基于 N-Gram 的 SQL 注入检测研究

万卓昊 徐冬冬 梁 生 黄保华
(广西大学计算机与电子信息学院 南宁 530004)

摘 要 SQL 注入攻击是 Web 面临的主要安全威胁,文中针对 SQL 注入难以检测的问题,提出基于 N-Gram 的 SQL 注入检测方法。该方法基于 N-Gram 将 SQL 语句转换成固定维数的特征向量,并采用改变不同特征子序列权重的方法改进距离,将改进距离和卡方距离通过 BP 神经网络计算得到的模糊距离作为向量间的距离标准。首先计算安全 SQL 语句的平均特征向量,然后计算各 SQL 语句与平均特征向量的距离以确定距离的阈值,接着将据待测 SQL 语句与平均特征向量的距离与阈值进行对比,以判断待测 SQL 语句的安全性。实验结果表明,与直接使用单词构成的特征向量相比,所提方法能有效提高检测率、降低误报率。

关键词 SQL 注入,N-Gram,特征向量,神经网络
中图法分类号 TP309 **文献标识码** A **DOI** 10.11896/j.issn.1002-137X.2019.07.017

Study on SQL Injection Detection Based on N-Gram

WAN Zhuo-hao XU Dong-dong LIANG Sheng HUANG Bao-hua
(School of Computer and Electronic Information, Guangxi University, Nanning 530004, China)

Abstract SQL injection attack is the main security threat faced by Web. Aiming at the problem that SQL injection is hard to detect, this paper proposed an SQL injection detection method based on N-Gram. The method transforms the SQL statements into the feature vectors with fixed dimension based on N-Gram, and the distance is improved by changing the weights of different feature subsequences. The fuzzy distance obtained from the improved distance and chi-square distance through BP neural network is used as the distance criterion between vectors. Firstly, the average feature vector of the secure SQL statements is calculated. Then, the distances between every SQL sentence and average feature vector are calculated to determine the distance threshold. The distance between the unknown SQL statement and the average feature vector is compared with the distance threshold to judge the safety of the unknown SQL statement. The experimental results show that the proposed method can effectively improve the true positive rate and reduce the false positive rate in terms of detection compared with the feature vector directly composed by words.

Keywords SQL injection, N-Gram, Feature vector, Neural network

1 引言

注入类漏洞在 OWASP(Open Web Application Security Project)于 2013 年和 2017 年发布的最严重的前 10 名 Web 安全漏洞中都排名第一,而其中最典型的就是 SQL 注入漏洞。SQL 注入利用现有应用程序,将恶意的 SQL 命令注入到后台数据库,使之不能按照设计者意图去执行 SQL 语句。

针对 SQL 注入的隐蔽性非常高导致普通网络防火墙很难发现的特点,国内外提出了很多检测策略。文献[1-5]使用预先定义好的常用关键词将 SQL 语句向量化,然后对向量进行学习和分类。文献[6-11]分别使用 Word2vec, N-Gram 等方法从训练集中提取特征向量,然后用 SVM 进行分类。文献[12]提出一个 N-Gram 特征有向选择公式,选择特征浓度高的特征子序列作为特征向量。文献[13]将模式匹配和特征过滤两种方法混合使用,在特征过滤时,先对不同特征赋予不同权重,再进行累加计算判断安全性。文献[14-17]提出通过集成两种不同方法形成两道防线来判断 SQL 语句的安全性。文献[18]提出替换非 SQL 关键字的字符串、算术表达式、组合运算符等,根据替换后的语法结构串判断安全性。

文献[1-5]根据关键词的出现与否,将 SQL 语句转化为可以具体量化的特征向量以便于后续分类,然而该方法使用预先自定义的特征词进行检测,只能检测出已知类型的注入语句,并且只能通过词的出现与否进行判断,没有考虑词与词之间的联系。文献[6-11]指出,根据训练集动态生成的特征

到稿日期:2018-06-04 返修日期:2018-10-22 本文受国家自然科学基金项目(61262072)资助。
万卓昊(1993-),男,硕士生,主要研究方向为数据库安全;徐冬冬(1993-),男,硕士生,主要研究方向为数据库安全;梁 生(1992-),男,硕士生,主要研究方向为数据库安全;黄保华(1973-),男,博士,副教授,CCF 高级会员,主要研究方向为数据库安全等,E-mail:bhhuang66@gxu.edu.cn(通信作者)。

