文章编号: 1672 - 4348 (2007) 04 - 0392 - 05

# 基于随机森林方法的异常样本检测方法

邱一卉, 林成德

(厦门大学信息科学与技术学院自动化系, 福建 厦门 361005)

**摘要**: 提出一种基于随机森林方法的异常样本 (outliers)检测方法。仿真实验表明,与其他 2种基于距离的异常样本检测技术相比,这种方法可以更好地提高模型的准确率,且具有较强的鲁棒性,在处理大规模数据集时还能显著地减少计算时间。

**关键词**: 异常样本检测; 随机森林; 马氏距离

**中图分类号**: TP274 **文献标识码**: A

## Outlier detection based on random forest

Qiu Yihui, Lin Chengde

(Automation Department, Information Science and Technology School, Xiamen University, Xiamen 361005, China)

**Abstract:** It introduces an outliers detection method based on random forest Compared with the other two common outliers detection methods based on distance, the proposed method can improve the performance and robustness of the model and can also reduce the computation time.

**Keywords:** outlier detection; random forest; Mahalanobis distance

## 0 引言

如何识别出数据集中的异常样本已经成为近年来数据挖掘研究的热点,例如在大量的银行企业信用数据中挖掘异常,去除建立信用评估模型时的噪声,提高模型的准确率。另外,数据挖掘技术还可以发现可能的异常信用欺诈行为。本文引入一种新的能较好容忍噪声的学习算法 ——随机森林 (random forests,简称: RF),介绍一种衡量样本相似程度的方法,结合样本的相似度 (proximity),提出异常点尺度 (outlier measure)的概念来量化度量样本的异常程度,并根据这一尺度筛选出异常样本。

## 1 异常样本检测

### 1.1 概述

异常样本是远离其他观测数据的样本,由于离得太远以致于产生怀疑,可能由一个不同的机制产生的。通常,异常样本定义为:集合中严重偏离大部分数据所呈现趋势的那些数据点。常规的异常样本检测有基于统计的检测方法,该方法通常需要预知样本的分布以及异常样本的数量;还有在实践中应用较多的基于距离的检测方法,该方法建立在距离或者核方法的基础上,可利用的距离概念有欧氏距离及马氏距离等。马氏距离能较好的消除不同数量级属性值差异的影响,广泛应用于各种基于距离的检测方法。下面介绍基于马氏距离的异常样本检测方法。

### 1.2 马氏距离方法

对于一个 $P$维的样本总体 G,假设其数学期望为 μ、协方差矩阵为 Σ,定义某一样本 $x$到样本总体的马氏距离为: $d(x, G) = [(x - \mu)^T \Sigma^{-1} (x - \mu)]^{1/2}$。全体数据集按不同类别可分为不同的样本总体 $G_i$ ($i = 1, 2, \cdots$),不同类的 $\mu_i$和 $\Sigma_i$不全相同,且 $\mu_i$和 $\Sigma_i$未知,此时可用它们的无偏估计代替,设 $x_1^{(k)}$, $x_2^{(k)}$, …, $x_{n_k}^{(k)}$为来自 $G_k$的样本:

收稿日期: 2007 - 06 - 20

第一作者简介: 邱一卉 (1983 - ), 女 (汉), 福建仙游人,硕士研究生,研究方向为智能计算方法及其应用.

$$\mu_k = \frac{1}{n_k}\sum_{i=1}^{n_k} x_i^{(k)} = \bar{x}^{(k)};$$

$$S_k = \frac{1}{n_k - 1}\sum_{i=1}^{n_k}(x_i^{(k)} - \bar{x}^{(k)})(x_i^{(k)} - \bar{x}^{(k)})^{\mathrm{T}}$$

利用异常样本的定义也可以对马氏距离方法进行改进,此时将所有的数据不分类别的当成样本总体,计算每个样本与样本总体的马氏距离,并以马氏距离的大小来衡量某个样本的异常程度。传统的马氏距离方法中由于异常样本的存在导致样本总体的期望和方差不稳定。最新的技术采用各种方法增加数值计算的稳定性,常用方法有最小协方差行列式法 (MCD)、多变量修剪法 (MVT)、最小半球体积法 (SHV)[1]、以及稳健的马氏距离方法[2]。以下简要介绍稳健的马氏距离方法。

