**Обнаружение аномалии (AD), используемое в области обнаружения вторжений в сеть, или сетевое вторжение AD (NIAD), зависит от правильного использования методов сходства и расстояния, но используемые меры часто не документируются в опубликованных исследованиях. В результате, в то время как объем исследований NIAD значительно вырос, знание полезности сходства и дистанционных мер в поле не выросло соответственно. Исследования NIAD охватывают множество доменов и используют разнообразные методы из простой кластеризации k-средних через расширенные распределенные системы AD с множеством агентов. В этом обзоре представлен обзор использования методов сходства и дистанции в исследованиях NIAD. Анализ дает теоретические основы дистанционных измерений и обсуждения различных типов дистанционных мер и их использования. Приводится примерное использование дистанционных измерений в опубликованных исследованиях, равно как и общее состояние строгой дистанционной меры в поле. Наконец, представлены области, которые требуют дальнейшего сосредоточения на улучшении строгости дистанционной меры в поле NIAD. Индексные термины - компьютерные сети, обнаружение аномалий, обнаружение вторжений, машинное обучение, измерение расстояния.**

**I. Введение (I. INTRODUCTION)**

ЦЕЛЬ системы обнаружения вторжений сети (NID) - автоматизировать процесс обнаружения, когда интрузии происходят в сети. В более общем смысле обнаружение вторжений может быть сформировано как подзадача в рамках проблемы обнаружения аномалии сети: определить, является ли трафик нормальным (хорошим) или аномальным (плохим). Автоматизированные системы, которые различают нормальные и аномальные, часто используют некоторые методы машинного обучения, такие как классификация или кластеризация, чтобы отличать нормальные от аномального трафика. В основе этих систем лежит сравнение модели нормали и модели аномального. Эти сравнения часто полагаются на способность измерять сходство или расстояние между мишенью и известным типом, чтобы определить, следует ли объявлять новую цель аномальной или нет. Таким образом, эффективность многих систем обнаружения аномалий (и многих систем обнаружения вторжений сети (NIDS)) зависит от выбранных дистанций или критериев подобия и того, как они используются.

К сожалению, большая часть опубликованных работ по NID не дает адекватной информации о мерах дистанции и сходства, используемых в исследовании. Во многих случаях дистанционные меры не упоминаются. В результате будущие исследования будут испытывать трудности с воспроизведением или сравнением результатов с предыдущей работой. Также трудно исследовать торговое пространство выбора дистанционной меры, когда оно плохо документировано в исследовании.

Некоторые авторы представляют подробную информацию о применяемых дистанционных мерах, а некоторые даже сравнивают эффективность нескольких вариантов дистанционных мер. Мы представляем эти публикации в качестве примеров в этой области и выносим рекомендации о том, какие авторежиссеры должны включать в свои публикации для улучшения прозрачности, дублирования и сопоставимости исследований.

Чтобы обеспечить всесторонний обзор того, как меры применяются к наборам данных Network Intrusion (NI) в исследованиях NID, в этой работе рассматривается, насколько хорошо определены меры, типы дистанционных мер, используемых в полевых условиях, и как они используются. Остальная часть этой статьи организована следующим образом. В разделе II обсуждаются несколько примеров других исследований использования дистанционной меры как внутри, так и вне области обнаружения аномалий. В разделе III представлен учебник по классификации, кластеризации и обнаружению аномалий, в котором основное внимание уделяется важности дистанционных мер в вычислительных рамках машинного обучения. В разделе IV мы предлагаем подробный учебник по теории дистанционных мер, типы используемых исследователями мер и сравнение этих мер. В разделе V представлен обзор публикаций с использованием дистанционных мер во время выбора, классификации и кластеризации признаков обнаружения аномалий. Мы излагаем уроки, извлеченные из литературы в разделе VII, и вывод в разделе VIII.

**II. Связанные работы (II. RELATED WORK)**

Было проведено много исследований в области обнаружения аномалий (AD), в том числе тех, которые рекомендуют различные методы для конкретных применений. В частности, Chandola et al. [1] обеспечивают всесторонний обзор области AD, организованный методами (например, статистическими или классификационными), используемыми для обнаружения аномалий. Несмотря на тщательный подход, их подход основан на типе обнаружения, используемом в исследовании, а не на типе дистанционных мер, используемых при обнаружении. Эта работа должна предусматривать дополнительный взгляд на область распространения знаний, предоставляя руководство о том, как использовались измерения расстояния и сходства.

Внутри набора данных Network Intrusion (NI) наиболее близкой работой к этому является Chmielewski и Wierzchon [2], которая «рассматривает проблемы, присущие использованию lp-метрики (определенной в (5), где p = r ≥ 1), Дробное lp-расстояние (определенное в (5), где 0 <p = r <1) и косинус-подобие (определенное в (35)) для измерения расстояния между различными образцами высокомерных данных. Используя эксперименты с использованием различных значений p на lp-метрике и используя полученное расстояние при применении отрицательного отбора к набору данных NI, они заключают, что значения p на интервале [0,5, 1.0] должны обеспечивать улучшение скорости обнаружения По сравнению с другими значениями.

Вне поля наборов данных NI наиболее близкая работа с этим - Ча [3], а также дает синтаксическую и семантическую классификацию измерений расстояния и подобия применительно к функциям распределения вероятностей, а также анализ корреляции между различными мерами с использованием кластеризации И представлены в иерархических кластерах. Несмотря на то, что Cha не является обзором того, как используются эти меры, работа Cha является полезной ссылкой для дистанционных мер с интересным разделением, основанным на том, насколько хорошо меры расстояния коррелируют друг с другом. Подобная работа с различными намерениями Дезы и Дезы [4] дает исчерпывающее перечисление основных мер дистанции, используемых в различных областях. Междисциплинарный подход, при котором учитывается список дистанционных мер, особенно полезен при попытке определить меры, используемые в опубликованных работах, поскольку на них ссылаются синонимы и аналогичные формулировки.

В областях, непосредственно не связанных с измерением различий и сходства в многомерных наборах данных, но которые могут быть полезны при рассмотрении многомерных наборов данных, существуют и другие значимые работы. Стааб [5] рассматривает отношения между онтологиями и мерами сходства, особенно в свете использования мер в рамках логических систем рассуждений. Каннингем [6] разрабатывает полезную структуру для рассуждений о том, какие меры подобия следует использовать при первом исследовании проблемы путем предоставления таксономии механизмов подобия. Chung et al. [7] разработать новую меру безопасности для узлов в облачной системе. Многие факторы, включая число, уязвимость и эксплуатационную пригодность каждой виртуальной машины, используются при расчете индекса безопасности виртуальной машины, а затем при выборе контрмеры для конкретной атаки.

**III. Руководство по классификации, кластеризации и обнаружению аномалий**

**(III. PRIMER ON CLASSIFICATION, CLUSTERING AND**

**ANOMALY DETECTION)**

В обнаружении аномалий сетевой интрузии (NIAD) целью является определение того, является ли конкретное наблюдение (активность в сети) аномальным. Обнаружение аномалий требует маркировки наблюдений. Во всех областях исследования обнаружения сетевых вторжений (NID) существуют тонкие различия между терминами, используемыми при обращении к различным фазам обнаружения аномалий (AD). Этот праймер дает объяснения фаз системы NIAD и определения терминов, используемых в этих фазах.

* Наблюдение: единый объект данных. В Network Intrusion (NI) объект может представлять собой пакет данных компьютерной сети или статус конкретного сервера в определенное время.
* Особенность: конкретный тип информации. Наблюдения обычно имеют несколько особенностей. В сетевой системе обнаружения вторжений (NIDS) функции могут включать в себя длину пакета, IP-адрес получателя и отметку времени.
* Набор данных: набор наблюдений, каждый из которых содержит значения для каждой из функций. Часто набор данных выражается в матричной форме со строками, представляющими наблюдения, и столбцами, представляющими функции.
* Предварительная обработка: любые манипуляции с набором данных, которые необходимы для того, чтобы инструменты AD исследователей могли работать в наборе данных. Предполагается, что предварительная обработка не влияет на результат эксперимента. Например, может потребоваться преобразование из формата данных, разделенных запятыми, в таблицу базы данных в реляционной базе данных, оставив значения компонентов в наборе данных неизменными.
* Трансформация: любое изменение применяется к данным с целью масштабирования, нормализации или взвешивания данных до их использования. В то время как значения функций изменяются, количество функций и порядок значений признаков сохраняются. Например, некоторые методы маркировки более эффективны, если отмеченные наблюдения нормализуются, чтобы все значения признаков лежали между 0 и 1 перед маркировкой.
* Генерация функций: любое создание новых функций на основе исходных или производных наборов данных. Например, преобразование функции с семью возможными номинальными значениями в семь двоичных признаков или создание новой функции, которая является квадратным корнем из суммы квадратов двух других признаков.
* Выбор: операция, в которой используется только подмножество всех доступных функций или наблюдений для использования при маркировке.
* Контролируемый метод: метод, который требует обучения с использованием обозначенных примеров обучения (учебного набора данных) перед выполнением обучаемой системы по немеченым наблюдениям. Обозначенные примеры обучения могут быть помечены вручную исследователем или могут быть результатом некоторых ранее запущенных процессов.
* Неконтролируемый метод: метод, который не требует обучения до выполнения по немеченым наблюдениям.
* Классификация: акт маркировки каждого наблюдения в данных как члена определенного класса. Большинство методов классификации контролируются.
* Кластеризация: разделение наблюдений на группы по подобию. Часто кластеризация не контролируется: метки групп выбираются после сгенерирования групп.

Во время AD процесс маркировки, который использует трансформацию, создание, выбор, классификацию или кластеризацию, часто зависит от измерений расстояния между значениями признаков двух или более наблюдений в наборе данных. Например, для генерации генерации может потребоваться эвклидовое расстояние для создания новой функции от двух других. Кластеризация зависит от некоторого понятия расстояния, чтобы определить, какие наблюдения близки друг к другу в пространстве, предназначенном для обнаружения аномалий. Поскольку измерения расстояния настолько важны для AD, мы подробно обсудим их в следующем разделе.

**IV. Меры расстояния  
(IV. DISTANCE MEASURES)**

**А. Теория мер расстояния (A. Theory of Distance Measures)**

В настоящем документе рассматривается использование дистанционных и сходных мер, используемых при обнаружении сетевых вторжений (NID). Существует ряд фундаментальных определений дистанционных мер, которые широко определены в математической литературе [4]. Определение дистанционной меры включает три требования. Четвертое требование устанавливает подкатегорию дистанционных мер, называемых метками расстояния. Чтобы определить эти требования, мы будем использовать функцию dist (), которая принимает в качестве входных данных две различные переменные A и B и возвращает значение расстояния. Затем требования для дистанционной меры заключаются в следующем:

1. Неотрицательность: расстояние между A и B всегда является значением, большим или равным нулю.

Dist (A, B) ≥ 0 (1)

1. Идентичность неразличимых: расстояние между A и B равно нулю тогда и только тогда, когда A равно B.

dist (A, B) = 0, если A = B (2)

1. Симметрия: Расстояние между A и B равно расстоянию между B и A.

dist (A, B) = dist (B, A) (3)

Расстояния, которые соответствуют, по меньшей мере, этим трем требованиям, известны как Дистанционные меры. Эти дистанционные меры, которые также удовлетворяют одному дополнительному ограничению, также квалифицируются как метрика расстояния:

1. Неравномерность треугольника: учитывая наличие третьей точки C, расстояние между A и B всегда меньше или равно сумме расстояния между A и C и расстояние между B и C. Это требование также известно как неравенство треугольника.

Dist (A, B) ≤ (dist (A, C) + dist (B, C)) (4)

Хотя измерения расстояния обычно считаются мерами, которые должны применяться к точкам в физическом трехмерном пространстве, дистанционные меры также могут применяться к многомерным данным, которые могут не представлять местоположения в пространстве. Это позволяет применять дистанционные метрики к таким не пространственным данным, которые обычно встречаются в приложениях NID. Кроме того, существуют методы сравнения разницы между двумя точками, которые не удовлетворяют даже первым трем требованиям. Эти методы не квалифицируются как дистанционные меры, а вместо этого упоминаются как меры сходства. Затем мы рассмотрим меры, которые были использованы в недавней литературе по НИД.

**B. Типы мер (Types of Measurements)**

При изучении типов дистанционных мер, используемых в поле NI, полезно учитывать дистанционные меры как часть отдельных семейств или категорий. Семьи, отобранные для этой работы, относятся к тем, которые перечислены в «Энциклопедии расстояний» Деза и Деза [4]. Поскольку в поле NID нет определенной таксономии, рассмотренные меры и индексы будут упорядочены по их отношению к семьям мер.

1) Расстояния мощности: расстояния мощности - это меры расстояния, которые используют формулу, математически эквивалентную формуле мощности (p, r) в (5).

(5)

Сила (p, r) - расстояние измеряет расстояние между двумя векторами x и y длины n. Эта категория включает в себя несколько наиболее распространенных дистанционных мер, включая эвклидову дистанцию ​​и расстояние Манхэттена. Хотя обычно это считается физическим расстоянием между двумя точками в трехмерном пространстве, эти типы расстояний могут быть применены к векторам любой размерности, если тип данных является числовым. Определенная мера расстояния, указанная этой категорией, определяется значениями р и г. Например, где p = r = 2, степень (p, r) -distance определяет евклидову метрику, а где p = r ≥ 1 определяет lp-метрику. Другие значения p и r дают другие показатели. При 0 <p = r <1 мощность (p, r) -распространение известна как дробная lp-метрика.

Чтобы дать некоторую интуицию о том, какие значения p и r будут иметь значение при измерении расстояния между двумя точками, на рис.1 показаны диаграммы Вороного [8], построенные с использованием силы (p, r) -дружества при изменении значения p и r. На фиг.1 заштрихованные области показывают полный набор точек, для которых точка посева области (показанная как черная точка) ближе, чем точка посева другого региона. Этот тип диаграммы полезен для визуализации расхождения между нашей интуитивной концепцией физического расстояния и концепцией расстояния в соответствии с другими дистанционными измерениями. На рисунке 1 показано, что расстояния мощности, отличные от евклидова, не обязательно соответствуют нашему интуитивному пониманию физического расстояния. В контексте NID взвешенное евклидово расстояние может использоваться как мера несходства. Разница между взвешенным и невзвешенным евклидовым расстоянием является добавлением весового вектора (w) длины n к формуле в (5).