向量更具代表性，更能表示 SQL 语句的特征，但是使用 SVM 分类时不能为不同特征子序列赋予不同权重。文献[12]指出，在特征选择时，只有特征子序列在不安全特征中所占的比例越大，被选为有效特征的概率才越大，而实际上那些几乎只出现在安全 SQL 语句中的特征子序列同样也具有很强的代表性。文献[13]预先给每个特征字符串赋予特定的权重，但无法对根据训练集动态获取的特征子序列加权。文献[13-17]都是先后采用两种不同方法判断 SQL 语句的安全性，只能单方面提高检测率或降低误报率，无法对这两个指标同时优化。文献[18]替换了表达式、嵌套查询语句等可能存在注入的部分，影响了检测的准确性。

为解决这些问题，本文引入计算距离度量相似度和信息增益，在特征选择时以信息增益为权重值，在计算距离时为不同特征子序列添加不同的权重值。在分词时，该方法仅对数字和字符串进行替换。对于不同加权方式得到的距离，结合 BP 神经网络，将多种影响因素综合在一起，把 3 种距离统一成一种检测距离。本文通过对 SQL 语句进行 N-Gram 特征提取和信息增益特征选择得到特征向量，根据特征向量向量化 SQL 语句，并求出安全 SQL 语句的平均特征向量。利用 BP 神经网络分别计算安全 SQL 语句和不安全 SQL 语句到平均特征向量的模糊距离以确定阈值，然后计算待测 SQL 语句到平均特征向量的距离，最后将其与阈值进行比较来判断 SQL 语句的安全性。

2　SQL 语句向量化

2.1　分词

SQL 语句分词和英语文本分词类似，使用空格和一些常用符号（如逗号、括号、运算符、注释符等）进行切分即可。此外，SQL 语句中通常存在数字或字符串，包括用户输入的查询关键字、用户名称、id、页码等，无论其内容是各个数字，还是各种字符串，对该条 SQL 语句的安全性的影响都并无区别。如果直接统计原始 SQL 语句分词，那么那些具有相同结构、不同数字或字符串的子序列，如“or 1=1”与“or 2=2”，将会被算作 2 个不同的子序列进行统计。如此一来，它们对整体的影响将会被稀释，并且当待检测语句中含有“or 3=3”时，也无法准确判断其安全性。

本文统一替换了 SQL 语句中的数字和字符串，替换前，特征子序列共 26246 个，替换后为 11586 个，这在减少计算工作量的同时避免了特征向量中不同的具体数字或字符串对检测准确度的影响。

2.2　N-Gram 特征提取

N-Gram 模型主要应用于自然语言处理中，该模型基于马尔科夫假设。N-Gram 特征提取是指把长度为 N 个词的窗口，从 SQL 语句的第一个词开始，从左向右每次滑动 1 个词，将每次窗口中同时出现的 N 个词当作一个新的词提取，即把 N 个连续的词看作一个整体进行提取。当 $N=3$ 时，滑动窗口的切分示意图如图 1 所示。

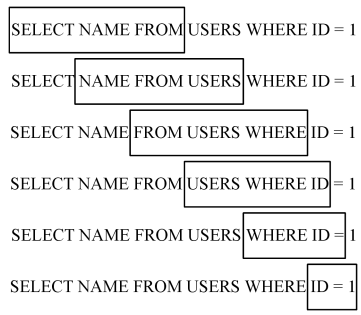

图 1　3-gram 切分示意图

Fig. 1　3-gram segmentation diagram

对于一个长度为 L 个单词的 SQL 语句，以单个词进行切分，即窗口大小为 1 的 1-gram 切分，会产生 L 个特征子序列；当窗口大小为 N 时，会产生 $L-N+1$ 个特征子序列。

以“SELECT NAME FROM USERS WHERE ID=1”为例，N-Gram 的切分结果如下：

1-gram：{SELECT，NAME，FROM，USERS，…}

2-gram：{SELECT NAME，NAME FROM，FROM USERS，…}

3-gram：{SELECT NAME FROM，NAME FROM USERS，…}

4-gram：{SELECT NAME FROM USERS，…}

2.3　特征选择

特征选择即从特征提取所得到的所有特征子序列中选择一部分构成一个可以代表整个 SQL 语句的特征向量。特征选择的好坏将直接影响分类的准确率和效果，如果选择不当，会导致存在过多不相关的特征子序列，不但会增加消耗，还无法达到预期的效果。

