练数据集, 再使用 SVM 和 ELM 算法对处理后的数据分类, 提高已知和未知攻击的检测效率[17]. 可见, 使用两种检测方式融合的入侵检测方法中, 通常要训练不止一个模型, 网络行为数据要经过多个模型判断.

3 模型

记忆神经网络 (Memory Neural Network) 由 Weston J 等人提出[18]. Sukhbaatar 等人在此基础上提出了端到端的记忆神经网络 (N2N Mem Network)[19]. 模型中, 背景知识或上下文信息被存储在外置记忆单元中, 通过神经网络循环完成输入数据和记忆项之间关联关系的计算, 过程中自动选择出 K 个最相关的记忆项融合到原始输入数据中, 用于对标签的训练.

本文使用端到端记忆神经网络的一种变体, 是一个基于领域知识的分类器, 其基本思想是利用领域知识辅助分类, 领域知识与当前网络行为的匹配度越高, 该知识项在分类器中发挥的作用越大, 据此设计模型架构, 主要由五个部分组成, 如图 1 所示. 网络流量数据首先与领域知识做匹配度计算, 然后将领域知识按照匹配结果融合进原始流量数据, 对融合数据分类得到最终判断结果, 同时匹配结果作为可解释信息输出.

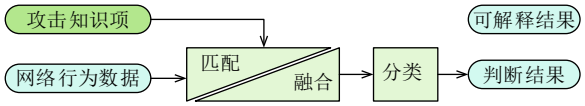

图 1 模型框架

3.1 输入输出

本文模型利用领域知识辅助对网络行为进行分类, 其输入包括领域知识和待分类的网络行为数据. 领域知识是关于攻击的知识, 可以是 IDS 规则、防火墙规则、安全人员经验数据等; 网络行为数据是反映当前网络行为或状态的数据, 一般从网络流量数据包中提取, 如连接方式、操作类型等. 模型中领域知识应当与网络行为数据具有关联关系, 才能使用领域知识辅助对网络行为分类.

领域知识由若干条彼此独立且具有实际意义的领域知识项组成, 在训练和使用前静态加载进入模型, 用来表征关于攻击的特点, 在模型中用一个矩阵M来表示:

$$M = \begin{pmatrix} m_1 & \cdots & m_n \end{pmatrix}^{\mathrm{T}} \tag{1}$$

其中, m_i表示第i条领域知识. 网络行为数据用来表征当前网络行为或状态, 用向量x表示.

模型的输出包括两部分, 一个是分类器对网络行为的分类结果, 为某一种的攻击类型, 用值$\hat{y}$表示; 另一组可解释信息, 用于向安全人员提供参考, 用值E表示.

3.2 匹配

匹配模块计算网络行为数据与知识项的匹配程度. 当网络行为与某条领域知识相匹配时, 认为网络行为倾向于是该知识项所表征的行为类别, 或者说当前网络行为具有该知识项表征行为类型的部分特点. 匹配模块中有一个匹配算法来计算网络行为与知识项的匹配度.

首先将x和M转换到统一的计算空间, 得到嵌入行为数据x'和嵌入知识矩阵M':

$$M, x \xrightarrow{V} M', x' \tag{2}$$

其中, V是一组转换矩阵.

假设g是匹配度计算函数, 则网络行为x和领域知识的匹配度为:

$$p = g(x', M') = (g_V(x, m_1), \cdots, g_V(x, m_n)) \tag{3}$$

网络行为x与知识项m_i所表征的攻击类型越接近, 相应的p_i的值越大, 表示x和m_i匹配度越高.

匹配模块的另一个作用是输出关联度高的知识项的类型标签, 作为最终分类结果可解释信息.

3.3 融合

融合模块试图在分类计算之前将原始输入和领域知识相结合, 从而使知识信息能够在分类中发挥作用. 融合的基本思想是知识项与网络行为之间的匹配度越高, 则该知识项在分类器中发挥的作用越大. 假设重构输入为o, 需要通过原始输入x, 知识项M和匹配度p来计算.

$$o = h_p(x', M') \tag{4}$$

这样, 重构输入o就同时包含了原始输入内容和领域知识的内容.