Другие расстояния, связанные с расстоянием мощности (p, r), также могут использоваться для NID, включая расстояние Махаланобиса и функцию гетерогенного расстояния. Расстояние Махаланобиса [9] определяется как показано в (6) [4]. Внутри (6) x и y - векторы, каждая из которых имеет n элементов, A - положительно определенная матрица (обычно ковариационная матрица x и y), det A - определитель A, а T указывает операцию транспонирования.

x − yA = (det A) 1 n (x − y)A−1(x − y)T (6)

Упрощенная версия расстояния Махаланобиса, разработанная Вангом и Столфо [10], может быть полезна для быстрых вычислений при тяжелых вычислительных нагрузках, требуемых в высокопроизводительных средах. Упрощенная мера mμ, σ определена в (7), где x - вектор, содержащий все размеры одного наблюдения, μ - вектор, представляющий центр масс всех наблюдений данных, n - число элементов в x И μ, а d (xi, μi) - разность между i-м элементом x и μ.

mμ,σ(x) = n i=1 d(xi, μi) σi (7)

Функция гетерогенного расстояния (H (x, y)) [11] формулируется следующим образом, где x и y - векторы из m элементов, которые поступают из наборов данных X и Y соответственно. Во-первых, расстояние между элементами, содержащими непрерывные значения, определяется в (8), где σj - дисперсия j-го атрибута набора данных X.

ddiff(xj , yj ) = |xj − yj | 4σj (8)

Затем расстояние между элементами, содержащими дискретные значения, определяется в (9), где Nj, x - количество записей, в которых значение j-го атрибута равно xj, Nj, x, i - количество записей, в которых значение J-ый атрибут - xj, а класс - i, а k - количество выходных классов.

dvdm(xj , yj ) = k i=1     Nj,x,i Nj,x − Nj,y,i Nj,y     (9)

Поскольку теперь определены непрерывные и дискретные расстояния, можно определить условное расстояние. Обратите внимание, что формулировка в (10) учитывает недостающие атрибуты, предоставляя значение, в котором отсутствует один или другой атрибут (xj или yj).

dj (xj , yj ) =  1, xj or yj dvdm(xj , yj ), xj , yj discrete ddiff(xj , yj ), xj , yj continuous (10)

Наконец, можно определить гетерогенную функцию расстояния, которая использует модификацию мощности (p, r) -диск, где | xi - yi | Заменяется условным расстоянием, определенным в (10), dj (xj, yj), что приводит к (11).

H(x, y) = m i=1 d2 j (xj , yj ) (11)

2) Расстояния по законам распределения. Второй тип меры, расстояния от законов распределения, описывает эти меры, основанные на распределении вероятностей набора данных. К ним относятся большинство форм энтропии, а также распределения условных вероятностей. Эти меры расстояния основаны на законах распределения и применяются к распределению вероятностей по переменным с одинаковым диапазоном.

Одной из наиболее распространенных дистанционных мер в этой категории является коэффициент Бхаттачарья [4], который можно использовать для ранжирования признаков в зависимости от способности каждой функции отличать один класс от других [12]. Коэффициент Бхаттачарья показан в (12), где P1 и P2 - вероятностные распределения по области X, p1 (x) - вероятность x, входящая в P1, и p2 (x) - вероятность x, встречающаяся в P2.

ρ(P1, P2) = x∈X p1(x)p2(x) (12)

Дальнейшим расстоянием закона распределения является χ2-расстояние. Уравнение (13) является стандартным χ2-расстоянием, где x и y - векторы длины n, p (xi) - вероятность появления i-го элемента x, а p (yi) - вероятность наступления i-го Элемент y.

d(x, y) = n i=1 (p(xi) − p(yi))2 p(yi) (13)

Модифицированное χ2-расстояние [13] иногда может быть полезно для использования NID, для измерения расстояния между строками и столбцами в матрице соответствия между исходными сетевыми данными и сгенерированными наборами данных. Модифицированная версия выполняется следующим образом: Предположим, что E представляет собой матрицу m × n, как показано в (14) в нижней части следующей страницы, с координатами в каждом элементе (вместо значений) для ясности. Каждая пара векторов, которые нужно сравнить, это либо векторы столбцов, либо строки, но обе они должны иметь одинаковую ориентацию. Каждый вектор столбца имеет длину m - 1, а каждый вектор строки - n - 1. n-й элемент каждой строки содержит сумму значений в элементах от 1 до n - 1 этой строки, тогда как m-й элемент каждого столбца содержит Сумма значений в элементах от 1 до m - 1 этой колонки.

Для сравнения двух векторов строк или столбцов используется уравнение, аналогичное (13). Например, для сравнения двух векторов строк, затененных в (14), rk и rl, необходимая формула находится в (15). В (15) rki - элемент i в строке k, rkv - сумма всех значений в векторе строк rk, а cui - сумма всех значений в векторе столбца ci. Если i = s, то столбец ci будет заштрихованной частью столбца s в (14).

(15)

Уравнение (15) очень похоже на (13), и сходство становится более очевидным с некоторыми заменами. Пусть x = rk, y = rl, p (xi) = rki / rkv, p (yi) = rli / rlv и n = v - 1, затем применим эти подстановки к (15), чтобы получить результат (16).

d(x, y) = n i=1 (p(xi) − p(yi))2 cui (16)

Разница между ними состоит в том, что χ2-расстояние использует p (yi) в качестве делителя, а в модифицированной версии используется cui.

В пределах вероятностных мер расстояния есть энтропия, которая может быть вычислена из любой числовой случайной величины. Энтропия может использоваться в качестве общего суммарного показателя функций, генерируя векторы энтропии набора данных для выбранных функций. Энтропия случайной величины вычисляется следующим образом: вероятность случайной величины X, удерживающей значение x, равна P (X = x) или p (x), а a - база для использования, энтропия случайной величины X, H (X) рассчитывается по формуле (17).

H(X) = − x∈X p(x) loga p(x) (17)

Многие варианты энтропии могут быть полезны в NID. Эти вариации включают стандартизованную энтропию, условную энтропию и расстояние Дженсена. Стандартизованная энтропия [14] - метод, который компенсирует изменения энтропии из-за числа значений случайной величины. Используя определение энтропии в (17), стандартизованная энтропия (Hs) определяется, как показано в (18), где m - общее число значений в X, a - база, в которой вычисляется энтропия.

Hs(X) = H(X) loga m (18)

Условная энтропия - это мера расстояния, которая допускает две случайные величины. Где X и Y - дискретные переменные, формула условной энтропии находится в (19).

H(X|Y ) = x∈X y∈Y p(x, y) loga p(x|y) (19)

Расстояние Йенсена [15] использует энтропийно подобную функцию H. Переменные внутри H - две из суммирующих переменных a, b и c; (20) где П † w (x) - частота появления w в x.

a = w∈L min (φw(x), φw(y)) b = w∈L (φw(x) − min (φw(x), φw(y))) c = w∈L (φw(y) − min (φw(x), φw(y))) (20)

Переменные суммирования содержат различные совпадения и несоответствия: a положительные совпадения, b - несоответствия слева, c - неправильные несоответствия. Уравнение (21) определяет H через a, b и базу, для которой выполняются вычисления, o.

H (φw(x), φw(y)) = a logo 2a a + b (21)

Используя H, расстояние Йенсена формулируется так, как показано в (22).

djens(x, y) = w∈L H (φw(x), φw(y)) + H (φw(y), φw(x)) (22)

Помимо энтропии, также в категории вероятностей меры расстояния являются популярным расстоянием Kullback-Leibler (KLD) или увеличением информации. KLD формулируется так, как показано в (23), где P1 и P2 являются вероятностными распределениями по области X, p1 (x) является вероятностью x, входящей в P1, p2 (x) - это проницаемость x, происходящая в P2, и A - база, в которой вычисляется KLD.

KLD(P1, P2) = x p1(x) loga p1(x) p2(x) (23)

Другим вероятностным расстоянием является расстояние Хеллингера, связанное с более известной метрикой Хеллингера. Метрика Хеллингера показана в (24), где P1 и P2 - вероятностные распределения по области X, p1 (x) - вероятность x, входящая в P1, а p2 (x) - вероятность x, встречающаяся в P2.

(24)

Расстояние Хеллингера, как определено Sengar et al. [16], существенно отличается от метрики Хеллингера. Во-первых, множественность изменяется от 2 до 1/2. Во-вторых, квадратный корень всей формулы не вычисляется. Уравнение (25) показывает расстояние Хеллингера.

(25)

Наконец, мы имеем тест отношения правдоподобия. Это расстояние похоже на отношение правдоподобия в (26), где H0 не предполагает аномалии, H1 предполагает наличие аномалии, X - дискретная переменная, N - общее число наблюдений, p (xk | H0) - вероятность Xk без аномалии, p (xk | H1) - вероятность xk при наличии аномалии, а LN (X) - отношение правдоподобия X.

LN (X) = N k=1 p(xk|H1) p(xk|H0) (26)

Отношение правдоподобия может быть расширено для вычисления отношения правдоподобия дискретной переменной (X), сформулированной в (27).

L = p1(x) p2(x) (27)

Коэффициент правдоподобия может использоваться для функций пакетной передачи и -size для обнаружения атак или обычного трафика. В этом приложении пороговые значения задаются для верхней и нижней границ, где, если отношение правдоподобия превышает верхнюю границу, обнаружена атака, и если коэффициент правдоподобия падает ниже нижней границы, то обнаружение атаки не происходит.

3) Корреляционные сходства. Корреляционные сходства и расстояния - это меры, которые пытаются охарактеризовать взаимосвязь между двумя наборами данных и рассматривать это как меру сходства или расстояния, а не использовать распределение вероятностей или величину векторов.

Соотношение р-ранга Спирмена является одной из таких мер сходства. Уравнение (28) дает ранговую корреляцию Спирмена, где Xr и Yr содержат ранжировки дискретных переменных X и Y, xi и yi содержат i-й ранг в Xr и Yr (соответственно), Xr и Yr имеют одинаковое количество элементов и N - число элементов в Xr.

ρ(Xr, Yr)=1 − 6 n i=1 (xi − yi)2 n(n2 − 1) (28)

Корреляция рангов Kendall τ является еще одной мерой сходства этого типа. Корреляция рангов Kendall τ определяется как в (29), где функция sgn используется для вычисления числа несогласных пар рангов, вычитаемых из согласных пар рангов.

τ (Xr, Yr) = 2 n −1 i=1 n j=i+1 sgn(xi − xj ) · sgn(yi − yj ) n(n − 1) (29)

Функция sign или signum определена в (30).

(30)

Чтобы применить корреляцию рангов Kendall τ к NID, в стандартную формулировку в (29) можно внести небольшую модификацию, которая позволяет захватывать как сходства, так и различия. Результат показан в (31), где функция eq определена, как показано в (32).

(31)

(32)

Третий метод корреляционных коэффициентов, который может быть применен к NID, является линейным коэффициентом корреляции момента продукта Пирсона или r. Формула для r приведена в (33), где X¯ - среднее дискретной переменной X.

(33)

Эквивалентная формулировка линейного коэффициента корреляции момента продукта Пирсона, который использовался для приложений NID, определен в (34).

(34)

Четвертый метод в этой категории - это метод, основанный на квантовании вектора обучения, и использует сходство косинусов с искусственной нейронной сетью. Определение подобия косинуса показано в (35), где φ - угол между векторами x и y.

(35)

При применении подобия косинуса и квантования векторного обучения к NID определяется сходство, определяющее, какой из нейронов искусственной нейронной сети «выигрывает» [17]. Формула дана в (36), где U - один из обучающих наборов образцов атаки, V нейрон, Wu и Wv - весовые векторы, связанные с U и V (соответственно), Wu и Wv имеют одинаковое количество элементов , А n - общее число элементов в Wu.

(36)

Сходство косинусов также можно использовать в качестве меры самоподобия [18]. В этой реализации сходства косинусов, показанного в (37), # »Gt - это моментальный снимок состояния системы в течение временного интервала t, а #» 1 - вектор всех тех, которые имеют ту же размерность, что и # »Gt.

(37)

Среднее значение St берется за последние k периодов времени (где k выбирается с использованием методов, описанных в статье), то каждый St сравнивается со средним, чтобы определить, находится ли он в пределах, установленных исследователями. Если St не находится в пределах двух стандартных отклонений пробега, то некоторые авторы предполагают, что произошла атака или нарушение безопасности.

4) Другие сходства и расстояния: некоторые меры сходства, которые полезны для NID, не вписываются в три основные категории выше. Например, один метод использует сходство Dice [19], определенное в (38), где X и Y - строки из каналов интернет-ретрансляции, n-грамм - это подпоследовательность из n элементов из данной последовательности, а число N-граммы в X задаются | ngrams (X) |.

(38)

Изменение расстояния χ2, известное как квадрат χ2 (также известный как χ-квадрат) [20], также может использоваться для вычисления расстояний с n-граммами. Квадратная мера χ2 показана в (39), где S - множество всех возможных n-граммов, x и z - байтовые последовательности, взятые из полезной нагрузки пакета, φs (x) - частота появления s в x.

(39)

Некоторые из этих сходств и расстояний, которые не вписываются в три основные категории, особенно полезны в применении к NID, который не основан на необработанном сетевом трафике или функциях, построенных из пакетов, но на других характеристиках данных. Один из таких методов основан на линейном ядре (dlk, см. (40) [15].

(40)

Заметим, что это эффективно точечное произведение вероятностных распределений w внутри x и y.

Вариантом геодезического расстояния является другой метод, который может быть применен к NID. Для такого применения геодезическое расстояние можно сформулировать как dgd = arccos (dlk), что несколько отличается от стандартного определения геодезического расстояния.

**C. Сравнение типов мер (Comparison of Types of Measures)**

Различные типы измерений расстояния и сходства демонстрируют различные свойства, которые следует учитывать, прежде чем применять их к приложению NID. Эти различия делают определенные меры более или менее полезными для определенных типов данных, степеней размерности и других соображений. Например, в то время как все приведенные здесь измерения расстояния удовлетворяют требованиям (неотрицательность) и двум (идентичность различимых) для дистанционных мер, не все удовлетворяют требованиям три (симметрия) и четыре (неравенство треугольника). Здесь типы мер расстояния и меры сходства, описанные в части B, будут сравниваться в соответствии с определениями, представленными в части A, и в соответствии с другими ключевыми отличиями.