本文采用信息增益法进行特征选择，信息增益是针对每个特征子序列而言的，对于某一特征，包含该特征和不包含该特征时整体的信息熵的差值即为该特征为整体带来的信息量。对于一个 N-Gram 特征 T，其信息增益 $IG(T)$ 的计算式如式(1)所示：

$$IG(T)=H(C)-H(C|T) \tag{1}$$

其中，$H(C)$ 为类别 C 的熵；$H(C|T)$ 是特征为 T 时，类别 C 的条件熵。假设类别 C 有 n 个不同取值 C_1，C_2，C_3，…，C_n，则 $H(C)$ 的计算式如式(2)所示：

$$H(C)=-\sum_{i=1}^{n}P(C_i)\log_2 P(C_i) \tag{2}$$

其中，$P(C_i)$ 为 C_i 出现的概率。根据条件概率式可得 $H(C|T)$ 的计算式如式(3)所示：

$$\begin{aligned}H(C|T)&=P(t)H(C|t)+P(\bar{t})H(C|\bar{t})\\&=-P(t)\sum_{i=1}^{n}P(C_i|t)\log_2 P(C_i|t)-\\&\quad P(\bar{t})\sum_{i=1}^{n}P(C_i|\bar{t})\log_2 P(C_i|\bar{t})\end{aligned} \tag{3}$$

其中，$P(t)$ 为 T 出现的概率；$P(\bar{t})$ 为 T 不出现的概率；$P(C_i|t)$ 为特征 T 出现的情况下 C_i 出现的概率；$P(C_i|\bar{t})$ 为特征 T 不出现的情况下 C_i 出现的概率。

综上，N-Gram 特征 T 的信息增益计算式如式(4)所示：

$$IG(T)=-\sum_{i=1}^{n}P(C_i)\log_2 P(C_i)+P(t)\sum_{i=1}^{n}P(C_i|t)\log_2 P(C_i|t)+P(\bar{t})\sum_{i=1}^{n}P(C_i|\bar{t})\log_2 P(C_i|\bar{t}) \quad (4)$$

对于 SQL 语句而言,样本类别只有安全 C_1 和不安全 C_2 两种,$P(C_1|t)$和 $P(C_2|t)$分别为该特征值出现时 SQL 语句安全和不安全的概率。同理,$P(C_1|\bar{t})$和 $P(C_2|\bar{t})$分别为该特征值未出现时 SQL 语句安全和不安全的概率。本次实验使用的安全 SQL 语句与不安全 SQL 语句的数量相同,可在一定程度上减少计算所需工作量。

特征向量的生成步骤如下:

1)提取所有样本的 *N*-Gram 特征子序列;

2)计算每个特征子序列的信息增益,并按其信息增益由大到小排序;

3)选取前 n 个 *N*-Gram 特征子序列作为特征向量。

对于每一条 SQL 语句,如果语句中存在该特征子序列,则对应属性值为 1,否则为 0。最终将一条 SQL 语句转化为一个 n 维向量。

3 SQL 语句安全性判断

3.1 距离度量

根据特征向量遍历每一条 SQL 语句,将语句中包含的子序列记为 1,不含的子序列记为 0,每条 SQL 语句可转化为与特征向量维度相同的 0,1 序列。对每条安全 SQL 语句的特征向量取平均值,即可得到安全 SQL 语句的平均特征向量。针对每一条安全 SQL 语句和不安全 SQL 语句分别计算它们与安全 SQL 语句的平均特征向量的距离,确定安全与不安全的分界阈值。对于每一条待测 SQL 语句,只需计算其与安全 SQL 语句平均特征向量之间的距离,再将其与阈值比较,即可确定其安全性。具体步骤如下:

1)将安全 SQL 语句转化为 *N*-Gram 特征向量,计算所有安全 SQL 语句的平均特征向量;

2)分别计算所有安全 SQL 语句和不安全 SQL 语句与平均特征向量之间的距离,确定距离阈值;

3)将待测 SQL 语句转化为 *N*-Gram 特征向量,计算该特征向量和平均特征向量之间的距离,若该距离小于阈值,则该待测 SQL 语句为安全,否则为不安全。

3.2 卡方距离

卡方距离(Chi-square measure)利用列联表分析的方法得到一个卡方统计量来衡量两个个体之间的差异性,卡方统计量越大表明个体的选择对变量的取值的影响越显著,即两个个体之间差异越大,其计算式如式(5)所示:

$$d(x,y)=\sqrt{\sum_{i=1}^{n}\frac{(x_i-E(x_i))^2}{E(x_i)}+\sum_{i=1}^{n}\frac{(y_i-E(y_i))^2}{E(y_i)}} \quad (5)$$