3.4 分类

分类是模型的主要部分. 它负责将流量记录分类到预定义的类别中, 确定它们是正常流量还是其他特定的攻击类型:

$$\hat{y} = f_W(o) \tag{5}$$

3.5 模型细节

本文将采用余弦相似度作为匹配算法, 使用线性加权作为融合算法, 模型细节如图 2 所示.

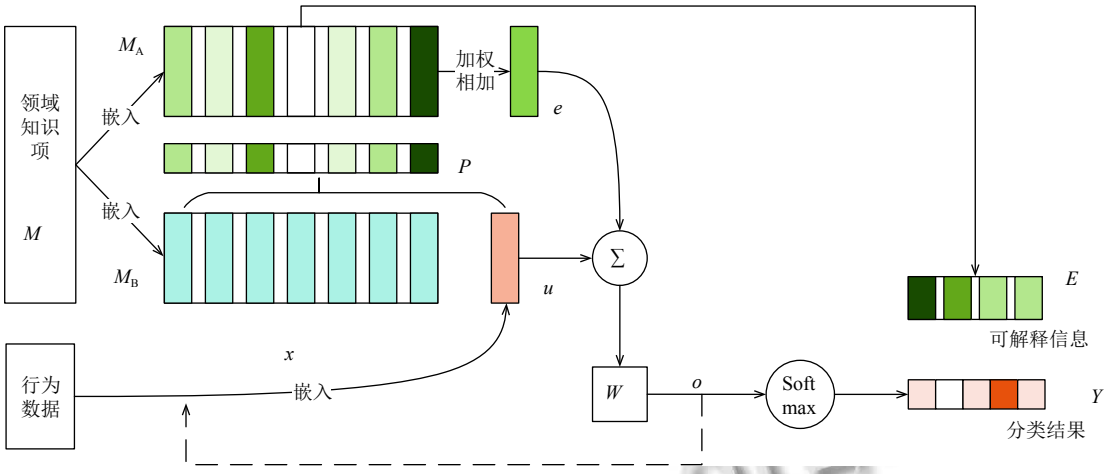

图 2 N2N Mem-IDS 模型

首先使用一组转换矩阵V将知识项M和行为数据x映射到维度相同的空间, 分别得到M_A, M_B和u:

$$V = \{A, B, C\} \tag{6}$$

$$\begin{cases} M_A = M \cdot A \\ M_B = M \cdot B \\ u = x \cdot C \end{cases} \tag{7}$$

假设向量a_i是M_A中第i个嵌入知识项, 使用余弦相似度计算每个嵌入知识项与输入网络行为的匹配程度:

$$p_i = g_V(x', m_i') = \frac{u^{\mathrm{T}} \cdot a_i}{\|u\|_2 \cdot \|a_i\|_2} \tag{8}$$

矩阵M_B中的嵌入知识项b_i根据匹配度p_i, 用线性加权算法与u融合, 然后使用转换矩阵W得到重构输出o:

$$o = h_p(x', M') = W \cdot (u + \sum p_i b_i) \tag{9}$$

本文模型是一个堆叠网络, 本层的输出将作为下一层的输入, 即:

$$x_{j+1} = o_j \tag{10}$$

最后一层的输出用来得到最终的分类结果:

$$\begin{cases} Y = \mathrm{softmax}(o) \\ \hat{y} = \arg\max(Y) \end{cases} \tag{11}$$

训练过程中, 参数V和W将会被模型自动学习.

4 实验设计

本章首先介绍实验使用的数据集, 接着介绍数据处理方法和攻击知识项提取方法, 然后介绍实验方案设计, 最后给出实验结果与分析.

