KDDCUP99 数据集的数据分析研究

吴建胜 张文鹏 马 垣
(辽宁科技大学软件学院 辽宁 鞍山 114051)

摘 要 kddcup99 数据集的网络连接数据量很大,各特征属性的取值范围较广,决策类型的种类也很多。因此,如果直接在原数据集上进行数据预处理或是数据挖掘都将是一件十分困难的事情。通过对 kddcup99 进行数据分析,提出一种对其按照 service 属性的不同进行分块的新思路,在不失真的前提下,将大问题转化成小问题,从根本上解决了数据集过大的难题。

关键词 kddcup99 数据集 分块 数据分析

中图分类号 TP3 **文献标识码** A DOI:10.3969/j.issn.1000-386x.2014.11.081

DATA ANALYSIS AND STUDY ON KDDCUP99 DATA SET

Wu Jiansheng Zhang Wenpeng Ma Yuan
(*School of Software Liaoning, University of Science and Technology Liaoning, Anshan 114051, Liaoning, China*)

Abstract There is a mass of network connections in kddcup99 dataset. The values of the feature attributes are widely ranged, and the descriptions of decision types are also rich. For this reason, it will be very difficult to pre-process or mine the data on original data set. In this paper, through analysing the data of kddcup99, we put forward a new idea of blocking the dataset according to service attribute, which converts the big problem into some small problems without distortion, and fundamentally solves the difficulty of too large the dataset to be.

Keywords Kddcup99 dataset Blocking Data Analysis

0 引 言

近些年,网络安全研究人员从各个不同的角度对入侵检测技术进行了全方位的综合研究分析,结合不同应用技术的优点,提出了许多空间时间复杂度较低并且误检率漏检率较低的高效率的新颖思想和算法。其中基于数据挖掘的入侵检测也逐渐成为人们研究的热点方向。kddcup99 数据集是网络安全研究人员公认的数据集,作为研究入侵检测系统及各种应用算法的好坏的评价标准,kddcup99 数据集为计算智能的网络入侵检测研究奠定了坚实的基础[1-4]。

如今研究人员对 kddcup99 的研究还仅限于对其决策类型以及各个属性的取值及分布情况的介绍,并没有准确地给出一个处理的具体办法。Kddcup99 的 10% 训练数据集中共有 494 021条连接记录,每一条连接记录都是由 41 个特征属性和一个决策类属性组成的。特征属性用来描述这条连接记录的各项指标,其中包含 9 个离散型属性和 32 个连续型属性。其中,1 ~9 是属于网络连接的基本特征属性,10 ~22 是属于网络连接的内容特征属性,帮助对 U2R 和 R2L 攻击进行检测,23 ~41 是属于网络连接的流量特征属性(2 s 时间单元)。而决策类属性则是用来标识这条记录是正常的或是属于哪种已知攻击类型的。决策属性共包含 1 种 normal 的正常行为和 22 种攻击行为。而 kddcup99 的测试集中共包含 20 种已知攻击类型和 14 种未知攻击类型。与此同时,哪怕是共同出现在测试集和训练集中的攻击类型,其数量的多少和概率分布情况可能也会有些许的不同。但对于某一种攻击类型来说,在这庞大的数据集中都隐含着它独有的数据特征[5-9]。

从数据的总体统计分析数据,不难看出,各特征属性的取值千变万化,各不相同,甚至有的属性取值情况相当复杂,再加上训练数据集的连接记录的数量之大,如果直接在原始数据集之上进行数据挖掘算法操作的话,过程将十分复杂,并且会耗费大量的人力物力,哪怕这样,有时也可能并不能得到想要的正确决策规则,最后只能是事倍功半徒劳无功。因此,在进行数据挖掘操作之前,进行一定的数据预处理是相当必要的,只有在进行数据挖掘之前,对数据进行一定的分析和一定的数据预处理操作,才能事半功倍,快速高效地获取所需要的知识信息。