其中,n 是 x 和 y 的维度;x_i 是个体 x 的第 i 个变量的取值(在第 i 类上的频数),$E(x_i)$是个体 x 在第 i 类上的期望频数;y_i 是个体 y 的第 i 个变量的取值,$E(y_i)$是个体 y 在第 i 类上的期望频数。

由于其中一个个体 y 为平均特征向量,即 $y_i=E(y_i)=E(x_i)$,因此卡方距离公式可简化为:

$$d(x,y)=\sqrt{\sum_{i=1}^{n}\frac{(x_i-y_i)^2}{y_i}} \quad (6)$$

当待测 SQL 语句中存在安全 SQL 语句中没有的 *N*-Gram 特征子序列时,分母 y_i 的值为 0。此时,令分母为该 *N*-Gram 特征子序列仅出现 1 次时的频数,进行平滑处理,以避免出现分母为 0 的情况。

3.3 改进距离

由于不同长度的特征子序列对结果的影响不同,本文参考欧氏距离公式(即式(7))对距离计算公式做出修改。

$$d(x,y)=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\cdots+(x_n-y_n)^2}=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \quad (7)$$

其中,将不同维度的距离平方后再累加,当 $n=4$ 时,1-gram,2-gram,3-gram,4-gram 可分别看作欧氏距离中 4 种维度的距离。先分别计算出每种 *N*-Gram 的距离,再取平方和,这样,按照 *N*-Gram 的种类,原本独立的子序列被分成了 4 个部分,在后续研究中,不仅可以对每个子序列加权,还可以根据 *N*-Gram 的种类对它们再次进行加权。改进后的原卡方距离可表示为:

$$d(x,y)=\sqrt{\sum_{j=1}^{4}\Big(\sum_{x_i,y_i\in j\text{-gram}}\frac{|x_i-y_i|}{y_i}\Big)^2} \quad (8)$$

其中,$x_i,y_i\in j$-gram 表示 x_i 和 y_i 所对应的特征子序列是 j-gram 子序列,$j=1,2,3,4$。

使用卡方距离计算距离时,各维度的权重相同,而实际上不同特征子序列对安全性判断结果的影响不同,越靠前的子序列对结果的影响越大。因此,可进一步改进式(8),在累加前,使每一维度先乘以对应的权重。由于特征选择时已经将特征子序列按照信息增益从大到小排序,本文选择的权重为关于顺序的一次函数和对应子序列的信息增益值。

顺序权重表示为:

$$d(x,y)=\sqrt{\sum_{j=1}^{4}\Big(\sum_{x_i,y_i\in j\text{-gram}}(-ki+b)\cdot\frac{|x_i-y_i|}{y_i}\Big)^2} \quad (9)$$

其中,$-ki+b(k>0,b>0)$为关于 i 的一次函数,i 越小函数值越大,即顺序靠前的子序列占更大的权重。

信息增益权重表示为:

$$d(x,y)=\sqrt{\sum_{j=1}^{4}\Big(\sum_{x_i,y_i\in j\text{-gram}}IG(x_i)\cdot\frac{|x_i-y_i|}{y_i}\Big)^2} \quad (10)$$

其中,$IG(x_i)$为 x_i 对应特征子序列的信息增益。

3.4 BP 神经网络

BP(Back Propagation)神经网络是 1986 年由 Rumelhart 等科学家提出的概念,是一种按照误差逆向传播算法训练的多层前馈神经网络[19],是目前应用最广泛的神经网络。

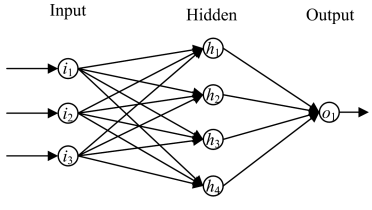

图 2 3 层 BP 神经网络结构图

Fig. 2 Diagram of three-layer BP neural network structure

本文选用一种输入层有 3 个输入单元，隐藏层有 4 个单元，输出层有 1 个单元的神经网络结构进行模糊计算，其结构如图 2 所示。首先将归一化后的 3 种距离作为输入，SQL 语句的安全与否（安全 SQL 语句为 0，不安全 SQL 语句为 1）作为输出，进行训练。然后再次以这些 SQL 语句为输入计算输出值，确定分类阈值，通常为 0.5 左右。最后，以待测语句的 3 种距离为输入，计算输出的模糊距离，并将其与阈值进行比较并判断其安全性。