稳健的马氏距离方法其实是改进的快速 MCD方法,从 $n$行 $p$维的数据中随机抽取 $h$个样本 ($h$通常取样本总数 $n$的一半),计算这 $h$个样本的均值 $\mu_1$和协方差矩阵 $\Sigma_1$,然后通过以下式子计算全体 $n$个样本到中心 $\mu_1$的马氏距离: $d_1(i) = \sqrt{(x_i - \mu_1)\Sigma_1^{-1}(x_i - \mu_1)}$;选出其中距离最小的 $h$个样本,重新计算它们的均值 $\mu_2$和协方差矩阵 $\Sigma_2$、以及每个样本距 $(\mu_2, \Sigma_2)$的马氏距离。可以证明 $\det(\Sigma_2) \leq \det(\Sigma_1)$,当且仅当 $\mu_1 = \mu_2$, $\Sigma_1 = \Sigma_2$ 时等号成立[3]; 重复上述过程, 直至 $\det(\Sigma_m) \leq \det(\Sigma_{m-1})$或者二者充分接近时停止迭代。根据此时得到的均值 $\mu_m$和协方差矩阵 $\Sigma_m$算出的马氏距离为稳健马氏距离。这个方法需要计算属性值的协方差矩阵的逆阵,如果协方差矩阵奇异,则使用伪逆来代替逆阵。

另一种较为简便的基于距离的方法是半数重采样方法 (RHM)。预先设定程序运行迭代次数 $N$。从原始 $n$行 $p$维的数据中随机抽取 $h$个样本 ($h$通常取样本总数 $n$的一半),计算这 $h$个样本的均值 $\mu_1$和协方差矩阵 $\Sigma_1$,利用这个均值和方差阵对原始矩阵进行标准化,计算每个样本的向量长度: $l(i) = \sqrt{\sum_{k=1}^{p}[(x_k - \mu_1(k))/\Sigma_1(k)]^2}$;对全体数据的向量长度进行排序,将长度较大的那些样本 (例如取距离最大的前 5%样本)进行异常标记;重新抽样,重复上述过程。迭代 $N$次之后检查每个样本被标记异常的次数,将被标记次数多的那部分样本视为异常样本。

## 2　利用随机森林方法检测异常样本

### 2.1　随机森林

随机森林[4]是一种组合分类器方法,构成随机森林的基本分类器是决策树。决策树是一种由结点和有向边组成的层次结构,树中包含 3种结点:根结点、内部结点、终结点。决策树仅有一个根结点,是全体训练数据集合。树中的每个内部结点是一个分裂问题,它将到达该结点处的样本按某个特定属性分块。每个终结点 (又称为叶结点)是带有分类标签的数据集合。从决策树的根结点到叶结点的一条路径就形成一个判别规则。决策树算法采用自顶向下的贪婪算法,每个内部结点选择分类结果最好的属性将到达该结点的数据分成 2块或者更多块,继续这个过程直至这棵树能准确的分类全部训练数据。决策树算法的核心问题是选择较优的分裂属性。选择分裂属性的标准很多,例如信息增益、信息增益比、*Gini*索引等,对应不同的属性选择方法决策树算法有 ID3、C4.5、CART等。

本文中决策树算法与 CART算法相似,其分裂属性的选择以 *Gini*指数为指标。*Gini*指数是一种不纯度分裂方法,它能适用于类别、二进制、连续数值等类型的字段,具体算法思想是:假设某结点 $t$处的数据样本集合 $T$包含 $k$个类别的记录,那么其 *Gini*指标为:

$$Gini(t) = 1 - \sum_{j=1}^{k}[p(j|t)]^2$$

其中 $p(j|t)$为类别 $j$在 $t$结点处的概率,当 $Gini(t)$最小为 0时, 即在此结点处所有样本都属于同一类别,表示能得到最大的有用信息;当此结点中的所有样本对于类别字段来说均匀分布时, $Gini(t)$最大,此时的有用信息最小。如果集合分成 $l$个部分, 那么进行这个分割的 *Gini*指数就是:

$$Gini(T) = \sum_{i=1}^{l}(n_i/n)Gini(i)$$

其中 $l$是子结点的个数, $n_i$是在子结点 $i$处的样本数, $n$是在母结点处的样本数。*Gini*指数的基本思想是:对于每个属性都要遍历所有可能的分割方法,若能提供最小的 $Gini_{split}$,就被选择作为此结点处分裂的标准;此时再按对应的属性值来分裂,并且根据每一个属性值创建树枝;进一步向下划分样本,直到满足停止条件,例如单个叶结点上的样

