文章编号:1001 - 9081(2009)09 - 2499 - 03

必要规则对分类影响的研究

李英杰[1,2],尹怡欣[1]
(1. 北京科技大学 信息工程学院,北京 100083; 2. 浙江林学院 信息工程学院,浙江 临安 311300)
(comliy@163.com)

摘 要: 基于规则分类方法的主要计算依据是形如"$A \rightarrow C$"的规则(称为充分规则)及其置信度。其中:"A"代表数据集中决策属性取值的集合,"C"代表某个类标号。那么,形如"$C \rightarrow A$"的规则(称为必要规则)是否可以在分类算法中起到积极的作用呢?依据规则分类方法原理设计了简单的实验,实验只考虑单个决策属性的不同取值与类之间的关联。根据实验目标,分类测试采用了两种方法:方法 1 只考虑充分置信的影响;方法 2 考虑充分置信和必要置信的影响。通过在几个典型的分类集上测试,结果表明:在分类计算时适当利用必要规则置信度可以提高分类精度。

关键词: 分类;置信度;充分规则;必要规则
中图分类号: TP312 **文献标志码:** A

Research of necessary rules' influence on classifying

LI Ying-jie[1,2], YIN Yi-xin[1]
(1. *School of Information Engineering, University of Science and Technology Beijing, Beijing* 100083, *China*;
2. *School of Information Engineering, Zhejiang Forestry University, Linan Zhejiang* 311300, *China*)

Abstract: The computing gist of algorithms based on rules involves the rules like "$A \rightarrow C$" and their confidences. Here, "A" represents the set of decision attributes and their values, and "C" represents a kind of class label. Can the rules like "$C \rightarrow A$" act positively in classifying algorithms? A simple experiment was designed, which considered the associations between single attribute values and class label. Two testing methods were made according to the experiment goals. By the first method, confidences of "$A \rightarrow C$" were used. By the second method, the confidences of both "$A \rightarrow C$" and "$C \rightarrow A$" were used. The experiments were made on several typical classifying data sets. The results show the higher classifying precision by using the double confidences.

Key words: classifying; confidence; sufficient rule; necessary rule

0 引言

分类是数据挖掘的一项重要任务,其目标是从已知类标号的训练集中学习规则,并用该规则对类标号未知的记录进行分类。针对分类问题,人们开发了很多算法,较经典的有:神经网络方法、支持向量机方法、关联规则分类算法、K 近邻分类算法、决策树分类算法和贝叶斯分类算法等。

设数据集 D 由记录组成,记录数为 $|D|$,属性有 $\{A_1, A_2, \cdots, A_n, C\}$,其中,$\{A_1, A_2, \cdots, A_n\}$ 是决策属性,C 是类标号。仔细研究各算法就会发现,决策树分类算法、关联规则分类算法、贝叶斯分类算法都是基于规则"$A \rightarrow C$"和其统计特性的,此处,A 表示 $\{A_1, A_2, \cdots, A_n\}$ 全部或部分属性的一些取值组合,C 表示某个类标号。

C4.5 是决策树分类算法的代表[1]。该方法首先生成一棵未经剪枝的决策树,将决策树转换为规则集合,此集合中规则形如"$A \rightarrow C$",但其中存在大量的冗余和错误。C4.5 按悲观错误率降低的原则逐渐剪除规则中的项,使规则简化。简化后的规则按类进行分组,对每组规则按照 MDL 原则选择编码最短的规则子集,形成最终的分类规则集。

CBA 是典型的关联规则分类算法[2],算法基于关联规则计算的两个指标:属性集的支持度 *Support* 和规则的置信度 *Confidence*,由式(1)和(2)来计算。首先,根据指定的支持度阈值和置信度阈值,在训练集中找出所有形如"$A \rightarrow C$"的关联规则,这类规则在文献[2]中称为"类关联规则"(Class Association Rules, CARs),其特点是后件只包含类标号,这样产生的 *CARs* 作为初始的规则集;将 *CARs* 按置信度、支持度和产生顺序排序,之后按悲观出错率和数据覆盖率进行剪枝,得到最终的规则集。

$$Support(X) = \frac{D \text{ 中包含 } X \text{ 的记录数}}{|D|} \tag{1}$$

$$Confidence(A \rightarrow C) = \frac{D \text{ 中同时包含 } A \text{ 和 } C \text{ 的记录数}}{D \text{ 中包含 } A \text{ 的记录数}} \tag{2}$$

