信息熵在网络设备智能修复中的应用研究

唐启涛
(湖南信息学院 电子信息学院,湖南 长沙 410105)

摘 要:在日常的网络设备维护中,很多问题往往是由配置文本命令出错或者配置不当引起的,为了能快速找出配置文本命令出错的地方,需要对设备的配置文本命令进行智能检查。要在繁杂的信息中快速地找到用户需要定位的信息并及时修改配置文本命令,配置命令的智能匹配起着非常重要的作用,它可以有效地组织和管理这些信息,从而提高信息搜索的效率。文中以信息熵理论知识为基础,应用信息熵在文本分类中的可适应特性,提出了一种基于信息熵的网络设备配置命令分类算法。在该算法中网络设备的配置命令的分类依据是信息熵的大小,它只处理信息熵大于给定阈值的信息特征向量。然后将该算法应用到网络设备的配置命令智能修复系统中。仿真实验结果表明,在网络设备配置命令的智能修复系统中应用信息熵方式处理网络设备的配置命令文本是一种效率更高、精度更准的配置命令分类处理算法。

关键词:网络设备;特征选择;分类算法;信息熵;配置命令
中图分类号:TP301 **文献标识码**:A **文章编号**:1673-629X(2019)12-0116-06
doi:10.3969/j.issn.1673-629X.2019.12.021

Research on Application of Information Entropy in Intelligent Repair of Network Equipment

TANG Qi-tao
(School of Electronics and Information, Hunan Institute of Information Technology, Changsha 410105, China)

Abstract: In the daily maintenance of network equipment, many problems are often caused by errors in configuration text commands or improper configuration. In order to find out the errors in configuration text commands quickly, it is necessary to check the configuration text commands intelligently. To quickly find the information users need to locate and modify the configuration text commands and configure the commands in time with the complicated information, intelligent configuration plays a very important role, which can organize and manage these information effectively, so as to improve the efficiency of information search. Based on the theory of information entropy and the adaptability of information entropy in text categorization, we propose an information entropy-based classification algorithm for network equipment configuration commands. In this algorithm, the classification basis of network equipment configuration commands is the size of information entropy, which only deals with information eigenvectors whose information entropy is greater than a given threshold. Then the proposed algorithm is applied to the intelligent repair system of network equipment configuration command. The simulation experiment shows that the application of information entropy in the intelligent repair system of network equipment configuration commands is a more efficient and accurate classification algorithm for network equipment configuration commands.

Key words: network equipment; feature selection; classification algorithm; information entropy; configuration commands

0 引 言

配置命令归类的过程实质是一个模式识别的过程,是在给定的分类体系下,根据配置命令的内容确定配置命令所属类别的过程,在实际操作过程中,存在向量维数高与训练样本数量巨大两个显著特征。常见的配置命令分类步骤总结如下[1]:首先是配置命令的预处理,通过预处理,把配置文本中一些不必要的冗余信息删除;第二步是配置命令的模型表示;在确定配置命令的表示方式后,第三步是选取配置命令的独有特征;第四步根据特征选择合适的分类器进行训练。由于在配置命令的分类中,并非所有高维的特征集都具备相同的重要性,而通过配置命令的特征选择可以有效降

收稿日期:2019-01-27 **修回日期**:2019-05-28 **网络出版时间**:2019-09-24
基金项目:湖南省教育科学技术研究项目(14C0114)
作者简介:唐启涛(1975-),男,硕士,副教授,研究方向为计算机网络及信息安全。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20190924.1535.040.html

低分类维数从而简化计算的难度。同时,对分类结果的准确性、样本的训练时间均有着非常重要的影响,因此在配置命令分类过程中特征选择成为了最关键的步骤。目前在配置命令分类中比较常见的特征选择算法有词频方法、信息增益、期望交叉熵、互信息等[2]。这些特征选择算法在文本分类中针对不同的文本类型,其应用效果也有所不同,针对网络设备中的配置命令进行分类处理时,简单地应用其中某一种算法,往往会存在配置命令过度冗余、计算配置命令权重繁杂等问题。为了更好地进行网络设备中的配置命令分类处理,在对日常应用中常见的特征算法进行分析比较的基础上,引入了基于信息特征向量的信息熵对传统的配置命令分类算法进行优化,进而提出一种基于信息熵的网络设备配置命令分类算法。