1) Расстояния мощности: все расстояния мощности соответствуют всем четырем требованиям для дистанционных измерений и квалифицируются как показатели расстояния. При применении к приложениям NID евклидова дистанция показывает явные различия между обычными образцами и образцами атаки и предоставляет характеристики разных типов записей трафика. Эвклидовое расстояние также может быть применено косвенно, сначала используя меру распределения, такую ​​как дискретное преобразование Фурье, затем применяя евклидовую или другую меру расстояния мощности к частотной области. Этот подход может потенциально решить некоторые проблемы с применением евклидова расстояния непосредственно к функциям (или к самому трафику).

В приложениях NID данные трафика часто могут иметь большое количество функций. В этом случае с большой размерностью дробный lp-метод может быть более эффективным, чем евклидово в наборе данных более высокого порядка, а полезность дробного lp-расстояния может возрастать по мере увеличения размерности. Это может иметь место, хотя дробное lp-расстояние не приводит к знакомому определению «ближайшего» в двух- и трехмерном пространстве, как показано на рисунке 1. Обратите внимание, что дробные значения lp, такие как 0,75, дают разрывные области ближайших пространств на рисунке. 2) Расстояния по законам распределения. Только некоторые из методов измерения расстояний по законам распределения квалифицируются как дистанционные меры и показатели. Эти расстояния часто называются расхождениями, а не расстояниями в случае, когда они не удовлетворяют неравенству треугольника и симметрии. Например, KLD, описанный в части B этого раздела, на самом деле является расхождением, а не расстоянием, из-за его нарушения как неравенства треугольника, так и симметрии. Расстояние Бхаттачарья нарушает неравенство треугольника, но проявляет симметрию. Расстояние Хеллингера соответствует как неравенству треугольника, так и симметрии. Условная энтропия также квалифицируется как мера дивергенции или подобия из-за ее нарушения как симметрии, так и неравенства треугольника.

Помимо их свойств, связанных с определением расстояний, другие свойства расстояний по законам распределения могут мотивировать их использование. Например, расстояние Махаланобиса особенно полезно, поскольку рассчитанное расстояние учитывает дисперсию в данных выборки, что позволяет более точно идентифицировать выбросы. Условная энтропия может использоваться в NID для целей, отличных от дистанционных мер, таких как измерение индивидуального потока трафика.

3) Корреляционные сходства. Корреляционные сходства отличаются от расстояний по законам распределения и степенным законам тем, что они способны работать не только на непрерывных или дискретных типах данных, как и в законах распределения и степенном праве, но также на ранжировании или порядковых Типы данных. Таким образом, четыре требования к дистанционным мерам не обязательно применяются к этим методам; Все они считаются мерами сходства, а не дистанционными мерами. Тем не менее все приведенные здесь корреляционные сходства удовлетворяют по меньшей мере двум требованиям для дистанционных мер (идентичность неразличимых).

Корреляционные сходства - это типы тестов гипотез, которые проверяют гипотезу о корреляции между двумя наборами данных. Корреляция рангов Kendall τ является наиболее общей из методов, рассмотренных в этой статье. Корреляция рангов Kendall τ является типом непараметрического теста; Он используется для проверки степени статистической зависимости между двумя переменными, не полагаясь ни на какие предположения о распределениях двух переменных.

Продукт-момент корреляции Пирсона и Спирмена р ранговой корреляции очень похожие виды мер: как используются для оценки того, насколько хорошо переменная может быть описана третьей опорной функции. Разница между этими двумя измерениями заключается в том, что корреляция момента продукта Пирсона дает меру линейной корреляции между двумя переменными, в то время как корреляция ранга Спирмена р оценивает, насколько хорошо можно описать связь между двумя переменными с помощью монотонной функции.

Косинус-сходство - более упрощенный тип корреляционного сходства, который измеряет степень сходства между двумя векторами с точки зрения ориентации, а не величины.

В приложениях NID иногда полезно использовать множественные коэффициенты корреляции в пределах одной дистанционной меры, чтобы обнаруживать сходства, которые одна мера не может уловить.

В этом разделе сравнивались три различных типа доступных дистанционных мер: расстояния мощности, расстояния от законов распределения и сходства в корреляции. В следующем разделе мы определяем использование этих дистанционных мер в области исследований NID.

**V. Использование меры расстояния**

**(V. Distance Measure Use)**

Изучение текущего состояния использования расстояния и сходства требует обзора последних работ. Цель состоит в том, чтобы изучить, как исследования используют измерения в области обнаружения аномалий (AD). Мы анализируем исследования, опубликованные с 2005 года, для поиска статей, в которых указаны имена и формулы для используемых мер. В дополнение к открытию, которое многие авторы не предоставляют достаточной информации о мерах, используемых для тиражирования их работы, мы также обнаружили, что, хотя меры используются повсеместно в этой области, они используются только в поднаборе фаз, описанных в разделе III.

В этом обзоре были найдены меры, используемые в фазах выбора объектов (раздел V-A), классификации (раздел V-B) и кластеризации (раздел V-C). При объяснении применяемых мер работы разделяются фазой использования, и рассматриваются только примерные статьи, предоставляющие отличное описание и идентификацию используемой меры (ов). Чтобы обеспечить согласованность ярлыков, все меры, названные в этой работе, обозначаются стандартными именами, описанными в «Энциклопедии расстояний» Деза и Деза [4].

**A. Выбор признаков (Feature Selection)**

Первая фаза AD для явного использования мер в литературе - это выбор функции. Меры, используемые при выборе признаков, как правило, связаны с вероятностью, поскольку вероятное появление признака является полезным механизмом, позволяющим уменьшить размерность больших наборов данных.

Eid et al. [21] разработать метод выбора функции для обнаружения сетевых вторжений (NID), в котором первый слой использует расстояние Kullback-Leibler (KLD) для ранжирования функций в наборе данных. KLD или информационный выигрыш формулируется, как показано в (23), где P1 и P2 являются вероятностными распределениями по области X, p1 (x) - вероятность x, возникающая в P1, p2 (x) - вероятность появления x В P2, a - база, в которой вычисляется KLD.

Функции оцениваются KLD, затем набор данных классифицируется с использованием классификатора J48 (вариация с открытым исходным кодом алгоритма дерева решений C4.5). После классификации идентифицируются локальная и глобальная точность максимумов, а функции выбираются для включения при выполнении окончательной классификации на основе идентифицированных максимумов. В результате уменьшенный набор функций привел к увеличению классификации F-меры 99,2%, используя только 20 функций в наборе данных NSL-KDD. Вклад статьи не ограничивается новой структурой для выбора функций, но также является повторяемостью работы. Eid et al. [21] являются точными в своих определениях используемых мер и методах, используемых для достижения результатов, так что подтверждение их экспериментов будет относительно простым.

В другой заметной недавней работе Хэнкок и Ламонт [12] выполняют функцию выбора в системе классификации сетевых атак с несколькими агентами, используя коэффициент Бхаттачарья (12), чтобы оценивать функции в соответствии с способностью каждой функции отличать один класс от других.

Три функции с наибольшим перекрытием (наибольшее значение ρ (P1, P2)) выбираются после отклонения любой функции, которая сильно коррелирует с функцией с более высоким рейтингом, чтобы уменьшить избыточность среди выбранных функций. Выбор функции является частью каждого агента в их системе. Затем агенты распределяются по всей сети, пытаясь обеспечить эффективную многоагентную сеть обнаружения вторжений (NIDS) с использованием репутации.

В других работах используются расстояния, связанные с вероятностью в задачах выбора объектов, и способами, которые могут быть явно не связаны с расстоянием. Wang et al. [14] используют вектор энтропии, рассчитанный с использованием стандартизованной энтропии, как определено в (18). Затем Entropy используется для вычисления значения для каждой функции перед обучением определенного типа классификатора.

При выборе функции использование мер ограничено теми, которые связаны с вероятностью. Тем не менее, способ, которым меры используются при выборе функции, варьируется от простых вычислений энтропии до функций упорядочения рангов на основе KLD наряду с несколькими прогонами классификации. Использование мер при вычислении признаков ограничивается вычислением с использованием единственной функции (измерение для будущего сравнения) или вычислением с использованием двух или более функций (прямое сравнение), но эти два могут использоваться с мерами по предоставлению множества различных методов для Выбор функций.

**В. Классификация (Classification)**

После выбора функции классификация набора данных с использованием выбранных функций является следующим шагом в этом процессе. Некоторые предшествующие исследования очень эффективно сравнивали результаты классификации эффективности при использовании нескольких дистанционных мер. Chmielewski и Wierzchon [2] исследуют использование «lpmetric ((5) с p = r ≥ 1), дробное lp-расстояние ((5), с 0 <p = r <1) и косинус-подобие (35 ) Для измерения расстояния между различными образцами высокоразмерных данных. Используя эксперименты с использованием разных значений p на lpmetric и используя полученное расстояние при применении отрицательного отбора к набору данных Network Intrusion (NI), они заключают, что значения p на интервале [0,5, 1.0] должны обеспечивать улучшение Скорость обнаружения по сравнению с другими значениями при применении к высокоразмерным наборам данных.

В другой работе Chmielewski и Wierzchon [22] предлагают «использовать как двоичные, так и вещественные детекторы для обнаружения несамоприемников. Они демонстрируют, что результаты согласуются с теорией о том, что полезность дробного lp-расстояния возрастает по мере увеличения размерности и что использование двух разных типов детекторов значительно увеличивает охват несамочувствительной области с помощью генерируемых детекторов. Обе работы демонстрируют применение мощности (p, r) -distance способами, которые бросают вызов общей интуиции относительно расстояния.

Tan et al. [23] генерируют евклидовы карты расстояний для характеристик каждого образца в наборе данных, а затем преобразуют расстояния в каждой ячейке карты в цвет для целей быстрого визуального сравнения. Визуализация показала явные различия между обычными образцами и образцами атаки и обеспечила новые характеристики различных типов записей трафика компьютерной сети. Хотя эвклидово расстояние не используется как сингулярное различие между двумя наборами данных, генерация матрицы характеристик объекта является новым методом классификации.

Gu et al. [19] разработал метод обнаружения Botnets, который использует сходство Dice (см. (38)), где X и Y являются строками из каналов интернет-ретрансляции, n-грамм является подпоследовательностью n элементов из данной последовательности, а число N-граммов в X задается | ngrams (X) |.

Большая часть анализа, обнаруженного в этом обзоре, основана на необработанном сетевом трафике или функциях, генерируемых сетевыми пакетами. Использование данных из абстракции более высокого уровня, чем обычно, с помощью NID (потоков символов) дает представление о всей области поля, в которой мало что делается для сравнения эффективности различных мер.

Zhao et al. [24] использовать единую меру расстояния, которая включает в себя один из трех коэффициентов корреляции для обнаружения атак наступающих камней, когда один компьютер используется злоумышленником для достижения другого. Авторы используют альтернативу ранговой корреляции Kendall τ, ранее описанную в (29), (31), (33) и (34). Каждая мера применяется к двум потокам трафика, и каждый результат вычитается из числа один, чтобы рассчитать минимальное расстояние между двумя потоками (σ (X, Y)), как показано в (41).

σ(X, Y ) = min (1 − ρ(Xr, Yr), 1 − τ  (Xr, Yr), 1 − r(X, Y )) (41)

Если σ (X, Y) меньше заданного порога (задано исследователем), то сравниваемая пара является достаточно подобной, чтобы считаться ретранслированным трафиком. Использование множественных коэффициентов корреляции в пределах одной меры расстояния является хорошим примером использования нескольких мер для обнаружения сходств, которые одна мера не может уловить.

Классификация полезна, но она предполагает априорное знание классов, к которым относятся аномалии, а также уникальные характеристики, которые определяют этот класс. С другой стороны, кластеризация может быть полезной при определении количества классов и идентификации каких-либо уникальных характеристик данного класса, когда эти две части информации неизвестны.

**C. Кластеризация (Clustering)**

Некоторые опубликованные работы эффективно используют более одного семейства дистанционных мер для кластеризации данных. Одна из таких работ - Лахина и др. [25], в котором используются меры как из семейств степеней (p, r), так и с вероятностями. Первая используемая мера - это энтропия трафика, как показано в (17).

Авторы используют энтропию как общую суммарную меру признаков, генерируя векторы энтропии набора образцов для выбранных признаков. Вторая используемая мера - это квадратичное евклидово расстояние (мощность (p, r) - расстояние, где p = 2 и r = 1), которые авторы используют для расчета величины аномальной составляющей вектора энтропии. Авторы демонстрируют, что энтропия является эффективным методом обнаружения необычных изменений распределения функций трафика, вызванных аномалиями, и что их многопоточный метод подпространства является эффективным методом извлечения аномальных изменений трафика.

Четкое объяснение методологии как причины использования нескольких мер из разных семей, так и качества, необходимые для решения проблемы, чрезвычайно полезно для понимания того, как и почему использовалась каждая мера. Авторы не представили формулу для квадратичного евклидова расстояния, но они идентифицировали ее как 2-мерную, а затем как xLњ 2, где xLњ - вектор аномальных компонент. Хотя формула не указана, уровень ясности был достаточен для понимания и повторения эксперимента.

Рик и Ласков [15] используют множество измерений расстояния при оценке влияния выбора меры на точность обнаружения аномалий. Первая определенная мера - это линейное ядро ​​(dlk, см. (40), где L - это язык, соответствующий всем последовательностям длины n, доступным в полезных нагрузках входящего соединения, w представляет собой последовательность, которая является частью L, x и y, являются двумя входящими Полезная нагрузка соединения, а φw (x) - частота появления w в x. Это фактически является точечным произведением вероятностных распределений w внутри x и y.

Вторая определенная мера идентифицируется авторами как геодезическое расстояние, которое авторы формулируют как dgd = arccos (dlk). Следует отметить, что это не соответствует стандартному определению геодезического расстояния. Третья определяемая мера - расстояние Канберры (dcan), как показано в (42).

dcan(x, y) = w∈L |φw(x) − φw(y)| φw(x) + φw(y) (42)

Четвертой мерой, определяемой авторами, является Jensen Distance, другая мера, не определенная в «Энциклопедии расстояний», в которой используется энтропийно подобная функция H. Переменные внутри H - две из суммирующих переменных a, b и c; (20) – (22).

