doi:10.3969/j.issn.1000-436x.2012.z1.008

分布式传感器网络混合探测信号分类方法

李侃[1]，许航[1]，黄忠华[2]
（1. 北京理工大学 计算机学院，北京 100081；2. 北京理工大学 机电工程学院，北京 100081）

摘 要：针对分布式传感器网络的局限性特征，研究分布式传感器网络混合探测信号的分类算法。提出了基于属性重要度的贝叶斯分类算法，该算法继承了朴素贝叶斯分类算法结构简单、运算快捷的特点，同时弥补了类条件独立假设带来的缺陷，在实践中具有较高的分类精度，其特点符合混合探测信号的分类要求。实验结果表明，该算法分类效果优于同类分类算法，可以有效地完成混合探测信号的分类任务。

关键词：朴素贝叶斯分类器；属性重要度；分布式传感器网络；混合探测信号
中图分类号：TP181 **文献标识码**：A **文章编号**：1000-436X(2012)Z1-0053-05

Classification method for mixed detection signal in the distributed sensor network

LI Kan[1], XU Hang[1], HUANG Zhong-hua[2]
(1. School of Computer, Beijing Institute of Technology, Beijing 100081, China;
2. School of Mechano-Electronics Engineering, Beijing Institute of Technology, Beijing 100081, China)

Abstract: Taking into account the limitations of the distributed sensor networks, a simple and efficient classification method was found. According to the main idea of naïve Bayes classification (NBC) algorithm, a new naïve Bayes classification based on attribute significance (NBCBAS) was proposed. The algorithm inherited the characteristics of NBC algorithm that was simple and fast computation. At the same time, the algorithm made up for the defects of conditional independence assumption. It had high classification accuracy in practice. The characteristics of the NBCBAS met the classification requirements of the mixed detection signal. At last, the NBCBAS was tested on UCI datasets and mixed detection signal datasets. The results illustrate that our algorithm improves the classification performance.

Key words: naïve Bayes classification; attribute significance; distributed sensor networks; mixed detection signal

1 引言

近年来，随着计算机、无线通信、微电子和传感器技术的飞速发展，传感器网络成为国内外研究的一个热点[1]。它最早起源于军事领域，最为典型的是由声音、地震动、红外、电磁、化学等多种传感器组成的无人值守地面传感器（UGS, unattended ground sensor）网络[2]。随着传感器网络技术的不断发展，其应用范围开始向民用扩展，它已经应用在商业、环境、卫生保健等诸多领域[3]。很多研究者根据需要解决的具体问题提出具体的方法，但仍然缺少一种通用且分类效果较好的分类方法。

近期，一些学者在民用领域先后提出了基于多传感器的分类方法。Brooks 提出了一种传感器网络中的分布式目标分类与跟踪算法[4]。Marco 等针对运动的车辆，提出了传感器网络中的车辆分类算法[5]。

收稿日期：2012-08-06
基金项目：国家自然科学基金资助项目（60903071）
Foundation Item: The National Natural Science Foundation of China (60903071)

Commault 等采用一种结构化方法对传感器分类，从而进行故障检测和隔离[6]。Gong 等提出一个优化的人工免疫网络分类模型，完成基于遥感的土地利用/土地覆盖分类方法[7]。Liu 等提出了一种基于二分类树解决多媒体传感器网络分布式目标分类方法[8]。这些方法在各自解决的问题上取得了明显的效果，但方法都仅适用特定的应用以及环境。本文的研究针对分布式混合探测信号的分类，既要保证分类精度，同时要有实时性的考虑。对于其他分类问题，例如文献[9]尝试解决入侵探测分类问题中，文中所提出的人工神经网络方法平衡了效率与精度在分类中的地位，但并没有将两者提高到令人满意的效果。文献[10]所提出的粒子群优化算法存在着同样的矛盾，尽管该方法在追求高分类精度的前提下，已经很大程度地提高了运算效率，但依然不能满足很多分类工作的实时性要求。

本文的工作旨在建立一种能够得到较高分类精度且满足系统实时性要求的混合探测信号分类器。使分类效果提高且满足系统的实时性要求是十分有难度的，这也是本文工作具有挑战性的原因。