朴素贝叶斯分类器是基于贝叶斯概率理论的算法[3]。算法假定在给定类变量的条件下各个属性变量之间条件独立。首先从训练集中计算出每个属性取不同值时各类记录出现的概率,以此作为先验概率。设类标号 $c_j \in C$,分类器利用式(3)来判断无类标号记录 $(a_1, a_2, \cdots, a_n)$ 属于类 c_j 的后验概率,并且将此记录判定为属于后验概率最大的类。贝叶斯分类器是一种简单高效的分类方法,但其条件独立性和属性权重相同的假设并不符合客观实际,后续的文献对其进行了多方改进,而其思路仍是依据规则"$A \rightarrow C$"的统计数据的[4]。

$$P(c_j \mid a_1, a_2, \cdots, a_n) = \frac{P(a_1, a_2, \cdots, a_n \mid c_j) P(c_j)}{P(a_1, a_2, \cdots, a_n)} \tag{3}$$

在一阶逻辑中如果有"$A \rightarrow C$",则称 A 蕴含 C,或称 A 为 C

收稿日期:2009 - 03 - 20;**修回日期:**2009 - 05 - 18。

作者简介:李英杰(1968 -),女,浙江临安人,副教授,博士研究生,主要研究方向:数据挖掘、知识工程、计算机视觉; 尹怡欣(1957 -),男,北京人,教授,博士生导师,博士,主要研究方向:人工智能。

的充分条件。对应地，如果有式“$C \rightarrow A$”，则称A是C的必要条件。由于现实事件很难用全部和精确的数据来描述，并且规则“$A \rightarrow C$”在D中的置信度一般不会达到1，所以在数据挖掘的文献中不称“A为C的充分条件”。但是认为A的出现增加了结果为C的概率或可能性。可以看出，C4.5算法、CBA算法和朴素贝叶斯分类器都是基于此原理的。从另一个角度讲，如果D中规则“$C \rightarrow A$”的置信度达到一定的阈值，则可以认为A不出现减少了结果为C的概率或可能性。在分类算法中如果同时考虑“$C \rightarrow A$”的作用，会提高分类精度。事实是不是这样的呢?本文将通过实验来说明。

1 问题与方法描述

1.1 问题描述

分类问题描述为：设D_1和D_2是针对相同领域的两个数据集，是二维数据表形式，有相同的属性，属性结构如引言所述，分为决策属性和类标号。D_2中的类标号的集合包含于D_1中类标号的集合。以D_1为训练集获取分类规则，用来对D_2的记录进行分类。

本文的实验目标是测试两种方法的分类效果。

1）方法1，只考虑规则“$A \rightarrow C$”的影响。

2）方法2，同时考虑“$A \rightarrow C$”和“$C \rightarrow A$”影响。

分类效果有很多评价指标，本文简单地采用分类精度来评价，其计算式如下：

$$\text{分类精度} = \frac{\text{被正确分类的记录数}}{\text{待分类的记录总数}} \tag{4}$$

基于规则的形式，本文中将形如“$A \rightarrow C$”的规则称为充分规则，形如“$C \rightarrow A$”的规则称为必要规则。为了叙述方便，将由式(2)定义的“$A \rightarrow C$”的置信度*Confidence*称为充分置信度，将“$C \rightarrow A$”的置信度称为必要置信度，记为*N_Confidence*，并由式(5)来计算：

$$N\_Confidence = \frac{D\text{中同时包含}A\text{和}C\text{的记录数}}{D\text{中包含}C\text{的记录数}} \tag{5}$$

1.2 实验方法

本文实验目标是测试必要规则对分类的影响，将采用最简单的方法。训练过程只采集各属性不同取值与各类之间规则的置信度，不采集组合属性与类之间规则的置信度。实验有两个过程。1）训练过程。从D_1中生成分类规则集R。2）测试过程。以一定的计算方法，用R来对D_2中记录进行分类。

