文章编号: 1000-5862(2012)03-0284-04

基于马氏距离的 *K* 均值聚类算法的入侵检测

易 倩, 滕少华*, 张 巍

(广东工业大学计算机学院, 广东 广州 510006)

摘要: 经典的 *K* 均值聚类算法是基于欧式距离的, 它只适用于球形结构的聚类, 而且在处理数据时不考虑变量之间的相关性和各变量的重要性差异. 针对以上问题改进了 *K* 均值聚类算法, 将马氏距离与 *K* 均值相结合, 并在目标函数中增加变量权重因子和协方差矩阵调节因子, 利用马氏距离优点有效地解决了 *K* 均值聚类算法的缺陷, 最后通过实验证实了该方法的可行性和有效性.

关键词: *K* 均值; 马氏距离; 聚类; 入侵检测

中图分类号: TP 393.08 文献标志码: A

0 引言

聚类分析是一种重要的入侵检测方法, 它利用某种相似性度量, 把一个未知类别的样本集组织成若干个有意义的子集, 相似度较高的样本归为一类, 相似度较小或不相似的样本则在不同的类中[1]. 通过这样的划分, 可将网络流量样本集中的正常数据和异常入侵数据区分开来. 当前的聚类算法大多采用距离作为样本间的相似性度量, 这是一种样本间的模糊关系, 反映样本间的相似程度.

近年来, 多位学者对聚类分析算法进行了研究, 陈媛媛等[2]建立了一种优化初始中心点的相似度概率模型, 利用一种二分法快速确定*K*的最优值, 获得了较好的聚类结果. 田彦山[3]提出一种基于山峰聚类确定聚类数目上限的算法, 提高了聚类效率. 马氏距离聚类分析方法也引起了众多学者的关注. 朱惠倩[4]提出了加权马氏距离, 并证明了其在理论上的有效性. Xiang Shiming 等[5]论述了马氏距离在聚类与分类中的优点, 可以使相似数据点的距离较近. 吴相华等[6]对聚类分析中马氏距离的协方差矩阵的估算进行了改进, 提高了聚类准确率. 张翔等[7]构造了一种特殊的基于马氏距离的可能性聚类方法, 并通过图像分割实验和标准数据集实验验证了算法的优越性. 马氏距离聚类方法的发展使之在聚类分析及图像处理等领域有着广泛的应用.

经典的 *K* 均值聚类算法采用欧式距离度量不同样本间的相似程度. 欧氏距离将样本的不同属性(即各指标或各变量)之间的差别等同对待, 这一点有时不能满足实际要求. 因为对于每种攻击来讲, 不同的属性对于入侵的支持度不同, 即一种属性可能比另一种属性更能说明某种攻击的发生[6]. 例如单位时间某连接出现 RST 包的数目比单位时间某连接传输的数据量的变化更能说明系统是否发生了攻击. 在网络入侵检测中, 有多达 37 种不同的攻击类型. 每条连接记录用同样的特征属性描述, 只是属性取值不同, 很明显, 同一属性对不同类型的攻击贡献不同, 不同属性对同一种攻击类型的支持度也不同, 一种属性可能比另一种属性更能说明发生了某种异常[8]. 比如, 在应对 UDP 攻击检测时, 分片错误数和紧急包数比连接持续时间更能说明是否出现了攻击行为, 但按照算法中的欧式距离进行度量时, 这 2 个属性在判断攻击时的作用度是一样的. 同时欧式距离忽略变量相关性的影响, 将使某些信息被重复计算, 从而使聚类结果与实际情况产生偏差. 针对上述缺点, 本文对*K*均值聚类算法进行改进, 将欧氏距离用马氏距离替代, 并在目标函数中增加变量权重因子. 实验结果表明该方法提高了入侵检测率,

收稿日期: 2012-01-22

基金项目: 广东省自然科学基金(06021484, 915100900100007)和广东省科技计划(2008A0602011)资助项目.

作者简介: 滕少华(1964-), 男, 江西南昌人, 教授, 博士, 主要从事数据挖掘和网络安全的研究.

是可行有效的.