由于数据量过大,许多研究者会选择随机抽取其中的一千或者是一万条记录进行实验,虽然有时也可能在所选的样本数据集中获取一定的决策规则,但是其随机性较大,选取的连接记录不同得到的规则可能也会有较大的不同。这样很难全方位地产生隐含在数据集中的全部有用知识,有时甚至可能会产生一些无用规则。因此,虽然数据集十分庞大,但是尽量还是要在对整个数据集分析的基础上进行一定的约简及挖掘工作。文献[1]已经对 kddcup99 数据集的各攻击类型和特征属性及其大致分布进行了一定的介绍。通过这些数据,可以发现许多数据表面所隐含并能通过简单分析快速发现的知识。而本文的创新点就在于给出了一个在不失真的条件下分块处理 kddcup99 数据

收稿日期:2013 -04 -01。**吴建胜**,教授,主研领域:信息安全,数字图像处理。**张文鹏**,硕士生。**马垣**,教授级高工。

集的新思路,解决了 kddcup99 数据集由于数据集过大而难以处理的难题。根据数据集中可以快速发现的知识,通过按照 service 网络服务属性进行数据集的分块操作,一方面在不失真的前提下减少了一次数据挖掘过程所需要处理的数据集大小,另一方面,每个小数据集中所包含的攻击类型较原训练数据集来说有很大的减少,一次数据挖掘所需要区分的决策类型减少了,相应地用来帮助决策的针对性特征属性,即属性约简得到的属性集合大小也会相应地减少。此外,随着数据集的分块操作,对于一个确定的属性,在某一指定的数据子集中的取值范围也会比该属性在整个训练测试集中的取值范围有较为明显的减少,这样一来,可以明显地地降低后期连续数据离散化还有属性约简操作的复杂程度。同时,由于只是将原训练集进行了分块细化,只要最后将所有数据子集挖掘出的规则进行归并总结便可以正确地得到原始训练集的全部有用知识。

1 特殊属性与攻击的删除

1.1 属性取值完全相同的属性删除

在 kddcup99 训练集中,由于每一条连接记录在 num_outbound_cmds 和 is_hot_login 这两个属性上的取值相同均为0。如果某个属性在全体数据集中取值均相同,即无论当决策结果取何值时,这个属性值均为同一个数,则说明通过这个属性的取值对进行决策的判定没有任何帮助作用。同时,根据信息论中关于信息量的定义也可以得到相同的结论。

定义 1 如已知事件 Xi 已发生,则 Xi 所包含或所提供的信息量为:

$$H(Xi) = -\log P(Xi) \tag{1}$$

如果某个属性的取值均为同一个数,则说明“这个属性取得这个属性值”这个事件发生的概率为 1,根据定义 1 可知如果一个事件的概率为 1,即是一个必然事件的话,这个必然事件的信息量为 0,即这属性不能给后期决策带来任何信息,对它的删除操作并不会造成决策表的不一致性,因此在后期处理之前可以首先将此类属性进行删除,以达到降维的目的。

1.2 属性取值情况较少的属性删除

通过分析统计,可以发现在 kddcup99 训练数据集中有一部分属性的取值情况很特殊,大多数连接记录中该属性的取值均相同,只有少数几条记录的取值较为特殊,只要将这些特殊情况加以统计标识,余下的记录条目在该属性上取相同的值,这样一来该属性对于后期的决策分类便起不到任何作用了,因此可以予以删除。

(1) land 属性及 land 攻击的删除

在训练测试集中,所有的 land 属性取值为 1 的 22 条连接记录中有 21 条均属于 land 攻击,且 land 攻击和 normal 之间,除 land 属性之外的其他属性之间并无明显区别,因此可以忽略。因此,根据定义 1,可以将所有 land = 1 的记录删除并删除 land 属性。与此同时需要在默认异常规则库中加入相应的规则。