4.1 数据介绍

本文实验使用的 NSL KDD 数据集源于 KDD CUP99 数据, 是目前最重要的公开入侵检测数据集之一. 1998 年美国空军发起 DARPA'98 入侵检测系统评估计划, 在模拟真实空军网络的局域网内收集网络流量数据, 用于进一步研究[20,21]. 1999 年从原始数据中提取出 KDD CUP99 数据集. 2009 年, Tavallaee 等人通过去重、重新设置数据比例等操作改进了 KDD CUP99 数据集存在一些内在问题, 形成 NSL KDD 数据集[22].

NSL KDD 数据集中每一条数据表示一个网络连接记录, 其类别包括正常类型和 4 大类、40 子类攻击类型. 数据集由一个训练集 train+和一个测试集 test+组成, 测试集中的攻击类型都是训练集中的攻击类型或者训练集中攻击的变体类型.

4.2 数据预处理

NSL KDD 数据有 41 个特征, 包括离散特征和连续特征. 对于连续特征, 我们使用 Z 分数 (*z-score*) 标准化进行处理, *z-score* 标准化计算方法为:

$$Z = \frac{X - \mu}{\sigma} \tag{12}$$

其中, $\mu = E(X), \sigma = Var(X)$. 对于离散特征, 我们使用独热编码 (One-Hot Encoding) 对特征值进行处理. One-Hot Encoding 是使用 N 维特征来对原始特征中的 N 个取值进行编码, 即原始特征中的每一个取值都由新特征组中对应位置的 0/1 状态来表示. 经过处理后的数据共有 122 个特征.

数据集中共有 41 个行为类型, 我们将这 41 个类型归纳为五大类攻击类型, 同样适用 One-Hot Encoding 对其编码, 使用五维向量作为数据标签, 用于五分类实验.

4.3 攻击知识项提取

模型要求领域知识与网络行为数据具有关联关系, 本实验中我们采用从行为数据集中直接提取攻击知识项的方式来获取领域知识. 根据数据集特点, 我们设计如下算法, 在数据集全集中提取攻击知识项:

算法 1. 攻击知识项提取算法

1) 选取一个行为类别, 将该类别数据单独取出, 形成子数据集;
2) 使用随机森林算法确定子数据集中所有特征的重要程度 (feature importance), 并将特征按重要程度降序排列;
3) 将重要程度从大到小依次累加, 在累加和超过阈值$\alpha(0<\alpha<1)$的位置截断, 选择截断位置之前的特征作为局部特征;
4) 保留该子集中局部特征的特征值, 其它特征值置 0, 形成的新数据成为近似行为数据;
5) 使用 K 均值聚类算法, 根据 Calinski Harabaz 分数, 从近似行为数据中计算出若干个聚类中心, 作为该类别的知识项.
6) 回到 1), 选择提取其它类别的知识项.

步骤 3) 中, α表示特征重要程度累加值所占比例, 经过预实验$\alpha=80\%$时有较好的效果, 因而采用$\alpha=80\%$, 即知识项保留重要程度前 80% 的特征. 这使得新特征能够排除非重要因素的影响, 降低数据复杂度, 同时在一定程度上保留该行为类别的特点.

4.4 实验设计

使用检测率 (*Detection Rate*, *DR*) 和精确度 (*Precision*) 作为入侵检测模型的评价指标, 把记忆神经网络模型和其它相关工作中的结果进行对比. 其中:

$$\begin{cases} DR = TP/(TP+FN) \\ \text{Precision} = FP/(TP+FP) \end{cases} \tag{13}$$

记忆神经网络模型在输出分类结果的同时, 输出可解释信息. 可解释信息为与网络行为数据相关度最高的三个攻击知识项的类别. 定义 *ExpV* 和 *MemV* 用于评价记忆神经网络中记忆模块的作用.

$$\begin{cases} ExpV = ExpScore/count \\ MemV = MemScore/count \end{cases} \tag{14}$$

ExpV 表示可解释结果对研究人员的可参考价值, *ExpV* 越大, 可参考价值越大. *MemV* 表示攻击知识项对判断结果所起作用, *MemV* 越大说明模型中攻击知识项发挥的作用越大. *ExpScore* 和 *MemScore* 取决于可解释结果和输入值的真实类别, 计算方式如表 1 所示.