Из четырех указанных мер две из них не найдены в Энциклопедии расстояний: dgd и djens. В дополнение к измерениям расстояния авторы описывают четыре коэффициента подобия, также используемые для сравнения двух полезных нагрузок соединения, которые используют переменные суммирования, определенные в (20). Описанные коэффициенты подобия - это Jaccard (sj), Czekanowski (sc), Sokal-Sneath (ss) и Kulszynski (sk), как показано в (43).

(43)

Рик и Ласков [15] также используют упрощенную версию расстояния Махаланобиса, разработанную Вангом и Столфо [10] для использования в высокопроизводительных средах. Упрощенная мера mμ, σ определена в (7).

Примечательны два аспекта обработки ими этих мер. Во-первых, авторы целенаправленно включают измерения расстояния и сходства в той части эксперимента, которую нужно изменить, изучая влияние различных мер на их результаты. Это особенно заметно, поскольку в этом обзоре обнаружено только пять работ, в которых основное внимание уделяется влиянию выбора измерения на результаты. Во-вторых, упрощенное расстояние Махаланобиса используется авторами как механизм обнаружения аномалий: оно не включено в список дистанционных мер.

В этом разделе мы представили примерные исследования, показывающие, как авторы эффективно идентифицировали и представили четкую параметризацию выбора дистанционной меры.

TABLE I QUANTITY OF EXPLICITLY NAMED DISTANCE OR SIMILARITY MEASURES WITHIN SAMPLED WORKS

Мы также показали, как несколько разных методов используются для сравнения производительности системы при изменении параметров расстояния или параметров. Наконец, важно отметить, что несколько авторов успешно оценили новые применения дистанционных мер. Несмотря на эти несколько хороших примеров надлежащего лечения дистанционных мер, которые мы можем найти, к сожалению, следует отметить, как будет описано в следующем разделе, что подавляющее большинство исследований в этой области не имеет сходного строгого подхода к определению И объяснение их использования дистанционных мер в НИД.

**VI. Анализ использования мер расстояния в области**

**(VI. AN ANALYSIS OF THE USE OF DISTANCE**

**MEASURES IN THE FIELD)**

Чтобы обеспечить объективное представление о состоянии идентификации и объяснения дистанционной меры в литературе, мы случайно опробовали сечение публикаций в исследованиях обнаружения сетевых вторжений (NID). Во-первых, мы определили 556 статей, опубликованных в период с 2008 по 2012 год (включительно), содержащих термины «сетевое вторжение» в заголовок, реферат или ключевые слова, используя функцию поиска обратных ссылок Google Scholar на глубину до трех. Из этих работ мы случайно выбрали образец из 100 статей для ручного обзора.

В нашем опросе используется «Энциклопедия расстояний» [4] в качестве стандартного списка определений мер. Любые имена и формулы, найденные в образце, переводятся в соответствии с энциклопедией. Чтобы количественно определить, насколько хорошо название и формулировка каждой дистанционной меры соответствует стандарту, мы классифицировали документы на три группы:

* Стандарт: название или формула измерения указаны в документе и идентифицируются в «Энциклопедии расстояний».
* Роман: название или формула измерения указаны в документе и не идентифицируются в Энциклопедии расстояний.
* Не указано: название или формула измерения не указаны в документе.

Категориальная информация из 100 работ в этом примере изображена в двух таблицах: в таблице I показано количество документов, в которых явно указаны используемые меры расстояния, а в таблице II классифицируются документы, основанные на том, предоставляются ли конкретные формулы измерения расстояния или сходства.

Среди работы, отобранной для этого обзора, 65 из этих документов не указали название меры, и 74 из этих документов не дали четкой формулировки.

Полезно понять, какие типы мер используются в области NID, а какие нет. Ориентируясь только на работу с указанными дистанционными измерениями, есть 33 статьи с обозначенными расстояниями,

TABLE II

QUANTITY OF DISTANCE OR SIMILARITY MEASURES WITH

EXPLICIT FORMULATIONS WITHIN SAMPLED WORKS

TABLE III

FREQUENCY OF DISTANCE OR SIMILARITY MEASURES TYPES

USED WITHIN SAMPLED WORK

TABLE IV

FREQUENCY OF DISTANCE MEASURE COMPARISON

некоторые с более чем одной мерой, и три типа расстояний, идентифицированных в рамках выборочной работы. Три различных типа наблюдаемых мер определены в «Энциклопедии расстояний: мощность (p, r) - расстояние, расстояния от законов распределения и расстояния на строках и перестановках. Распределение дистанционных мер в 33 документах приведено в таблице III.

Большинство (57% работ с указанной мерой) измеренных дистанционных мер, используемых в выборке, основаны на мощности (p, r) -состоянии и расстояниях от законов распределения. Образец показывает, что мало изучено возможное использование других мер в этой области, поскольку большинство мер основано на мощности (p, r) -состоянии или расстояниях от законов распределения.

В этом обзоре из 100 работ в этой области также было обнаружено, что подавляющее большинство (98%) документов не исследуют более одной дистанционной меры в течение любой отдельной фазы NID. Таблица IV показывает, что только две статьи сравнивали более одного измерения расстояния на этапе исследования. Эта поразительная находка показывает, что есть еще больше возможностей для изучения в области торговых возможностей дистанционной выборки в исследованиях.

Учитывая эти данные, мы ясно видим, что необходимо проделать работу по идентификации, спецификации, обоснованию и сравнению использования дистанционных мер. Содержание примера 100-бумаги, классифицированного в таблице IV, суммируется в следующих записях. Эти записи организованы в первую очередь, определяется ли используемая мера (ов) в работе,

TABLE V

MEASURES USED IN SELECTION

а затем подразделяется на фазу NID, в которой используется эта мера.

**А. Мера определена (Measure Defined)**

В системе обнаружения вторжений (IDS) существуют три фазы, в которых измерения расстояния или сходства наблюдаются в выборке: выбор, кластеризация (неконтролируемая классификация) и контролируемая классификация. Одним из первых возможных применений меры расстояния или сходства в IDS является этап выбора. Разбивка различных мер, указанных в образце, приведена в таблице V. Две категории показателей выделяются при их использовании в выборе: мощность (p, r) расстояние и расстояния по законам распределения.

1) Выбор: Есть три примера, в которых используется выбор мощности (p, r) в качестве основы для выбора функции. Li et al. [27] предлагают использовать кластеризацию k-средних для уменьшения числа оцениваемых наблюдений, и авторы обсуждают использование расстояния между данными и кластерами. Хотя авторы никогда точно не определяют расстояние, использование кластеров k -средств предполагает, что Euclidean Distance является разумным предположением. Tan et al. [46] генерируют карты Махаланобиса для совершения выбора функции, а затем анализируют результаты сопоставлений между нормальными и атаками. Затем анализ карты используется для выбора функций для классификации. Lu et al. [43] выбирают функции на основе алгоритма Isomap, который используется для уменьшения размерности, сохраняя при этом близость точек данных, не изменяется при уменьшении размерности.

Дифференциалы мощности (p, r) являются очень полезными инструментами и часто используются практически во всех работах, но при использовании мощности (p, r) необходимо учитывать трудности. Например, установка p = r = 2 приводит к евклидову дистанции, которая при использовании на высокоразмерных данных может иметь встречные интуитивные результаты.

Существует шесть примеров использования расстояний по законам распределения на этапе выбора признаков. Первые четыре используют KLD. Sindhu et al. [28] используют KLD для вычисления коэффициента усиления каждого атрибута в процессе выбора функции. Коэффициент усиления затем используется для вычисления атрибута, наиболее удаленного от коэффициента усиления нуля (и, следовательно, наиболее полезного для классификации), и неявно используется в качестве меры подобия. Wang et al. [51] и Сингх и Силакари [50] используют исходный KLD для определения выбранных функций: сохраняются функции с высокими значениями KLD. Во всех трех примерах KLD используется как базовая мера сходства с порогом, установленным исследованием, используемым для определения коэффициента усиления. Lima et al. [35] представлено сравнение различных мер энтропии (Реньи, Цаллис и Шеннон) для выбора признаков. Три определения энтропии используются для уменьшения признаков на основе коэффициента усиления, затем выбираются 10 атрибутов с наибольшими коэффициентами усиления.

В последних двух статьях вместо KLD используются энтропия на этапе выбора. Devarakonda et al. [33] используют энтропию как метод выбора признаков. Функции ранжируются в порядке убывания по энтропии, тогда лучшие 15 функций выбраны для использования в классификации. Chen et al. [48] ​​используют энтропию потоков сетевого трафика для обучения Vector Vector Machine (SVM). Они сравнивали скорость обнаружения атак и точность классификаторов SVM, обучаемых по полному набору набора данных Knowledge Discovery и Data Mining Cup 1999 (KDD99), функции, выбранные с помощью теории грубого набора и энтропии. Энтропия используется как мера каждого потока, но не используется как расстояние между потоками и является хорошим примером категории расстояний по категориям распределения.

Во всей выборке только одна статья рассмотрела влияние различных мер на эксперимент при применении к селекции и наиболее просто используемый коэффициент усиления информации. Меры, используемые в выборке, четко указывают на отсутствие опубликованных исследований в этой области, а также подразумеваемую пригодность KLD в качестве меры по выбору признаков.

2) Кластеризация: неконтролируемая классификация или кластеризация предоставляет метод обнаружения для групп, которые неизвестны пользователю. Эти группировки частично зависят от метода кластеризации и меры, используемые для определения расстояния между экземплярами. В частности, тип меры имеет решающее значение для определения эффективности кластеризации. Разбивка различных мер, указанных в образце, приведена в таблице VI. В фазе кластеризации среди документов в выборке используются три разных типа меры: мощность (p, r), расстояния на строках и перестановки и переменная расстояния.

Среди тех методов кластеризации, которые используют мощность (p, r), наиболее распространенным методом кластеризации является k-средство и его производные. В литературе существует широкий спектр реализаций при изучении использования кластеров k -средств: Brahmi et al. [40] используют стандартную методологию k-средств, Borah et al. [32] хэш-входы перед кластеризацией, Ян и Ми [47] используют функцию пригодности для максимизации межкластерного расстояния и минимизации межкластерного расстояния, а Bharti et al. [45] предлагают модификацию алгоритма k-средних для решения проблем без класса и класса.

Помимо кластеризации k-сред, некоторые другие методы неконтролируемой классификации реализуются с использованием расстояния мощности (p, r) в качестве меры выбора. Devarakonda et al. [33] используют метод кластеризации k-ближайшего соседа как один из многих в ансамблех голосования. Сингх и Силакари [50] также используют k-ближайший сосед в качестве метода классификации. Использование расстояния находится в контексте классификатора k-ближайшего соседа, и авторы указывают использование евклидова расстояния, но обратите внимание, что вместо этого можно использовать другие расстояния (например, расстояние Манхэттена). Chou et al. [53] классифицируются с помощью метода нечеткой c-меры кластеризации. Этот метод использует функцию оптимизации, которая определяется как основанная на l2-норме (евклидовом расстоянии). Хаят и Хашеми [42] демонстрируют кластеризацию с ограничением ограниченной памяти и вводят метод кластеризации на основе дискретного косинусного преобразования. Чэн и Вэнь [41] используют самоорганизующуюся карту для классификации. Палмиери и Фиоре [44] рассматривают потоки трафика как динамическую систему и поэтому используют анализ количественной оценки повторения как метод неконтролируемой классификации. Однако авторы признают, что преобразование между размерностями может привести к искажению расстояний между точками и попытка компенсировать путем определения «ложных ближайших соседей» и использовать квадратичное евклидово расстояние как метрику. Наконец, Obimbo et al. [37] исследуют использование самоорганизованных карт функций на основе голосования. Обучаются несколько самоорганизующихся карт функций, и каждый голосует за полученную форму данных. В эксперименте авторы сравнивают использование евклидовой дистанции с использованием индивидуального измерения расстояния. Чжэн и Ван [39] исследуют кластерный подход, состоящий из двух фаз: фаза кластеризации, которая задается с использованием евклидова расстояния и фаза оптимизации частиц, которая использует кластеры из фазы кластеризации в качестве исходных частиц. Евклидовое расстояние точки от центра кластера используется в фитнес-функции оптимизации роицы частиц.

Используемые дистанции или методы сходства не всегда основаны на физическом расстоянии. Gogoi et al. [36] демонстрируют подход кластеризации к обнаружению аномалий, в котором сходство между двумя объектами определяется как количество атрибутов, которые имеют одинаковые значения между двумя объектами. В другой работе Zhuang et al. [52] внедрить подход IDS с поддержкой близости для идентификации червей. Кластерный подход используется на основе «близости». Представленный алгоритм широко использует переменную, представляющую расстояние, но эта переменная не определена.

3) Контролируемая классификация: контролируемая классификация - это широкое поле со многими возможными алгоритмами и мерами, которые будут использоваться. Разбивка различных мер, указанных в образце, приведена в таблице VII. В следующих работах основное внимание уделяется возможным улучшениям классификации с использованием контролируемых методов и использованию вариаций по следующим категориям мер: мощность (p, r) расстояние, расстояния по законам распределения, расстояния на строках и перестановки и новые меры, которые ранее не были определены.

Среди контролируемых методов классификации, использующих расстояние мощности (p, r), существует более широкое разнообразие типов классификации. Gong, et al. [26] изучают модификацию алгоритма отрицательного выбора с использованием дополнительной фазы обучения, чтобы свести к минимуму количество самопроверки, необходимых для покрытия саморегиона, и значительно уменьшить частоту ложных срабатываний и скорость обнаружения. Ferreira et al. [34] исследуют использование вейвлет-анализа и искусственной нейронной сети (ANN) для IDS.

TABLE VII

MEASURES USED IN SUPERVISED CLASSIFICATION

Yi et al. [38] вводит улучшенный инкрементный алгоритм поддержки векторной машины с модифицированной функцией ядра, которая весит точки данных в соответствии с их расстоянием до гиперплоскости, чтобы указать вероятность того, что точка будет опорным вектором.

Однако не все использование мощности (p, r) расстояния строго евклидова. Kou et al. [49] обращаются к случаю классификации нескольких классов (более двух). Для классификации данных с нелинейностями используется метод ядра, называемый многокритериальным математическим программированием. Концепция расстояния до гиперплоскости требуется для метода ядра и задается как формула, содержащая l2-норму.

