文章编号:1001-506X(2012)05-1046-05

基于 Markov blanket 和互信息的集成特征选择算法

姚 旭,王晓丹,张玉玺,权 文
(空军工程大学导弹学院,陕西 三原 713800)

摘 要:针对大量无关和冗余特征的存在可能降低分类器性能的问题,提出一种基于近似 Markov blanket 和动态互信息的特征选择算法并将其应用于集成学习,进而得到一种集成特征选择算法。该集成特征选择算法运用 Bagging 方法结合提出的特征选择方法生成基分类器,并引入基分类器差异度进行选择性集成,最后用加权投票法融合所选基分类器的识别结果。通过仿真实验验证算法的有效性,以支持向量机(support vector machine, SVM)为分类器,在公共数据集 UCI 上进行试验,并与单 SVM 及经典的 Bagging 集成算法和特征 Bagging 集成算法进行对比。实验结果显示,该方法可获得较高的分类精度。

关键词:特征选择;集成;Markov blanket;互信息
中图分类号:TP 391 **文献标志码**:A **DOI**:10.3969/j.issn.1001-506X.2012.05.34

Ensemble feature selection algorithm based on Markov blanket and mutual information

YAO Xu, WANG Xiao-dan, ZHANG Yu-xi, QUAN Wen
(*Missile Institute, Air Force Engineering University, Sanyuan 713800, China*)

Abstract: To resolve the poor performance of classifiers owing to the irrelevant and redundancy features, a feature selection algorithm based on approximate Markov blanket and dynamic mutual information is proposed, then it is introduced to an ensemble feature selection algorithm. In the ensemble algorithm, a base classifier is trained based on Bagging and the proposed feature selection algorithm, and the base classifier diversity is introduced to selective ensemble. Finally, the weighted voting method is utilized to fuse the base classifiers' recognition results. To attest the validity, experiments on data sets with support vector machine (SVM) as the classifier are carried out. The results have been compared with single-SVM, Bagging-SVM and AB-SVM. Experimental results suggest that the proposed algorithm can get higher classification accuracy.

Keywords: feature selection; ensemble; Markov blanket; mutual information

0 引 言

集成学习和特征选择是近年来统计学、机器学习和数据挖掘等领域中的经典研究问题。文献[1]将集成学习和特征选择融合在一起,通过选择多个特征子集生成多个个体分类器来提高集成中个体差异度,从而提高学习系统的性能。如何有效地生成属性子集是该类方法需要解决的核心问题。文献[2]中的随机属性子空间方法就是这类方法的典型代表。目前最常采用的是通过特征选择方法来获得属性子集,如文献[3]的基于遗传算法(genetic algorithm, GA)的特征选择方法以及文献[4]的基于多目标优化的 GA 的特征选择方法。文献[5]提出的特征 Bagging(attribute Bagging, AB)集成算法通过随机方法产生属性子集以获得集成个体之间更大的差异度,并成功地应用于手势识别。

本文在分析 Markov blanket 原理的基础上,以动态互信息为度量标准,提出一种基于近似 Markov blanket 和动态互信息的特征选择算法,并将其应用于集成学习,从而得到一种集成特征选择算法。为验证算法的有效性,以支持向量机(support vector machine, SVM)作为分类器,在公共数据集 UCI 上进行实验,最后给出实验结果和分析。

收稿日期:2011-09-06; **修回日期**:2012-02-21。
基金项目:国家自然科学基金(60975026)资助课题
作者简介:姚旭(1982-),女,博士研究生,主要研究方向为智能信息处理和机器学习。E-mail:ffxy132@163.com

1　Markov blanket 相关知识

1996 年,文献[6]将 Markov blanket 引入到特征选择中,在删除特征的过程中以特征中是否存在 Markov blanket 为标准。下面给出 Markov blanket 的一些基本概念,关于 Markov blanket 的更多知识可参见文献[7]。

定义 1　给定一个特征 f_i,设特征子集 $M_i \subset F(f_i \notin M_i)$,称 M_i 是 f_i 的 Markov blanket,当且仅当在给定 M_i 的条件下 f_i 和 $F-M_i-\{f_i\}$ 是独立的,即 $P(F-M_i-\{f_i\}|f_i,M_i)=P(F-M_i-\{f_i\}|M_i)$。

推论 1　如果特征子集 M_i 是 f_i 的 Markov blanket,那么在给定 M_i 的条件下 f_i 与类别 C 也是独立的,即 $P(C|f_i,M_i)=P(C|M_i)$。