1　几种特征选择方法的比较

1.1　词频方法

该方法是根据给定文本中相应词语出现的次数,记录各个词汇的词频,当词频大于给定阈值时,才保留其在语料库中的数据,对于低于给定阈值的词汇,直接忽略[3]。在实际应用中,往往需要对给定文本进行样本训练,其目的是为了统计各词汇的词频数,排除低词频的词汇,只保留对语料库中文本分类影响较大的词汇。

词频方法的应用简化了文本分类的流程,在词频计算中其计算复杂度较低,是一种最简单的降维方法,但由于其对词汇的处理只是简单地去除低词频的词汇,对文本分类的准确率不高,通常作为文本分类处理的一种辅助方法[4]。

1.2　信息增益方法

该方法是根据给定文本中相应特征向量的信息熵之差来确定特征项是否入选,基于此,需要先计算特征项的信息熵差,计算公式如下[5]:

$$IG(T)=P(T)P(C_i\mid T)\log_2(P(C_i\mid T)/P(C_i))+P(\bar{T})\sum P(C_i\mid \bar{T})\log_2(P(C_i\mid T)/P(C_i)) \quad (1)$$

其中,$P(C_i\mid T)$ 表示给定文本中有特征项 T 时,该文本属于 C_i 的概率;$P(C_i\mid \bar{T})$ 表示给定样本中没有特征项 T 时,样本属于 C_i 的概率;$P(C_i)$ 表示类别 C_i 在给定样本中的概率;$P(T)$ 表示特征项 T 在样本中出现的概率,它的大小直接影响相应类别的预测[6]。

信息增益方法本质上是根据给定样本中特征项的信息增益值的大小,决定其对文本中的分类的影响程度。该方法在应用过程中没有考虑一些特殊情况,导致当特征项或者类别的分布信息不均衡时,应用信息增益的效率会大大减弱[7]。

1.3　互信息方法

互信息方法是通过计算相关词汇与给定样本中类别的关联程度,来确定词汇在文本分类中的重要性,从而确定其被选入的可能性[8]。

词汇 t 的 MI 评价函数为[9]:

$$MI(t,C_j)=\log_2(P(t/C_j)/P(t)) \quad (2)$$

其中,t 表示相关的文本特征项;C_j 表示第 j 种文本类别。

通过该公式可以计算出相关特征项与给定文本类别的关联性,其 MI 值越大,说明两者之间相关程度越高,当 MI 为 0 时,说明两者之间相互独立[10]。

1.4　交叉熵方法

交叉熵方法中词汇 t 的 CE 评价函数为[11]:

$$CE(t)=P(t)\sum_{j=1}^{k}P(C_j\mid t)\log_2(P(C_j\mid t)/P(C_j)) \quad (3)$$

该方法的基本原理是根据文本特征项在相关类别中出现的频率计算出评价函数值。在计算过程中,只考虑文本特征项在样本中出现的概率,把一些不出现的词汇作为噪声来处理,与信息增益方法相比,其在文本分类应用中更优越[12]。

2　信息熵

先从随机过程的角度介绍熵的概念。

定义 1:设 X 表示一组随机事件:$x_1,x_2,\cdots,x_n$,其中 $p(x_i)=P_i(0\leqslant P_i\leqslant 1)$ 分别是它们出现的概率,且 $P_1+P_2+\cdots+P_n=1$,则 X 的信息熵 $H(X)$ 定义为其自信息量的统计平均值,即[13]:

$$H(X)=E(I(x_i))=E(-\log_2P_i)=-\sum_{i=1}^{n}P(x_i)\log_2P(x_i)=-\sum_{i=1}^{n}P_i\log_2P_i \quad (4)$$

通过上式计算随机变量 X 的信息熵值,该值的大小直接决定随机变量对给定样本的影响程度。在进行随机试验中,有效获取随机变量的熵就能确定各随机变量的概率分布,在这里,熵作为信息的一种度量,熵值的确定也表明信息的不确定性完成。

熵最早是由 Shannon 提出的,最初是应用在统计热力学中,通过计算熵值,确定一个系统混乱度,熵值越小,其混乱度越弱,表明系统越稳定[14]。在统计热力学中,应用的是熵增原理,而对于通常的文本样本信息,其信息熵值根据要求只能减少,应用的是熵不增性原理。

