针对信息缺失的复杂系统的特征选择

宋家勇，杨 杰

（上海交通大学 图像处理与模式识别研究所，上海 200030）

摘 要：特征选择是数据挖掘中的重要研究内容。在现实中，许多待研究系统都很复杂，其中还存在噪声，信息缺失等问题。通过几种特征筛选方法：样本可分类性的评价、对特征集各元素的评价，找出一个信息缺失的复杂系统几个可测特征中对系统性能有较大影响的特征。从而正确指出了系统优化的改进方向，实验结果验证了方法的有效性。

关键词：数据挖掘； 特征选择； 信息缺失

中图分类号：TN919 **文献标识码**：A **文章编号**：1007-2276(2004)05-0516-04

Feature selection for complex information-absent system

SONG Jia-yong, YANG Jie

(Institute of Image Processing & Pattern Recognition, Shanghai Jiaotong University, Shanghai 200030, China)

Abstract: Feature selection is one of important research areas of data mining. In real world, many systems to be investigated are very complicated, some of them have problems of noise and information-absent. Through several feature selection methods——evaluation of samples classification and evaluation of each element of feature set, some features with great influence to the performance of the complex information-absent system can be found. Then a practical and correct direction to improve this system is pointed out. The improvement has been proved by the experimental results.

Key words: Data mining; Feature selection; Information absence

0 引 言

特征选择是数据挖掘中的一种重要方法，它能在不降低性能的前提下简化所研究的问题的复杂性。而在现实中，许多待研究系统都很复杂，其中多数还存在噪声、信息缺失等问题。本文探讨了特征选择在改进一个复杂的现实系统中的作用。

1 存在信息缺失的复杂系统

本文所研究讨论的对象是一个结构复杂，又存在信息缺失的系统。

如图 1 所示，左侧的第 I 部分是系统的实际体系：数据的读取与检测装置由数据存储媒质获取数据，经过其本身的校验、更正，输出最终的数据流。研

收稿日期：2003-12-03； **修订日期**：2004-01-12

作者简介：宋家勇(1979-)，男，上海人，硕士，主要研究方向为数据挖掘、模式识别。

究考察这一系统的目的是提高数据读取与检测装置的性能，以使其对某些受损存储媒质和有错误的存储媒质也能具备一定的读取能力。

右侧的第Ⅱ部分是笔者的研究途径：对于读取出的数据流，得出一定的主观与客观的评价结果；对于某个数据存储媒质的集合，抽取出它们的可测特征；再由可测特征集和评价结果集分析得出读取与检测装置的弱势，以便对此装置做进一步的改进。

此系统的整个步骤存在着信息缺失的环节：在①处，仅有存储媒质的可测特征被抽取出来，存在着某些有益特征的缺失；在②处，由于无法获得读取与检测装置的性能细节，只能把这一装置作为黑箱对待；在③处，由于评价方式的局限，存在着遗漏错误的可能，而对于观察到的错误，也无法得出自何处始，至何处终的结果。

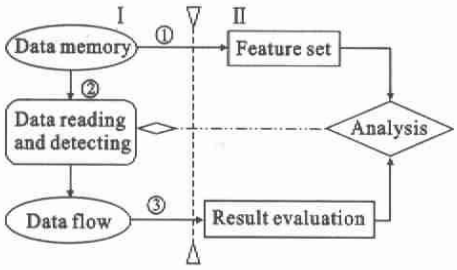

图 1　系统的结构与研究方法

Fig. 1　Structure and study method of the system

2　样本可分类性的评价

由于研究对象存在上述局限性，对于所得到的数据集并不能有过高的期望。但是，所提取的数据存储媒质的特征集，如其确实包含了影响着数据读取与检测装置的部分特征，也一定会如实地反映在所得到的数据集中。基于此，如果所得到的数据集具有一定的可分类性，就可以认为存储媒质的特征集中存在着具有影响力的因素。

通过尝试发现，线性分类器、二次分类器以及神经网络分类器在此数据集上的效能并不明显[1,2]。表 1 给出线性和二次分类器针对训练集的结果。表 2 给出神经网络分类器的结果。

表 1　线性与二次分类器结果

Tab. 1　Results of linear and quadratic classifier

Classifier	Correctness of classification
Linear classifier	67.2%
Quadratic classifier	78.6%

表 2　神经网络分类器结果

Tab. 2　Results of neural network classifier

Samples	Classified as GOOD	Classified as BAD
Real GOOD	81	4
Real BAD	29	17

再对此数据集做 PCA 分析，其结果如表 3 所示。

表 3　PCA 分析结果

Tab. 3　Results of PCA analysis

Feature	Ratio value	Feature	Ratio value
F_1	0.3646	F_5	0.10629
F_2	0.1718	F_6	0.09215
F_3	0.13616	F_7	0.07375
F_4	0.11819		