通过 Markov blanket 方法,可以有效地去除无关和冗余特征,但是在高维空间中求 Markov blanket 要对所有的特征子空间进行搜索,计算复杂度很高,其时间复杂度为 $O(2^d)$(d 为特征的维数)。因此,可以利用近似 Markov blanket 来解决此问题。用 $R(f_i;C)$ 表示特征 f_i 与类别 C 的相关性;$R(f_i;f_j)$ 表示特征 f_i 与 f_j 之间的相关性;$R(f_j;C|f_i)$ 表示给定特征 f_i 的条件下,特征 f_j 与类别 C 的相关性。

定义 2　设特征 f_i 和 f_j,C 为类别,R 为一种度量准则,$R(f_i;C)>R(f_j;C)$,称 f_i 是 f_j 的一个近似 Markov blanket,当且仅当 $R(f_j;C|f_i)>R(f_j;C)$。

在 Markov blanket 理论中,当去除某些冗余特征后在开始阶段被剔除的特征仍然是冗余的,但对于近似 Markov blanket 则不一定成立。但是强相关特征不存在近似 Markov blanket,因此在任何阶段强相关特征都不会被删除。所以利用近似 Markov blanket 同样可以得到一个最优子集。

2　基于近似 Markov blanket 和动态互信息的集成特征选择算法

2.1　特征相关性度量

信息熵和互信息等信息测度是目前普遍采用的特征相关性评价准则。因为它能以量化的形式度量特征间的不确定程度,并且能有效的度量特征间的非线性关系。信息测度已多次被引入到特征选择的过程中。如文献[8]提出了基于互信息的特征选择算法并将其应用于故障诊断和识别,取得了良好的效果;文献[9]用互信息来度量两个变量之间的统计相关性,用来进行步态识别;文献[10]将正规化互信息和遗传算法相结合,提出了一种混合式的特征选择算法等。本文在特征的相关性和冗余性分析中也采用互信息作为度量标准。当两变量完全无关或互相独立时,它们的互信息为 0,意味着两者之间不存在相同的信息;互信息值越大,意味着所包含相同的信息也越多。因此,如果互信息 $I(f_i;C)$ 越大,表示类别 C 对特征 f_i 的依赖性越大。可见,用互信息可以很好的度量特征和类别的相关性。

同样,可以根据互信息来度量两个特征之间的相互关联程度。但是,以此来确定特征是否冗余也存在一定的困难。例如当两个特征彼此不完全相关时,很难判断哪一个特征是冗余的,而且很可能要对全部特征计算 $d(d-1)/2$(d 为特征的维数)个特征间的相互关联度,这在高维数据集中效率是极低的。因此,直接计算特征间的互信息来判断冗余特征是不可行的。由 Markov blanket 的定义可以知道,Markov blanket 方法可以有效地去除冗余特征。很多文献也将 Markov blanket 应用于特征选择中,如文献[11-14]。由于 Markov blanket 通过特征子集来计算,其计算量比较大,因此可以利用近似 Markov blanket 来近似的确定冗余特征。本文用互信息和条件互信息作为度量准则,用 $I(f_i;C)$ 表示特征 f_i 与类别 C 的相关性;$I(f_i;f_j)$ 表示特征 f_i 与 f_j 之间的相关性;$I(f_j;C|f_i)$ 表示给定 f_i 的条件下,f_j 与类别 C 的相关性。利用近似 Markov blanket 确定冗余特征的判断准则如下:如果 $I(f_i;C)>I(f_j;C)$,$I(f_j;C|f_i)>I(f_j;C)$,那么 f_i 是 f_j 的一个近似 Markov blanket,即特征 f_j 是冗余的,应该删除。