1　K 均值聚类算法

K 均值聚类算法的基本思想是: 对于给定的数据对象集 X 和指定的聚类个数 K, 首先随机选取 K 个点作为初始聚类中心; 对于剩下的每个数据对象, 计算其到各聚类中心的距离, 并将该数据对象赋给离它最近的聚类中心所在的类; 然后重新逐一计算每个类的平均值, 从而得到新的聚类中心. 重复指派和更新步骤, 直到准则函数收敛使得平方误差值最小, 至此算法结束, 得到最终的数据聚类结果. 若需把数据集 X 划分为 K 类, 首先从数据集中随机选取 K 个样本点作为第 1 次迭代的聚类中心, 然后计算其余各个样本到聚类中心的距离, 把样本标记为距离它最近的那个聚类中心所属的类. 重新计算每个聚类的样本平均值, 得到新的聚类中心. 比较相邻2次的聚类中心是否发生变化, 若聚类中心相同, 说明样本调整已完成, 算法结束, 否则修改聚类中心, 进入下一次迭代, 直到聚类中心不再发生变化, 即平方误差准则函数趋向于收敛, 此时所有样本被划分到正确的类, 聚类算法结束.

平方误差准则函数为

$$J_C=\sum_{j=1}^{k}\sum_{x_i\in C_j}\left\|\boldsymbol{x}_i-\boldsymbol{c}_j\right\|^2, \tag{1}$$

其中 $\boldsymbol{x}_i$ 表示样本对象, $\boldsymbol{c}_j$ 表示聚类簇 C_j 的质心, J_C 则表示数据集中所有对象的误差平方和, 该目标函数采用欧氏距离. 具体算法描述如下:

(1) 算法的输入: 聚类个数 K 及包含 n 个数据对象的样本集.

(2) 算法的输出: 满足方差最小标准的 K 个聚类.

(3) 算法的具体过程:

(i) 任意选择 K 个对象作为初始聚类中心;

(ii) repeat

① 根据簇中对象的平均值, 将每个对象(重新)赋给距离最近的族(即最相似);

② 更新簇的均值;

(iii) until 聚类中心不再发生变化.

2　基于马氏距离的 K 均值聚类算法

马氏距离是由印度统计学家马哈拉诺比斯(P. C. Mahalanobis)提出的, 表示数据的协方差距离. 它是一种有效的计算 2 个未知样本的相似度的方法. 该距离计算仅涉及协方差矩阵的求逆, 不再和特征矢量的维数有关, 而和样本数目有关, 因此在高维特征空间中带来计算上的优势(马氏距离包含一个 $N\times N$ 的协方差矩阵的逆矩阵, N 表示样本数目, 在高维特征空间中有计算优势), 与欧式距离不同的是它考虑到各种特性之间的联系并且是独立于测量尺度的.

对于一个样本集 X_i $(i=1,2,\cdots,n)$, 其中的 2 个样本 $\boldsymbol{x}_i$ 和 $\boldsymbol{x}_j$ 之间的马氏距离[9]定义为

$$d_{ij}=(\boldsymbol{x}_i-\boldsymbol{x}_j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{x}_j), \tag{2}$$

其中 $\boldsymbol{\Sigma}$ 是样本的协方差.

基于马氏距离的 K 均值聚类算法是在度量样本之间的相似度时用马氏距离代替欧式距离, 并把属性的权重引入目标准则函数, 同时在马氏距离的基础上增加协方差矩阵调节因子 $-\ln\left|\boldsymbol{\Sigma}^{-1}\right|$, 从而改进后的目标函数为

$$J_C(\boldsymbol{\Sigma},X)=\sum_{j=1}^{k}\sum_{i=1}^{n}\boldsymbol{w}_i(\boldsymbol{x}_i-\boldsymbol{c}_j)^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}_i-\boldsymbol{c}_j)-\ln\left|\boldsymbol{\Sigma}^{-1}\right|, \tag{3}$$

$$\boldsymbol{w}_i=(n_{x_{i1}},n_{x_{i2}},\cdots,n_{x_{ip}})/n, \tag{4}$$

其中 $\boldsymbol{x}_i=(x_{i1},x_{i2}\cdots,x_{ip})$, x_{if} 表示 x_i 的属性分量 f, $\boldsymbol{x}_i$ 共有 p 个分量, $x_{1f},x_{2f}\cdots,x_{nf}$ 是分量 f 的 n 个测量值. $\boldsymbol{c}_j=\dfrac{1}{n_j}\sum_{x_i\in\boldsymbol{c}_j}x_i(j=1,2,\cdots,k)$ 是类 C_j 中样本的平均值, 代表第 j 类的聚类中心, n_j 是属于类 C_j 的样本数量. $\boldsymbol{w}_i$ 是关于向量 $\boldsymbol{x}_i$ 的各个属性分量的权重, 以各分量 x_{if} 在 $(x_{1f},x_{2f}\cdots,x_{nf})^{\mathrm{T}}$ 中出现的数目 $n_{x_{if}}$ 及样本总数目来估计.