(2) urgent 属性的删除

经过统计,在训练集中,该属性值几乎全部为 0,只有 4 条记录取值不为 0。只要将这四条记录加以标记,便可以删除该属性。通过数据分析,可以发现在 urgent 属性不为 0 的情况下,如果网络服务类型为 login 则该网络连接属于 ftp_write 攻击。如果网络服务类型为 telnet,hot 属性不为 0 并且 dst_host_srv_count 不为 0 的时候,则该网络连接属于 rootkit 攻击。如果网络服务类型为 telnet,hot 属性为 0,则该网络连接属于 normal 正常行为。这样,根据定义 1,便可以删除所有 urgent! =0 的记录同时删除 urgent 属性。与此同时,需要在默认异常规则库和默认正常规则库中加入相应的决策规则。

(3) su_attempted 属性的删除

经过统计,在训练集中,该属性值几乎全部为 0,只有 6 条记录为 1 和 6 条记录为 2。只要将特殊情况进行统计,便可以将该属性删除。如果 su_attempted 不为 0,且 num_shells 为 0,则是一条 normal 正常行为。因此,根据定义 1,可以将所有 su_attempted 不为 0 的记录删除并删除 su_attempted 属性。与此同时在默认正常行为规则库中加入响应的记录条目。

(4) num_shells 属性的删除

经过统计,在训练集中,该属性值几乎全部为 0,只有个别连接记录的值为 1 或 2,将特殊值记录加以标识便可以删除该属性。通过分析可以发现,如果 num_shells 的值为 2,且 hot 的值 2,则该连接记录属于 loadmodule 攻击。如果 num_shells 的值为 2,且 hot 的值不为 2,则该连接记录属于 multihop 攻击。如果 num_shells 的值为 1,service 为 telnet 服务,hot 的值为 0,且 su_attempted 值为 0 则该连接记录属于 perl 攻击,并且通过这条规则已经分析出了所有的 perl 攻击,因此该攻击也被删除。如果 num_shells 的值为 1,service 为 telnet 服务,hot 的值为 0,且 su_attempted 值不为 0 则该连接记录属于 normal 正常连接记录。

这样,根据定义 1,便可以删除所有 num_shells 不为的 0 记录同时删除 num_shells 属性字段。

综上所述,通过数据集中的数据表层信息,已经可以删除其中的 6 个属性和 2 个攻击类型。此时,数据集中还剩余 35 个特征属性和 20 种攻击类型和 1 个 normal 正常行为。

2 数据库分块

通过对特殊属性的删除,虽然已经删除了一定的属性和攻击类型。但是,数据记录仅仅被删除不到 100 条,对于 49 万多条的数据集来说这只是杯水车薪,因此还要继续在不是真的前提下对其进行下一步的处理操作。

根据 kddcup99 训练数据库中的数据信息加上决策类型和网络服务有较大联系这一领域先验知识,可以将 service 这一属性作为数据库分块的标准。由于训练集中共有 66 种 service 类型,因此,初步将训练数据集分为 66 块,即将原数据集按照 service 属性分到 66 个数据库中,其中每一个数据库都已将该网络服务类型作为数据库名称,其次,由于每一个数据库中的所有数据记录均具有相同的网络服务类型,因此可以将该数据库中的 service 属性进行删除。最后,通过表面数据信息不难看出,对于某一种固定的网络服务类型来说,除 private 和 other 网络服务外,均只是对应着一个单一的网络协议类型,因此可以将分块后的除 private 和 other 数据库之外的数据库中的 protocol_type 属性进行删除。

经过分块以后,可以明显地看出其中一大部分的网络服务类型只对应着一种或几种攻击类型。

2.1 数据分块的作用

通过对 kddcup99 训练数据集按照网络服务类型(service)属性的不同进行数据分块,可以明显地减少单次分析的数据量大小,针对性较强,下一步只需通过一定的数据预处理和数据挖