本都属于同一类或者叶结点的纯度(即该结点处包含某类样本的频数)满足某个阈值范围。预先设定阈值,当叶结点纯度超过阈值时停止划分,这个过程相当于对树进行剪枝。

随机森林重复上述的建树过程构建多个决策树的组合。首先设定森林中有 $M$ 棵树,即有 $M$ 个决策树分类器,且全体训练数据的样本总数为 $N$。使用 bagging方法,即通过从全体训练样本中随机地有放回的抽取 $N$ 个样本,形成单棵决策树的训练集。重复 $M$ 次这样的抽样过程分别得到 $M$ 棵决策树的学习样本。除了单棵决策树的学习样本是随机产生的,随机森林还将随机性加入到每棵树的生成过程之中。设样本共有 $Q$ 个属性,事先给定 $q < Q$ ($q$通常取 $Q$的平方根),在选择每个结点的分裂属性时,并不对所有的属性进行比较,而是从全体属性中随机选择 $q$个属性进行比较,选择其中分类结果较好的属性进行分裂。这样可以增加每棵树之间的差异度,从而提高森林的泛化误差。单棵决策树建造过程不进行剪枝,森林形成之后,对于一个新的样本,每棵树都得出相应的分类结论,最后由所有树通过简单多数投票决定分类结果。与其它组合分类技术相比较,当树的数目相当大时,随机森林并不易出现过拟合的现象。可以证明,其泛化误差的上界小于 $\bar{\rho}(1 - s^2)/s^2$,其中 $\bar{\rho}$是树之间的平均相关系数(代表各分类树之间的相关程度),$s$是单棵树的分类效能。

### 2.2 样本的相似度

随机森林是多个决策树的组合,可以用 2个样本在每棵树的同一个结点上出现的频率大小来衡量这 2个样本之间的相似程度,或 2个样本属于同一类的概率大小。对于样本数为 $N$ 的训练集,首先生成一个 $N \times N$的零元素矩阵 Prox。用每次生成的树对所有样本进行判别,每个样本将到达该树的某个叶结点上;对于任意两个样本 $n$和 $k$,如果样本 $n$与样本 $k$出现在该树的同一个叶结点上,则在 Prox矩阵相应的第 $n$行第 $k$列上加 1;重复这个过程直至 $M$ 棵树全部建好,得到相应的矩阵,将矩阵中的每个元素都除以树的数目进行归一化处理,得到最后的 Prox矩阵,它是一个对称且对角线元素为 1的矩阵,其第 $n$行第 $k$列的元素 $\mathrm{Prox}(n, k)$ 可定义为样本 $n$与样本 $k$的样本相似度。在随机森林建成的同时,也得到了 Prox矩阵。不难看出,如果数据集中某一类的样本数较多,则该类中的样本所对应的行通常都包含较多接近 1的元素,而那些包含较多接近零元素的行所对应的样本,我们有较大理由相信它们和其他样本的相似程度较小。由此,我们很自然的引出一种对样本异常程度的衡量方法 ——异常点尺度。

### 2.3 样本的异常点尺度

对于样本 $n$,定义它的原始异常点尺度为:

$$\mathrm{raw\,om}(n) = \mathrm{nsample}/\text{琣}(n)$$

其中,$\text{琣}(n) = \sum_{cl(k)=cl(n)} [\mathrm{prox}(n, k)]^2$,nsample为与样本 $n$同类的样本总数。在同一个类内,如果某个样本的 $\text{琣}(n)$ 值较低,则它的 raw om 值将很大。对每一类的所有样本,计算出该类所有样本的原始异常点尺度的均值 $\overline{\mathrm{raw\,om}}$ 以及方差 $\sigma$,标准化后得到每个样本的最终异常点尺度:$\mathrm{outlier\,measure}(n) = [\mathrm{raw\,om}(n) - \overline{\mathrm{raw\,om}}]/\sigma$。