3　入侵检测实验验证

3.1　选取数据集

为了验证上述改进的K均值聚类算法在入侵检测中的可行性和有效性, 选取KDD Cup99数据集[10]进行试验. 由于整个数据集有近500万条记录, 通常选用一个10%的子集测试算法性能. 为了使测试数据更加符合真实环境的情况, 把数据集的异常数据比例控制在5%.

从样本集中选取了 2 组数据, 每组各 1 万条记录, 其中有 500 条攻击数据, 满足聚类假设: 正常数据远远多于入侵数据. 第 1 组样本数据选自 KDD99 数据集中的 10%训练集, 第 2 组则选自测试子集(Corrected)包含训练集中未出现过的一些攻击类型, 即未知攻击.

3.2 数据属性处理及标准化

KDD数据集的每条记录有41个属性, 其中有13个属性最为重要[11], 既包含了数值型也包含了符号型属性. 对符号性属性进行数值化处理, 用一组整数代表不同的符号类型, 这些整数只用于数据处理, 并不代表任何特定的顺序. 例如属性protocol_type存在3种属性值为tcp、udp、icmp, 分别用1、2、3代替; 对属性service, 字段取值有66种情况, 分别用1~66代替. 这样经过转化后数据集只包含数值型属性, 由于不同的属性特征有不同的量纲, 为了避免对度量单位选择的依赖, 对数据进行如下的标准化处理[12]:

(1)计算均值绝对偏差 (mean absolute deviation) s_f, 其计算公式为

$$s_f = \left(\left|x_{1f} - m_f\right| + \left|x_{2f} - m_f\right| + \cdots + \left|x_{nf} - m_f\right|\right)/n, \quad (5)$$

其中$x_{1f}, \cdots, x_{nf}$为 f 的n个度量值, m_f 为 f 的均值, 即

$$m_f = (x_{1f} + x_{2f} + \cdots + x_{nf})/n. \quad (6)$$

(2)计算标准度量值或 $z-score$:

$$z_{if} = (x_{if} - m_f)/s_f. \quad (7)$$

均值绝对偏差 s_f 比标准差 σ_f 对于孤立点具有更好的鲁棒性. 对每个样本记录都按照(5)~(7)式进行计算并得到新的数据. 这相当于利用统计特性将原始实例的属性映射到一个标准的属性空间上, 更有利于聚类.

3.3 实验结果及分析

实验环境配置: Win XP, VC++6.0, CPU 2.0 GHz, 2.0 GB内存.

对实验的结果以 2 个参数检测率(*DR*)和误检率(*Fdr*)作为评价标准. 在算法实现过程中忽略攻击类型标识属性, 其仅供算法结果分析之用. 仿真实验结果如表 1 所示.

表 **1** 基于马氏距离的 ***K*** 均值聚类实验结果

K	*K* 均值算法		基于马氏距离的 *K* 均值算法	
	DR/%	*Fdr*/%	*DR*/%	*Fdr*/%
8	48.4	1.3	52.2	1.1
12	57.8	1.9	61.4	1.5
16	69.3	2.6	73.5	2.3
20	76.1	3.7	78.1	3.5
24	80.2	5.4	84.9	4.7

由于 *K* 均值聚类算法需要预先输入聚类个数, 而不同的聚类个数对聚类的效果影响很大. 随着初始聚类个数的增加, 检测率和误检率也有所增加, 通过比较实验结果可以看到, 改进后的算法由于采用了马氏距离, 考虑变量的相关性并在计算时赋予属性变量不同权重, 提高了检测率, 可以得到更好的聚类效果. 虽然误检率随着检测率的提高有所上升, 但总体相对改进之前较低.