4　实验验证

4.1　评价标准

本文中 SQL 语句分为两种：1）安全 SQL 语句，记为 positive；2）不安全 SQL 语句，记为 negative。当安全性被正确判断时，用 true 标记；当安全性被判断错误时，标记为 false。如表 1 所列，检测安全性时，会出现以下 4 种情况：

1）安全 SQL 语句被判断为安全 TP；

2）安全 SQL 语句被错误地判断为不安全 FN；

3）不安全 SQL 语句被判断为不安全 TN；

4）不安全 SQL 语句被错误地判断为安全 FP。

表 1　SQL 语句分类标识

Table 1　SQL statement classification identification

	判断为安全(P)	判断为不安全(N)
实际安全(P)	TP	FN
实际不安全(N)	FP	TN

根据这 4 种标识，可以计算出各评价指标。

精确率（Precision，P）：指判断为安全且实际也安全的语句数除以被判断为安全的语句数，其计算公式如式（11）所示：

$$P=\frac{TP}{TP+FP} \tag{11}$$

检测率（True Positive Rate，TPR）：指判断为安全且实际也安全的语句数除以实际安全的语句数，其计算公式如式（12）所示：

$$TPR=\frac{TP}{TP+FN} \tag{12}$$

综合评价指标（F1-Measure，F1）：由于 Precision 和 Recall 有时是矛盾的，因此无法根据这两个值来对比模型的好坏。$F1$ 值是精确率和检测率的调和均值，当 $F1$ 较高时则比较说明实验方法比较理想，其计算公式如式（13）所示：

$$F1=\frac{2P\cdot TPR}{P+TPR}=\frac{2TP}{2TP+FP+FN} \tag{13}$$

误报率（False Positive Rate，FPR）：指被判断为安全的不安全语句数除以实际上不安全的语句数，其计算公式如式（14）所示：

$$FPR=\frac{FP}{FP+TN} \tag{14}$$

4.2　实验结果

本文使用的 SQL 语句语料库均来源于互联网，共包括 2018 条 SQL 语句，其中安全 SQL 语句和不安全 SQL 语句各 1009 条。选取其中 336 条安全 SQL 语句和 336 条不安全 SQL 语句用来计算平均特征向量和距离阈值，其余作为待测语句。

经过 N-Gram 特征提取后，产生的特征子序列个数与 N-Gram 窗口大小的关系如表 2 所列。

表 2　不同 N-Gram 窗口大小产生的特征子序列个数

Table 2　Number of feature subsequences generated by N-Gram with different window sizes

N-Gram	特征子序列个数
1-gram	884
2-gram	2565
3-gram	3709
4-gram	4428
5-gram	4825

由于 1-gram 特征子序列共 884 种，且部分子序列出现频数较低，无法作为特征代表所有 SQL 语句，因此本文特征向量维数的范围为 50 到 500。分别以 50 维和 500 维特征向量进行实验，结果如表 3 所列。

表 3　不同 N-Gram 窗口大小的检测结果

Table 3　Detection results of different N-Gram window sizes

N-Gram	特征向量维数	检测率/%	误报率/%
1-gram	50	97.02	2.67
2-gram	50	96.58	3.26
3-gram	50	98.36	3.71
4-gram	50	95.69	6.24
5-gram	50	93.46	8.61
1-gram	500	89.15	11.88
2-gram	500	95.83	5.64
3-gram	500	97.62	3.12
4-gram	500	97.17	4.60
5-gram	500	94.50	7.72

由表 3 可以看出，当特征向量维数较小时，4-gram 的检测效果比 1-gram 和 2-gram 差，但比 5-gram 检测效果好；当特征向量维数较大时，由于特征子序列个数较少，1-gram 和 2-gram 会有一些信息增益较小的子序列被选入特征向量中，此时 4-gram 的检测效果比 1-gram 和 2-gram 好，并且 4-gram 依然比 5-gram 的检测效果好。另外，随着 N-Gram 窗口的增大，提取的特征子序列种类数逐渐增加，运算所需的时间和空间也随之增加，因此，本文中的特征向量由 1-gram 特征子序列、2-gram 特征子序列、3-gram 特征子序列、4-gram 特征子序列构成。

