计算机病毒检测中 Bayes 分类技术探究分析

石 雷
（中国人民解放军 66483 部队，北京 100093）

摘 要：贝叶斯分类算法是利用概率统计知识进行分类的分类算法，因此如何获得概率的初始知识是该算法进行分类的一个难点。这类算法利用贝叶斯定理来计算未知类别样本所属类别的可能性。本文分析了计算机病毒检测中应用贝叶斯分类技术的方法。

关键词：计算机病毒；检测；Bayes 分类技术
中图分类号：TP309.5 **文献标识码：**A **文章编号：**1007-9599（2011）24-0097-01

Bayes Classification Techniques Analysis of Computer Virus Detection

Shi Lei
(66483 Troops of PLA,Beijing 100093,China)

Abstract: Bayesian classification algorithm is the use of probability and statistics knowledge classification algorithm,so the initial knowledge of how to obtain the probability that the algorithm to classify a difficult.The type of algorithm uses Bayes'theorem to calculate the possibility of unknown class sample category.This paper analyzes the application of Bayesian classification techniques in computer virus detection method.

Keywords: Computer viruses;Detection;Bayes classification techniques

贝叶斯分类是结合了统计学和贝叶斯网络的分类方法，它基于如下假定：待考察的变量遵循某种概率分布，且可以根据这些概率及已观察到的数据进行推理，以做出最优决策。贝叶斯分类器可以发现变量间的潜在关系，预测类成员变量的可能性，即给定样本属于某个类的概率。

一、贝叶斯分类的原理

贝叶斯定理成为依赖性很强的独立性假设前提，因此，属性之间的独立性是分类准确与否的另一个难点和重点。同时确定贝叶斯最优假设的计算代价比较大（与候选假设数量成线性关系）。

在朴素贝叶斯学习算法中一般用出现频度代替概率，则可知道概率的初始知识，即可进行分类。朴素贝叶斯算法分类的准确度取决于属性之间的独立性，独立性好的准确度高，否则偏低。另外和决策树相比，朴素贝叶斯没有分类规则输出。由于贝叶斯算法过于依赖属性相互之间的独立性，为了减少对独立性的依赖，TAN 算法被提出。TAN 算法通过发现属性间的关联来减少朴素贝叶斯中对任意属性间独立性的依赖。TAN 在朴素贝叶斯网络结构基础上通过增加属性对之间的关联来实现。由于 TAN 算法考虑了两两属性的关联性，该算法对属性间的独立性依赖有一定程度的减少，但是可能存在的其他方面的关联性并未涉及，因此适用范围有限。

二、基于朴素贝叶斯分类的异常检测方法

设 X 是类标识号未知的数据样本，如 X 可以表示为流量异常，H 为某种假定，p 可以表示为系统当前遭受人侵，则我们可以确定 p（H/X）。p（H/X）表示当流量异常时，当前系统遭受入侵的概率。p（H/X）是后验概率，即条件 X 下 H 的后验概率。p（H）表示系统遭受入侵的概率，p（X/H）表示系统遭受入侵时流量发生异常的概率。p（X）表示流量发生异常的概率。其中 p（X）、p（H）、p（X/H）是可以根据历史数据计算得到的，在这种情况下，贝叶斯定理非常有用，我们可以根据这个定理计算出后验概率 p（H/X）。

$$p(H/X)=\frac{p(x/H)p(H)}{p(X)}$$

贝叶斯算法本身比较简单，下面简单介绍一下贝叶斯算法在入侵检测系统中的实际应用。在入侵检测系统中，数据样本往往具有多个属性，每个属性表示系统不同方面的特征（如流量情况、磁盘 I/O 的活动情况，系统中页面出情况），如样本 X 有 A1，A2…An，Ai 的值为 0 或者 1，0 表示正常，1 表示异常．i 为整数且大于等于 1 小于等于 n。H1，H2 分别表示为当前系统正遭受人侵和当前系统正常。这样我们要判断当前系统是否遭受入侵，只要计算 p（H1/X）和 p（H2/X）的值，如果 p（H1/X）大于 p（H2/X），则表示当前系统正遭受入侵，如果 p（H1/X）小于 p（H2/X），则表示当前系统正常。根据贝叶斯定理，我们计算 p（H1/X）时，先要计算 p（X/H1），而要计算 p（A1，A2，…，An/H1），计算 p（A1，A2，…，An/H1）的开销非常大，为降低计算 p（A1，A2，…，An/H1）的开销，可以做条件独立的朴素假定。即假定条件 A1，A2，…，An 是彼此独立的，属性间不存在依赖关系。这样

$$p(X/H1)=\prod_{k=i}^{n}p(Ak/H1)$$

同理，我们可以计算出 p（XH/(-)，这样我们就可以根据贝叶斯定理计算出 p（H，/X）和 p（H2/X）了。需要注意的是，我们在使用这个方法时要对状态数据进行预处理，如网络流量超过一定值时就认为是异常（该属性值记为 1），否则为正常（值记为 o）。

三、方法评价

基于贝叶斯聚类异常检测方法通过在数据中发现不同类别数据集合，这些类反映-厂基本的因果机制（同类的成员比其他的更相似），以此就可以区分异常用 p 类，进而推断入侵事件发生来检测异常入侵行为。Cheeseman 和 Stutz 在 1995 年开发的自动分类程序(AutoClassProgram)是一种监督数据分类技术。AutoClass 实现了使用贝叶斯统计技术对给定的数据进行搜索分类。这种方怯尽可能地判断处理产生的数据，没有划分给定数据类别，但是定义了每个数据成员。其优点是：(1)根据给定的数据，AutoClass 自动地判断决定尽可能的类型数目；（2）不要求特别相似测量、停顿规则和聚类准则；(3)可以自由地混合连续的及离散的属性。统计入侵异常检测对所观测到的行为分类处理。到目前为止，所使用到的技术主要集中于监督式的分类，这种分类是根据观测到的用 p 行为建立起用 p 轮廓。而贝叶斯分类方法允许最理想化的分类数、具有相似的轮廓的用 p 群组以及遵从符合用 p 特征集的自然分类。但是，该方法比较新，在入侵检测系统中还没有实现测试。自动分类程序怎样处理好固有的次序性数据（如审计跟踪）以及将统计分布特性植入分类中等方面，效果并不是十分明显。当自动分类程序支持处理在线数据时，对新的数据能否递增式地分类或者是否立即需要全部输入数据等这些问题的处理还尚未定论。由于统计的固有的特性，自动分类程序还存在选定合适的异常阈值和用 p 逐步地影响类型分布能力的困难。

参考文献：

[1]李伟光,黄常全.Bayes 分类技术在计算机病毒检测中的研究[J].网络安全技术与应用,2010,(9):9-11

[2]张波云,殷建平,蒿敬波等.基于多重朴素贝叶斯算法的未知病毒检测[J].计算机工程,2006,32(10):18-21

[3]文琪,彭宏,徐志根等.基于粗糙集和贝叶斯分类器的病毒程序检测[J].西南交通大学学报,2005,40(5):659-662,672