2.2　基于近似 Markov blanket 和动态互信息的特征选择算法

在度量特征的相关性程度时,数据样本集给定之后,特征在这个样本集的概率分布就被确定下来。这种确定性使度量标准不能准确反映信息或不确定性的动态变化情况,因为特征选择是一个动态的过程,即随着已选特征增多,类别 C 的不确定性也逐渐降低,同时样本空间中不可识别的样本数量也呈减少的趋势,在某种程度上说明不发生变化的相关性度量标准包含部分“假”信息[15]。因此,如果在特征选择过程中,不断的删除已被识别的样本,使评价标准在未识别样本上动态估值,不仅可以更准确地反映相关程度,而且提高了计算效率。文献[15]提出一种基于动态互信息的特征选择(dynamic mutual information feature seletion, DMIFS)算法,该算法通过每次选择一个与类别相关性最大的特征加入到已选特征集合中,再不断删除该特征能识别的样本,使互信息在不断更新的样本集上动态估值,从而保证相关性度量的准确性。但是,该算法也存在一个问题,它只考虑特征与类别的相关性,没有考虑特征之间的相关性,因此所选择的特征中可能存在冗余。从 2.1 节的分析中可知,近似 Markov blanket 是一种去除冗余和无关特征的有效方法。因此,本文以动态互信息为评价标准,采用近似 Markov blanket 来去除冗余和无关特征,提出一种基于近似 Markov blanket 和动态互信息的特征选择(approximate Markov blanket and dynamic mutual information, AMBDMI)算法。算法分为两个阶段,第一阶段选择与类别有相关性的特征集合,第二阶段用近似 Markov blanket 去除冗余特征。设样本数据集为 D,类别为 $C=(c_1,c_2,\cdots,c_m)$,特征集合 $F=(f_1,f_2,\cdots,f_d)$,已选特征集合记为 S,算法具体步骤如下:

输入　训练数据集 $D=(x_1,x_2,\cdots,x_N)$,其中 $x_i=(x_{i1},x_{i2},\cdots,x_{id})$,类别 $C=(c_1,c_2,\cdots,c_m)$,$F=(f_i|i\in 1,2,$

…,d),阈值 p。

步骤 1 参数初始化,已选特征集合 $S=\varnothing$,$D_l=\varnothing$。

步骤 2 对 F 中所有特征值 f_i 计算 $I(f_i;C)$。如果 $I(f_i;C)=0$,则从 F 中去除 f_i。

步骤 3 对 F 中剩余的特征按照 $I(f_i;C)$ 的值降序排列,构成的特征集合记为 F'。

步骤 4 取 F' 中的第一个特征,即 $f=\arg\max I(f_i;C)$,将 f 加入到已选特征集合 S 中,即 $S=S+\{f\}$,$F'=F'-\{f\}$。

步骤 5 由特征 f 得到 f 所能识别的样本集合 D_l,更新数据集 $D=D-D_l$。如果 D 为空或者在总样本中所占比例小于阈值 p,算法停止。

步骤 6 取 F' 中的下一个特征 f,如果 $f=NULL$,算法停止。否则执行步骤 7。

步骤 7 对任意的 $f_i\in S$,如果 $I(f;C|f_i)>I(f;C)$,则从 F' 中将 f 删除,即 $F'=F'-\{f\}$,返回步骤 6。否则,将 f 加入到 S 中,并从 F' 中删除 f。即 $S=S+\{f\}$,$F'=F'-\{f\}$。返回到步骤 5。

输出 特征子集 S。

2.3 基于 AMBDMI 的集成特征选择

AMBDMI 利用互信息和条件互信息度量特征的相关性,这种度量方法依赖于样本分布,对样本分布较为敏感。因此,通过对训练集的样本分布进行扰动能产生不同的特征子集,从而构造基于特征扰动的集成方法。Bagging 是一种基于样本扰动的集成方法,但由于 SVM 是一种"稳定"的分类器,对样本扰动不敏感,导致单纯的基于 Bagging 集成的方法在遭遇 SVM 时效果不理想。本节对基本 Bagging-SVM 方法加以改进,引入基分类器差异度用于选择性集成。基分类器两两间的差异度和整体差异度为

$$BCD(h_i,h_j)=\frac{n_i+n_j-2n_{ij}}{n_i+n_j} \tag{1}$$

$$BCD(h_i)=\frac{1}{N-1}\sum_{j=1}^{N}BCD(h_i,h_j) \tag{2}$$

式中,$i,j=1,2,\cdots,N$,N 为基分类器集合规模;n_i,n_j 为第 i,j 个基分类器分别正确识别的样本数;n_{ij} 为第 i,j 个基分类器都正确识别的样本数。由式(1)和式(2)可知 $BCD(h_i,h_j)$,$BCD(h_i)\in[0,1]$,且取值越大,互补性越强,集成效果越好。

基于差异度选择性集成的步骤如下:

输入 由 Bagging 方法产生初始基分类器集合 RS,集成规模 L'。

步骤 1 初始化基分类器集合,$CS=\varnothing$。

步骤 2 RS 中各基分类器对验证集的分类正确率为 $CR_i(i=1,2,\cdots,L)$,如果 $CR_i\leqslant 0.6$,则从 RS 去除其对应的基分类器 h_i,更新 $RS=\{h_1,h_2,\cdots,h_M\}$,$M=|RS|\leqslant L$。