Существует два примера расстояний по законам распределения среди контролируемых методов классификации. Altwaijry и Algarny [29] рассматривают классификатор Naive Bayes, в котором обрабатываются данные обучения, чтобы найти вероятность того, что каждое значение функции встречается в нормальных данных. Вычисленная вероятность затем используется как порог, чтобы определить, являются ли новые данные атакой или нормальным. В этом случае вычисления вероятности используются как «дистанционная мера», хотя расстояние явно не обсуждается. Аршади и Джахангир [31] предлагают использовать энтропию времени между приемами пакетов в скользящем окне для обнаружения атак типа SYN, основанных на концепции, что пакеты атаки имеют более низкую энтропию, чем обычные пакеты. Если энтропия в текущем окне меньше средней энтропии минус три стандартных отклонения, то атака идентифицируется. Опять же, энтропия определяется и используется как мера расстояния, но расстояние явно не обсуждается.

Среди расстояний на строках и перестановках существует множество применений. Большинство идентифицировано этим образцом, чтобы попасть в те методы, которые слабо обозначены как искусственная иммунная система. Antunes и Correia [30] исследуют иммунологическую концепцию порогов настраиваемой активации. Это в отличие от общепринятых иммунологических концепций отрицательного отбора и теории опасности. При изучении искусственных иммунных систем концепция расстояния известна как сродство. В этом исследовании аффинность определяется как максимальная длина подстроки, общей между рецептором Т-клеток (элемент детектора) и пептидом (элемент данных, подлежащих классификации), или подобие общей общей подстроки между х и Y (LCS (x, y)), где x и y - строки. Zhang et al. [54] изучают искусственные иммунные системы и определяют аффинность как позволяющую «любой тип Хэмминга, r-непрерывное совпадение и т. Д.». В этом исследовании используется алгоритм соответствия Хэмминга для определения сродства. Он и Парамешваран [58] работают из-за того, что аномальные связи от одного атакующего похожи друг на друга. Они разрабатывают систему, которая проверяет множественные соединения для сходства в кластерных группах, сравнивает их с заданным порогом и помещает все выше порога как аномальные. Разработанная новая мера сходства называется Anomaly Metric SG и связана с измерениями расстояния на перестановках или строках.

Некоторые из рассмотренных примеров классификации не могут использовать ранее определенные меры дистанции и сходства и разрабатывают новые методы измерений, специфичные для темы исследования. Mabu et al. [55] предлагают структуру IDS для генерации правил ассоциации классов с использованием теории нечетких множеств в сочетании с программированием генетической сети. При обнаружении неправильного использования или аномалии генерируются два пула для хранения правил ассоциации для обычных и вторжений, и эти правила применяются к классификации набора данных. Новая мера сходства определяется автором под названием «Степень сопоставления», которая используется для определения того, соответствует ли вновь созданное правило известным правилам. Шью и Сайнани [56] предлагают структуру, которая будет использоваться при разработке IDS с использованием нескольких методов классификации и нескольких агентов. Цель состоит в том, чтобы уменьшить сложность, связанную с созданием IDS и распределить нагрузку по всей сети, вместо того чтобы иметь всю нагрузку в одной точке. Мера расстояния определена, но упоминается только как «мера расстояния» и использует собственные векторы и собственные значения для генерирования значения меры, которое затем сравнивается с порогом для определения аномалии или нормальности. Su et al. [57] предлагает метод сравнения наборов правил нечеткой ассоциации, сгенерированных с использованием инкрементного интеллектуального анализа данных. Набор правил генерируется из входящего трафика, а другой - из учебного трафика без атаки. Сходство между наборами правил используется для определения ненормальности трафика, из которого генерируются наборы правил, причем решение принимается каждые две секунды. Мера подобия - это новая формулировка, определяемая с целью оценки правил в этом исследовании.

**В. Мера не определена (Measure Undefined)**

Ниже приводятся те работы, которые не предусматривали конкретной идентификации или формулировки используемой меры. Документы будут сгруппированы, как в разделе VI-A, с добавлением еще одной категории: работы, которые сосредоточены на всей системе, представляя новую структуру или структуру вместо того, чтобы сосредоточиться на одной фазе.

1) Сфокусированная на системе: грубо говоря, работы, которые сосредоточены на всей системе, а не на одной фазе системы, можно сгруппировать следующим образом: ансамбль, мультиагент и сравнение существующих систем с использованием какого-либо нового метода для сравнения. В ряде работ основное внимание уделяется ансамблевым методам, в которых сильные стороны нескольких методов классификации, независимо от того, не контролируются или контролируются, объединяются для создания (надеюсь) более эффективного метода обнаружения вторжений. Время от времени эффективность измеряется скоростью обнаружения, ложноположительной скоростью или другими способами. Основное внимание, независимо от того, как оценивается улучшение, заключается в улучшении результатов классификации и уменьшении количества ложно классифицированных экземпляров. Методы ансамбля можно разделить на два общих типа: многоуровневое и голосование.

Среди многоуровневых методов существует ряд, которые используют два или более методов классификации для улучшения результатов. Али и Джантан [63] используют два уровня для обнаружения атак: первый распознает нежелательные характеристики, а затем передает любые неатак на второй уровень, который распознает желаемые характеристики. После того как количество обнаружений, сделанных вторым слоем, пройдет порог, они кластерируются с использованием k-средних, и первый уровень обучается результатам. Xu et al. [116] используют «дерево персептрона», которое является деревом решений, в котором каждый узел является нейронной сетью. В этом исследовании предлагается, чтобы такая комбинация отображала преимущества как символических, так и несимволических моделей. Мохаммед и Аваделькарим [71] предлагают структуру NID с использованием дерева решений для обнаружения известных атак и двухступенчатой ​​кластеризации для обнаружения новых атак. В оценке используются как KDD99, так и реальные сетевые данные, а структура сравнивается с системой обнаружения вторжений в Миннесоте, которая использует Snort и локальную кластеризацию факторов выбросов. Salama et al. [75] внедрить метод обнаружения вторжений с использованием сети глубоких убеждений (DBN) для выполнения функции сокращения, а затем классифицировать остальные функции с помощью SVM. Авторы также сравнивают метод DBN-SVM с каждым в качестве автономного классификатора и сравнивают DBN с другими общими методами сокращения признаков. Результаты более чем DBN-SVM более точны, чем один, и DBN более эффективен, чем основной компонентный анализ (PCA), коэффициент усиления и Chi-Square в качестве метода сокращения признаков. Чжан [84] предлагает нейронную сеть, в которой выбор признаков, структура сети и адаптация веса развиваются вместе с использованием генетического алгоритма как части улучшенной эволюционной нейронной сети. Оператор эвристической мутации используется для предотвращения завершения поиска в локальных оптимумах и позволяет полностью охватить пространство поиска. Сарвари и Кейха [92] предлагают использовать несколько методов машинного обучения для обнаружения атак в системе обнаружения вторжений комбинаторных методов машинного обучения (IDSCML). В предлагаемом методе используется k-ближайший сосед, деревья решений, нейронные сети и SVM, чтобы воспользоваться преимуществами методов обнаружения аномалий и неправильного использования. Наиболее полезным методом является комбинация дерева решений, 1 ближайшего соседа, 2 ближайшего соседа, 3 ближайшего соседа и SVM, которые параллельны по данным, затем объединяются с использованием нейронной сети. Wang et al. [97] представляет использование алгоритма искусственной пчелиной колонии для выбора как бесплатных параметров классификатора SVM, так и функций, используемых при классификации набора данных KDD99. В каждой категории нападений, использованных в экспериментах, метод искусственной пчелиной колонии свободного параметра и выбор признаков приводил к большей точности, чем оптимизация рога или генетического алгоритма. Folino et al. [90] используют ансамблевый подход, основанный на генетическом программировании. Этот подход применяется распределенным образом для создания сетевого профиля. Этот подход проверен в наборе данных KDD99 и показывает производительность, аналогичную выигрышной записи в конкурсе KDD99.

Несмотря на преобладание многослойных систем, существует несколько систем голосования, ориентированных на всю систему, а не на одну фазу. Panda и Patra [111] собирают классификаторы AdaBoost, MultiBoosting и Bagging каждый в сочетании с деревом решений, обрезанным с использованием обрезания с уменьшенной ошибкой, и результаты затем сравниваются с другими классификаторами, используемыми в ходе недавних исследований. Zainal et al. [117] объединяют отдельные классификаторы, предназначенные для обнаружения одного класса, и каждый из которых использует другой метод обучения. Затем классификаторы голосуют, чтобы определить окончательную классификацию. Выбор функции выполняется с использованием техники грубого набора и двоичной частичной частички в двухуровневом процессе. Показано, что система работает лучше, чем самый эффективный классификатор. Farid et al. [67] объединяет Boosting (AdaBoost) с наивным байесовским классификатором. Один классификатор создается для каждой функции в данных, тогда все классификаторы используются для вычисления вероятности появления невидимой точки данных и голосования классификаторов, чтобы определить, является ли точка нормальной или аномальной.

Хотя расслоение может быть общим решением для объединения нескольких методов обнаружения сетевых аномалий (NIAD), некоторые из них попробовали несколько агентов. Joldos and Muntean [69] представляют идею сравнения расстояний между векторами признаков наборов данных, образованных из подмножеств больших стандартных наборов данных: в этом исследовании основное внимание уделяется исследованию преимуществ выполнения функций сокращения и классификации задач для обнаружения вторжений с использованием грид-вычислений. Zeng et al. [101] включает в себя три основных компонента: узел обнаружения вторжений, координатор обнаружения вторжений и агент snooper. Координатор и агент обычно расположены на одном и том же хосте, в то время как на каждом сегменте локальной сети (ЛВС) есть по меньшей мере один узел. Узел фиксирует и анализирует трафик и передает новую информацию координатору. Координатор управляет базами данных и подписи, а также функцией оповещения. Агент запускается при получении новой информацией от узла координатором и собирает информацию, которая может понадобиться для новых атак. Barika et al. [105] и [106] предлагают архитектуру IDS с использованием распределенных мобильных агентов по всей сети с четырьмя типами агентов: сниффер, фильтр, анализатор и решение. Производительность MA\_IDS тестируется с использованием как сканирования портов, так и SYN-наводнений, а также путем передачи сообщений через все агенты. MA\_IDS также сравнивается с централизованной системой обнаружения, и система обнаружения агента имеет меньшую потерю пакетов и более короткие задержки обнаружения, так как количество пакетов в секунду и общие пакеты увеличиваются. Rehak et al. [124] использует мультиагентный подход для обнаружения аномалий и построения модели доверия. Каждый агент обнаружения использует один метод обнаружения аномалий и вносит вклад в модель доверия, собранную коллективно несколькими агентами. Атаки определяются пороговым значением «счет доверия», который в этом контексте может считаться дистанционной мерой. Yu et al. [83] представлена ​​модель обнаружения вторжений на основе модульных мобильных агентов. Они используют среду Aglets для создания имитационной среды мобильного агента и используют модель цепи Маркова в качестве классификатора вторжений. Положительная корреляция определяется как количеством детекторов, так и длиной данных, а значение True Positive Rate (TPR): увеличение количества детекторов или длины данных приводит к увеличению TPR. Gao et al. [109] предлагают распределенную структуру IDS с использованием как локальных, так и глобальных детекторов. Локальный детектор - это алгоритм множественного адабоуста, в котором используется метод максимизации ожидания для обновления параметров детектора. Глобальный детектор представляет собой комбинацию оптимизации роя частиц и SVM. Результаты показывают, что выбор параметра для локального детектора важен для минимизации False Positive Rate (FPR) и показывает, что глобальный детектор имеет скорость обнаружения 99,99% и ложную положительную скорость 0,3713%.

Существуют также работы, посвященные сопоставлению различных IDS или различным исследованиям конкретных наборов данных. Engen et al. [66] сравнивают результаты исследований с обнаружением вторжений с использованием деревьев решений и методов Найева Байеса для классификации набора данных KDD99. Сравнение используется для исследования расхождений в подмножествах KDD99, используемых в исследованиях, и определения того, полезен ли KDD99 для дальнейших исследований или слишком ошибочный, чтобы быть полезным. Эльджади и Отман [65] демонстрируют использование трех отдельных методов интеллектуального анализа данных для обнаружения аномального трафика в наборе данных сетевого трафика реального мира. Результаты используют пороговое значение для определения того, содержат ли сегменты исследуемого набора данных попытки вторжения. Каждый метод оценивается на предмет пригодности при обнаружении вторжений путем оценки времени, необходимого для оценки сегмента набора данных, и возможности генерации правил, в которых используется обновление системы обнаружения вторжений сети (NIDS). Пастрана и др. [74] предлагают основу для использования при моделировании существующих НИДС. Генетическое программирование используется для генерации экземпляра, который ведет себя как смоделированный NIDS, но гораздо проще, с целью определения слабых точек. Созданная модель позволяет исследователям открывать новые методы уклонения и предоставляет новый метод проверки эффективности коммерческих NIDS. Gogoi et al. [91] представляют ограниченный обзор NIDS, основанный на аномалии, несколько числовых и категоричных мер приближения и сравнение производительности некоторых контролируемых и неконтролируемых NIDS. В этой работе обсуждаются различные меры близости и их формулировки, но не упоминается описание или определение меры. Шафи и Аббас [77] предлагают метод генерации наборов данных NIDS с реальным фоновым трафиком и имитируемыми атаками. Затем они сравнивают ряд различных алгоритмов обнаружения вторжений с набором данных, созданным авторами в соответствии с предлагаемым методом. Day and Burns [64] сравнивают производительность Snort и Suricata NIDS с многоядерными компьютерными системами. Производительность каждого оценивается на основе точности, ложных срабатываний и негативов, использования системы и снижения трафика. Вывод заключается в том, что, хотя Suricata работает немного лучше, чем Snort, увеличение ресурсов, необходимых Suricata, делает Snort лучшим выбором для NIDS с открытым исходным кодом. Zhengbing et al. [125] предлагают алгоритм поиска новых сигнатур нападения на основе известных сигнатур атаки, используя вариацию алгоритма Apriori (Signature Apriori) для поиска часто встречающихся шаблонов потоков. Предлагаемый алгоритм протестирован против Signature Apriori с использованием SNORT. Обнаружено, что предлагаемый алгоритм более эффективен, чем Signature Apriori с одинаковыми скоростями обнаружения, в случае, если новая атака получена из более ранней атаки. Фанелли [88] экспериментирует с иммунной системой, вдохновляющей IDS, сравнивая результаты обнаружения Snort и распознавания сетевых угроз с обнаружением аномальной иммунной системы (NetTRIIAD). Два NIDS сравниваются с использованием набора данных KDD99 в категориях известных атак, неизвестных атак и тестов на абляцию.