定义 2:对于一个数据集合中的属性 X,它的属性域为 $\{a_1,a_2,\cdots,a_n\}$,对应的概率分布为 $\{p(a_1),p(a_2),\cdots,p(a_n)\}$,则 X 的熵定义为[15]:

$$X = \begin{pmatrix} a_1 & a_2 & \cdots & a_n \\ P(a_1) & P(a_2) & \cdots & P(a_n) \end{pmatrix} \tag{5}$$

$$H(X) = \sum_{i=1}^{n} P(a_i)\log_2 P(a_i) \tag{6}$$

在实际应用过程中,熵的使用对于无序结构比较实用。通过一个非负的熵值表述信息的不确定程度,熵作为样本中的概率分布函数,是一个全局变量,熵值的大小直接决定信息在样本中的重要程度。$H(X)$ 表达了属性 X 所包含的信息量,针对属性 X,可以通过熵来衡量其纯度,其值越小,属性 X 的纯度就越高。

在信息论中,通常通过信息熵表述信息量的多少。对于一个随机变量 X,如果用 $p(t)$ 表示 X 取值为 t 的概率,那么其信息熵表示为:

$$H(X) = \sum p(x)\log_2 p(x) \tag{7}$$

通过式 7 可以看出,信息熵的大小决定信息量的确定性,信息熵值越大,其不确性越大,反之,其不确定性越小。在实际应用过程中,信息熵值越小,对于信息量的处理越有效。

多个属性变量同时拥有的信息量称为联合熵,其表示如下:

$$H(X_1,X_2,\cdots,X_n) = -\sum x_1 \cdots \sum x_n p(x_1,x_2,\cdots,x_n)\log_2 p(x_1,x_2,\cdots,x_n) \tag{8}$$

其中,$p(x_1,x_2,\cdots,x_n)$ 表示变量 $X_1,X_2,\cdots,X_n$ 分别同时取 $x_1,x_2,\cdots,x_n$ 时的概率。

条件熵用于表示当某几个变量确定时其他变量所含的信息量,若已知变量 $X_1,X_2,\cdots,X_n$,变量 Y 的条件熵可表示为:

$$H(Y \mid X_1,X_2,\cdots,X_n) = H(Y,X_1,X_2,\cdots,X_n) - H(X_1,X_2,\cdots,X_n) \tag{9}$$

其中,$H(Y,X_1,X_2,\cdots,X_n)$ 表示变量 $Y,X_1,X_2,\cdots,X_n$ 的联合熵。

条件熵反映了成员元素之间的关联关系,其值越小,其对应的关联关系越紧密。当两个不同变量的条件熵与信息熵值相等时,则这两个变量具备完全独立性,此时,有以下关系式成立[16]:

$$H(Y \mid X) = H(Y) \tag{10}$$

3 基于信息熵的网络设备配置命令智能修复系统初步设计

3.1 基于信息熵的网络设备配置命令分类算法设计

(1)基于信息熵的网络设备配置命令分类模型。

在网络设备的配置命令分类中应用信息熵时,主要是通过信息熵值的大小来判断命令短语在已有配置命令集中是否存在相似或者相同的,当该短语在配置命令集中存在或相近时,其信息熵为 0。当检测到新的配置命令文本时,需先进行特征向量处理,在此基础上,再应用信息熵原理,把已提取的特征向量进行优化,并把得到的结果存储到配置命令集中。在整个优化过程中,信息熵值是决定一个命令短语是否存在于配置命令集合中的依据。

算法流程如下:

假设已定类别 C 中有 k 个配置文本命令集合,k 篇配置文本的特征词集合为 $\{f_{i1},f_{i2},\cdots,f_{in}\}$($i = 1,2,\cdots,k$),待加入的配置文本为 x_{k+1},其抽取的特征词集合为 $\{d_{(k+1,1)},d_{(k+1,2)},\cdots,d_{(k+1,n)}\}$。

(a)文本预处理。根据预处理流程,对配置命令文本抽取特征向量,得到相应的特征词集合 $\{f_{i1},f_{i2},\cdots,f_{in}\}$($i = 1,2,\cdots,k$)。