由表 3 可见，PCA 所得的新分量的比重并不明显。鉴于前两个分量的比重和超过了 50%，将这两个分量的新实例集以二维图的形式给出，如图 2 所示。

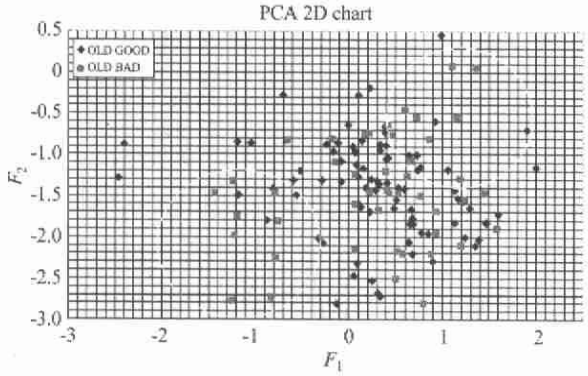

图 2　PCA 结果的二维图

Fig. 2　2D figure of PCA result

图 2 可以看出，虽然在大范围内两类盘错，但在圆圈围起的区域中，则明显以“BAD”类为主。这足以表明，尽管由于信息缺失致使此数据集无法具有显

著的可分类性，但在当前研究的范畴（数据存储媒质的可测特征集）中，必然存在某些重要的影响因素。

3 对特征集各元素的评价

笔者认为，在此系统中，不应满足于找到一个复杂的分类器；现实系统不是一个抽象的数字世界，需要确保所得的结果具有实际的意义。因此，需要在数据媒质的特征集中找出有影响的因素。这样做首先要找出此集合中的每个特征同类别信息的相关度。笔者使用了以下几种方法来比较它们各自的相关度。

(1) OneR

针对每个特征构建一个单层决策树，用该决策树的分类正确率作为该特征的评价值。这个单层决策树将选中的特征分为若干区间，使每区间中都有某个占多数的类别。

(2) 信息增益

若数据集中共有 k 类，分别为：$s_j(j=1,2,\cdots,k)$，每类含 $C_i(i=1,2,\cdots,k)$ 个样本，则类别信息的熵为：

$$H(Class)=-\sum_{j=1}^{k}(p(s_j)\log_2 p(s_j))=-\sum_{j=1}^{k}\left(\frac{C_j}{\sum_{i=1}^{k}C_i}\log_2\frac{C_j}{\sum_{i=1}^{k}C_i}\right)$$

若此属性（$Attribute$）取 m 个值，属性为 $t_i(i=1,2,\cdots,m)$，且所属类别为 $s_j(j=1,2,\cdots,k)$ 的样本数为 x_{ij}，则此属性（特征）下类别信息熵为：

$$H(Class\mid Attribute)=-\sum_{i=1}^{m}p(t_i)\sum_{j=1}^{k}p(s_j\mid t_i)\log_2(p(s_j\mid t_i))=-\sum_{i=1}^{m}\left\{\frac{\sum_{h=1}^{k}x_{ih}}{\sum_{g=1}^{m}\sum_{h=1}^{k}x_{gh}}\sum_{j=1}^{k}\left\{\frac{x_{ij}}{\sum_{h=1}^{k}x_{ih}}\log_2\frac{x_{ij}}{\sum_{h=1}^{k}x_{ih}}\right\}\right\}$$

$$InfoGain(Class,Attribute)=H(Class)-H(Class\mid Attribute)$$

以上式作为属性的信息增益的度量。

(3) 增益率

信息增益的量值会随此属性（特征）可取值的增多而骤增，从而引起偏差。

$$GainRatio=\frac{H(Class)-H(Class\mid Attribute)}{H(Attribute)}$$

上式为增益率，可避免信息增益的偏差。

(4) 对称不可靠性

$$SymmUncertain(Class,Attribute)=\frac{2\times InfoGain(Class,Attribute)}{H(Class)+H(Attribute)}$$

对称不可靠性是另一种补偿信息增益的偏差方法。同时，上式可变化为：

$$SymmUncertain(Attr1,Attr2)=\frac{2\times InfoGain(Attr1,Attr2)}{H(Attr1)+H(Attr2)}$$

此式可用来刻画两个属性（特征）之间的相关性。

(5) Relief

给每个属性（特征）赋一个相同的初始权值，从原始样本集中随机选出一个集合，遍历这个集合，找到距离每个样本最近的同类和异类的样本，来更新属性的权值。

$$W_X=W_X-\frac{diff(X,R,H)^2-diff(X,R,M)^2}{m}$$

式中 W_X 为属性 X 的权值；m 为随机集合的样本数；$diff(X,R,H)$ 为样本 R 与最近同类样本 H 在属性 X 上的差别，$diff(X,R,M)$ 为样本 R 与最近异类样本 M 在属性 X 上的差别。