训练过程首先将训练集中的属性进行分类。对于连续属性，将简单地进行等间隔离散，属同一间隔的数值被划分为一类；对于字符属性或离散属性，取相同值的属性被划分为一类。为了区分记录的类别，下文中属性的类将称为簇，记录的类别将称为类。接下来计算每个簇与类之间的相互支持的置信度，将符合阈值的规则作为分类规则。训练过程需要设定2个参数：充分置信度阈值和必要置信度阈值，另外要设置一整数K，作为离散连续属性时的区间数。具体训练步骤如下。

1）访问一次D_1，获取记录类集合和每类记录数；获取每个连续属性的最大值和最小值，由K计算出每个连续属性的离散区间大小；获取每个字符属性的所有不同取值。

2）再访问一次D_1，统计每簇与每类的关联关系数据，得到一中间集合。

3）依据中间集合中数据计算各簇与各类间的双向置信度，将满足阈值（≥）的规则记入集合R，R是训练结果，将在测试时作为分类计算的依据。

下面以一小训练集的训练过程来说明其中的细节。训练集D有$A1$和$A2$两决策属性和一个类别属性C，其中$A1$是连续的数值型，$A2$是离散的字符型。如表1所示。

表1 示例训练集D

序号	$A1$	$A2$	C
1	1855	优秀	$C1$
2	2016	良好	$C1$
3	1600	优秀	$C1$
4	1722	良好	$C2$
5	2588	良好	$C2$
6	2290	良好	$C1$

设充分置信度阈值为70%，必要置信度阈值为90%，$K = 3$。

第1次访问D得到类和簇的集合或划分，结果有：

1）类别C的集合为{0:$C1$,1:$C2$}，其中$C1$记录数为4，$C2$记录数为2；

2）$A1$的最大值2588，最小值1855，离散间隔 = (2588 − 1855)/3 = 244，则$A1$的离散区间为{0：[1855,2099)，1：[2099,2343)，2：[2343,2588]}；

3）$A2$取值集合为{0：优秀，1：良好}。

第2次访问D得到类和簇关联关系的统计数据，结果以属性簇为索引，如表2所示。其中的第一行表示：D中满足属性$A1$的0簇条件的记录有4条，其中属于$C1$的记录有3条，属于$C2$的记录有1条，后面的类似。

表2 第2次访问D后得到簇与类关联关系的统计数据

序号	属性标识	簇标识	簇中记录数	$C1$记录数	$C2$记录数
1	$A1$	0	4	3	1
2	$A1$	1	1	1	0
3	$A1$	2	1	0	1
4	$A2$	0	2	2	0
5	$A2$	1	4	2	2

依据表2数据计算双向置信度，得到的分类规则按类索引，结果如表3所示，表3就是训练得到的规则集R。以表2的第一行为例，属性$A1$的0簇共有4条记录，其中有3条记录属于$C1$，而D中属于$C1$的记录有4条，那么“$A1:0 \rightarrow C1$”的充分置信度为3/4 = 75%，大于阈值70%；必要置信度3/4 = 75%，小于阈值90%，记为0。

表3 D训练得到的分类规则表

序号	类标识	属性标识	簇标识	充分置信度	必要置信度
1	0($C1$)	$A1$	0	3/4 = 0.75	0
2	0	$A1$	1	1/1 = 1	0
3	0	$A2$	0	2/2 = 1	0
4	1($C2$)	$A1$	2	1/1 = 1	0
5	1	$A2$	1	0	2/2 = 1

测试时，分两种方法。

1）方法1只考虑充分置信度。从D_2中取出一条记录t，离散化其属性。分别以t的属性匹配R中各类的规则。t属于c_j的得分按式(6)来计算，即：匹配上的规则，计算其充分置信度平方和。将t判定为属于得分最高的类。

$$SCORE1(c_j,t) = \sum_{\text{匹配}} Confidence^2 \tag{6}$$

2）方法2考虑充分置信度和必要置信度。从D_2中取出一条记录t，离散化其属性。分别以t的属性匹配R中各类的规则。t属于c_j的得分按式(7)来计算，即：匹配上的规则，计算其充分置信度平方和；匹配不上的规则，计算其必要置信度的