(b)统计词频。根据词频方法统计各篇命令配置文本中特征词的词频,并设定一个阈值。低于该阈值的特征词直接忽略,只统计高于该阈值的特征词,然后再对选出的特征词排序,在此基础上,再应用信息熵原理,计算每个特征词的熵值,当其熵值为 0 或者接近 0 时,取消该特征词。特征词的熵值计算公式如下:

$$W(t,d) = (\mathrm{tf}(t,d)x\log_2(N/n_t + 0.01))/\sqrt{\sum_{t\in d}(\mathrm{tf}(t,d)x\log(N/n_t + 0.01)^2)} \tag{11}$$

(c)计算新文本信息熵。对于新加入的配置文本命令,按照上述流程选取特征词,在此基础上,再对所得到的特征词计算其信息熵,只保留符合要求的特征词,用于实现对配置命令进行分类。

(d)反馈。将分类好的配置命令文本中的特征词加入到特征词库中,然后重新计算各个特征词的信息熵,返回执行步骤(c),继续进行配置命令文本类别的分类。在操作过程中,对于每篇配置命令文本分类后,可根据实际需要,动态调整信息熵的阈值。

(2)基于信息熵的网络设备配置命令分类算法。

在该算法中,应用词频方法初步选取特征词,在此基础上,再应用信息熵原理,对已抽取的特征词计算其信息熵值,利用其值与给定的阈值进行比较,当其值低于阈值时,直接忽略掉,排除不重要的文本,保持文本的原有数据特性。

(a)关键短语的界定。

对在给定的文本信息,适当的划分关键短语至关重要,文本可以通过关键短语鲜明的表达文本内容,关键短语的界定可从三个方面来确定:在结构上要求凝固性好;在语义方面,能清晰明确地表达词意,具有一定的专指性和完整性;在统计方面,对于任何具有完整意义的文本信息,通常具有一定的可重用性,在表述上,往往通关键短语来实现,而对于网络设备的配置命

令文本而言,短语便于表达配置命令的特征,更适合作为其特征项。

(b)英文分词。

在网络设备的配置命令文本中,英文表述较多,对于英文分词的划分也就显得格外重要,但是,作为一个英文分词系统进行的是基础性检测划分工作,对于每个信息应用领域没有必要都建立一个英文分词系统,那样浪费资源,而且分词的效果也不一定是最优的。文中网络设备的配置文本命令英文分词的划分直接使用基于 DAG 思想的配置元集无关性算法,该算法在英文文本的分类中,特征项的提取、配置命令的表示更易于实现。

(c)配置命令预处理。

配置命令文本的预处理主要是实现对配置命令划分短语,为后期确定关键短语作辅垫,最终的处理结果实现了将非结构化的配置命令结构化,从而进一步提高短语的匹配速度。具体操作流程主要分两步:第一步利用英文的标点符号及非汉字符号对英文配置命令进行切分,切分成长度较短的英文短语;第二步利用汉字标点符号进行切分,同时结合连接词库,把诸如"and"、"or"这样的单词去掉,使得文本句子得到进一步的切分。

(d)关键词集抽取。

网络设备的配置命令文本经过文本预处理后,使得原有的命令文本变成了分割的短语。为了便于处理,把这些分割的短语存放至一个集合 T 中,通过对集合 T 不断进行扫描,最终把确定的关键短语存储到 keyset 集合中,配置文本命令的扫描处理过程如下:

· 通过对预处理的命令文本利用词频方法统计各个分词的词频,将达到给定阈值的分词提取出来,存储到集合 keyset 中。

· 对于未有完整意义的短语,查看停用词库,判断其是否包含不能用的单词,如果是,直接删除,如果否,则转入下一步。

· 对于给定的预处理配置命令文本,按照顺序依次自前往后扫描,直到预处理文本所有短语处理完毕为止。

· 对于集合 keyset 中短语应用信息熵原理,进一步优化,实现对文本配置命令的分类处理。

在此基础上分别对每类网络设备的配置文本命令进行比较统计,将不同的关键词集进一步组成每类的关键词组描述集合。网络设备的配置命令文本处理过程主要包括四大基本模块:特征库的建立、样本训练、短语测试、文本命令分类。系统框架如图 1 所示。