然后，根据最终的权值给出对各属性（特征）的评价。

(6) Chi-squared (χ^2)

基于类别的 χ^2 统计量。对于两个毗邻的区间，若类别数为 C，则：$\chi^2=\sum_{i=1}^{2}\sum_{j=1}^{C}\frac{(A_{ij}-E_{ij})^2}{E_{ij}}$，其中 A_{ij} 为第 i 个区间中属于第 j 类的样本数，E_{ij} 为 A_{ij} 的期望值，$E_{ij}=R_i\times C_j/N$，R_i 为第 i 个区间的样本数，C_j 为这两个区间中属于 j 类的样本数，N 为两区间中样本总数。

图 3 为各种方法的比较结果，横轴表示特征的序号，纵轴表示某特征在使用某方法时所得的重要性的

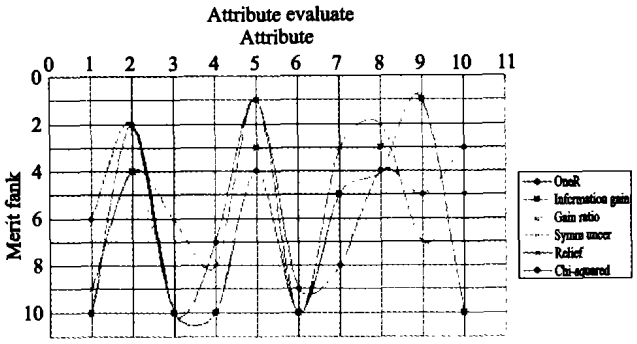

图 3　特征比较结果汇合图

Fig. 3　Composed figure of feature comparisons

位次(1 最高,10 最次)。

综合这些结果,得到各特征总的评价值,如表 4 所示。

表 4　特征比较汇总结果

Tab. 4　Combined results of feature comparisons

Feature No.	1	2	3	4	5	6	7	8	9	10
Evaluation	55	20	56	55	15	59	31	19	16	48

由这些结果可见,特征 5、9、8、2 在此集合中为较有影响力的几个因素。

4　寻找具影响力的特征集合

在上节得到的结果中,所找到的特征 5、9、8、2 是对数据读取与检测装置较有影响的几个因素,这并不表明,它们所组成的集合$\{F_5, F_9, F_8, F_2\}$是一个对读取与检测装置甚有影响的因素集。当前的存储媒质的可测特征集中,各元素并不是相互独立的;各元素彼此间的相互影响不可避免地影响着存储媒质的可读性。尚待获知的影响因素集需要剔除彼此相关的元素,以使得此集合的规模更小,同时还要尽可能地保持此集合的分类能力。在本系统中,考虑到本身的复杂性,当对数据读取与检测装置进行改进时,涉及的因素应当越少越好。笔者使用了下面的方法,来减少候选特征集的元素个数。

基于关联的特征子集评价(Correlation-based Feature Subset Evaluator,CFS)[3]方法,通过启发式函数计算特征子集内部各元素间的相关性,从而对特征子集作出性能评价,达到消减冗余特征的目的。一个特征能否被添加进这个子集,关键在于它能否提高此特征子集的预测性能,而与此特征与样本集的类别信息的相关性无关。

CFS 对特征子集的评价基于以下的相关性表达式:

$$M_s = \frac{k\,\bar{r}_{cf}}{\sqrt{k + k(k+1)\bar{r}_{ff}}}$$

式中　M_s 是特征子集同分类信息的相关性,也就是对特征子集的评价值;k 为特征子集的维数;$\bar{r}_{cf}$ 为子集中各元素同类别信息的相关性的均值;$\bar{r}_{ff}$ 为特征子集内部各成员间相关性的均值。对于两两元素(特征－特征,特征－类别)间的相关性,常用对称不可靠性和 Relief 方法来获得。

对原始的可测特征集使用 CFS 方法进行特征筛选之后,其结果显示:特征子集$\{F_2, F_5, F_9\}$是一个优解,它的各元素都属于较有影响的特征,同时没有冗余元素的存在。

70%现可正常读取为 GOOD,这也是对上述特征选择结果的实际验证。

5　结　论

本文对一个信息缺失的复杂系统进行了一些探讨。对于研究现实的复杂系统往往需要从简化入手,如文中使用的特征选择方法。但是对于所有的现实问题,其实际意义通常总会处于第一位,这一点经常容易被忽视。

文中提到的改进方向与实际结果的吻合是提高系统整体性能的重要步骤,对于后续的工作会有很大的帮助。

参考文献:

[1]　Tom Mitchell. Machine Learning[M]. McGraw Hill, 1997.

[2]　边肇祺, 张学工 . 模式识别(第二版)[M]. 北京:清华大学出版社,2000.

[3]　Mark A Hall. Correlation-based Feature Selection for Machine Learning[D] . Ph D Diss,1998.