平方和;以前者与后者的差作为 t 支持 c_j 的得分。将 t 判定为属于得分最高的类。

$$SCORE2(c_j,t) = \sum_{匹配} Confidence^2 - \sum_{不匹配} N\_Confidence^2 \tag{7}$$

以记录 $X = \{2300,良好\}$ 为例进行分类测试。首先识别 X 中属性的簇,结果为 $X = \{A1:1, A2:1\}$;扫描表3,X 匹配其中的2,5 两条规则,那么:

按方法 1,$SCORE1(C1,X) = 1^2 = 1$;$SCORE1(C2,X) = 0$;X 被判定为 $C1$ 类;

按方法 2,$SCORE2(C1,X) = 1^2 = 1$;$SCORE2(C2,X) = -1^2 = -1$;X 被判定为 $C1$ 类;

应该说明的是,此训练与分类方法有很多缺陷,不能成为实用的算法,但对于本文的实验目标是有效的。

2　在 UCI 分类集上的测试

本章实验数据集是来自 UCI 机器学习库[5] 的四个分类集:Mushroom, Wine, Zoo 和 Breast。以每个数据集作为训练集,再以自身作测试集。采用 1.2 节的方法,设置充分置信度阈值为 50%,必要置信度阈值为 80%,测试结果如表 4 所示。

表 4　UCI 分类集上的测试结果

序号	训练集	测试集	数据集规模	分类精度	
				方法 1	方法 2
1	Mushroom	Mushroom	8 124	0.582	0.597
2	Wine	Wine	178	0.567	0.567
3	Zoo	Zoo	101	0.703	0.851
4	Breast	Breast	699	0.961	0.973

从表 4 可知,在考虑了必要置信度后,分类精度普遍得到了提高。但 Wine 数据集的测试精度不变,研究其生成的规则集 R,发现其中根本没有满足必要置信度阈值的规则。再分析 Wine 数据集,它有 13 个属性和 178 条记录,其中 12 个是连续值的决策属性,属性值分布较均匀。另 1 个属性是类别,类别有"1"、"2"、"3"三类,记录数分别是 59、71、48。在实验时属性的离散化采用的是等间隔离散,间隔数是 10,由式(5)可以看出为什么没有采集到满足必要置信度阈值(≥80%)的规则。这组实验数据表明,如果能采集到合适的必要规则置信度,并让它们在分类时起作用,分类精度会提高。

3　KDDCUP99 数据集上的测试

非平衡数据集是指同一个数据集中某些类的样本数远大于其他类的样本数,其分类问题在医疗诊断、欺诈检测、故障预测、网络入侵检测等领域有应用。传统的分类方法直接应用在非平衡数据集上一般不会得到好的准确率。针对非平衡数据集分类问题研究主要集中在两个方面,一是直接改进传统分类算法;二是采用适当的方法重构训练样本数据集[6]。本章实验的数据集是网络访问数据集,属于典型的非平衡数据集。

KDDCUP99 是网络访问记录数据集,每行数据记录了一次网络访问的属性及访问类别[7]。访问类别反映了本次访问的性质,其中大部分记录属正常访问类(normal),例如:>90%,另外的类别均属于攻击类。我们从 KDDCUP99 中选择了 corrected 数据集,corrected 数据集共有 65 536 条记录。从中选择了正常类记录和 8 类攻击记录组成 4 个数据子集,Sub1、Sub2、Sub3、Sub4。各子集中攻击记录的比例均 <10%,Sub1 和 Sub4 中攻击类记录是均匀的,Sub2 和 Sub3 中攻击类记录是非均匀的。

设置充分置信度阈值为 50%,必要置信度阈值为 80%。考虑到不平衡数据集中起决定作用的是稀有属性(支持度小于等于指定阈值的属性);一记录被判定不是某个攻击类,那么它就是正常类,我们对 2.2 节的实验方法进行了如下修改。

1)设置一支持度阈值 10%,在生成规则集时只采集前件支持度≤10%,并且满足两置信度要求的规则,不采集正常类的规则。

2)在测试时,t 仍被判定属于得分最高的类。由于不采集正常类的规则,只要最高得分 >0,就判定是攻击类,当所有攻击类得分均小于等于 0 时,判定 t 为正常类。

分别以各子集为训练集和测试集,进行实验,结果如表 5 所示。表中错判记录数由三列合计得到,"攻—攻"表示其中攻击记录被错判为另一类攻击的记录数,"攻—正"表示其中攻击记录被错判为正常类的记录数,"正—攻"表示其中正常记录被错判为攻击类的记录数。