掘算法对各个数据库块分别进行处理便可以了。与此同时,在进行分块之后,每一块数据中包含的决策类型也会有相应的减少,有些数据块中甚至只存在两三种决策类型需要进行区分,这样一来辨别二三十种决策类型的大难题也便被随之简而化之。

每一种攻击类型或是正常行为均有它独特的行为特征描述,对于实验环境来讲,这个特征描述则表现在各个特征属性的取值情况。当对一条网络连接记录进行决策分类时也是通过这些特征属性的取值情况来决定的。但区分不同的决策类型可能需要用到的属性个数和类别也会有所不同。分块之后每个数据块中需要区分的决策类型少了,用来区分它们的特征属性自然而然地也会相应地减少。与此同时,进行分块之后,每个属性的取值范围可能会随之相应地减小,这样一来,后期进行离散化和特征属性约简时的时间复杂度也会同时有所降低。

2.2　特殊网络服务类型

在对 kddcup99 数据集进行分块之后,可以发现其中一大部分网络服务类型的数据集中只包含有一种或是两三种决策类型。

表 1 所示的是决策类型基本为 neptune 的网络服务类型。其中,每个网络服务类型中除 neptune 外,可能还存在很少的几条其他攻击类型。

表 1　决策类型基本为 neptune 的网络服务类型

supdup	whois	vmnet	uucp_path
uucp	systat	sunrpc	ssh
sql_net	shell	rje	remote_job
printer	pop_2	nntp	nnsp
netstat	netbios_ssn	netbios_ns	netbios_dgm
name	mtp	login	link
ldap	kshell	klogin	iso_tsap
http_443	hostnames	gopher	exec
efs	echo	domain	discard
daytime	ctf	csnet_ns	courier
bgp	Z39_50		

经过数据统计分析,上述 42 网络服务类型基本上对应着 neptune 攻击,但在这五六千多条网络连接中也掺杂着 38 条 portsweep 攻击,13 条 ipsweep 攻击,8 条 satan 攻击,5 条 normal 正常行为,3 条 nmap 攻击,2 条 ftp_write 攻击。

表 2 所示的是决策类型基本为 normal 的网络服务类型。其中,每个网络服务类型中除 normal 正常行为之外,可能存在很少的几条其他攻击类型。

表 2　决策类型基本为 normal 的网络服务类型

urp_i	urh_i	tftp_u	red_i
ntp_u	IRC	domain_u	

经过数据统计分析,上述 7 个网络服务类型基本上对着 normal 一种决策类型,但是在这近七千条网络连接记录中也掺杂着 3 个 satan 攻击。

对于上述这 49 种网络服务类型来说,由于特殊情况数量不多,所占的比例也较小,因此暂时将其忽略不计,后期也可以进行一些相应的研究,对这些存在于特殊网络服务类型数据集中的攻击类型进行标识,产生一定的默认决策规则,在检测前先进行辨识。

通过上述数据分析,在这 66 个数据子集中,由于上述的 49 个数据子集几乎只对应这一种攻击类型,且特殊情况所占的比例很低,并不会对整体地决策率产生太大的影响,可以忽略不计,因此对于这 49 个数据集中的记录,可以仅针对网路服务这一个特征属性建立相应地决策规则。这样一来,下一步需要解决的小数据集的个数就从 66 个减少为 17 个。解决 kddcup99 训练数据集 49 万多条记录的数据挖掘工作,已经被分解为攻击类型相对较少且特征属性取值范围较小,特征属性针对性较强的小数据集的数据挖掘。这样便可以化整为零、化繁为简,化大问题为小问题了。只需要针对某一网络服务,对其对应的几种攻击类型进行决策规则的挖掘工作便可。

3　分块结果及应用算法分析

3.1　分块结果分析

通过分块操作之后,除上述 49 种特殊网络服务类型对应的数据集块之外,还剩余 17 个小数据集,针对各个小数据集,需要对其进行特征属性的二次删除操作。根据定义 1 可知,如果在整个数据集中,对于所有的网络连接记录来说,某个特征属性的取值均相同,则可以在这个小数据集块中将该属性进行删除操作。