步骤 3 如果 $M\leqslant L'$,跳转执行步骤 4。否则,用式(2)计算 RS 中每个基分类器的整体差异度 $BCD(h_i)(i=1,2,\cdots,M)$,从 RS 中去除整体差异度最小的基分类器,更新 $RS=RS-\{h_i\}$,$M=|RS|$,返回步骤 3。

步骤 4 $CS=RS$。

输出 用于测试样本分类的基分类器集合 CS。

基于以上分析,本节提出一种基于 AMBDMI 的 SVM 集成特征选择算法即 Bagging-AMBDMI 算法,简称 B-AMBDMI 算法。算法的基本思路:运用 Bagging 方法中的 bootstrap 技术产生多个训练样本子集,在每个训练子集上应用 AMBDMI 算法进行特征选择,得到相应的特征子空间,将每个训练样本子集投影到相应的特征子空间得到基分类器的训练集训练基 SVM,对测试样本,先分别投影到各特征子空间,然后输入相应的基分类器,最后对基分类器的输出结果进行融合输出。算法描述如下:

输入 训练集 D,测试样本 x,训练样本子集规模 N_{sub},基分类器 SVM 的初始数目 L,集成规模为 L'。

步骤 1 运用 bootstrap 抽样方法产生 L 个规模为 N_{sub} 的训练样本子集。

步骤 2 对每个训练子集运用 AMBDMI 进行特征选择,得到一个特征子空间,并将训练子集投影到该特征子空间上。

步骤 3 用每个投影后的训练子集训练一个 SVM,作为集成的基分类器。

步骤 4 将测试样本投影到每个特征子空间上,并用相应的基分类器进行分类。

步骤 5 利用差异度选择用于集成的基 SVM,得到基分类器集合 CS。

步骤 6 把测试样本映射到选择后的每个特征子空间,用 CS 中对应的基 SVM 对测试集进行分类,采用加权投票法融合 CS 中各基分类器的分类结果。权系数由步骤 4 中 CS 包含的基分类器对验证集的分类正确率进行归一化处理后获得。

输出 测试样本的判决类标签。

3 实验结果及分析

3.1 实验数据

实验数据均来自 UCI 数据库[16],选择其中 8 组数据(特征维数范围为 8~60,样本范围为 208~6 435),详细描述如表 1 所示。

表 1 UCI 数据集各数据描述

数据集	训练集	测试集	维数	类别
Sonar	208	—	60	2
Glass	214	—	10	7
Soybean	307	—	35	19
Ecoli	336	—	8	8
Ionosphere	351	—	34	2
Segment	2 310	—	19	7
Waveform	3 000	2 000	21	3
Satimage	4 435	2 000	36	6

测试集中的数据 1/2 用于测试,1/2 用于验证。对表中没有测试集的数据,做如下处理:随机抽取数据集的 2/3 样

本(向上取整)用于训练,其余 1/6 用于测试,1/6 用于验证。实验前,对数据进行归一化处理。

3.2　实验结果和分析

为验证算法 B-AMBDMI 的性能,在公共数据集 UCI 上进行实验,并与单分类器(single)及著名的 Bagging 集成算法和特征 Bagging(AB)集成算法进行对比实验。在估计分类正确率时采用五重交叉验证,并利用双边估计 t 检验法来计算置信水平为 0.95 的分类正确率置信区间作为最终结果。实验分两部分:第一部分为固定基分类器下 4 种算法的分类精度比较;第二部分研究集成规模与集成精度之间的关系。

实验中以 SVM 为分类器,来自 PRTool(http://www.prtools.org)工具箱,采用多项式核函数(polynomial)的 SVM,$q=1$。实验机器处理器主频为 2.82 GHz,内存 2 G,算法基于 Matlab7.10(R2010a)实现。

3.2.1　固定基分类器下的分类精度比较

实验采用五重交叉验证,进行 10 次实验,所有结果取 10 次实验的平均值。实验中训练子集规模等于训练集规模,基分类器数取 25。实验结果如表 2 所示。

表 2　4 种算法的分类精度比较(基分类器数为 25)