2) Выбор. Многие исследователи фокусируются на фазе выбора как основной метод снижения сложности в области NIAD: если метод кластеризации или классификации имеет меньше возможностей или наблюдений для обработки, то общий результат достигается быстрее. Хор и др. [59] предлагают метод разбиения наборов данных NID на основе частоты атаки. Обычные записи включены в оба, но в один только редкие атаки, в то время как другие включают в себя только редкие атаки. В обоих наборах данных авторы сократили количество записей, включенных для предотвращения того, чтобы какой-то класс был подавляющим другим из-за количества включенных записей. Экспериментальные результаты показывают, что разделение наборов данных может улучшить точность классификации некоторых классов, но не всех. Farid et al. [108] исследуют алгоритм наивного байесовского классификатора и алгоритм Итеративного Дихотомизера 3 для сокращения признаков, где KLD используется для выбора лучших атрибутов. Ahmad et al. [61] и [62] предлагают два метода выбора объектов, которые применяют генетические алгоритмы к результату PCA. Функция пригодности, используемая для генетического алгоритма, использует многослойную точность персептрона в одной статье, точность SVM в другом и количество функций, не выбранных для конкретного подмножества признаков в обеих документах. Точность 0,99 достигается с помощью 12 из 38 функций KDD99, выбранных с использованием многослойного персептрона, и с SVM достигается точность 99,6%. Сен и Кларк [76] рассматривают использование эволюционных вычислений для разработки программ обнаружения вторжений для мобильных сетей ad hoc. Генетическое программирование и грамматическая эволюция используются для развития программ обнаружения вторжений, когда программы выполняют свой собственный выбор функций (все функции предоставляются) для отдельных атак, множественной атаки и, наконец, совместного обнаружения атаки. Пригодность любого данного решения определяется скоростью обнаружения-ложноположительной, а в экспериментах рассматриваются меры безопасности и энергопотребления из-за присущих ограничений мобильных устройств. Das et al. [86] предлагают использовать несколько методов машинного обучения для классификации попыток вторжения в сеть. Препроцессор извлекает 14 функций из сетевого трафика каждые 4 секунды. Выделенные функции затем исследуются для использования в классификации двумя методами: PCA и грубой теорией множеств. Основываясь на сравнительном исследовании в рамках этой работы, в качестве более эффективного метода выбора объектов выбирается теория грубого набора. Выбранные функции отправляются в SVM для изучения и классификации. Аль-Шарафат и Наум [104] изучают важность выбора признаков при обнаружении сетевых атак и проводят эксперименты для определения наиболее эффективной комбинации функций, используемых в исследованиях. Набор функций, выбранных в четырех разных работах, используется для классификации сетевых атак. Используемые классификаторы генерируются генетическим алгоритмом машинного обучения на основе устойчивого состояния. Из четырех классов функций один из них привел к скорости обнаружения выше, чем другие три: 97,5%. Шанмугам и Идрис [112] вносят вариацию алгоритма интеллектуального анализа данных Куком и Априори с использованием нечеткой логики для создания правил, выражаемых как логические последствия. Этот подход используется для сокращения функций для определения функций, которые обеспечивают максимальный KLD для классификации атак. Сингх и Силакари [114] вводят подход обобщенного дискриминантного анализа к сокращению признаков, затем используют ANN для классификации. Обобщенный дискриминантный анализ (GDA) представляет собой форму линейного дискриминантного анализа (LDA), в котором набор данных сначала преобразуется в пространство с более высоким размером до сокращения функции, чтобы более успешно обрабатывать нелинейные наборы данных. Подход ANN сравнивается с подходом классификации дерева решений C4.5, и оба они рассматриваются как для LDA, так и для GDA для сокращения возможностей. GDA-редуктор работает немного лучше, чем LDA, как с C4.5, так и с ANN-классификаторами. Заман и Каррей [118] представляют собой исследование по уменьшению признаков, в котором делается попытка улучшить метод поддержки векторного решения (ESVDF), интегрируя его с моделью нечеткого вывода. Этот подход упрощает сложность проектирования и сокращает время выполнения. Этот подход проверен на KDD99, сравнивая нечеткие ESVDF с другими типами ESVDF в качестве селекторов функций. Классификаторы векторной машины Neural Network и Support используются для классификации сокращения пост-функции данных. Было обнаружено, что точность достигается с помощью предлагаемого метода с уменьшением времени обучения. Zargar и Kabiri [120] применяют анализ основных компонентов для уменьшения возможностей набора данных KDD99 для обнаружения атак типа smurf. Метод k-ближайшего соседа используется в качестве классификатора для данных с уменьшенными характеристиками. Abdulla et al. [85] предлагают набор функций, которые следует использовать при применении искусственных нейронных сетей к проблеме обнаружения вторжений. Изучается влияние, вызванное использованием разных типов и чисел векторов в качестве входа в нейронную сеть, а также влияние изменения количества эпох для прогона. Авторы демонстрируют, что скорость обнаружения возрастает по мере увеличения числа эпох или векторов.

3) Кластеризация. Кластеризация - это, как правило, эффективный метод обнаружения групп среди данных, которые исследователь может не знать, и он поддается использованию явных мер расстояния и подобия. Это может быть причиной, из-за которой мало (относительно) выборочных работ сосредоточено на кластеризации, но не указывается дистанционная мера. Niemalä [72] классифицирует, используя k-метод кластеризации. Song et al. [78] представляют способ кластеризации, за которым следует использование машины носителя поддержки. В этом исследовании ранее невидимая точка данных определяется как аномальная, если она «находится внутри гиперсферы». Zhenying [121] сравнивает одноуровневую и многоуровневую самоорганизующуюся карту и показывает, что карта с несколькими слоями неспособна повысить производительность. Предполагается, что причина отсутствия улучшения связана с перекрытием между различными классами данных. Taranum и Lamble [79] имитируют гибридную IDS, работающую на мобильных устройствах. Комбинация кластеризации по запросу (с использованием протокола маршрутизации векторной рекламы Ad Hock On Demand) и сбора информации о соседях (с использованием протокола маршрутизации маршрутизации по маршруту назначения) используется в дополнение к NIDS на основе хоста для обнаружения вторжений. Мобильные рекламные сети (MANET) группируются с каждым кластером, у которого есть голова, которая отвечает за обнаружение вторжений для этого кластера. В случае, когда глава подозревает вторжение, но нуждается в дополнительных доказательствах, он привлекает узлы в кластере для совместного обнаружения вторжений. Модификация гибридного кооперативного подхода была продемонстрирована симуляцией.

4) Контролируемая классификация: контролируемая классификация является наиболее надежным методом классификации, так как позволяет исследователю или оператору определять, что является нормальным или аномальным, и обучать классификатор результатам этих решений. Одним из методов, используемых исследователями, является дерево. Farid et al. [89] пытаются уменьшить скорость ложных срабатываний, используя взвешивание атрибутов на основе дерева решений для выбора объектов и адаптивное наивное байесовское дерево для классификации. Srinivasulu et al. [95] используют частый алгоритм обучения шаблону дерева деревьев, чтобы изучить обычное поведение клиентов в базе данных транзакций. В исследовании определены две меры для правил ассоциации: поддержка и уверенность, которые служат для целей дистанционных мер. Поддержка - это отношение транзакций, содержащих все указанные значения признаков, к общему количеству проанализированных транзакций, а доверие - отношение транзакций, содержащих одно указанное значение функции, к общему количеству проанализированных транзакций. Hu et al. [123] применяют Adaboost, алгоритм обучения, который объединяет слабые классификаторы, с обнаружением вторжений. Они используют «решающие пни», деревья решений с одним корнем и двумя листовыми узлами, как слабые классификаторы. Показано, что этот подход является менее сложным по сравнению с классификаторами ANN и SVM, тогда как скорость обнаружения и скорость ложной тревоги сопоставимы. Visumathi и Shunmuganathan [82] предлагают новую архитектуру IDS, которая использует обнаружение неправильного использования (эффективное распознавание сигнатур) для выявления атак, использует алгоритм Apriori для известных сигнатур атаки для создания набора шаблонов, сравнивает известные шаблоны атаки с вероятной атакой Затем добавляет подпись атаки для любого вероятного шаблона атаки, который имеет сходство более 0,9 с любым известным шаблоном атаки. Авторы получили набор результатов с использованием множества методов машинного обучения, включая SVM, и обнаружили, что алгоритм случайного леса дал наивысшую точность с 99,97, но не предоставляет никакой информации о скорости ложной тревоги.

Другой метод моделируется после иммунной системы, иногда обозначаемой как искусственная иммунная система. Фанелли [122] представляет гибридную иммунную систему IDS: NetTRIIAD. Система использует два слоя для предотвращения неправильного использования и обнаружения аномалий. Врожденный слой выполняет обнаружение неправильного использования для существующих угроз, обеспечивая обнаружение известных атак. Адаптивный слой выполняет самонесущую дискриминацию, имитируя реакцию организма на злоумышленников с использованием Т-клеток. По сравнению со Snort, система NetTRIIAD имеет эквивалентный TPR и значительно более низкий FPR, поэтому его положительное прогностическое значение (0,65) почти удваивается от Snort (0,38). Xiao-Pei и Hou-Xiang [99] предлагают и демонстрируют экспериментальную IDS, которая использует модель обнаружения, основанной на иммунной системе, с детекторами, генерируемыми как случайным образом, так и с использованием генетического алгоритма для создания незрелых детекторов. Когда обнаружены атаки, функции извлекаются и кодируются для использования в детекторе вакцины. Функция фитнеса с использованием TPR и FPR используется для определения вероятности того, что конкретный детектор станет родителем. Показано, что предлагаемый метод более эффективен как в TPR, так и в FPR при обнаружении всех типов атак в наборе данных KDD99 по сравнению с классической моделью обнаружения иммунитета. Zamani et al. [119] используют концепцию «теории опасностей» из биологической иммунологии, в которой клетки идентифицируют чужеродные клетки, которые необходимо атаковать на основе сигналов, отправленных умирающими самоэлементами. Они определяют «уровень концентрации коллимуляции» как взвешенную сумму «отправляет» и «получает» от других сетевых узлов. Этот подход используется для обнаружения атак распределенного отказа в обслуживании (DDoS) и тестируется в имитируемой сети.

Имитация человеческих систем для обнаружения атак не является новой, так как нейронные сети моделируются в течение многих лет. Использование ANN привело к достижениям в области обнаружения вторжений и многим другим. Seliya и Khoshgoftaar [93] представляют собой активную процедуру обучения с использованием нейронных сетей. Затем производительность новой процедуры сравнивается с результатами применения дерева решений C4.5 к одному и тому же набору данных (KDD99). Общие результаты показывают, что метод активного обучения лучше обобщается при обнаружении вторжений, чем метод С4.5. Norouzian и Merati [73] используют двухслойную нейронную сеть для классификации атак как отдельных типов на основе сетевого трафика. Многие работы рассматривают классификацию как проблему с двумя слоями, когда запись является либо атакой, либо нормальной. Вместо этого классы атаки используются, как записано создателями набора данных KDD99. Sheikhan et al. [94] эксперимент с многослойными персептронами и нейронными сетями Эльмана для классификации части набора данных KDD99. Классификаторы оцениваются согласно скорости обнаружения и скорости ложной тревоги, и авторы также предлагают формулу «Стоимость за образец» для использования при оценке классификаторов. Показано, что многослойный персептрон с 15 входами работает лучше, чем эквивалентная нейронная сеть Эльмана и две лучшие исполнители в конкурсе 2000 KDD. Wang et al. [98] эксперимент с использованием нейронной сети с нечеткой синхронизацией для обнаружения атаки «Отказ в обслуживании» (DoS). Нечеткие множества используются для определения восьми функций, которые будут использоваться при обнаружении DoS-атак, а затем для преобразования части набора данных KDD99 в форму новых нечетких функций. Часть преобразованных функций используется для обучения двухслойной нейронной сети в течение 276 эпох. Ahmad et al. [103] демонстрируют использование нейронной сети обратного распространения как IDS для DoS-атак. Корневая среднеквадратичная ошибка используется для оценки производительности различных конфигураций скрытого слоя с наименьшим значением, выбранным в качестве конфигурации для эксперимента. Шейхан и Шабани [113] используют подход нейронной сети (многослойный персептрон), в котором они пытаются улучшить скорость обучения, используя алгоритм обучения оптимизации «оптимизированный весовой вес». Результаты показывают увеличение скорости обучения без потери точности классификации по сравнению с победителем 2000 года конкурса KDD99. Тиан и Гао [115] применяют генетический алгоритм к процессу обратного распространения в многослойном персептроне для повышения скорости сходимости сети. Этот подход применяется к набору данных, основанному на наборе данных лаборатории MIT Lincoln Laboratory. Обратное распространение с использованием генетического алгоритма приводит к уменьшению средней квадратичной ошибки обнаружения аномалии по сравнению со стандартным алгоритмом обратного распространения. Orfila et al. [110] исследовать применение генетического программирования к проблеме NID, в частности, к автоматическому созданию правил и шаблонов для использования при обнаружении атак. Генетический программный подход к построению правил сравнивается с использованием C4.5 и оказался более простым и требует меньше операций для пакета протокола управления передачей (TCP).

SVM - еще один полезный инструмент в контролируемом наборе инструментов классификации. SVM позволяет экстраполировать низкоразмерные данные в более высокие размеры с целью классификации каждого наблюдения с помощью гиперплоскости в качестве разделителя для групп данных. Abdulla et al. [60] предлагают использовать SVM для классификации данных NetFlow и предупреждения о червях. Они демонстрируют ложные положительные и отрицательные значения от 0.09 до 0.00 после уточнения данных для удаления несуществующих узлов, тогда как ставки между 0.10 и 0.28 без уточнения. Он [68] предлагает использовать машину векторных реле (RVM) с использованием логистической хаотической карты вместо стандартного гауссова в качестве оценки шума в выходном сигнале. Результаты классификации набора данных KDD99 как с RVM, так и с SVM сравниваются с использованием рабочих характеристических кривых приемника и требуемого количества векторов. Результаты показывают, что RVM имеет более низкие значения ложной тревоги при определенной вероятности обнаружения и генерирует меньше векторов для обработки одинакового количества записей.