2 基于属性重要度的贝叶斯分类算法

本文基于朴素贝叶斯分类算法的思想，提出了一种基于属性重要度的贝叶斯分类算法(NBCBAS)。该算法的主要思想是将数据集划分为决策属性和条件属性，再通过条件属性与决策属性的依赖关系来计算它们之间的相关系数，最终依据该相关系数来为每一个属性赋予权值。

首先给出属性重要度的定义。

定义 1 属性重要度描述一个属性参与分类决策时的重要程度。属性重要度也可以视为该属性与决策属性关联的紧密程度。属性重要度的大小将决定该属性在参与分类时对决策工作影响的大小。

设训练数据集可以由多个条件属性和几个决策属性来描述，用属性相关度 δ 来描述条件属性与决策属性之间的关系，最终根据 δ 来确定赋予属性的重要度。

用随机变量 X_i 来表示 n 个条件属性，用随机变量 Y_a 来表示 m 个决策属性。

当决策属性多于一个时，需要首先给每个决策属性赋予权值，表示决策属性 Y_a 的决策地位，决策地位参数的和为 1，即

$$\sum_{a=1}^{n}\varphi_a = 1 \tag{1}$$

通常可以认为决策属性的地位是相同的，则设定

$$\varphi_1 = \varphi_2 = \cdots = \varphi_m = \frac{1}{m} \tag{2}$$

由于数据集可以表示为若干随机变量的集合，而各随机变量的取值按一定的概率分布，因此可以分别计算每个随机变量 X_i 与每个决策属性 Y_a 的协方差为

$$\mathrm{Cov}(X_i, Y_a) = \mathrm{E}[(X_i - \mathrm{E}(X_i))(Y_a - \mathrm{E}(Y_a))]^{\mathrm{T}} \tag{3}$$

X_i 与 Y_a 的方差分别为

$$\mathrm{D}(X_i) = \mathrm{E}(X_i - \mathrm{E}(X_i)) \tag{4}$$

$$\mathrm{D}(Y_a) = \mathrm{E}(Y_a - \mathrm{E}(Y_a)) \tag{5}$$

再计算 X_i 与 Y_a 的相关系数的绝对值来表示它们之间的属性相关度。

$$\delta_{ia} = \left| \frac{\mathrm{Cov}(X_i, Y_a)}{\sqrt{\mathrm{D}(X_i)}\sqrt{\mathrm{D}(Y_a)}} \right| \tag{6}$$

算法利用 δ 描述每个条件属性对每个决策属性的关联程度。条件属性重要度用 w_i 来表示，则

$$w_i = \sum_{a=1}^{m}\delta_{ia}\varphi_a, i = 1, 2, \cdots, n \tag{7}$$

其中，n 为条件属性的个数，m 为决策属性的个数。权值 w_i 越大，属性 X_i 对决策的影响越大。决策属性的重要度总为 1。

考虑更加一般的情况，传感器网络运行在有环境影响的情况下，可以设置属性可靠度参数 ε_i 进一步调整属性在分类算法中的影响。通过在算法中设置状态表，记录传感器节点的工作状态，可以得到属性的可靠度参数。

$$\varepsilon_i = \sqrt{\frac{i\text{类别正常级节点}}{i\text{类别总节点}}} \tag{8}$$

显然，在无环境因素影响时，可靠度参数 ε_i 总为 1。

在训练过程中，计算获得条件 X_i 的后验概率 $P(X_i \mid H)$ 后，利用赋予 X_i 的重要度和可靠度对 $P(X_i \mid H)$ 进行调整：