	Single-SVM	Bagging-SVM	AB-SVM	B-AMBDMI
Sonar	85.40±3.85	85.46±3.56	86.52±3.24	87.45±3.12
Glass	**80.28±3.08**	80.19±2.99	81.42±3.05	82.57±2.86
Soybean	91.86±3.52	92.47±3.41	93.05±3.26	94.25±2.98
Ecoli	80.66±2.34	81.02±2.56	81.98±2.47	82.53±2.84
Ionosphere	**92.32±2.40**	92.24±2.12	92.84±1.95	93.62±2.01
Segment	93.82±0.24	94.84±0.35	94.98±1.02	95.23±1.25
Waveform	83.36±1.92	89.15±1.77	90.31±1.84	91.29±1.76
Satimage	84.28±2.32	84.85±1.74	85.22±2.06	86.09±1.98
Average	86.50±2.46	87.53±2.31	88.29±2.36	89.13±2.35

表 2 中黑体表示 Single-SVM 分类精度好于 Bagging-SVM 分类精度的数据集。从表 2 可以看出,相比于其他 3 种算法,B-AMBDMI 算法在 8 个数据集上的分类精度均有所提高,其中,较 Single-SVM 提高约 2.63%,较 Bagging-SVM 提高约 1.6%,较 AB-SVM 提高约 0.84%;AB-SVM 算法介于 B-AMBDMI 算法和 Bagging-SVM 算法之间,而 Bagging-SVM 算法在 Ionosphere 和 Glass 数据集上较 Single-SVM 的分类精度分别下降了 0.08%和 0.09%,在其他 6 个数据集上,也较 Single-SVM 算法提升的幅度不大(在 Sonar 数据集上也仅提高约 0.06%),可见,Bagging 集成方法应用于 SVM 分类器时效果不理想。

3.2.2　基分类器数与分类精度之间的关系

第一部分只比较固定基分类器数为 25 的情况下 3 种集成算法的分类精度,为了验证基分类器数对集成精度是否有影响,对 3 种集成算法在不同基分类器数下的分类精度进行比较,在数据集 Sonar,Ecoli,Segment 和 Satimage 上进行实验。实验结果如图 1 所示。

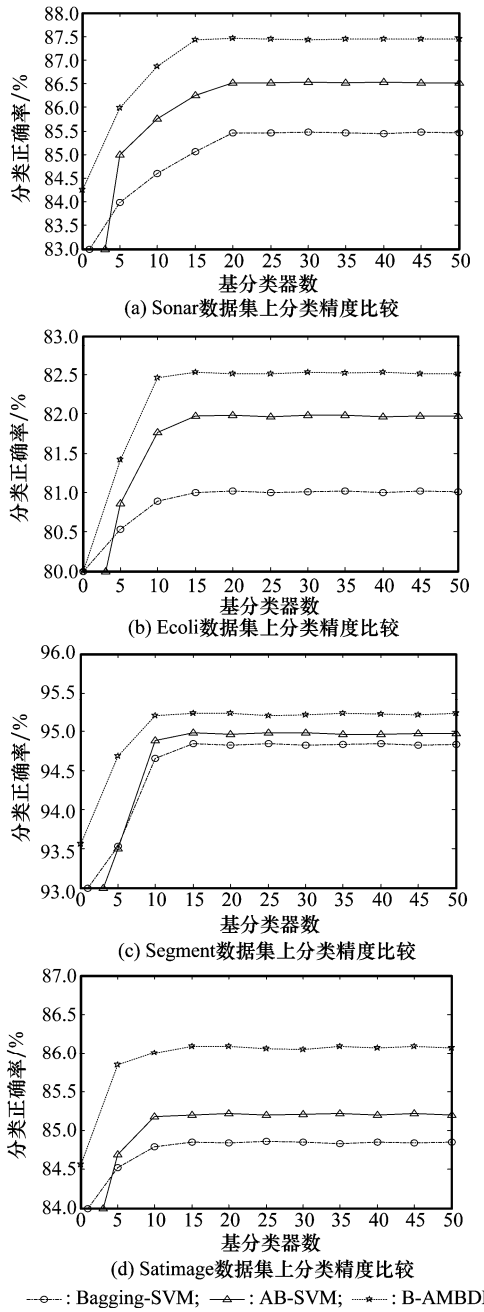

图 1　3 种算法在不同基分类器数下的分类精度比较

由图 1 可以看出,在 4 个数据集上,B-AMBDMI 算法在同等基分类器数下均取得了较 Bagging-SVM 算法和 AB-SVM 算法高的分类精度,Bagging-SVM 算法在基分类器数较少(<5)时分类精度较 AB-SVM 算法好,而当基分类器数达到一定数目以后,AB-SVM 算法的分类精度明显较 Bagging-SVM 算法好。分析原因,AB-SVM 算法对特征空间的随机抽样可能导致重要特征缺失,因此其基分类器精度不高,而随着基分类器的增加,基分类器之间的差异性使它们的错误形成互补,从而提高分类精度。而 Bagging-SVM 算法基于样本随机抽样,随着样本集的增大,随机抽

