batter Than All [J]. Artificial Intelligence, 2002, 137 (1-2):239～263

5 Yu L A, Wang S Y, Lai K K. A Novel Nonlinear Ensemble Forecasting Model Incorporating GLAR and ANN for Foreign Exchange Rates [J]. Computer & Operation Research, 2005 (32):2523～2541

6 Yu L A, Wang S Y, Lai K K, et al. A Bias-variance-complexity Trade-off Framework for Complex System Modeling [J]. Lecture Notes in Computer Science, 2006, 3980:518～527

7 Lin J, Zhu B Z. Improved Principal Component Analysis and Neural Network Ensemble Based Economic Forecasting [J]. Lecture Notes in Computer Science, 2006, 4113:135～145

8 林健，彭敏晶. 基于神经网络集成的 GDP 预测模型[J]. 管理学报, 2005, 2(4):434～436

9 Lin J, Zhu B Z. Neural Network Ensemble Based on Forecasting Effective Measure and Its Application [J]. Journal of Computational Information Systems, 2005, 1 (4):781～787

10 程其云，王有元，陈伟根. 基于改进主成分分析的短期负荷预测方法[J]. 电网技术, 2005, 29(3):64～67