Те, которые не вписываются в одну из опрятных категорий выше, также заслуживают внимания. Kahn и Burney [70] предлагают систему обнаружения вторжений, состоящую из машины конечного состояния (FMS), которая использует Push Down Automata (PDA) для выполнения хранения экземпляров атаки. Результаты использования FSM как NID не ясно продемонстрированы, так как точность и ложноположительные ставки не указаны. Vijayasarathy et al. [81] предлагают легкую структуру классификатора DoS для работы как с протоколами TCP, так и с протоколами пользовательских дейтаграмм (UDP). В структуре используется пакетная обработка окон для разделения входного трафика на подмножества, используется флаги TCP для определения шести категорий с T1 по T6 и кросс-валидация для определения порога, по которому предполагается, что атака происходит. Точность повысилась, а уровень ложной тревоги уменьшился по мере увеличения размера окна, однако авторы отмечают, что порог должен быть установлен опытным сетевым администратором. Faizal et al. [87] предлагает метод обнаружения аномалий на основе количества соединений в течение одного второго периода. Результаты этого обнаружения сравниваются при условии, что весь трафик нормальный. Скорость обнаружения увеличивается до 85,9%, а ложноположительная - до 3,2%. Torrano-Gimenez et al. [96] предлагают новый брандмауэр веб-приложений для обнаружения атак на веб-приложение. Брандмауэр снабжен файлом расширяемого языка разметки (XML), содержащим подробное описание нормального поведения веб-приложения и пороговые значения, обеспечивающие некоторую гибкость в определении нормального поведения для веб-приложений. Любой трафик, превышающий пороговые значения, считается атакой. Производительность брандмауэра с использованием описания нормальной работы XML превосходна, обнаруживая все попытки веб-атак. Однако авторы отмечают, что автоматическое описание нормальных описаний операций будет необходимо для реализации в больших масштабах. Ye et al. [100] предлагают систему обнаружения аномалий, используя простую скрытую марковскую модель (HMM). Предполагается, что все сетевые поведения являются нормальными в течение заданного временного окна. Если поведение сети отклоняется от нормального поведения, то поведение считается атакой. Обнаружено, что HMM способен обнаруживать атаки, но не тип атаки. Tuncer и Tatar [80] предлагают встроенную систему для обнаружения DoS-атак в режиме реального времени. Встраиваемая система использует программируемую систему на чипе для обучения по трафику и генерации аварийных сигналов на этапе тестирования. Логические правила ассоциаций выводятся на этапе обучения на основе пяти атрибутов трафика, и тогда микросхема запрограммирована на распознавание трафика, соответствующего этим правилам, как атаки DoS. Changguo et al. [107] модифицировать стандартный алгоритм Apriori с помощью разработки правил нечетких ассоциаций и применить этот метод для обнаружения вторжений беспроводной сети. Эксперимент показывает уменьшение числа «наборов предметов» кандидата с предлагаемым методом. Zhu et al. [102] представляют атаку на полуконтролируемые классификаторы, вводя поддельные немеченые экземпляры. Предложен алгоритм генерации вводящих в заблуждение экземпляров, показано влияние на классификатор и предложена возможная защита, включающая самообучение путем сравнения помеченных и немаркированных экземпляров. Обнаружено, что вводящие в заблуждение экземпляры уменьшают точность наивных байесовских и самоназванных классов Naivve Bayes, но самообучение уменьшило снижение точности.

В разделе VII представлены подводные камни и уроки, извлеченные в ходе этого обзора области. Он также предоставляет примерные документы, которые читатель может использовать в качестве шаблонов для улучшения своей собственной будущей работы.

**VII. Уроки выучены  
(VII. LESSONS LEARNED)**

Тенденции в публикации исследований в области обнаружения сетевых вторжений (NID), изученных в разделе VI, показывают несколько областей для улучшения. Четыре распространенных недостатка:

1. Неисправность выбора выбранных расстояний
2. Невозможность предоставить адекватные данные (например, режим или параметрирование) на выбранных дистанционных измерениях
3. Невозможность объяснить, почему выбрана дистанция (ы)
4. Неисправность измерения расстояния и параметризации как Еще один экспериментальный фактор оценки эффективности

Когда авторы становятся жертвами этих ловушек, экспериментальная повторяемость и валидация подвергаются риску. Кроме того, выгоды, обнаруженные в отношении дистанционных мер во время исследования, не могут быть полностью реализованы полем. В оставшейся части этого раздела эти ловушки описаны подробно, и обсуждаются стратегии смягчения. Для каждой категории мы также предлагаем примеры в области, которые демонстрируют рекомендуемые рекомендации для будущих исследований.

Первая ошибка заключается в том, что идентификация дистанционных мер часто отсутствует. В полевой выборке из 100 статей только 40% идентифицировали название используемой дистанционной меры. Поскольку выбор расстояния измеряет центральную роль алгоритма в обнаружении аномалий, исследователи должны четко определить, какие дистанционные меры они используют, и определить, как они их используют. Исследователи также должны использовать стандартизованные имена для дистанционных мер, и мы рекомендуем «Энциклопедию расстояний» [4], которые помогут в этом. Примерами работ, которые делают это очень хорошо, является статья Гунга с описанием суммы квадратов расстояния, используемого для порогового значения [26], и статьи Борах, описывающей использование мощности (p, r) -состояния для алгоритма K-ближайших соседей [ 28]. В каждой из этих статей авторы четко идентифицируют используемые расстояния и предоставляют формулы.

Во-вторых, хотя некоторые авторы указывают, какие измерения расстояния используются, гораздо меньше дают детали реализации, такие как формулы или параметры. В полевом образце только 32% дали конкретную математическую формулировку для их дистанционной меры. Параметры параметров также важны в некоторых измерениях расстояния, но часто игнорируются, когда авторы представляют свои исследования. Например, при использовании мощности (p, r) - расстояния p и r могут быть выбраны независимо, но часто предполагаются одинаковыми, как в евклидовой дистанционной формулировке (p = r = 2). Другой пример потенциальной двусмысленности раскрывается при расчете различных расстояний, связанных с Энтропией, где расстояние зависит от выбранной базы лога. В большинстве случаев предполагается, что используется база данных по умолчанию 2, но можно выбрать другую базу. Другим недостатком этой категории является использование «настроек по умолчанию» готового алгоритма с предположением, что эти настройки по умолчанию останутся неизменными на бесконечность. Библиотеки машинного обучения часто обновляются, а настройки по умолчанию, используемые во время исследования, могут быть изменены к моменту опубликования исследования. Еще более опасная практика возникает, когда исследователи не понимают, что, когда они выбирают готовый алгоритм без полного понимания деталей реализации, они могут не понимать, что для них были выбраны настройки по умолчанию - эти настройки могут быть даже суб- Оптимальным для изучаемого явления.

Типичным документом в этой категории являются Аршади и Джахангир [31], которые идентифицировали меру расстояния, предоставили формулу, и представили подробные описания параметров и переменных для использования энтропии в качестве меры расстояния для определения случайности скорости прибытия Пакетов. Этот уровень документации значительно облегчает воссоздание эксперимента для проверки и сравнения с будущими исследованиями.

В-третьих, даже когда детали реализации известны и предоставляются, исследователи часто не могут указать, почему делаются выборы. Невозможность объяснить эти варианты приводит к упущенной возможности передать важные возможности обучения читателю. Это может замедлить общий прогресс в области продвижения.

Палмиери служит примером этой категории: он объясняет, почему он использует квадрат-евклидовое расстояние в качестве меры для определения ближайших и ложных ближайших соседей, поскольку размеры пространственных объектов увеличиваются один за другим [44].

Наконец, в отличие от иллюстративных статей, описанных в Разделе V, образец показывает, что в подавляющем большинстве исследований выбрана единая дистанция или мера подобия, и авторы не исследуют торговое пространство альтернатив дистанционной меры или даже альтернативные параметры. В нашем примере поля только два автора (2%) исследовали выбор дистанционной меры как фактор в дизайне эксперимента.

В то время как в большинстве исследований не было проведено более одного измерения расстояния для одной фазы обнаружения аномалии, два рассмотренных нами лечения выглядели многообещающими, поскольку они обеспечивали структуру, которую можно было бы повторить с помощью дистанционных мер, которые авторы не использовали. Одна статья Обембо исследует эффективность классификации с использованием двух мер - евклидовой дистанции и индивидуальной меры, основанной на системе голосования [37]. В другом примерном сравнительном исследовании была рассмотрена эффективность трех мер энтропии для определения наилучших возможностей использования [35]. Эти авторы прокладывают путь, используя дистанционную меру как еще один фактор в дизайне эксперимента. Мы рекомендуем будущие исследования следовать этому примеру.

**VIII. Заключение  
(Conclusion)**

В каждом эксперименте, использующем детектирование аномалии (AD) в поле обнаружения сетевых вторжений (NID), используются дистанционные меры, большинство из которых не задумывались о том, какая дистанционная мера будет наиболее подходящей. Тем не менее, ясно, что иногда сходство и дистанционные меры используются с осторожностью в исследованиях. Есть отличные примеры для руководства и расширения. Мы рекомендуем использовать следующие навыки для улучшения качества исследования:

* Четко назовите и опишите все дистанционные меры и параметры, используемые во время исследовательских усилий. Когда в «Энциклопедии расстояний» [4] стандартизируется название меры.
* Методы измерения расстояний между другими областями, которые имеют схожие проблемы, такие как обработка естественного языка, и исследование решений, которые не используются в поле NID.
* Рассмотрите возможность использования различных измерений расстояния и сходства в рамках эксперимента, чтобы определить, как они влияют на скорость обнаружения.
* Включить гибкие методы для сбора и выражения значений данных для сопоставления расстояний. Например, постройте матрицу расстояния графа (обсуждается в разделе VII), как используется Tan et al. [23] как метод сравнения.
* Разработать методы для сравнения матриц расстояний графа без визуализации и определить, какие пороговые значения наиболее полезны при определенных условиях.
* Получите лучшее понимание того, как дистанционные меры бросают вызов интуитивному пониманию ближайшего слова и развивают визуализацию и моделирование, чтобы помочь в понимании того, как устанавливать параметры, подходящие для проблемного пространства.

ACKNOWLEDGMENT

Мнения, выраженные в этом тезисах, относятся к авторам и не отражают официальную политику или позицию Военно-воздушных сил Соединенных Штатов, Министерства обороны или правительства США.

REFERENCES

[1] V. Chandola, A. Banerjee, and V. Kumar, “Anomaly detection: A survey,”

ACM Comput. Surv., vol. 41, no. 3, pp. 15:1–15:58, Jul. 2009.

[2] A. Chmielewski and S. Wierzchon, “On the distance norms for detecting ´

anomalies in multidimensional datasets,” Zeszyty Naukowe Politechniki

Białostockiej, vol. 2, pp. 39–49, 2007.

[3] S.-H. Cha, “Comprehensive survey on distance/similarity measures

between probability density functions,” Int. J. Math. Models Methods

Appl. Sci., vol. 1, no. 4, pp. 300–307, 2007.

[4] M. Deza and E. Deza, Encyclopedia of Distances. Berlin, Germany:

Springer-Verlag, 2009.

[5] S. Staab, “Ontologies and similarity,” in Case-Based Reasoning Research

and Development, vol. 6880. Berlin, Germany: Springer-Verlag,

2011, ser. Lecture Notes in Computer Science, pp. 11–16.

[6] P. Cunningham, “A taxonomy of similarity mechanisms for case-based

reasoning,” IEEE Trans. Knowl. Data Eng., vol. 21, no. 11, pp. 1532–

1543, Nov. 2009.

[7] C.-J. Chung, P. Khatkar, T. Xing, J. Lee, and D. Huang, “NICE: Network

intrusion detection and countermeasure selection in virtual network

systems,” IEEE Trans. Dependable Secure Comput., vol. 10, no. 4,

pp. 198–211, Jul./Aug. 2013.

[8] F. Aurenhammer, “Voronoi diagrams—A survey of a fundamental

geometric data structure,” ACM Comput. Surv., vol. 23, no. 3, pp. 345–

405, Sep. 1991.

[9] P. C. Mahalanobis, “On the generalised distance in statistics,” Proc. Nat.

Inst. Sci. India, vol. 2, no. 1, pp. 49–55, Jan. 1936.

[10] K. Wang and S. Stolfo, “Anomalous payload-based network intrusion detection,”

in Recent Advances in Intrusion Detection, vol. 3224. Berlin,

Germany: Springer-Verlag, 2004, ser. Lecture Notes in Computer Science,

pp. 203–222.

[11] S. Teng, H. Du, W. Zhang, X. Fu, and X. Li, “A cooperative network intrusion

detection based on heterogeneous distance function clustering,”

in Proc. 14th Int. Conf. Comput. Supported Coop. Work Des., Apr. 2010,

pp. 140–145.

[12] D. Hancock and G. Lamont, “Multi agent system for network attack classification

using flow-based intrusion detection,” in Proc. IEEE Congr.

Evol. Comput., Jun. 2011, pp. 1535–1542.

[13] J. Cho, K. Choi, T. Shon, and J. Moon, “A network data abstraction

method for data set verification,” in Secure and Trust Computing, Data

Management and Applications, vol. 186. Berlin, Germany: SpringerVerlag,

Jun. 2011, ser. Communications in Computer and Information

Science, pp. 54–62.

[14] Y. Wang, Z. Zhang, L. Guo, and S. Li, “Using entropy to classify traffic

more deeply,” in Proc. 6th IEEE Int. Conf. Netw., Architecture Storage,

Jul. 2011, pp. 45–52.

[15] K. Rieck and P. Laskov, “Language models for detection of unknown

attacks in network traffic,” J. Comput. Virol., vol. 2, no. 4, pp. 243–256,

Feb. 2007.

[16] H. Sengar, X. Wang, H. Wang, D. Wijesekera, and S. Jajodia, “Online

detection of network traffic anomalies using behavioral distance,” in

Proc. 17th Int. Workshop Qual. Serv., Jul. 2009, pp. 1–9.

[17] Z. Lu and T. Peng, “The VoIP intrusion detection through a LVQ-based

neural network,” in Proc. Int. Conf. Internet Technol. Secured Trans.,

Nov. 2009, pp. 1–6.