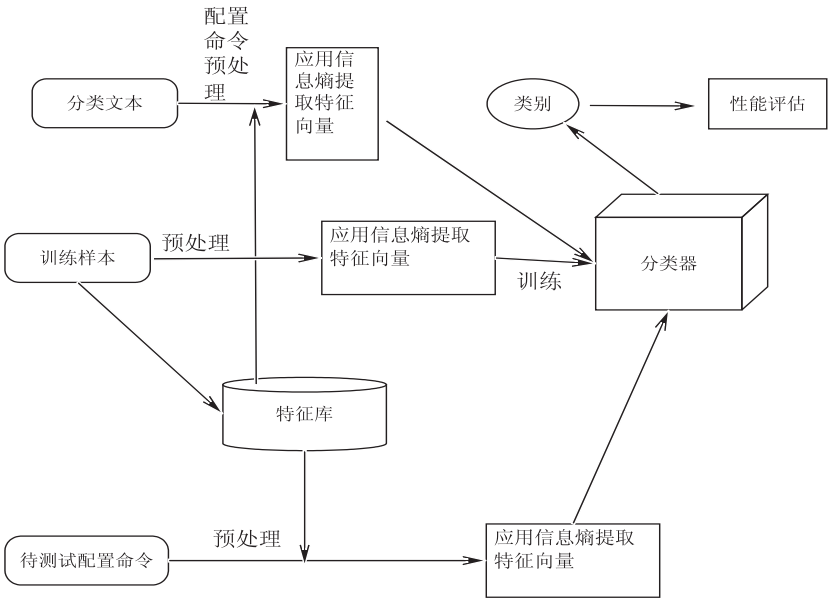

图 1　结构框架

系统的处理过程如下:

系统对于训练文本,先按照词频方法选取特征向量,将选取的特征向量存储到特征库;对于配置命令分类文本,根据特征库的信息划分关键短语,在此基础上,再应用信息熵原理对优化的命令文本进行分类处理,最后对分类结果进行评估;对于测试命令文本首先根据特征向量库提取文本中的特征向量,在此基础上,再应用信息熵原理,对已提取的特征向量进行优化,以提高分类器中进行分类的准确率。

3.2　基于信息熵的网络设备配置命令智能修复系统总体设计

该系统的总体结构包括数据采集、处理、传输以及基本的数据更新功能,其结构如图 2 所示。

(1)自动配置子系统。

主要用于实现管理员在自己所属的网络管理区域对相应的网络设备进行配置及查询其相应的设备配置

信息。该子系统主要利用配置命令知识库生成子系统实现自动配置,在配置过程中,可能由于使用的是最新的网络设备,在这种情况下,配置命令知识库生成子系统会找不到匹配的设备配置文本。如果在配置过程找不到相对应的配置文本,则需要管理员手动输入正确的网络设备配置命令文本,然后通过网络设备的配置命令分类算法将其在配置命令知识库自动生成相应的网络设备配置文本。

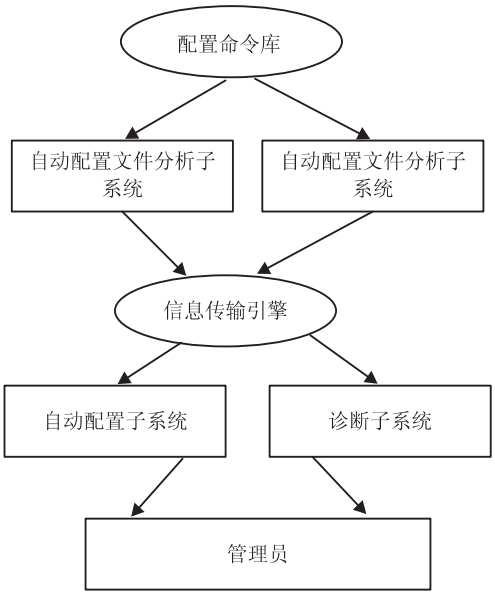

图 2 配置命令智能修复系统结构

(2)诊断子系统。

该子系统主要用于实现对所监控的网络设备的配置命令文本进行智能匹配,具体的操作通过在知识库的命令集中搜索与源配置文件相一致的配置命令文本,在启用诊断子系统时,需要设置匹配差异度。在智能诊断中,当智能匹配结果高于所设的差异度值时,则其相应的网络设备配置命令文本被诊断出错。