上述变换,避免了由于各类样本数目相差较大而产生的数值上的差异,以便于比较各类样本的异常点尺度。随机森林建成后,按上述计算过程我们得到所有样本的异常点尺度,如果某样本的异常点尺度较大,这个样本与其他样本的相似程度较小,就可能存在异常。如果预知置信区间,可把异常点尺度超过某个阈值的样本视为异常点。例如本文的仿真实验中所使用的每个数据集都已知包含 5%的异常样本,因此可以通过对全体样本的异常点尺度进行排序,把其中最大的前 5%的样本视为异常样本。

## 3 仿真实验

### 3.1 对照实验介绍

UCI(University of Califomia · Irvine)机器学习数据库是著名的机器学习检验数据库,广泛应用于学习算法的建模与检验。本文使用常用的 6个标准数据集进行仿真实验对照,已知这 6个数据集的异常样本个数占全体样本总数的 5%。将本文提出的基于随机森林技术的异常样本检测方法与上面介绍过的 2种基于距离的检测方法(RHM和稳健马氏距离)进行对照,比较 3种方法剔除异常样本之后对所建模型的预测准确率的提高程度;此外,还通过采用支持向量机(SVM)[5]技术所建的模型比较了 3种方法的鲁棒性。

首先,分别用 3种方法在每个数据集中删除占总数 5%的“异常样本”,再将删除后的数据集用于建立随机森林模型。根据 6个数据集样本总

数不同,建立包含 500～1 000棵树规模的随机森林,并将每个结点处候选分裂属性个数 $q$设为该数据集属性总数的平方根,分别进行 5重交叉检验,得到如表 1所示的模型准确率实验结果。

表 1　用 3种不同方法删除异常样本所建模型的准确率比较

Tab. 1　The accuracy comparison between models based on 3 outlier-discarding methods

%

| 数据集 | 不删异常样本 | RF方法 | RHM方法 | 稳健马氏距离 |
|---|---|---|---|---|
| Sonar | 79. 03 | 87. 01 | 83. 03 | 80. 80 |
| Wine | 97. 49 | 99. 35 | 98. 67 | 97. 61 |
| Zoo | 95. 08 | 98. 49 | 97. 07 | 98. 07 |
| Heart | 81. 37 | 89. 35 | 83. 84 | 83. 22 |
| Breast-cancer | 96. 37 | 97. 67 | 96. 95 | 96. 72 |
| Waveform | 85. 42 | 87. 45 | 85. 90 | 85. 61 |

为了进一步比较 3种不同检测方法的鲁棒性,我们还进行了如下实验:分别对每个数据集进行 5重交叉,得到 [训练集 $i$], [测试集 $i$] ($i$ = 1, 2, …, 5)。用训练集建立 SVM模型,再用测试集进行测试,将所有 5次测试结果平均,得到 5重交叉的准确率;接下来对每个 [训练集 $i$] 分别用 3种方法剔除异常样本,用删除后的 [训练集 $i$ - ] 建立 SVM模型,再用未经删除异常样本的 [测试集 $i$] 进行测试,得到 5重交叉的准确率。SVM建模使用 libSVM工具箱,工具箱采用高斯核函数,并且用网格法选择最优的惩罚系数 $C$。实验结果如表 2所示.

表 2　3种不同删除异常样本方法的鲁棒性比较

Tab. 2　Robustness comparison among 3 outlier-discarding methods

%

| 数据集 | 不删异常样本 | RF方法 | RHM方法 | 稳健马氏距离 |
|---|---|---|---|---|
| Sonar | 82. 69 | 96. 08 | 83. 77 | 85. 01 |
| Wine | 97. 78 | 98. 89 | 97. 78 | 98. 68 |
| Zoo | 96. 00 | 99. 01 | 98. 61 | 97. 00 |
| Heart | 91. 18 | 92. 65 | 88. 24 | 88. 31 |
| Breast - cancer | 97. 14 | 98. 29 | 97. 24 | 97. 41 |
| Waveform | 86. 96 | 87. 04 | 86. 04 | 86. 72 |

### 3. 2　结果分析

3种方法在删除数据集中的异常样本之后,模型的准确率都有不同程度的提高,说明 3种异常样本检测方法均为有效。比较发现,基于 RF的异常样本检测方法较另 2种方法有更大的优越性,较大程度地提高了模型的准确率,同时具有更强的鲁棒性。

