基于集成学习的网络流量分类方法

网络流量分类和特征提取研究内容小结

Wu You

2018 年 10 月 20 日

目录

摘要

0.1. 当前网络环境分析

流量分类是许多其他网络监管活动的基础，在特别是在网络流量审计工作中，因此流量的准确分类对于网络安全具有十分重要的意义。在利用机器学习的流量分类中，存在的一大难题是数据的不均衡问题。本文中首先对不同机器学习算法的效果进行了对比，之后使用了效果较好的集成机器学习与数据重采样相结合的方法，对不均衡的流量数据进行了分类。

根据结果，我们发现集成机器学习算法具有良好的分类效果，并且具有良好的健壮性，同时对计算资源的占用较小。直接应用集成机器学习算法进行流量分类可达到 95% 以上的准确率，但对于占比较小的类别，并不能达到较好的准确率和召回率。在结合了重采样方法之后，整体准确率保持不变，但对于少数类样本的精度和召回率均得到了提高。

1. 引言

网络流量分类是网络监管的基础工作，随着网络环境的不断扩张发展，网络信息监管、服务质量控制以及异常检测等工作都需要基于流量分类来提高效率、降低成本。传统的流量分类方法基于深度包检测（Deep Packet Inspection, DPI），通过人工提取数据包中的特征序列来形成特征库，而随着流量种类的增多、数量的增加，特征库的更新维护成本在不断增加，同时这种方法也无法应对未知和加密流量。因此，基于机器学习的流量分类方法开始受到关注。机器学习是一类分类或回归算法的总称，其核心思想是通过现有数据形成一个分类或回归模型，从而实现对之后数据的准确处理。机器学习算法已经在多个领域得到了应用，并取得了不错的成果。

基于机器学习的网络流量分类研究开始于 2004 年，。2005 年，Moore 等人完成了流量数据的收集处理，从流量数据中提取出了 248 中用于训练的数据特征，并使用改进的朴素贝叶斯算法对流量进行了分类，这一系列工作为此后的很多流量分类研究提供了参考。之后的研究中更多着眼于解决流量分类中特定的的问题，如算法提速、新型协议的识别、以及分类机制的改进等，其中还包含了流量数据的不平衡问题。数据不平衡问题指的是全部数据中不同类别的样本数量相差巨大，该问题在多个领域内存在。数据不平衡的出现会使得分类算法在训练过程中产生偏差，进而导致其分类结果偏向于多数类，少数类的分类准确率降低。而在网络安全领域内，由于网络行为和数据中的大部分均属于正常范畴，异常和恶意数据样本较少，因此数据不平衡的问题也十分突出。

数据不平衡是机器学习领域的十大问题之一。目前在网络流量分类中主要使用特征选择的方法来处理样本不均衡的问题，另外还有基于代价敏感的方法和重抽样的方法。相比另外两种方法，特征选择的方法没有明显的缺点，但目前也不能很好地从理论上说明其有效性，并且，特征选择

的方法也用于优化训练结果，在这一步很难同时兼顾两方面。基于代价敏感的方法是一种十分有效的方法，但其效果取决于对代价本身的定义，而目前对于代价的定义并没有可遵循的规则，因此实际应用较少。

重抽样的方法则回避了以上二者的缺点，同时在一些实验中也证明了其有效性。Zhong 等人于 2009 年将重采样同决策树和神经网络分别结合，对 P2P 流量进行了分类，研究结果证明了重采样方法在网络流量分类问题中的有效性。Liu 等人则对包括重采样在内的三种数据不平衡的修正方法进行了对比，并以此说明了重采样方法相对另外两种方法的优势。而这些研究工作都只应用了简单的机器学习算法，本文中则选择将重采样和新兴的集成学习算法相结合，得到了更加精确的分类效果。

2. 基于集成学习的流量分类方法

采用 RES-LGBM 对网络流量进行分类的核心思想是对数据集进行预处理和特征筛选，之后采用重采样算法修正数据的不平衡性，再利用梯度增强树算法对处理后的数据进行训练，最终完成对流量的分类。

2.1. 网络流量数据集

为了确保研究工作的有效性，需要使用可靠的数据集。本文所使用的数据集是于 2005 年由剑桥大学的 Moore 等人采集的流量数据。作者采集了 1000 个用户在 24 小时内的流量数据，对其中每条 TCP 双向流进行特征提取，最终得到 377526 个数据样本。其中样本共分为 12 类，每个样本拥有 249 个属性，其中最后一项属性为样本的类别。样本的分布信息如表所示。