(3)配置文件分析子系统。

在该子系统中结合基于信息熵的网络设备配置命令分类算法实现对设备的源配置命令文本进行分析,生成有规则的命令集与配置实例供诊断子系统使用。

(4)配置命令知识库生成子系统。

该子系统通过对配置命令库中的命令文本使用基于信息熵的配置命令分类算法进行初步处理,实现对设备的配置命令分类,同时,分析各分类后的命令文本的相关性及其生成规则,最终形成系统的配置命令知识库,它主要用在诊断子系统和自动配置子系统中。

整个系统的工作流程主要包括以下步聚:首先对网络设备的配置命令文本进行提取,经过基于信息熵的配置命令文本分类算法进行分类处理后,进入到配置命令知识库,然后配置命令知识库生成子系统,使用配置命令知识库进行样本训练,实现对配置命令知识库进行优化处理。配置命令知识库可供配置文件分析子系统和诊断子系统使用,在诊断子系统中,使用配置文件分析子系统所得到的结果与命令知库中的配置实例进行快速匹配,根据所得结果的差异度值,判断设备的配置命令文本是否出错,当诊断结果显示有错误时,可通过直接重写配置文件方式避免定位出错位置的操作。

4 仿真实验结果

Simulink 是 MATLAB 软件支持的一个包,该软件包的特点是图形用户界面使用灵活,用户只需要通过简单的鼠标操作即可完成建模任务,不再需要使用复杂的命令。在建模过程中,通过选择合适的模块,把上述算法中的各个功能一一实现,然后再把各个模块结构连接起来,最后进行调试与模拟仿真。实验结果用 MATLAB 进行仿真描述,图 3 和图 4 为网络设备配置命令文本使用信息熵原理分类后与未归类在网络设备的诊断响应时间与诊断正确率方面进行的对比。