我们设计另外两种知识项提取方式, 方法 1 为不使用攻击项, 方法 2 为随机从数据集中选取 50 条数据作为攻击知识项, 方法 3 为直接对不同类别数据进行聚类. 将本文方法与这三种方法相比较, 验证知识项有效性.

表 1 ExpScore 和 MemScore 计算方式

真实子类	可解释结果	包含 A 大类	*ExpScore*	包含 A1 或主要类别为 A	*MemScore*
A1	A1, A2, B1	是	+1	是	+1
A1	B1, A2, A3	是	+1	是	+1
A1	B1, A1, B2	是	+1	是	+1
A1	B1, A2, B2	是	+1	否	+0
A1	B1, B2, B3	否	+0	否	+0

5 实验结果

本文中模型的输出包括分类结果和可解释信息. 根据 4.4 的评价指标计算方式, 采用不同攻击知识项提取方式对模型输出的可解释信息质量进行比较. 比较结果如图 3(a) 所示. 发现使用本文知识项提取方式在 *ExpV* 和 *MemV* 上具有最佳表现. 这是因为输出的可解释信息是与网络行为数据匹配度最高的 3 条攻击知识项标签. 而匹配度越高的攻击知识项越能够表征网络行为特点, 因而将其输出后能够给安全人员参考, 使其了解当前网络行为所具有的部分特征. 另一方面, 在模型设计中, 匹配度越高的攻击知识项对分类器的辅助作用越大, 因而匹配度能够表征分类器的部分分类依据.

图 3(b) 为使用攻击知识项和不使用攻击知识项的对比结果, 可以看出攻击知识项确实在模型中发挥了辅助分类的作用.

将本文模型与几个模型比较, 验证模型有效性, 结果如图 4 所示. 对比显示本文模型相比传统单模型机器学习算法, 在检测率和精确度上有明显的提升.

6 在不同类型数据上的扩展

由于数据来源限制, 本文实验中从训练数据中提取近似的聚类中心, 作为领域知识的替代项. 如 3.1 节中提到的, 领域知识的来源可以是 IDS 规则、防火墙规则、安全人员经验数据等, 本节以 Snort 规则为例, 提出不同类型数据在模型中使用方式的设想.