$$P'(X_i \mid H) = w_i\varepsilon_i P(X_i \mid H) \tag{9}$$

得到一系列调整后的后验概率 $P'(X_i \mid H)$，再

使用该后验概率进行类别 H 后验概率 $P(H \mid X_i)$ 的计算。

算法　NBCBAS

步骤 1　混合探测信号加工处理。将有效混合探测信号进行合并、清除冗余、规范化规约等操作。获取质量较优的目标特征向量。

步骤 2　系统进行节点检测，刷新无效信号接收表和状态表中的节点状态。

步骤 3　根据待分类特征向量的属性查询概率列表，获取各属性的先验概率和后验概率。

步骤 4　根据待分类特征向量的属性查询重要度列表，获取各属性的重要度值。

步骤 5　查询类别概率列表，获取各类别先验概率。

步骤 6　根据该属性类别传感器中正常级所占比例，计算属性的可靠度。

步骤 7　计算每个类别对于各个属性的后验概率，并计算该待测目标属于各类别的概率。

步骤 8　通过比较待测目标属于各类别的概率大小，得出分类结果。

步骤 9　分类结束。

3　实验分析

3.1　标准数据集分类实验

十折交叉验证是常用的分类算法精度测试方法，它的基本思想是将数据集分成 10 份，轮流将其中的 9 份作为训练数据，1 份作为测试数据进行实验。每次实验得出相应的正确率，将 10 次实验正确率的平均值作为对算法精度的估计。

本文使用 UCI 机器学习数据库中的 16 个标准数据集，采用十折交叉验证方法分别对贝叶斯网络分类器(BN)、朴素贝叶斯分类器(NBC)、AdaBoost 方法获得的朴素贝叶斯分类器(A-NBC)以及本文提出的基于属性重要度的朴素贝叶斯分类器(NBCBAS)进行测试，测试结果如表 1 所示（加粗表示最高精度）。

根据表 1 中的实验数据，对本文算法做出如下分析。

1) 表 1 中，本文算法的准确率最高的有 13 个数据集。从它们的特点来看，本文算法适合于处理类别个数较少的数据集；同时适合处理含有较多条件属性且属性间依赖关系强的数据集。对于有此特点的数据集，本算法的分类效果具有明显优势。

表 1　　**算法准确率比较**

序号	数据集（样本数-属性数）	算法的准确率/%			
		BN	NBC	A-NBC	NBCBAS
1	Annealing(798-38)	92.427 6	75.835 2	79.064 6	**92.761 7**
2	Balance Scale(625-4)	72.320 0	90.400 0	91.040 0	**91.200 0**
3	Lenses(24-4)	70.833 3	70.833 0	70.833 3	**83.833 3**
4	Statlog-Heart(270-13)	81.111 1	83.703 7	82.592 6	**86.296 3**
5	Hayes-Roth(160-5)	84.090 9	84.090 9	76.515 2	**88.636 4**
6	Hores Colic(368-27)	76.550 4	76.302 1	73.828 1	**77.864 6**
7	Iris(150-4)	92.666 7	96.000 0	93.333 3	**97.333 3**
8	Liver Disorders(345-7)	56.811 6	56.811 6	67.246 4	**71.884 1**
9	Lymphography(148-18)	87.837 8	87.162 2	80.405 4	**94.594 6**
10	Lung Cancer(32-56)	93.750 0	87.500 0	65.625 0	**100.000 0**
11	ORH Digits(5 620-64)	91.774 9	91.794 0	91.334 5	**92.437 7**
12	Page Blocks(5 473-10)	**93.400 0**	90.846 0	90.846 0	90.969 0
13	Segmentation(2 310-19)	**90.400 0**	81.066 7	80.216 5	81.200 0
14	Sonar(208-60)	80.288 5	67.788 5	80.769 2	**82.211 5**
15	SPECT Heart(267-22)	83.750 0	83.000 0	71.250 0	**90.000 0**
16	Waveform(5 000-21)	79.840 0	**80.000 0**	**80.000 0**	79.960 0

2) 在 lung Cancer 数据集上，由于属性个数较多而样本数量较少，所以出现了 100%准确的情况。在 Page Blocks、Segment 和 Waveform 数据集上，属性间的依赖关系较弱，根据本文算法得到的各条件属性与决策属性之间的相关系数较小，导致各条件属性的重要度基本相同，因此分类效果未得到显著提高。

3) 从整体结果上看，本文算法比朴素贝叶斯算法的分类正确率高，这充分说明了本文在属性重要度上的改进对分类效果有较强的影响。它避免了噪声较强的属性与其他属性在分类中具有相同地位的弊端。

综上所述，NBCBAS 算法确实改善了 NBC 算法的分类准确率，并且在具有属性多、属性间依赖关系强等特点的数据集上效果更加显著。

3.2　混合探测信号的目标分类

3.2.1　实验数据

实验数据集来自实验室搭建的分布式传感器网络的探测信号。传感器网络由空气传感器（化学气体传感器）、温度传感器、声音传感器和振动传感器组成。网络中部署每种传感器各 10 个节点。