1) 3种方法删除全体数据集异常点的模型准确率和计算时间比较

在 sonar, wine, zoo小数据集的建模中, 3种方法都提高了模型的准确率,其中 RF方法对模型准确率的提高最为显著,分别提高了 2%到 6%不等。

在 breast-cancer, heart中等容量的数据集上, RF方法突显其优越性,比其他 2种方法的用时短,并且更加准确地定位异常样本。如表 1中所示,在 heart数据集中,模型的准确率提高了 8%,明显优于其他 2种方法,说明 RF方法识别出数据集中的大部分异常样本。

对于 waveform这样的大数据集,涉及较大的数值计算任务,要求算法尽量简单快捷。然而基于距离的检测方法都存在计算时间长,内存消耗多的问题。由于异常样本的存在,大数据集均值和协方差矩阵的计算需要更多的迭代次数,例如稳健马氏距离方法中,为得到稳健的均值和协方差矩阵,算法运行的迭代次数超过 150次,所需计算时间长。另外马氏距离的计算不仅需要较长的计算周期,而且其庞大的矩阵计算占用大量内存空间。仿真实验过程中发现, RHM算法运行程序时所用时间达到了运行 RF算法时的 $10^2$ 数量级

倍数,占用的内存也远比 RF算法大。如果数据集的属性值较多,这样的问题就更加突出。RF方法的主要运行时间在建模过程,而这个过程所需的时间通常较少,例如对 5 000个样本容量的数据集构建 1 000棵树规模的随机森林只需要十几秒。随机森林模型一旦建立,就可以很快得到每个样本的异常点尺度,它所涉及的计算都是建模后简单的计数和标准化计算,与属性值矩阵大小无关。因此 RF算法在处理大数据集上比其他 2种方法能显著的减少计算运行时间。

2) 3种方法删除异常样本的鲁棒性比较

通过只删除训练集中的异常样本而不删除测试集中的异常样本来建模和测试,这对模型的泛化能力有较高要求。RHM方法和稳健的马氏距离方法需要计算协方差的逆阵,然而如果协方差矩阵奇异,则需要用到伪逆,伪逆使算法的鲁棒性受损。如表 2所示, RF方法的鲁棒性较 RHM和稳健马氏距离方法更好,在全部 6个数据集上 RF都保证了模型的泛化能力。另外 2种方法的鲁棒性都不如 RF算法,甚至在 heart数据集上还出现准确率下降的现象。与 1)相仿,大数据集的内存占用和耗时成为影响异常样本检测的一个瓶颈问题,而 RF算法不存在这样的问题。这样的优点使基于随机森林的异常样本检测方法的广泛应用成为可能。

## 4 结语

本文将随机森林算法引入异常样本的检测中,结合样本相似度提出异常点尺度的概念来衡量样本的异常程度,并根据这一尺度筛选出异常样本。仿真实验证明,与其他 2种基于距离的异常样本检测技术相比,基于随机森林的异常样本检测不论在提高模型的准确率方面,还是在减少计算时间方面都比其他 2种方法有更大优势。同时,该方法具有较其他 2种方法更强的鲁棒性。

样本相似度除了用在异常点检测之外,还可以用于建立数据集原型 (prototype)、数据集坐标描述、以及训练集和测试集缺失值的补充等方面。随机森林提供的样本相似度在挖掘数据集自身特性方面的研究还有更多更广的潜力。但是,异常样本的尺度阈值选择仅有实验性的结果,还没有定量的判断标准,值得做进一步研究。

**参考文献**:

[1] 刘蓉,陈文亮. 奇异点快速检测在牛奶成分近红外光谱测量中的应用 [J]. 光谱学与光谱分析, 2005, 25(2): 207 - 210.

[2] 王斌会,陈一非. 基于稳健马氏距离的多元异常值检测方法 [J]. 统计与决策, 2005, 3(6): 4 - 6.

[3] Rousseeuw P J, Driessem K V. A Fast Algorithm for the Minimum Covariance Determinant Estimator[J]. Technimetrics, 1999, 41: 212 - 223.

[4] Breiman L. Random forests[J]. Machine Learning, 2001, 45(1): 5 - 32.

[5] 刘闽,林成德. 基于支持向量机的商业银行信用风险评估模型 [J]. 厦门大学学报:自然科学版, 2005, 44(1): 29 - 32.

(责任编辑: 陈雯)