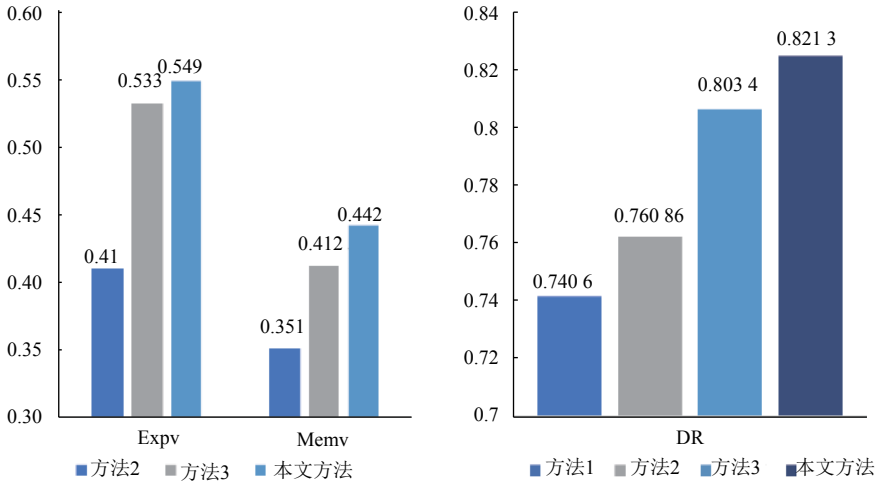

图 3 不同知识项提取方式实验结果对比

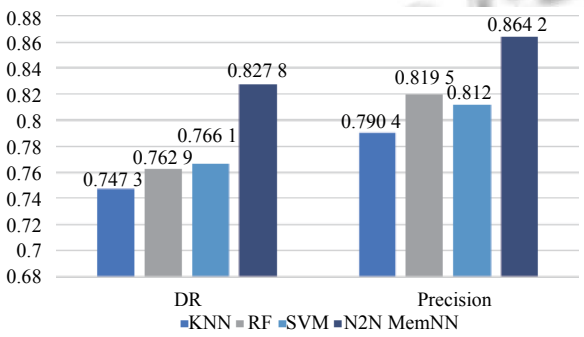

图 4 与其它模型实验结果对比

一条 Snort 规则对应一种网络威胁和应对方式, 由规则头和规则选项组成, 规则头指定待筛选的连接和规则动作, 规则选项包含若干匹配条件, 一旦待筛选连接满足全部的匹配条件, 则触发规则动作. 当使用 Snort 规则作为知识领域时, 本文提出的方法中, 匹配模块和融合模块应作出相应的修改.

匹配模块判断 Snort 规则与网络连接数据的匹配程度. 一种可行的方案为, 将 Snort 规则中的匹配条件与原始网络连接数据相比对, 计算被满足的匹配条件所占比例, 作为规则与连接的匹配程度. 即

$$p_i = g(x, m_i) = \frac{\sum I_k(x)}{K} \tag{15}$$

其中, $I_k(x)$表示连接x是否满足第i条规则m_i的第k个条件, K为m_i中的条件总数.

融合模块将按照匹配程度将 Snort 规则融合到原始数据中. 一种可行方案为, 使用匹配度作为特征, 将其以连接 (concatenate) 方式与原始数据融合. 连接方式如下图 5 所示.

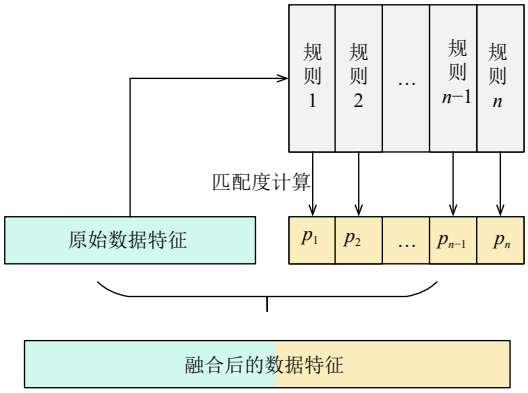

图 5 连接 (concatenate) 融合方法

在这种情况下, 网络连接数据与知识项的相似度直接作为新的数据特征放入分类器中, 能够为分类器提供更多信息以达到辅助分类的目的. 而每条 Snort 规则对应一种网络威胁, 可以用于可解释信息输出.

此外, 使用 IDS 规则的方式还有其它扩展方式, 如, 针对规则中的匹配条件, 按照专家知识设置匹配权重; 将描述形式的规则以某种方法转化为可计算的向量形式等. 同理, 防火墙规则或经验数据等都能够通过设计合适的匹配方法和融合方法, 使之在模型中发挥作用.

7 结论与展望

本文提出一种基于端到端的记忆神经网络的入侵检测方法, 这种方法利用神经网络将领域知识整合到算法中, 令神经网络分类器在领域知识的辅助下进行的端对端训练和预测, 并给出最终预测结果的可解释

依据. 之后本文通过实验对模型进行评估证明领域知识在模型中发挥了作用, 并且模型在检测率和精确度上有良好表现. 最后本文以 Snort 规则为例, 描述了模型在其它类型数据上的可行扩展方案. 今后, 将尝试诸如 IDS 规则等各种更加通用的领域知识数据, 并从神经网络优化角度考虑优化模型以降低模型训练时间.