1）使用 N-Gram 特征子序列并以单词为特征子序列，根据卡方距离进行 SQL 安全性检测，结果如图 3 所示。

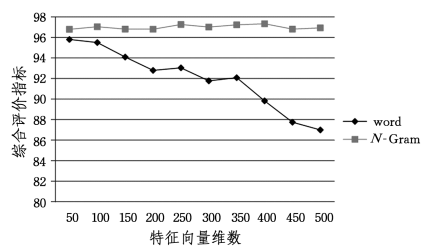

图 3　N-Gram 和单词特征向量的检测效果对比

Fig. 3　Comparison of detection effect for N-Gram and word feature vector

由图 3 可以看出，使用 *N*-Gram 特征提取的特征向量比直接使用分词后的单词构成的特征向量的检测效果好，尤其是特征向量维数较大时效果提升更为明显。由于 *N*-Gram 特征子序列中除了单词以外还包含连续多个单词组成的词组，因此与单纯的单词特征子序列相比，其更具代表性。1 个 4-gram 子序列不仅包含了 4 个 1-gram、3 个 2-gram、2 个 3-gram，还意味着这些 *N*-Gram 子序列以特定的顺序出现，所以多元的子序列构成的特征向量更能代表各 SQL 语句的特征。因此，使用 *N*-Gram 特征向量比不使用 *N*-Gram 特征向量的检测效果好。随着特征向量维数的增加，被选入特征向量的特征子序列个数也增加，能够入围特征向量的特征子序列的信息增益要求也逐渐降低，这就导致越来越多的低信息增益的特征子序列被加入到向量间距离的计算中来。使用卡方距离计算向量间的距离时，所有特征子序列的权重相同，对最终距离的影响也相同，实际上对结果影响小的子序列在计算过程中与影响大的子序列具有相同大小的影响力，这就会使得检测效果随着特征向量维数的增加而越来越差。而采用 *N*-Gram 特征子序列之后，特征子序列总数剧增，各个相邻特征子序列的信息增益差值被缩小，*N*-Gram 特征子序列第 500 个子序列的信息增益仅与单词特征子序列中第 100 个子序列的信息增益相等，因此，当特征向量维数在 50 至 500 之间波动时，使用 *N*-Gram 特征向量的检测效果并无太大波动。

2）使用单词构成的特征向量采用各种距离进行检测，结果如图 4 所示。

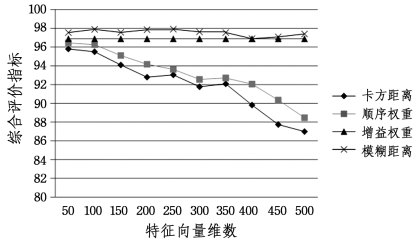

图 4 单词特征向量卡方距离和改进距离的比较

Fig. 4 Comparison of chi-square distance and improved distance of word feature vectors

由图 4 可以看出，对于单词构成的特征向量，使用顺序权重改进距离检测后，检测效果虽然有所提升，但随着特征向量维数的增加，*F*1 值依然持续降低。这是因为随着特征向量维数的增加，顺序权重线性降低，而特征子序列的信息增益变化幅度不稳定，单词特征子序列总个数少，降低幅度更大，检测效果依旧随着特征向量维数的增加而变差。当使用信息增益权重改进距离时，随着特征向量维数的增加，新增的特征子序列对结果的影响与其信息增益直接相关，且序号越大的特征子序列影响越小，以至于无法改变检测效果。因此，使用信息增益权重改进距离后，检测效果并未随特征向量维数的变化而变化。BP 神经网络能够根据输出调节各权重，因此，根据模糊距离检测的效果略好于其他距离检测效果。根据模糊距离检测的效果略好于其他距离检测效果，且 *F*1 值在信息增益权重检测效果附近浮动。

3）对 *N*-Gram 特征向量采用各种距离检测，结果如图 5 所示。

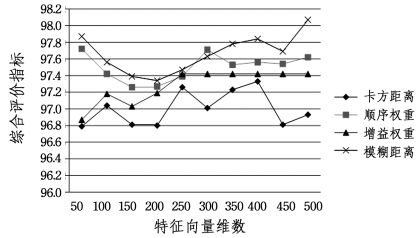

图 5 *N*-Gram 特征向量卡方距离和改进距离的比较

Fig. 5 Comparison of chi-square distance and improved distance of *N*-Gram feature vectors