可以看出，该数据集中 WWW 类占据了 86.906% 的比例，而其他类别的样本则相对较少，存在数据不均衡的问题。样本量不均衡的问题在很多领域内存在，由于某类样本占比较大，即使在训练过程中将其它类别误分为此类，也不会对优化过程产生影响。所以种现象的发生会导致利用数据训练得到的模型产生偏向性，即对多数类的分类效果较好，甚至倾向于将其它类别的样本划分到此类中。在网络安全及管理问题中，我们常常需要更多地关注少数类别的流量，因此需要针对数据不平衡的问题进行优化，以提高少数类样本的识别率。

2.2. LightGBM 算法

集成学习算法的原理是将多个弱分类器进行结合，只要每个分类器都具有一定的可信度，就能形成一个效果较强的分类器，从而达到较好的分类效果。由于集成学习算法的决策结果是多个分类器共同形成的，因此能够有效避免传统算法存在的过拟合问题，多个分类器共同决策的机制也有效地避免了部分噪声对于整体的影响。

LightGBM（LGBM）基于梯度增强树算法，于 2017 年提出，属于集成学习算法的一种，该算法主要优化了运行速度，同时几乎没有降低算法准确率。梯度增强树算法集成了多个回归树，回归树由决策树算法衍生而来，其节点的分裂方式和决策树相同，但对每个叶子节点赋予了分值。通过将集合中多个回归树的分值相加，即可得到最终的结果。但由于需要对多个子树进行训练，梯度增强树算法的运行速度低于传统的决策树，而 LGBM 则通过一系列优化手段，使其运行速度得到了很大提升，主要的优化手段是基于梯度的单边采样（Gradient-based One Side Sampling，GOSS）和互斥特征捆绑（Exclusive Feature Bundling，EFB）以及随机森林算法中所使用的列采样方法。在训练过程中，需要利用梯度对模型进行更新，而梯度较小的样本对模型的影响也相对较小。GOSS 就是在计算梯度后，根据一定的梯度阈值，在梯度较小的样本中只抽取部分进行训练，减少了需要训练的样本量，而舍弃的样本对模型的更新影响不大，因此最终模型仍能够保持与原来相当的分类效果。为了能更快地进行训练，在 GBDT 算法中还使用了直方图算法来加快节点分裂指标的计算，EFB。。。。。

2.3. 重采样算法

重抽样方法分为过采样、欠采样以及将二者相结合的方法。其中欠采样方法指的是通过随机的方式或根据某种规则，剔除部分多数类样本，从而使各类样本的比例接近平衡，这种方法的优点在于它可以在修正数据分布比例的同时降低运算量。它改变了样本的分布，同时在欠采样的过程中可能丢失部分重要信息，一些先进的欠采样算法可以减少一定的信息丢失。与欠采样相对的过采样方法，指的是重复使用部分少数类样本或根据原有样本生成相似数据来增加少数类样本的比例。虽然一些研究者认为这种方法会改变样本的分布，也有可能造成过拟合，但是它却十分适用于 LGBM 算法。首先，虽然样本的分布被改变，但树算法的理论是直接对分类可能性进行建模，而无需对样本分部进行假设，故样本分部的改变并不影响 LGBM 的分类效果。其次，LGBM 算法在生成每一棵树时会选取样本的部分属性，并从样本中随机抽取一部分用于训练以增加子树的多样性，在此过程中若某类样本所占比例较小，则有可能在大部分树的训练过程中该类样本很少或并未得到训练，进而造成整体误差。使用过采样的方法能够使得少数类的比例增加，这样就能够保证该类样本在多数子树的生成过程中得到足够的训练。虽然这样可能导致部分样本在训练过程中重复出现，但由于 LGBM 算法中使用了 GOSS 来减少过拟合并提升运算效率，即使重复出现已经得到训练的样本也不会对模型造成很大影响，因此可以很好地避免过拟合的问题。

3. 网络流量分类中的主要问题

网络流量分类是一项明确的工作，但由于网络环境的复杂性，再加上在这一工作中使用了多学科的应用技术，使得这一工作中仍可细分为很多部分，这里列举一些更加细化的研究问题，以及相关的研究工作总结