空气传感器（MQ 系列）主要用于测量空气中 CO 浓度（测量范围 0~300ppm）和 CH_4（测量范围 0~5 000ppm）浓度，同时可辅助测量 SO_2 和 CO_2 浓度。

温度传感器（DS 系列）主要用于测量空气温度（测量范围−55℃~125℃），同时可辅助测量空气

湿度和空气压力。

声音传感器（TZ 系列）主要用于测量空气中声音及其他噪声频率（测量范围 20~20 000Hz）。

振动传感器（M3 022）主要用于测量地面振动频率（测量范围 3~20 000Hz）和振动加速度（测量范围 0.02 m/s^2 ~316 m/s^2 ）。

数据集中每个样本都由上述传感器网络采集获得。根据实验的不同要求，每次提取每种类别目标的信号样本 2 500 条。再采用十折交叉验证法进行实验。

3.2.2 实验设计

为达到全面检测分类效果的目的，本文共设计 5 个实验，实验的相关参数如表 2 所示。

表 2 分类实验的主要参数

序号	数据样本数量	类别数量	决策属性数量	条件属性数量	类别先验概率
1	10 000	4	4	2	0.25
2	10 000	4	4	2	0.25
3	20 000	8	4	2	0.125
4	10 000	4	4	6	0.25
5	20 000	8	4	6	0.125

表 2 中各组实验都具有不同的测试目的。

1) 实验 1 主要测试属性和类别都较少且样本差异较大时的分类效果。

2) 实验 2 主要测试属性和类别都较少但样本差异较小时的分类效果。

3) 实验 3 主要测试属性较少但类别较多时的分类效果。

4) 实验 4 主要测试属性较多且类别较少时的分类效果。

5) 实验 5 主要测试属性和类别都较多时的分类效果。

3.2.3 实验结果

使用 NBCBAS 算法采用十折交叉验证法进行上述实验并对比结果。分类精度效果如表 3 所示。

表 3 对比实验的分类结果

序号	分类精度/%	
	NBC	NBCBAS
1	96.85	99.95
2	92.30	96.25
3	88.15	92.75
4	94.00	98.95
5	93.70	97.50

根据实验结果，可以做出如下分析。

1) 5 组实验中，NBCBAS 算法的分类精度均高于 NBC 算法的分类精度。由此看出，本文算法通过对属性重要度的分析以及对属性概率的调整，确实提高了分类器的分类精度。

2) 实验 1 中本文算法具有较高的精度是由于本文算法适用于类别较少的数据集，同时该数据集类别间差异较大，分类效果的提高十分明显。

3) 实验 4 和实验 5 中本文算法同样得到较高的精度是由于本文算法适用于条件属性较多的数据集，同时条件属性与决策属性的相关性越强，分类效果的提高越明显。

4) 实验 3 中分类精度略低于其他实验是由于该实验类别较多但条件属性较少，使得对于条件属性重要度的学习在整个分类过程中的影响程度减小，无法显著提高分类效果。

3.3 有环境影响时混合探测信号的目标分类

3.3.1 实验设计

在有环境因素影响时，算法添加状态表，并在每次有效信号采集的同时计算正常级节点所占比例。其中，非正常级节点的增加会带来数据噪声的增大。

依照无环境影响时的实验设计，在此设计 10 组实验，实验参数如表 4 所示。

表 4 考虑环境影响时分类实验的主要参数

序号	数据样本数量	类别数量	决策属性数量	条件属性数量	类别先验概率	正常级节点比例
1	10 000	4	4	2	0.25	较高
2	10 000	4	4	2	0.25	较高
3	20 000	8	4	2	0.125	较高
4	10 000	4	4	6	0.25	较高
5	20 000	8	4	6	0.125	较高
6	10 000	4	4	2	0.25	较低
7	10 000	4	4	2	0.25	较低
8	20 000	8	4	2	0.125	较低
9	10 000	4	4	6	0.25	较低
10	20 000	8	4	6	0.125	较低

上述实验可以看作前后 5 组实验的对照实验。前 5 组实验针对正常级节点比例较高时的探测信号数据集进行分类；后 5 组实验针对正常级节点比例较低时的探测信号数据集进行分类。前后 5 组实验的其他参数均与无环境影响时参数相同。