由图 5 可以看出，对于 *N*-Gram 特征向量而言，使用顺序权重改进距离检测时，随着特征向量维数的增加，检测效果的变化趋势与使用卡方距离进行检测时的变化趋势相似，虽然幅度不大，但检测效果仍然得到了提高。较使用卡方距离而言，使用顺序权重改进距离检测对 *N*-Gram 特征向量的提升与对单词特征向量的提升基本相同；其检测效果与卡方距离检测效果随特征向量维数的变化而变化的趋势相似，仅在具体数值上稍有提升。使用信息增益权重改进距离检测时，由于 *N*-Gram 特征子序列之间信息增益差值不大，当特征向量维数小于 250 时，检测效果随着特征向量维数的增加而提升。当特征向量维数达到 250 时，后续增加的特征子序列的信息增益较小，不足以改变检测效果。使用模糊距离进行检测时，检测效果随顺序权重的检测效果上下浮动。

综上，使用信息增益作为权重，特征向量维数达到一定数目后，检测效果不再随着特征向量维数的变化而变化，且效果优于卡方距离检测，尤其是对于特征子序列的信息增益差值较大的单词特征向量的提升更为显著。使用顺序作为权重时，检测效果提升幅度较为稳定，检测效果随特征向量维数变化而变化的效果也与使用卡方距离的检测效果相似。使用通过 BP 神经网络综合各种距离计算出的模糊距离检测时，检测效果优于其他几种距离的检测效果，且 F1 值在其他距离检测的最大值附近浮动。

4）特征向量维数为 500 时将模糊距离检测效果与文献[3]、文献[5]、文献[6]和文献[14]中的方法进行对比，结果如表 4 所列。

表 4 检测效果对比

Table 4 Comparison of detection results

（单位：%）

检测方法	检测率	误报率
本文方法	98.95	2.82
文献[3]	90.00	4.10
文献[5]	96.39	1.70
文献[6]	94.60	4.10
文献[14]	97.64	0.18

由表 4 可以看出，文献[3]的检测率最低、误报率最高，文献[5]和文献[6]的检测效果较好，本文方法的检测率最高，文献[14]的误报率最低。由于文献[3]使用的特征向量维数较低，一个特征值代表了多个同类关键字或符号的出现频率，当

其中某一个关键字或符号出现频率过高时，就会导致整个特征值过高，影响检测结果。文献[5-6]通过 SVM 进行训练分类，由于训练阶段并未考虑到信息增益等因素的影响，检测率不如本文方法。文献[14]的设计目标是减少处理时间并降低误报率，仅对第一次检测被判断为不安全的 SQL 语句进行第二次检测，并以第二次检测结果为准，而第一次检测被判断为安全的 SQL 语句则直接被认为是安全的。因此其检测率并非最高，而误报率极低。

结束语　本文采用 *N*-Gram 特征提取和特征选择得到可以代表 SQL 语句特征的特征向量，根据特征向量将 SQL 语句向量化，把 SQL 语句间的比较转化为向量间的比较。通过计算各向量与平均特征向量的距离确定距离阈值，对比待测 SQL 语句转化的向量与平均特征向量之间的距离和距离阈值来判断其安全性。实验结果表明，使用 *N*-Gram 特征向量检测比不使用 *N*-Gram 的检测效果好，尤其是特征向量较大时。改进距离后，检测效果得到进一步提升，使用 *N*-Gram 特征向量进行检测时，用顺序权重进行改进的提升效果更明显；不使用 *N*-gram 特征向量时，用信息增益权重进行改进的提升效果更明显；使用模糊距离检测时效果最好。

本文在特征选择时仅通过信息增益排序，未对 1-gram，2-gram，3-gram，4-gram 加以区分。在接下来的工作中，将着重研究特征选择的方法，结合 *N*-Gram 中 *N* 的取值和不同 *N* 值之间的比例，选择出更为合适的特征向量。在计算距离时，可对不同 *N* 值施加不同的权重，增加 3-gram 和 4-gram 对结果的影响。此外，本文仅统计了某个特征子序列是否出现，未对在同一语句中多次出现的子序列加以区分，在后续工作中，可以把出现次数作为特征向量对应维度的值，以替代目前的 0，1 值。

参 考 文 献

[1] LI H L，ZOU J X. Research of SQL Injection Detection Based on SVM and Text Feature Extraction[J]. Netinfo Security，2017，17(12)：40-46. (in Chinese)
李红灵，邹建鑫. 基于 SVM 和文本特征向量提取的 SQL 注入检测研究[J]. 信息网络安全，2017，17(12)：40-46.