表 5　KDDCUP99 数据集上的测试结果

序号	训练集	测试集	测试集规模(正常记录+攻击记录)	方法 1 错判记录数及精度				方法 2 错判记录数及精度			
				攻—攻	攻—正	正—攻	精度	攻—攻	攻—正	正—攻	精度
1	Sub1	Sub1	1080(1000+80)	0	0	38	0.965	0	0	0	1.000
2	Sub2	Sub1	1080(1000+80)	5	0	36	0.962	3	0	12	0.986
3	Sub3	Sub1	1080(1000+80)	8	0	33	0.962	0	0	6	0.994
4	Sub4	Sub1	1080(1000+80)	0	0	18	0.983	0	1	5	0.996
5	Sub1	Sub2	1588(1500+88)	3	0	66	0.957	0	1	0	0.999
6	Sub2	Sub2	1588(1500+88)	0	0	44	0.972	0	0	4	0.997
7	Sub3	Sub2	1588(1500+88)	5	0	45	0.969	0	0	1	0.999
8	Sub4	Sub2	1588(1500+88)	0	0	30	0.981	0	1	1	0.998
9	Sub1	Sub3	2119(2000+119)	9	0	86	0.955	0	6	0	0.997
10	Sub2	Sub3	2119(2000+119)	8	0	69	0.964	6	3	14	0.989
11	Sub3	Sub3	2119(2000+119)	2	0	65	0.968	0	1	6	0.997
12	Sub4	Sub3	2119(2000+119)	2	0	43	0.979	0	4	5	0.996
13	Sub1	Sub4	6548(6308+240)	21	1	175	0.970	1	16	0	0.997
14	Sub2	Sub4	6548(6308+240)	25	0	140	0.975	17	11	33	0.990
15	Sub3	Sub4	6548(6308+240)	25	0	132	0.976	2	8	17	0.996
16	Sub4	Sub4	6548(6308+240)	3	0	80	0.987	1	8	19	0.996

由表5可知,所有测试结果中方法2的精度均高于方法1的精度;从4项一组的实验中可以看出训练集的规模、训练集中小类记录的均衡程度对测试结果影响较小;由于不采集大

(下转第 2526 页)

```
            <Weight>100Kg</Weight>
            <Country>US</Country>
        </PackingList Title>
        <PackingList Detail>                    //装箱单表体信息
            <GoodsName>Computer</GoodsName>
            <Quantity>40</Quantity>
            <CurrencyCode>Dollar<CurrencyCode>
            <UnitPrice>1000</UnitPrice>
        </PackingList Detail>
    </PKInfo>
</xmlSendedInfo>
");
PortalWS.MessagePreDeal MessagePreDeal = new Client.PortalWS.
  MessagePreDeal();
XmlDocument xmlPKInfo = MessagePreDeal.GetXmlInfo(xmlSendedInfo)
```

代码片段2 该段代码是企业服务总线接收消息后，消息处理器解析消息，检索数据适配器，调用数据适配器进行数据转换，再通过Web服务调用，将转换后的符合服务接口要求的数据传送给Web服务，由Web服务生成进口报关单EDI报文。

```
//消息处理器解析消息，从供应商请求消息里解析出参数(如数
//据适配器号: p_Adapt_No 和装箱单数据(xmlPackingList)
String p_Adapt_No = root.SelectSingleNode("\\request\\header\\
  AdaptNo").Value.Trim();
XmlNodeList notes = root.SelectNodes("\\request\\PKInfo");
XmlNode xmlPackingList = notes[0];
//消息处理器声明注册中心实例
InfoCenter.LocateAdapt pkInfoCenter = new PortalWebService.
  InfoCenter.LocateAdapt();
//通过注册中心，获取处理装箱单数据转换的数据适配器
IAdapter DataAdapter = pkInfoCenter.GetDataAdapter(p_ AdaptId);
//注册中心调用相关的数据适配器，将装箱单数据转换成标准
//接口数据
xmlDocument xmlInterface = DataAdapter.BillConvert(xmlPackingList);
//数据适配器调用进口报关单EDI报文生成服务，生成XML格
//式的EDI进口报关单报文
xmlDocument xmlImportEDI = EdiService.CreateEDIBill(xmlInterface);
```