参考文献

1 Kim G, Lee S, Kim S. A novel hybrid intrusion detection method integrating anomaly detection with misuse detection. Expert Systems with Applications, 2014, 41(4): 1690–1700. [doi: 10.1016/j.eswa.2013.08.066]

2 Scarfone KA, Mell PM. Guide to intrusion detection and prevention systems (IDPS). Gaithersburg: NIST, 2007.

3 Griffin K, Schneider S, Hu X, *et al*. Automatic generation of string signatures for malware detection. Proceedings of the 12th International Symposium on Recent Advances in Intrusion Detection. Saint-Malo, France. 2009. 101–120. [doi: 10.1007/978-3-642-04342-0_6]

4 Snort network intrusion detection system. https://www.snort.org.

5 The bro network security monitor. https://www.bro.org/.

6 Gu GF, Perdisci R, Zhang JJ, *et al*. BotMiner: Clustering analysis of network traffic for protocol- and structure-independent botnet detection. Proceedings of the 17th USENIX Security Symposium. Berkeley, CA, USA. 2008. 139–154.

7 Kumar S, Spafford EH. A pattern matching model for misuse intrusion detection. Proceedings of the 17th National Computer Security Conference. Baltimore, MD, USA. 1994. 11–21.

8 杨忠明, 秦勇, 蔡昭权. 一种策略分流的入侵防御及恢复系统架构. 计算机系统应用, 2017, 26(2): 83–87. [doi: 10.15888/j.cnki.csa.005620]

9 Tajbakhsh A, Rahmati M, Mirzaei A. Intrusion detection using fuzzy association rules. Applied Soft Computing, 2009, 9(2): 462–469. [doi: 10.1016/j.asoc.2008.06.001]

10 Blowers M, Williams J. Machine learning applied to cyber operations. Pino RE. Network Science and Cybersecurity. New York: Springer, 2014. 155–175.

11 Singh R, Kumar H, Singla RK. An intrusion detection system using network traffic profiling and online sequential extreme learning machine. Expert Systems with Applications, 2015, 42(22): 8609–8624. [doi: 10.1016/j.eswa.2015.07.015]

12 Chong D. Learning automata based SVM for intrusion detection. arXiv: 1801.01314.

13 Agarap AF. A neural network architecture combining gated recurrent unit (GRU) and support vector machine (SVM) for intrusion detection in network traffic data. arXiv: 1709.03082.

14 Bhuyan MH, Bhattacharyya DK, Kalita JK. Network anomaly detection: Methods, systems and tools. IEEE Communications Surveys & Tutorials, 2014, 16(1): 303–336.

15 任晓芳, 赵德群, 秦健勇. 基于随机森林和加权 k 均值聚类的网络入侵检测系统. 微型电脑应用, 2016, 32(7): 21–24. [doi: 10.3969/j.issn.1007-757X.2016.07.007]

16 王锋. 一种误用和异常技术结合的网络入侵检测模型. 计算机光盘软件与应用, 2012, (12)–84.

17 Al-Yaseen WL, Othman ZA, Nazri MZA. Multi-level hybrid support vector machine and extreme learning machine based on modified K-means for intrusion detection system. Expert Systems with Applications, 2017, 67: 296–303. [doi: 10.1016/j.eswa.2016.09.041]

18 Weston J, Chopra S, Bordes A. Memory networks. arXiv: 1410.3916.

19 Sukhbaatar S, Szlam A, Weston J, *et al*. End-to-end memory networks. arXiv: 1503.08895, 2015.

20 Stolfo SJ, Fan W, Lee W, *et al*. Cost-based modeling and evaluation for data mining with application to fraud and intrusion detection: Results from the JAM project. 1999. 130–144.

21 Lippmann RP, Graf I, Wyschogrod D, *et al*. The 1998 darpa/afrl off-line intrusion detection evaluation. Proceedings of the 1st International Workshop on Recent Advances in Intrusion Detection. Louvain-la-Neuve, Belgium. 1998.

22 Tavallaee M, Bagheri E, Lu W, *et al*. A detailed analysis of the KDD CUP 99 data set. Proceedings of 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications. Ottawa, ON, Canada. 2009. 1–6.