样对分类器的影响不大，因此它的基分类器的差异性小，使得集成效果不如 AB-SVM。分析 B-AMBDMI 算法，由于它是在 Bagging 方法形成的训练子集上再进行特征选择而得到最终的训练子集，因此，它能得到比 Bagging-SVM 算法和 AB-SVM 算法都高的差异性基分类器，而且由于 B-AMBDMI 的特征子空间通过特征选择算法得到，因此特征子集具有较好的分类能力，保证基分类器的分类精度，所以可以获得更好的集成性能。

4 结 论

本文在研究近似 Markov blanket 和动态互信息的基础上，提出 AMBDMI 算法。该算法利用互信息作为特征相关性的度量准则，在未识别的样本上对互信息进行动态估值，利用近似 Markov blanket 原理准确的去除冗余特征。然后将此算法应用于集成学习，提出 B-AMBDMI 算法——运用 Bagging 方法中的 bootstrap 抽样技术结合 AMBDMI 特征选择方法生成基分类器，并引入基分类器差异度进行选择性集成，最后用加权投票法融合所选基分类器的识别结果。仿真实验结果显示，该方法可获得较高的分类精度。

参考文献：

[1] Opitz D. Feature selection for Ensembles[C]// *Proc. of the American Association for Artificial Intelligence*, 1999: 379 - 384.

[2] Ho T K. The random subspace method for constructing decision forests[J]. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 1998, 20(8): 832 - 844.

[3] Opitz D, Shavlik J. A genetic algorithm approach for creating neural network ensembles[C]// *Proc. of the Combining Artificial Neural Nets*, 1999: 79 - 97.

[4] Oliveira L S, Morita M, Sabourin R, et al. Multi - objective genetic algorithms to create ensemble of classifiers[C]// *Proc. of the 3rd International Conference on Evolutionary Multi-Criterion Optimization*, 2005: 592 - 606.

[5] Brylla R, Gutierrez O R, Queka F. Attribute Bagging: improving accuracy of classifier ensembles by using random feature subsets[J]. *Pattern Recognition*, 2003, 36(6): 1291 - 1302.

[6] Koller D, Sahami M. Toward optimal feature selection[C]// *Proc. of the International Conference on Machine Learning*, 1996: 284 - 292.

[7] Pearl J. *Probabilistic reasoning in intelligent systems*[M]. Washington: American Sociological Association, 1988:449 - 484.

[8] Sylvain V, Teodor T, Abdessamad K. Fault detection and identification with a new feature selection based on mutual information[J]. *Journal of Process Control*, 2008, 18(5): 479 - 490.

[9] Guo B F, Mark S N. Gait feature subset selection by mutual information[J]. *IEEE Trans. on Systems, Man and Cybernetics—Part A: Systems and Humans*, 2009, 39(1): 36 - 46.

[10] Estévez P A, Michel T, Perez C A, et al. Normalized mutual information feature selection[J]. *IEEE Trans. on Neural Networks*, 2009, 20(2): 189 - 201.

[11] Zhao H, Xiao M, Xiao Y. Optimal feature selection based on Bayesian networks[C]// *Proc. of the International Conference on Wavelet Analysis and Pattern Recognition*, 2007: 597 - 601.

[12] Pablo A D, Pablo A D, Fernando J V. Learning Bayesian networks to perform feature selection[C]// *Proc. of the International Joint Conference on Neural Networks*, 2009: 467 - 473.

[13] Sandeep Y, Dimitris M. Speculative Markov blanket discovery for optimal feature selection[C]// *Proc. of the 5th IEEE International Conference on Data Mining*, 2005: 809 - 812.

[14] Theo A K, Marcel J T R, Lodewyk F A W. Artifacts of Markov blanket filtering based on discretized features in small sample size applications[J]. *Pattern Recognition Letters*, 2006, 27(7): 709 - 714.

[15] 刘华文. 基于信息熵的特征选择算法研究[D]. 吉林：吉林大学，2010.(Liu H W. A study on feature selection algorithm using information entropy[D]. Jilin:Jilin University, 2010.)

[16] Hettich S, Bay S D. The UCI KDD archive [DB/OL]. [2011 - 08 - 30]. http://kdd.ics.uci.edu/.