经过上述操作,不论是在网络连接记录的数量上、决策类型的数目上还是在特征属性的数目上都有所减少,有时变化特别显著。在网络连接记录的数量方面,较好的情况下,可能出现一个 7 条记录的数据集,最坏的情况下,会出现一个二十多万条的数据集。对于攻击类型来说,较好的情况下,攻击类型可能从 23 种决策类型减少到两三种决策类型,最坏的情况下,也能相应地减少到 12 种决策类型。至于特征属性方面,在最好的情况下可能从原本的 41 个减少为 7 个。而最坏的情况下也可以减少到 31 个。表 3 所示的是分块后的具体统计数据。

表 3　分块后各数据集块的数据统计

service	连接记录数目	决策类型数	特征属性个数
auth	328	2	26
ecr_i	281 400	6	13
eco_i	1 642	4	12
finger	649	5	24
ftp	798	12	28
ftp_data	4 679	12	28
http	64 293	7	29
other	7 236	6	24
private	110 893	7	24
pop_3	202	5	26
pm_dump	1	1	35
imap4	117	2	27
smtp	9 721	5	29
tim_i	7	2	7
time	157	3	21
telnet	493	12	31
X11	11	2	22

原 kddcup99 数据集中共有 494 021 条记录,通过数据集分块操作,经过表 3 的统计分析,上面所述的 3 种最坏的情况并没

有同时出现。记录条目较多时其属性个数及攻击类型不会太多,且规律性较强。ecr_i 网络服务类型占了一大半的数据量,但是通过数据集分块,该数据集中只有 6 个决策类型,同时通过属性的二次删除,其特征属性的个数也只剩下 13 个。而对于决策类型较多的三个数据集块 ftp、ftp_data 和 telnet 来说,其特征属性个数也相对较多,但是其连接记录条目一般并不多,最多也只有四千多条而已。对于特征属性个数较多的 telnet、ftp_data、http、smtp 等如其连接记录个数较多则决策类型则较少,反之,若决策类型较多则连接记录个数则较少。而对于较为特殊的 pm_dump 数据块来说,因为此种只包含一条记录,所以并不能通过传统的数据预处理和数据挖掘算法对其产生相应的决策规则。但是通过对数据的分析,可以发现 count 属性对其有较大的辨识能力,因此可以以 count 属性作为重要属性建立默认规则加入默认规则库。

3.2 应用算法对比分析

对于一般的决策表离散化,属性约简及决策规则挖掘算法来说,决策表中的记录个数越多,特征属性的个数越多,决策类型越多,相应算法的难度也会较大,甚至有时由于数据量过大超过了现有机器的承载力而无法得到结果。想要在保证得到全部隐含知识的前提下进行数据挖掘,同时又要保证算法在现有计算机上运行的可行性,只有进行化整为零的分块操作。

通过上面的分块结果,可以明显地发现无论是连接记录个数,特征属性个数或是决策类型种类都有了较大的减少。单次数据挖掘过程的复杂度要远远地小于对整个原始数据集进行数据挖掘的过程的复杂度。

对于决策表的离散化操作来讲,一般情况下,对全体训练集进行离散化得到的各个属性的断点数目较多,而分块后的单个数据子集离散化得到的断点数目相对较少。由于原始数据集数量较大本文采取复杂度较低的 Naive Scaler 算法进行离散化对比实验。

图 1 所示的是经过 Naïve Scaler 离散化算法后,kddcup99 原始训练集及分块后各数据子集对应属性的断点个数对比图(如果该数据子集中不包含该属性,则将其该数据子集中该属性的对应的断点个数设置为 0)。图 1 中虚线表示的是原 kddcup99 训练集离散化后各个属性的断点个数,各实线表示的是各个数据子集离散化后的各个属性的断点个数。从图 1 中可以明显地看出,除个别子集的个别属性的断点数要稍微大于原始训练集外,大部分的数据子集的各个属性对应的断点数均要小于甚至远小于 kddcup99 原始训练集。