3.3.2 实验结果

使用 NBCBAS 算法采用十折交叉验证法进行上述实验并对比结果。分类精度效果如表 5 所示。

表 5　考虑环境影响时对比实验的分类结果

序号	分类精度/%	
	NBC	NBCBAS
1	95.35	98.95
2	92.50	96.45
3	86.45	91.60
4	94.15	98.75
5	95.20	97.65
6	90.85	98.55
7	84.60	94.15
8	79.00	86.85
9	87.05	96.95
10	90.70	94.50

根据实验结果，可以得到如下分析。

1) 在有环境影响的实验中，本文算法得到的分类精度均高于 NBC 算法。再次验证了本文算法较 NBC 算法在分类精度上有显著提高。

2) 对照前后 5 组实验可以发现，NBC 算法在正常级节点减少，噪声增大时会受到较为明显的影响；本文算法虽然也会受到一定影响，但仍能够保持较高的分类精度。可以证明有环境影响时，本文算法的可靠度调整具有比较明显的效果。

NBCBAS 算法在对混合探测信号进行分类时能够表现出较好的精度水平，尤其在对条件属性与决策属性关系较紧密、属性数量较多的数据集中体现明显。

在有环境因素影响时，NBCBAS 算法中的属性可靠度参数能够使分类精度保持在较高的水平，与 NBC 算法相比，没有出现很大的精度下降现象。

4　结束语

本文提出了 NBCBAS 算法，该算法继承了 NBC 算法的优势，同时利用属性重要度修正概率的方式弥补了类独立假设带来的缺陷。该方法在完成对混合探测信号的分类时获得了较高的分类精度。同时，NBCBAS 算法设置属性可靠度参数可以在有环境因素影响时减小精度的损失，保持较好的分类效果。

在本文现有工作的基础上，未来还将对以下内容进行进一步研究：继续研究不同类别的传感器受到不同类型环境影响的问题；不断改进方法的适用性和一般性；在降低开销的前提下，可以更有效地达到监控传感器节点的效果。

参考文献：

[1] GUEVARA J, BARRERO J, VARGAS F, *et al*. Environmental wireless sensor network for road traffic applications[J]. Intelligent Transport Systems, 2012, 6(2):177-186.

[2] SEKHAR V C, SARVABHATLA M. Security in wireless sensor networks with public key techniques[A]. International Conference on Computer Communication and Informatics[C]. Coimbatore, 2012. 1-16.

[3] JAWHAR I, MOHAMED N, AQRAWAL D P. Linear wireless sensor networks: classification and applications[J]. Journal of Network and Computer Applications, 2011,34(5):1671-1682.

[4] BROOKS R R. Distributed target classification and tracking in sensor networks[J]. Proceedings of the IEEE,2003,91(8):1163-1171.

[5] DUARTE M F, HU Y F. Vehicle classification in distributed sensor networks[J]. Journal of Parallel and Distributed Computing, 2004, 64(7): 826-838.

[6] COMMAULT C, DION J M, TRINH D H. *et al*. Sensor classification for the fault detection and isolation a structural approach[J]. International Journal of Adaptive Control and Signal Processing, 2011, 25(1): 1-17.

[7] GONG B, IM J, MOUNTRAKIS G. An artificial immune network approach to multi-sensor land use/land cover classification[J]. Remote Sensing of Environment, 2011,115(2):600-614.

[8] LIU L. A binary-classification-tree based framework for distributed target classification in multimedia sensor networks[C]. INFOCOM, Orlando, 2012.594-602.

[9] ABOELELA E H, KHAN A H. Wireless sensors and neural networks for intruders detection and classification[A]. International Conference on Information Networking[C]. 2012.138-143.

[10] GHARAIBEH K M, YAQOT A. Target classification in wireless sensor network using partical swarm optimization[A]. Sensors Applications Symposium(SAS)[C]. 2012.1-5.

作者简介：

李侃（1975-），男，辽宁大连人，博士，北京理工大学教授、博士生导师，主要研究方向为数据挖掘、机器学习与分布式系统。

许航（1990-），男，北京人，北京理工大学硕士生，主要研究方向为机器学习。

黄忠华（1965-），男，吉林延吉人，博士，北京理工大学副教授，主要研究方向为智能信息系统。