4 结束语

本文在 *K* 均值聚类算法的基础上针对欧式距离的缺点提出一种改进的算法, 将欧式距离用马氏距离替代, 并在聚类准则函数中加入变量权重因子. 马氏距离在聚类时记录了各个样本间的相互关系, 在数据相关的数据集中可以提高聚类精度, 减少聚类中心的误差. 通过在目标函数中增加协方差矩阵调节因子, 从而能完成非球型或椭圆型分布的数据集的聚类, 尤其在处理相关性比较大的数据集时具有比欧式距离更好的扩展性. 可以减少样本相关性的干扰, 提高聚类准确率, 马氏距离与属性的量纲无关, 即两点之间的距离不受样本的测量单位的影响, 因此可以减少样本与聚类中心的误差. 协方差矩阵包含样本集总体分布的相关性, 能更准确地反映不同性质的数据之间的关系, 因此改进的算法对椭圆形等非球形数据集有很好的聚类能力, 适合处理相关性比较大的数据集. 由于聚类时需要计算一个 N 维特征向量的协方差矩阵, 有 $N(N-1)/2$ 个参数, 在样本数目特别多的情况下, 聚类速度受到影响. 如何改进协方差矩阵, 提高聚类精度和速度是需要继续研究的方向.

5　参考文献

[1] 罗军生, 李永忠, 杜晓. 基于模糊 *C* 均值聚类算法的入侵检测 [J]. 计算机技术与发展, 2008, 18(1): 178-180.

[2] 陈媛媛, 屈志毅, 张恒龙, 等. 一种初值优化的 *K*-均值文档聚类算法 [J]. 江西师范大学学报: 自然科学报, 2008, 32(2): 206-210.

[3] 田彦山. 基于山峰聚类的聚类上限确定方法 [J]. 江西师范大学学报: 自然科学报, 2007, 31(2): 134-137.

[4] 朱惠倩. 聚类分析的一种改进方法 [J]. 湖南文理学院学报: 自然科学版, 2005, 17(3): 7-9.

[5] Xiang Shiming, Nie Feiping, Zhang Changshui. Learing a mahalanobis distance metric for data clustering and classification [J]. Pattern Recognition, 2008, 42(12): 3600-3612.

[6] 吴香华, 牛生杰, 吴诚鸥, 等. 马氏距离聚类分析中协方差矩阵估算的改进 [J]. 数理统计与管理, 2011, 30(2): 240-245.

[7] 张翔, 王士同. 一种基于马氏距离的可能性聚类方法 [J]. 数据采集与处理, 2011, 23(8): 86-88.

[8] 王晓峰, 沈庆浩. 利用聚类算法找出新的攻击 [J]. 华东理工大学学报: 自然科学版, 2004, 30(3): 288-291.

[9] 张尧庭, 方开泰. 多元统计分析引论 [M]. 北京: 科学出版社, 1982.

[10] KDD99 Cup Dataset [EB/OL]. [2011-12-11]. http: //kdd. ics. uci. edu/databases/ kddcup99/kddcup99.html.

[11] Mukkamala S, Janoski G, Sung A H. Intrusion detection using suppoa vector machines and neural networks [EB/OL]. [2011-12-20]. http: //www. cs. uiuc. edu/class/fa05/cs591han/papers/mukkCNN02.pdf.

[12] Han Jiawei, Micheline Kambe. 数据挖掘概念与技术 [M]. 2 版. 北京: 机械工业出版社, 2006: 252-264.

Mahalanobis Distance-Based *K*-Means Clustering Algorithm for Intrusion Detection

YI Qian, TENG Shao-hua*, ZHANG Wei

(College of Computer ,Guangdong University of Technology, Guangzhou Guangdong 510006, China)

Abstract: The classic *K*-means clustering algorithm is based on the Euclidean distance, it applies only to spherical structure clustering and in the processing of data without regard to the correlation between variables and differences in the importance of each variable. To solve the above problem, this paper propose a feasible clustering algorithm, it combines Mahalanobis distance with the *K*-means and adds a variable weighting factor and a regulating factor of covariance matrix to each class in the objective function. Using the advantage of Mahalanobis distance, it effectively solves the shortcomings of *K*-means clustering algorithm. Experimental results of date clustering illustrate its feasibility and effectiveness.

Key words: *K*-mean; Mahalanobis distance; clustering; intrusion detection

(责任编辑: 冉小晓)