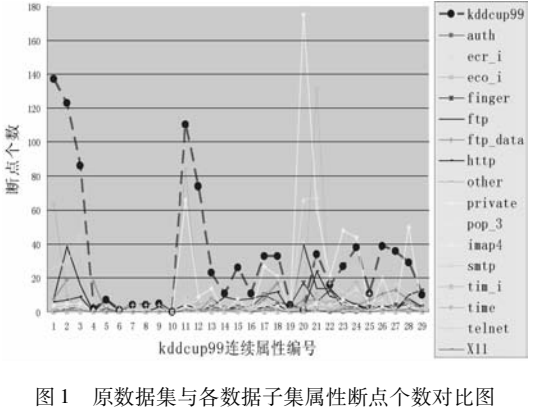

图 1 原数据集与各数据子集属性断点个数对比图

除此之外,数据集庞大的记录个数也会严重影响和制约后期属性约简算法的可选性及其运算效率。有的算法由于本身的原因并不能完成对过大数据集的约简操作,哪怕是那些可以承载大数据集的约简算法,也往往会因为数据量过大而效率极其低下。

由于原始数据集过大,文献[11]在保持了异常连接记录尽量平均的前提下随机选取了数据集中的 150 条作为训练集,368 条作为测试集,采用以互信息为标准的分布特征选择算法,将原始数据集中的 41 个特征属性减少到 21 个。文献[12]随机抽取 30 000 条记录作为训练集,60 000 作为测试集,采用基于粗糙集和遗传算法的特征选择算法,将原始数据集中的 41 个特征属性减少到 12 个。文献 13 选取 60 000 条记录作为训练数据集,120 000条记录作为测试数据集,采用对种群聚类的类内随机特征选择算法,将原始数据集中 41 个特征属性减少到 11 个。关于这方面的研究已经很多,但都是通过对原始数据集的随机选取一定的数据子集作为新的训练集和测试集,针对性不强。本文在进行数据集分块之后,采用 Hu 的基于可辨识矩阵属性约简算法,对 time_i、time、eco_i、ecr_i、auth 等数据子集进行约简,约简后的属性个数如表 4 所示。

表 4 参考文献同本文分块后各数据集属性约简结果对比

名称	训练个数	约简后属性编号
文献[11]	150	1、3、4、5、6、8、11、12、13、23、25、26、27、28、29、30、32、33、34、36、39(共 21 个属性)
文献[12]	30 000	2、3、5、6、7、12、15、28、30、34、37、40(共 12 个属性)
文献[13]	60 000	1、2、3、5、6、12、23、24、33、36、40(共 11 个属性)
tim_i	7	23、35(共 2 个属性)
time	157	34、35、40(共 3 个属性)
eco_i	1 642	5、24、32、37(共 4 个属性)
ecr_i	281 400	5、8、23(共 3 个属性)
auth	328	1、23、31(共 3 个属性)

从表 4 中可以明显看出,按照本文分块思想得到的约简效果要明显好于文献从原始数据集中随机抽取的小数据集的约简效果。那是因为对于分块后的据子集来讲,由于需要区分的决策类型较少,属性的针对性比全体数据集中属性的针对性要强,因此用于区分该数据子集中各个记录的特征属性的个数要比文献中的少很多。

4 结 语

kddcup99 数据集作为入侵检测研究领域的一个公认测试标准,从提出至今已有 14 年之久,可以帮助研究人员很好地测试与评价各种研究算法。但是由于其数据条目过多,特征属性与决策类型的个数也较大,直接操作十分困难。本文在对 kddcup99 数据集分析研究的基础之上,首次提出了通过 service 属性对原训练数据集进行数据集分块操作,在保持原数据集各种特征规则不丢失的前提下,很大程度地降低了基于数据挖掘的入侵检测系统规则库建立的难度。对后期研究有很大的帮助作用。

对于在训练集中没有而在测试集中存在的 icmp 网络服务,由于其只有 2 条记录,因此,并不会对我们的实验精度造成太大的影响。而对于 2.2 节中所提高的特殊网络服务类型,下一步

可以对其中掺杂的七十余条记录进行早期研究,排除其特殊性。

参 考 文 献

[1] 章金熔,刘峰,赵志宏,等. 数据挖掘方法在网络入侵检测中的应用[J]. 计算机工程与设计,2009,30(24):5561 -5566.