3.1. 针对特定协议或应用类型的分类问题

网络流量分类工作中，根据特定的场景需求，需要有针对性地进行某种流量的有效识别。例如对 P2P 流量、Tor 流量、VPN 流量或者视频、音频流量的识别等。其中针对 P2P、Tor、VPN 等的识别工作是为了加强网络监管，防止技术滥用，而对流媒体的识别则主要用于合理分配带宽资源，从而提高 QoS。文献 [?]的作者进行了 VPN 流量数据的采集，并对 VPN 和非 VPN 流量进行了分类，同时也对流量的应用类型进行了分类，之后，作者又在 2017 年收集了 Tor 流量数据，并完成了类似研究工作。文献 [?]中用 ANN 和 SVM 对 Tor 数据进行了分类，并对比了两种算法的效果，最后通过分析结果来说明 Tor 对于互联网用户的保护作用。[?]中的作者于 2018 年进行了 VPN 数据的采集，并用利用多层感知器（神经网络）算法对其进行了分类。由于网络中部分攻击手段具有固定的模式，因此流量分类也可直接用于部分威胁的检测，Wang 等人在 [?]中利用网络流量进行了恶意流量的检测，并应用了近几年备受关注的卷积神经网络（Convolutional Neural Network），通过截取数据包的前 784 个字节，并将其转化为灰度图片，作者发现相同应用的数据包的灰度图之间具有较强的相似性，根据这一特点，作者直接使用 CNN 来进行流量的分类。但由于没有使用时间特征，该方法仍可改进，如使用 RNN 进行训练。Wang 和 Yang 在 [?]中研究了加密 P2P 流量的识别，使用了隐式马尔科夫模型，但只分析了较为简单的流量行为。Mujtaba 则对通过隧道传输的流量进行了识别 [?]，作者指出，对于隧道流量，PSD 是一个有价值的特征。

[?]文献 [?]的作者则研究了 SSL 流量的分类，同时使用 SMOTE 算法解决数据不均问题。[?]中，作者对 P2P 流量进行了分类，其中也使用了重抽样的方法来解决数据不均衡的问题。高长喜等人在文献 [?]中结合了 DPI 和 DFI 的方法，进行了加密流量的应用类型识别。

3.2. 数据不平衡问题的处理

数据不平衡是机器学习领域的十大问题之一，在流量分类中这个问题尤其明显，特别是当研究目标为某类特殊流量时，因为网络中最常用几类的应用和协议如 Web 浏览等所产生的流量占据了一半以上，而流量分类问题常关注的流量类型则占比较小，不均衡问题严重影响了分类效果。为解决此问题，目前主要有以下解决方案：重抽样、Cost-sensitive 方法以及特征选择方法，三种方法各有其优劣。

重抽样方法可以有效降低由数据不平衡带来的分类器偏差，但从理论上来说，这种方法改变了数据的抽样分布，对于基于概率假设的分类器有较大影响，但对于树分类器、神经网络等影响并不大。重抽样方法分为过采样、欠采样、以及组合方法，过采样方法从少数类中抽取样本，并加入训练集中，这种方法会增加训练时间，同时可能造成过拟合；欠采样方法将多数类中的部分数据从训练集中剔除，使数据均衡化，这一方法降低了训练时间，但存在丢失重要信息的风险；组合方法结合二者的特点，但具体应用仍有待研究（没怎么看到过）。

Cost-sensitive 方法不会改变数据分布，且消耗的计算资源也较少，但是目前没有成熟的理论来确定如何定义合理的 Cost，特征选择方法可以在一定程度上提升分类器性能但由于特征选择是

数据分析中重要的一步，其主要目的是提升算法的整体性能，难以与数据不均问题的优化目标相结合，且该方法需消耗大量计算资源。针对重抽样方法存在的问题，可以通过结合梯度增强决策树（Gradient Boosting Decision Tree，GBDT）算法来解决。GBDT 结合了多个简单决策树，通过一定的决策集成策略，使得多个弱决策树组合成一个强分类器。这类方法的有以下几个优势：首先，组合分类器的算法可以有效防止过拟合的发生，因为在训练过程中，每个基分类器都只利用训练集的部分数据，且最终决策需要综合全部基分类器的结果，少数过拟合的现象并不影响整体的性能；其次，GBDT 算法基于决策树，树算法无需对样本分布进行假设，所以可以有效结合采样方法；第三，这类方法对噪声不敏感，分类精度高；最后，Ensemble 的性能受到基分类器的影响，且每个基分类器的分类效果越好，则整体效果也越好，所以在特征选择时，可以用决策树来代替 Ensemble，而 GBDT 的基分类器是决策树算法，决策树是一种简单快速的算法，这将显著提高特征选择的效率。树算法的另一个优势在于，特征的有效性可以被明确地定义，因为在这类算法中，有效的特征就是信息增益最大的特征，因此可以首先利用全部特征对 GBDT 进行一次训练，再从训练结果中对特征的重要性进行排序，选取一定数量的特征后再应用特征选择算法