由于系统框架的开放性、数据适配器的可配置性以及数据交换标准的统一性，进口企业可以引用供应商发票、装箱单或运单等不同单证信息作为进口报关的数据来源，同时支持Excel、TXT、CSV、XML等不同格式文件之间的数据转换，基于SOA的ESB方法为快速通关业务中异构系统间的数据交换提供了一个理想的系统集成框架，使通关效率得到了显著地提高。

4 结语

本文分析了当前企业实施供应链快速响应系统集成过程中，实现异构系统之间数据交换和共享时遇到的主要问题。将ESB作为实现基于SOA架构系统集成的具体方法，提出了一个遵循SOA基本原则、应用ESB为系统集成方法的供应链快速响应系统架构，重点对该系统架构组成和执行机制进行了详细说明，并应用此框架解决一个具体应用问题。这种松散耦合的系统集成框架提高了系统的可扩展性和可维护性，SOA和ESB共同为分布式、异构的系统集成提供更高效的、可扩展的平台和工具，使供应链快速响应集成系统的柔性大大提高。

参考文献：

[1] 马俊，丁晓明. 基于SOA的异构系统集成研究[J]. 计算机工程与设计，2008，29(14)：3638－3641.

[2] 梅立军，付小龙，刘启新，等. 基于SOA的数据交换平台研究与实现[J]. 计算机工程与设计，2006，27(19)：3601－3603.

[3] 曹晓叶，王知衍，许晓伟，等. 基于SOA的企业应用集成研究与应用[J]. 微计算机信息，2007，23(12)：16－19.

[4] 刘敏，严隽薇. 基于面向服务架构的企业间业务协同服务平台及技术研究[J]. 计算机集成制造系统，2008，14(2)：306－313.

[5] 顾天竺，沈洁，陈晓红，等. 基于XML的异构数据集成模式的研究[J]. 计算机应用研究，2007，24(4)：94－97.

[6] 王胜娟，江水. 企业集成中的企业服务总线技术[J]. 计算机工程，2006，32(13)：251－253.

[7] 叶宇风. 基于SOA的企业应用集成研究[J]. 微电子学与计算机，2006，23(5)：211－213.

[8] KEEN M. Patterns: Implementing an SOA using an enterprise service bus [EB/OL]. [2009－01－15]. http://www.ibm.com/redbooks.

(上接第2501页)

类的规则，实验中生成的规则集R规模较小。

4 结语

大部分基于规则的分类方法均考虑了充分规则“$A \rightarrow C$”及其置信度的作用，本文的实验表明适当采集与利用必要规则“$C \rightarrow A$”及其置信度可以有效提高分类精度。特别地，在网络访问数据集上只采集稀有属性的规则，分类精度很高，并且有规则集小等优势。本文的实验方法也许不是实用的分类方法，但希望我们的工作能为分类算法的研究提供新思路。

参考文献：

[1] QUINLAN J R. C4.5: Programs for machine learning [M]. Los Altos: Morgan Kaufmann, 1993.

[2] LIU B, HSU W, MA Y. Integrating classification and association rule mining [C]// Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining. New York: [s. n.], 1998: 80－86.

[3] DOMINGOS P, PAZZANIL M. Beyond independence: Conditions for the optimality of the simple Bayesian classifier [C]// Proceedings of the 13th International Conference on Machine Learning. San Francisco: [s. n.], 1996: 105－112.

[4] 邓维斌，黄蜀江，周玉敏. 基于条件信息熵的自主式朴素贝叶斯分类算法[J]. 计算机应用，2007，27(4)：888－891.

[5] BLAKE C L, MERZ C J. UCI machine learning repository of machine learning databases [EB/OL]. [2009－01－05]. http://archive.ics.uci.edu/ml/.

[6] 高嘉伟，梁吉业. 非平衡数据集分类问题研究进展[J]. 计算机科学，2008，135(4)：10－13.

[7] HETTICH S, BAY S D. The UCI KDD archive [EB/OL]. [2009－01－05]. http://kdd.ics.uci.edu.