[2] 刘云,刘学诚,朱峰. 数据挖掘技术在入侵检测中的应用[J]. 计算机应用与软件,2011,28(5):117 -119.

[3] 蒋建春,马恒太,任党恩,等. 网络安全入侵检测:研究综述[J]. 软件学报,2000,11(11):1460 -1466.

[4] 李涛. 基于数据挖掘技术的自适应入侵检测系统模型[J]. 计算机工程与设计,2010,31(6):1209 -1211,1229.

[5] 张新有,曾华燊,贾磊. 入侵检测数据集 KDD CUP99 研究[J]. 计算机工程与设计,2010,31(22):4809 -4812,4816.

[6] Kang Zhang. KDD CUP99 数据集之背景知识[EB/OL]. 2010. http://xifage.com/kdd -cup -99 -dataset -1/.

[7] Hettich S,Bay S D. KDD cup 1999data[EB/OL]. 1999. http://kdd.ics.uci.edu/databases/kddcup99.html.

[8] 王洁松,张小飞. KDDCup99 网络入侵检测数据的分析和预处理[J]. 科技信息,2008(15):79 -80.

[9] Haines J W,Lippmann R P, Fried D J, et al. Boswell,1999 DARPA Intrusion Detection Evaluation: Design and Procedures[C]. MIT Lincoln Laboratory: Lexington, MA, 2001.

[10] 王国胤. Rough 集理论与知识获取[M]:西安:西安交通大学出版社,2001.

[11] 肖立中,刘云翔. 适合于入侵检测的分布特征选择算法[J]. 计算机工程与应用,2010,46(11):81 -84,87.

[12] 陈路莹,姜青山,陈黎飞. 一种面向网络入侵检测的特征选择方法[J]. 计算机研究与发展,2008,45(S):156 -160.

[13] 赵新星,姜青山,陈路莹,等. 一种面向网络入侵检测的特征选择方法[J]. 计算机研究与发展,2009,46(S):477 -482.

(上接第 192 页)

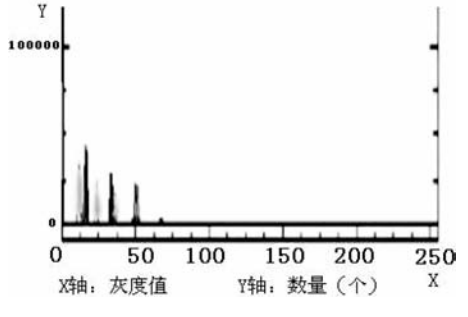

图 5　GLCM 对比度直方图

比较分析发现,前者的直方图中灰度值分布比较均匀,大体符合正态分布,所以纹理提取效果较好,而后者灰度值比较集中且较小,因此处理结果整体偏暗。依照此方法将九种纹理描述子的处理结果进行比较,发现 GLCM 相关性、GLCM 角二阶矩、GLDV 角二阶矩这三种纹理描述子对纹理的提取效果较好。

3.3　傅里叶频谱比较分析

对处理结果进行傅里叶变换,如图 6 为 GLCM 对比度傅里叶频谱图像,图 7 为 GLDV 角二阶矩傅里叶频谱图像。

比较分析可以发现,后者高频信息比较多,图像中的边缘信息、纹理信息比较丰富,纹理提取效果较好,而前者在水平和垂直方向上的边缘、纹理信息明显要少。依照此方法,将九种纹理描述子的处理结果进行傅里叶变换,比较其傅里叶频谱,发现 GLDV 反差、GLCM 角二阶矩、GLDV 角二阶矩这三种纹理描述子对纹理的提取效果较好。