[2] KAMTUO K，SOOMLEK C. Machine Learning for SQL injection prevention on server-side scripting[C]// Computer Science and Engineering Conference. IEEE，2017：1-6.

[3] WU S H，CHENG S B，HU Y. Web Attack Detection Method Based on Support Vector Machines[J]. Computer Science，2015，42(S1)：362-364. (in Chinese)
吴少华，程书宝，胡勇. 基于 SVM 的 Web 攻击检测技术[J]. 计算机科学，2015，42(S1)：362-364.

[4] SHEYKHKANLOO N M. A Learning-based Neural Network Model for the Detection and Classification of SQL Injection Attacks[C]// International Conference on Information Systems Security(ICISS 2014). 2015：16-41.

[5] CHOI J H，CHOI C，KO B K，et al. Detection of cross site scripting attack in wireless networks using n-Gram and SVM[J]. Mobile Information Systems，2012，8(3)：275-286.

[6] CHEN Z，GUO M. Research on SQL injection detection technology based on SVM[C]// MATEC Web of Conferences. EDP Sciences，2018：01004.

[7] KAR D，SAHOO A K，AGARWAL K，et al. Learning to detect SQLIA using node centrality with feature selection[C]// International Conference on Computing，Analytics and Security Trends. IEEE，2017：18-23.

[8] KAR D，PANIGRAHI S，SUNDARARAJAN S. SQLiGoT：Detecting SQL injection attacks using graph of tokens and SVM[J]. Computers & Security，2016，60：206-225.

[9] PRIYAA B D，DEVI M I. Hybrid SQL injection detection system[C]// International Conference on Advanced Computing and Communication Systems. IEEE，2016：1-5.

[10] KIM M Y，DONG H L. Data-mining based SQL injection attack detection using internal query trees[J]. Expert Systems with Applications，2014，41(11)：5416-5430.

[11] CHOI J，KIM H，CHANG C，et al. Efficient Malicious Code Detection Using N-Gram Analysis and SVM[C]// International Conference on Network-Based Information Systems. IEEE Computer Society，2011：618-621.

[12] YANG Y，JIANG G P. Improved Method of Computer Virus Signature Automatic Extraction Based on N-Gram[J]. Computer Science，2017，44(S2)：338-341. (in Chinese)
杨燕，蒋国平. 基于 N-Gram 的计算机病毒特征码自动提取的改进方法[J]. 计算机科学，2017，44(S2)：338-341.

[13] SHI C C，ZHANG T，YU Y，et al. New Approach for SQL-injection Detection[J]. Computer Science，2012，39(S1)：60-64. (in Chinese)
石聪聪，张涛，余勇，等. 一种新的 SQL 注入防护方法的研究与实现[J]. 计算机科学，2012，39(S1)：60-64.

[14] APPIAH B，OPOKU-MENSAH E，QIN Z. SQL injection attack detection using fingerprints and pattern matching technique[C]// 2017 8th IEEE International Conference on Software Engineering and Service Science (ICSESS). IEEE，2017：583-587.

[15] TIAN Y J，ZHAO Z M，WANG L J，et al. Research on Double Layer Defense Model for SQL Injection Attack Based on Classification[J]. Netinfo Security，2015(6)：1-6. (in Chinese)
田玉杰，赵泽茂，王丽君，等. 基于分类的 SQL 注入攻击双层防御模型研究[J]. 信息网络安全，2015(6)：1-6.

[16] DOGBE E，MILLHAM R，SINGH P. A combined approach to prevent SQL Injection Attacks[C]// Science and Information Conference. IEEE，2013：406-410.

[17] RAIKAR D D，KULKARNI S，DANDANNAVAR P. Preventing SQL Injection Attacks Using Combinatorial Approach[J]. International Journal of Advanced Research in Computer Engineering & Technology，2012，1(8)：46-52.

[18] ZHOU J L，WANG X F，YU S S，et al. A New Policy to Defend against SQL Injection Attacks[J]. Computer Science，2006，33(11)：64-68. (in Chinese)
周敬利，王晓锋，余胜生，等. 一种新的反 SQL 注入策略的研究与实现[J]. 计算机科学，2006，33(11)：64-68.

[19] 闻新. 应用 MATLAB 实现神经网络[M]. 北京：国防工业出版社，2015.