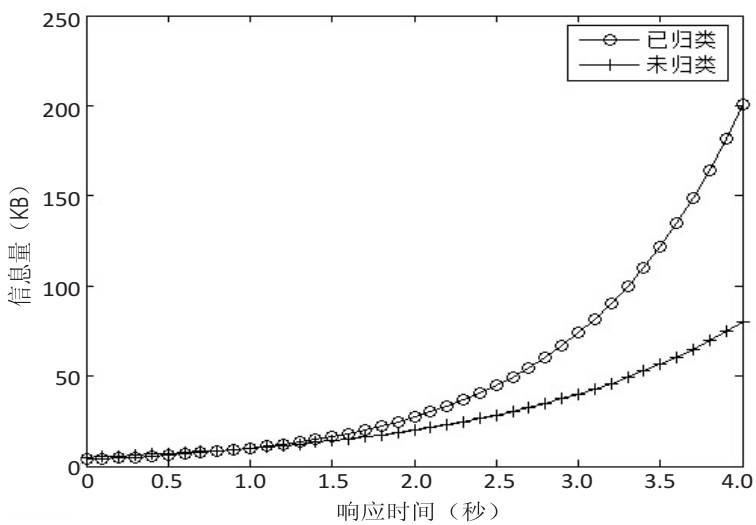

图 3 诊断响应时间比较

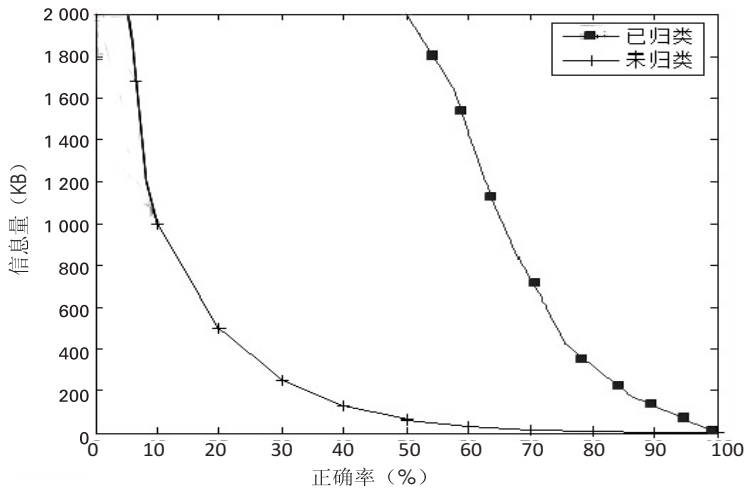

图 4　诊断正确率比较

通过以上两图可以看出,应用信息熵原理对网络设备配置命令分类后,在诊断正确率方面优势更明显,在诊断的响应时间方面,当配置命令文本信息量不大的情况下,两者相差不大,随着信息量越大,两者之间的差别也越大。文中算法不能保证配置命令诊断完全正确,但相比较未分类之前的配置命令文本,应用信息熵原理进行分类处理后,其在诊断响应与诊断正确率方面均有了一定程度的提高。

5　结束语

在实际的网络设备故障诊断中,待处理的信息量非常大,而其中对于网络设备故障诊断有帮助的数据只是其中很小一部分,通过在网络设备配置命令智能修复系统中应用基于信息熵的配置命令文本分类算法,可以有效提高网络设备在智能检错中的效率。通过在 MATALAB 中安装 Simulink 软件包,仿真实现了一般网络设备配置文本命令与基于信息熵的配置文本命令在配置故障诊断方面的影响。仿真结果显示,基于信息熵的网络设备配置命令文本在分类处理后,在网络设备由于配置出错导致不能正常运行时,其故障诊断在响应时间和诊断结果正确性方面都有较为明显的优势,为后续的网络设备故障的智能修复提供了保障。

参考文献:

[1] 方　磊,马溪骏. 基于信息熵的改进型支持向量机客户流失预测模型应用研究[J]. 情报学报,2011,30(6):643-648.

[2] 宋洪涛,王小峰,王勇军,等. 基于信息熵的分布式拒绝服务攻击协同检测系统的设计与实现[J]. 小型微型计算机系统,2015,36(1):133-137.

[3] 冯　建,Starzyk Janusz,邱菀华. 一种基于信息熵的金融数据神经网络分类方法[J]. 控制与决策,2012,27(2):211-215.

[4] 王素立,余建国. 基于主分量变换的决策树模型构建方法[J]. 技术经济与管理研究,2015(1):8-12.

[5] 李　波,陆海燕,毕雪梅. 基于土地利用结构信息熵模型的城市边界遥感识别研究[J]. 国土资源情报,2011(12):42-45.

[6] 窦丹丹,姜洪开,何毅娜. 基于信息熵和 SVM 多分类的飞机液压系统故障诊断[J]. 西北工业大学学报,2012,30(4):529-534.

[7] 葛新民,范宜仁,唐利民,等. 基于信息熵-模糊谱聚类的非均质碎屑岩储层孔隙结构分类[J]. 中南大学学报:自然科学版,2015(6):2227-2235.

[8] 陈　黎,周海海. 基于信息熵的产品造型风格形成原理与设计决策顺序研究[J]. 工程图学学报,2012,33(1):31-37.

[9] 郭红钰. 基于信息熵理论的特征权重算法研究[J]. 计算机工程与应用,2013,49(10):140-146.

[10] 张云雷. 一种基于信息熵的 web 信息提取的方法研究[J]. 科技资讯,2012(22):12.

[11] 杨胜刚,朱　琦,成　程. 个人信用评估组合模型的构建——基于决策树-神经网络的研究[J]. 金融论坛,2013,18(2):57-61.

[12] 刘　勘,袁蕴英,刘　萍. 基于随机森林分类的微博机器用户识别研究[J]. 北京大学学报:自然科学版,2015,51(2):289-300.

[13] 钱文彬,杨炳儒,徐章艳,等. 基于信息熵的核属性增量式高效更新算法[J]. 模式识别与人工智能,2013,26(1):42-49.

[14] 刘　扬. 基于决策树的轨道电路故障诊断知识表示方法研究[J]. 邵阳学院学报:自然科学版,2014(4):18-23.

[15] 李英英,纪昌杰. 基于信息熵加权去噪的半监督 SVM 分类器[J]. 电脑知识与技术,2013,9(25):5705-5707.

[16] 郑向阳,何　倩. 基于信息熵原理的水果定位检测方法[J]. 计算机仿真,2012(4):279-281.