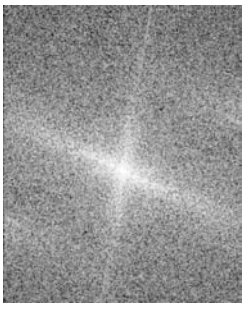

图 6　GLCM 相关性傅里叶频谱

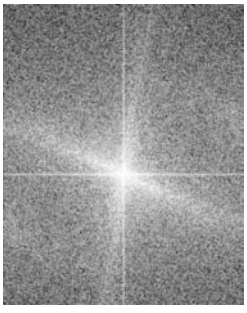

图 7　GLDV 角二阶矩傅里叶频谱

综上所述,通过目视观察、直方图比较、傅里叶频谱比较分析发现,对于此幅高分辨率无人机遥感影像,采用 GLCM 角二阶矩、GLDV 角二阶矩这两种纹理描述子对遥感影像的纹理提取效果较好,且其适应性更强,是较好的纹理提取方法。

4　结　语

本研究用 C ++ 编程实现了基于灰度共生矩阵的九种纹理提取算法,通过对高分辨率遥感影像处理结果的比较分析,得出 GLCM 角二阶矩、GLDV 角二阶矩是适应性、稳定性较好的纹理描述子。纹理特征作为遥感影像分类中的一个重要参数,对分类结果的精确度具有非常重要的作用。因此,在对高分辨率遥感影像进行纹理提取辅助分类或面向对象分类中纹理参数设置时,为了分类结果更加精确,纹理描述子可以选取 GLCM 角二阶矩或 GLDV 角二阶矩。GLCM 角二阶矩、GLDV 角二阶矩在高分辨率遥感影像的分类中将具有十分重要的应用价值。

如何设计灰度共生矩阵统计方向、统计距离、统计窗口大小、灰度压缩级数、纹理描述子才能又快又好地提取高分辨遥感影像的纹理特征,以及如何将其与其他纹理提取的方法更好地结合运用于遥感影像纹理提取将会是未来非常重要的研究方向。

参 考 文 献

[1] 杨凯陟,程英蕾. 基于灰度共生矩阵的 SAR 图像纹理特征提取方法[J]. 电子科学,2011,24(10):66 -69.

[2] 田艳琴,郭平,卢汉清. 基于灰度共生矩阵的多波段遥感图像纹理特征的提取[J]. 计算机科学,2004, 31(12):162 -164.

[3] 刘新华,舒宁. 纹理特征在多光谱遥感影像分类中的应用[J]. 测绘信息与工程,2006,31(3):31 -32.

[4] 潘洁,李明诗. 基于信息量的高分辨率影像纹理提取的研究[J]. 南京林业大学学报:自然科学版,2010,34(4):129 -134.

[5] Sklansky J. Image segmentation and feature extraction[J]. IEEE Transactions on Systems, Man, and Cybernetics, 1978,8(5):237 -247.

[6] 刘丽,匡纲要. 图像纹理特征提取方法综述[J]. 中国图象图形学报,2009,14(4):622 -635.

[7] Unser M. Texture classification and segmentation using wavelet frames [J]. IEEE Transactions on Image Processing, 1995, 4 (11): 1549 -1560.

[8] 高程程,惠晓威. 基于灰度共生矩阵的纹理特征提取[J]. 计算机系统应用,2010,19(6):195 -198.

[9] 郭德军,宋蛰存. 基于灰度共生矩阵的纹理图像分类研究[J]. 林业机械与木工设备,2005,33(7):21 -23.

[10] 冯建军,杨玉静. 基于灰度共生矩阵提取纹理特征图像的研究[J]. 北京测绘,2007(3):19 -22.