3.2.1. **重采样结合** GBDT

GBDT 的一个有效实现是 LightGBM（Light Gradient Boosting Machine），LightGBM 于 2017 年提出，主要运用了两种优化方法，这使得 GBDT 算法的运行效率大大提高，成为目前最有竞争力的算法。第一个方法是基于梯度的单边采样（Gradient-Based One Side Sampling），作者借鉴了 Adaboost 算法中的思想而提出了该方法。Adaboost 是最早提出的 Boosting 算法，该算法在训练新的基分类器时，会对之前之前未能正确分类的样本赋予更大的权值，强化其对后续分类器的影响，从而有效提高分类的准确率，但这样做的缺点是使得算法对噪声较为敏感。在 GBDT 中，由于树算法的特点，我们无法利用样本的权值来进行训练，但可以相应地使用梯度来代替，也就是 GOSS。GOSS 的解释如下：机器学习的本质是进行目标函数的最优化，而在优化过程中，若目标函数在某一处的梯度较小，我们可以认为该样本中的信息已经被“学习到了”，所以设定某一梯度阈值，将每次训练的样本分为两部分，对于梯度较小的那部分进行欠采样，从而弱化了这类数据的影响，强化了未被学习的样本。但我们还应该注意到，GOSS 是为了加快运行而提出的，它所带来的也只是效率的提升，因为它没有在训练过程中加强或削弱某一类的影响，而只是根据样本的梯度进行欠采样，而不是根据类别，故无法解决数据不平衡的问题。我们可以看到，GOSS 中包含了欠采样的机制，因此将 LightGBM 与欠采样相结合几乎不会带来任何提升。但正是由于 GOSS 的存在，以及 GBDT 算法本身的特点，即使训练集中存在冗余数据，也不容易造成分类器的过拟合，因此有效弥补了过采样方法的缺陷。而从 LightGBM 自身角度来看，由于每次训练都会随机在训练集中抽取样本，当少数类样本占比较小时，尤其在少数类别为多个时，可能会出现某些类别数据缺失或者训练不足的情况，这一影响在每个基分类器上都会发生，最终导致整体对少数类的训练不足。而通过使用过采样，我们提高了少数类样本的比例，使得每个基分类器都能得到足够的数据进行训练。

3.3. 特征选择算法

特征工程在所有数据分析的工作中都是十分重要的一环，在网络流量的分类中，可提取的特征有限，2005 年 Moore 提出的特征几何几乎涵盖了所有有效特征。[?] 因此，当前工作重点在于特征特征选择，网络流量分类中的特征选择包含两方面，首先从实际角度出发，一些网络特征是不容易获取的，例如完整流的统计特征需要等待整个流结束后才可以获取，为此部分研究提出利用发生于流量初期的数据包来提取特征，从而实现流量的及时分类，这对于一些持续时间较长的流量分类更有意义，如基于 UDP 协议的流量，由于 UDP 协议没有完整的确认机制，导致主机需要等待一段较长的时间才能够确认流的结束，故使用早期流量分析将有效解决这类流量的实时分类问题。其次，从算法的角度，过多的特征会降低运行速度，但随意删减特征会造成信息的丢失，所以需要有效的特征选择策略，根据需求来选取最有价值的特征，在保证分类性能的前提下，尽可能降低计算资源的消耗。

3.3.1. 特征选择算法

3.4. 其他方面

由于网络流量的高速性和多变性，分类模型的高效性、可扩展性等也是较为关键的问题。在一些常见的应用场景，对算法的实时性要求并不严格，更多的是强调算法的分类精度，而在网络中，若想要进行有效的监管和防护，就需要分类器及时对流量进行识别，这对算法的运行效率提出了一定要求。同时，网络流量不同于通常的识别对象的另一点在于，网络流量的特征可能会不断变化，产生概念漂移问题，或者在某一时刻出现新的流量类型等。解决流量多变性问题，可以从两方面入手，第一，定期进行模型的更新或重新训练算法，这就需要提高算法效率，减少运行时间，从而降低重新训练的代价，第二，通过提高算法的健壮性，使模型有效期延长，降低重新训练的频率。

4. 总结