[18] H. Kwon, T. Kim, S. Yu, and H. Kim, “Self-similarity based

lightweight intrusion detection method for cloud computing,” in Intelligent

Information and Database Systems, vol. 6592. Berlin, Germany:

Springer-Verlag, 2011, ser. Lecture Notes in Computer Science,

pp. 353–362.

[19] G. Gu, J. Zhang, and W. Lee, “Botsniffer: Detecting botnet command and

control channels in network traffic,” in Proc. 15th Annu. Netw. Distrib.

Syst. Security Symp., 2008, pp. 1–18.

[20] T. Krueger, C. Gehl, K. Rieck, and P. Laskov, “An architecture for

inline anomaly detection,” in Proc. Eur. Conf. Comput. Netw. Defense,

Dec. 2008, pp. 11–18.

[21] H. F. Eid, M. A. Salama, A. E. Hassanien, and T.-H. Kim, “Bi-layer

behavioral-based feature selection approach for network intrusion classi-

fication,” in Security Technology, vol. 259. Berlin, Germany: SpringerVerlag,

Dec. 2011, ser. Communications in Computer and Information

Science, pp. 195–203.

[22] A. Chmielewski and S. T. Wierzchon, “An immune approach to classify- ´

ing the high-dimensional datasets,” in Proc. Int. Multiconf. Comput. Sci.

Inf. Technol., Oct. 2008, pp. 91–96.

[23] Z. Tan, A. Jamdagni, X. He, P. Nanda, and R. Liu, “Multivariate correlation

analysis technique based on Euclidean distance map for network

traffic characterization,” in Information and Communications Security,

vol. 7043. Berlin, Germany: Springer-Verlag, Nov. 2011, ser. Lecture

Notes in Computer Science, pp. 388–398.

[24] G. Zhao, J. Yang, G. Hura, L. Ni, and S.-H. Huang, “Correlating TCP/IP

interactive sessions with correlation coefficient to detect steppingstone

intrusion,” in Proc. Int. Conf. Adv. Inf. Netw. Appl., May 2009,

pp. 546–551.

[25] A. Lakhina, M. Crovella, and C. Diot, “Mining anomalies using traf-

fic feature distributions,” in Proc. Conf. Appl., Technol., Architectures,

Protocols Comput. Commun., Aug. 2005, pp. 217–228.

[26] M. Gong, J. Zhang, J. Ma, and L. Jiao, “An efficient negative selection

algorithm with further training for anomaly detection,” Knowl.-Based

Syst., vol. 30, pp. 185–191, Jun. 2012.

[27] Y. Li et al., “An efficient intrusion detection system based on support vector

machines and gradually feature removal method,” Exp. Syst. Appl.,

vol. 39, no. 1, pp. 424–430, Jan. 2012.

[28] S. S. S. Sindhu, S. Geetha, and A. Kannan, “Decision tree based light

weight intrusion detection using a wrapper approach,” Exp. Syst. Appl.,

vol. 39, no. 1, pp. 129–141, Jan. 2012.

[29] H. Altwaijry and S. Algarny, “Multi-layer Bayesian based intrusion

detection system,” in Proc. WCECS, 2011, vol. II, pp. 918–922.

[30] M. Antunes and M. Correia, “Tunable immune detectors for behaviourbased

network intrusion detection,” in Artificial Immune Systems,  
WELLER-FAHY et al.: SURVEY OF DISTANCE AND SIMILARITY MEASURES USED WITHIN NIAD 89

vol. 6825. Berlin, Germany: Springer-Verlag, 2011, ser. Lecture Notes

in Computer Science, pp. 334–347.

[31] L. Arshadi and A. Jahangir, “Entropy based SYN flooding detection,” in

Proc. IEEE 36th Conf. LCN, Oct. 2011, pp. 139–142.

[32] S. Borah, S. P. K. Chetry, and P. K. Singh, “Hashed-K-means: A proposed

intrusion detection algorithm,” in Computational Intelligence and

Information Technology, vol. 250. Berlin, Germany: Springer-Verlag,

2011, ser. Communications in Computer and Information Science,

pp. 855–860.

[33] N. Devarakonda, S. Pamidi, V. Valli Kumari, and A. Govardhan, “Outliers

detection as network intrusion detection system using multi layered

framework,” in Advances in Computer Science and Information Technology,

vol. 131. Berlin, Germany: Springer-Verlag, 2011, ser. Communications

in Computer and Information Science, pp. 101–111.

[34] E. Ferreira, G. Carrijo, R. de Oliveira, and N. de Souza Araujo, “Intrusion

detection system with wavelet and neural artificial network

approach for networks computers,” IEEE Latin America Trans. (Revista

IEEE America Latina), vol. 9, no. 5, pp. 832–837, Sep. 2011.

[35] C. Ferreira Lemos Lima, F. Assis, and C. de Souza, “A comparative study

of use of Shannon, Rényi and Tsallis entropy for attribute selecting in

network intrusion detection,” in Proc. IEEE Int. Workshop Meas. Netw.,

Oct. 2011, pp. 77–82.

[36] P. Gogoi, B. Borah, and D. Bhattacharyya, “Network anomaly detection

using unsupervised model,” Int. J. Comput. Appl.—(Special Issue Netw.

Security Cryptogr.), no. 1, pp. 19–30, Dec. 2011.

[37] C. Obimbo, H. Zhou, and R. Wilson, “Multiple SOFMs working cooperatively

in a vote-based ranking system for network intrusion detection,”

Proc. Comput. Sci.—(Special Issue Complex Adaptive Syst.), vol. 6,

pp. 219–224, 2011.

[38] Y. Yi, J. Wu, and W. Xu, “Incremental SVM based on reserved set for

network intrusion detection,” Exp. Syst. Appl., vol. 38, no. 6, pp. 7698–

7707, Jun. 2011.

[39] H. Zheng, M. Hou, and Y. Wang, “An efficient hybrid clustering-PSO

algorithm for anomaly intrusion detection,” J. Softw., vol. 6, no. 12,

pp. 2350–2360, Dec. 2011.

[40] I. Brahmi, S. Yahia, and P. Poncelet, “MAD-IDS: Novel intrusion

detection system using mobile agents and data mining approaches,”

in Intelligence and Security Informatics, vol. 6122. Berlin, Germany:

Springer-Verlag, 2010, ser. Lecture Notes in Computer Science,

pp. 73–76.

[41] X. Cheng and S. Wen, “A real-time hybrid intrusion detection system

based on principle component analysis and self organizing maps,” in

Proc. 6th Int. Conf. Natural Comput., Aug. 2010, vol. 3, pp. 1182–1185.

[42] M. Hayat and M. Hashemi, “An adaptive DCT based intrusion detection

system,” in Proc. Int. Symp. Comput. Netw. Distrib. Syst., 2010, pp. 1–6.

[43] L. F. Lu, M. L. Huang, M. Orgun, and J. W. Zhang, “An improved

wavelet analysis method for detecting DDoS attacks,” in Proc. 4th Int.

Conf. Netw. Syst. Security, Sep. 2010, pp. 318–322.

[44] F. Palmieri and U. Fiore, “Network anomaly detection through nonlinear

analysis,” Comput. Security, vol. 29, no. 7, pp. 737–755, Oct. 2010.

[45] K. Bharti, S. Shukla, and S. Jain, “Intrusion detection using unsupervised

learning,” Int. J. Comput. Sci. Eng., vol. 2, no. 5, pp. 1865–1870, 2010.

[46] Z. Tan, A. Jamdagni, X. He, and P. Nanda, “Network intrusion detection

based on LDA for payload feature selection,” in Proc. IEEE

GLOBECOM Workshops, Dec. 2010, pp. 1545–1549.

[47] Y. Yang and J. Mi, “Design and implementation of distributed intrusion

detection system based on honeypot,” in Proc. 2nd Int. Conf. Comput.

Eng. Technol., Apr. 2010, vol. 6, pp. V6-260–V6-263.

[48] R. C. Chen, K.-F. Cheng, and C.-F. Hsieh, “Using rough set and support

vector machine for network intrusion detection,” Int. J. Netw. Security

Appl., vol. 1, no. 1, pp. 1–13, Apr. 2009.

[49] G. Kou, Y. Peng, Z. Chen, and Y. Shi, “Multiple criteria mathematical

programming for multi-class classification and application in network

intrusion detection,” Inf. Sci., vol. 179, no. 4, pp. 371–381,

Feb. 2009.

[50] S. Singh and S. Silakari, “An ensemble approach for feature selection

of cyber attack dataset,” Int. J. Comput. Sci. Inf. Security, vol. 6, no. 2,

pp. 297–302, 2009.

[51] J. Wang, Q. Yang, and D. Ren, “An intrusion detection algorithm based

on decision tree technology,” in Proc. Asia-Pac. Conf. Inf. Process.,

Jul. 2009, vol. 2, pp. 333–335.

[52] Z. Zhuang, Y. Li, and Z. Chen, “PAIDS: A proximity-assisted intrusion

detection system for unidentified worms,” in Proc. 33rd Annu. IEEE Int.

Comput. Softw. Appl. Conf., Jul. 2009, vol. 1, pp. 392–399.

[53] T. Chou, K. Yen, and J. Luo, “Network intrusion detection design using

feature selection of soft computing paradigms,” Int. J. Comput. Intell.,

vol. 4, no. 3, pp. 196–208, Jul. 2008.

[54] C. Zhang, J. Zhang, S. Liu, and Y. Liu, “Network intrusion active defense

model based on artificial immune system,” in Proc. 4th Int. Conf. Natural

Comput., Oct. 2008, vol. 1, pp. 97–100.

[55] S. Mabu, C. Chen, N. Lu, K. Shimada, and K. Hirasawa, “An intrusiondetection

model based on fuzzy class-association-rule mining using

genetic network programming,” IEEE Trans. Syst., Man, Cybern. C,

Appl. Rev., vol. 41, no. 1, pp. 130–139, Jan. 2011.

[56] M.-L. Shyu and V. Sainani, “A multiagent-based intrusion detection

system with the support of multi-class supervised classification,” in Data

Mining and Multiagent Integration. New York, NY, USA: SpringerVerlag,

2009, pp. 127–142.

[57] M.-Y. Su, G.-J. Yu, and C.-Y. Lin, “A real-time network intrusion

detection system for large-scale attacks based on an incremental

mining approach,” Comput. Security, vol. 28, no. 5, pp. 301–309,

Jul. 2009.

[58] X. He and S. Parameswaran, “MCAD: Multiple connection based

anomaly detection,” in Proc. 11th IEEE Int. Conf. Commun. Syst.,

Nov. 2008, pp. 999–1004.

[59] K.-C. Khor, C.-Y. Ting, and S. Phon-Amnuaisuk, “A cascaded classi-

fier approach for improving detection rates on rare attack categories in

network intrusion detection,” Appl. Intell., vol. 36, no. 2, pp. 320–329,

Mar. 2012.

[60] S. A. Abdulla, S. Ramadass, A. Altaher, and A. A. Nassiri, “Setting

a worm attack warning by using machine learning to classify

NetFlow data,” Int. J. Comput. Appl., vol. 36, no. 2, pp. 49–56,

Dec. 2011.

[61] I. Ahmad, A. Abdullah, A. Alghamdi, K. Alnfajan, and M. Hussain,

“Intrusion detection using feature subset selection based on MLP,”

Sci. Res. Essays, vol. 6, no. 34, pp. 6804–6810, Dec. 2011.

[62] I. Ahmad, A. Abdullah, A. Alghamdi, and M. Hussain, “Optimized

intrusion detection mechanism using soft computing techniques,”

Telecommun. Syst., vol. 52, no. 4, pp. 2187–2195, Apr. 2013.

[63] G. A. Ali and A. Jantan, “A new approach based on honeybee to improve

intrusion detection system using neural network and bees algorithm,”

in Software Engineering and Computer Systems, vol. 181. Berlin,

Germany: Springer-Verlag, 2011, ser. Communications in Computer and

Information Science, pp. 777–792.

[64] D. Day and B. Burns, “A performance analysis of Snort and Suricata

network intrusion detection and prevention engines,” in Proc. 5th Int.

Conf. Digit. Society, Feb. 2011, pp. 187–192.

[65] E. Eljadi and Z. Othman, “Anomaly detection for PTM’s network traf-

fic using association rule,” in Proc. 3rd Conf. Data Mining Optim.,

Jun. 2011, pp. 63–69.

[66] V. Engen, J. Vincent, and K. Phalp, “Exploring discrepancies in findings

obtained with the KDD cup ’99 data set,” Intell. Data Anal., vol. 15,

no. 2, pp. 251–276, Apr. 2011.

[67] D. M. Farid, M. Z. Rahman, and C. M. Rahman, “Adaptive intrusion detection

based on boosting and naïve Bayesian classifier,” Int. J. Comput.

Appl., vol. 24, no. 3, pp. 12–19, Jun. 2011.

[68] D. He, “Improving the computer network intrusion detection performance

using the relevance vector machine with Chebyshev chaotic

map,” in Proc. IEEE Int. Symp. Circuits Syst., May 2011, pp. 1584–1587.

[69] M. Joldos and I. Muntean, “Distributed investigations of intrusion detection

data on the grid,” in Proc. 10th RoEduNet Int. Conf., Jun. 2011,

pp. 1–4.

[70] D. A. Khan and D. Burney, “Efficient FSM techniques for IDS to reduce

the system attacks,” Int. J. Comput. Appl., vol. 29, no. 11, pp. 42–47,

Sep. 2011.

[71] R. Mohammed and A. Awadelkarim, “Design and implementation of

a data mining-based network intrusion detection scheme,” Asian J. Inf.

Technol., vol. 10, no. 4, pp. 136–141, 2011.

[72] A. Niemelä, “Traffic analysis for intrusion detection in telecommunications

networks,” M.S. thesis, Tampere Univ. Technol., Tampre, Finland,

Mar., 2011.

[73] M. Norouzian and S. Merati, “Classifying attacks in a network intrusion

detection system based on artificial neural networks,” in Proc. 13th Int.

Conf. Adv. Commun. Technol., Feb. 2011, pp. 868–873.

[74] S. Pastrana, A. Orfila, and A. Ribagorda, “A functional framework to

evade network IDS,” in Proc. 44th Hawaii Int. Conf. Syst. Sci., Jan. 2011,

pp. 1–10.

[75] M. Salama, H. Eid, R. Ramadan, A. Darwish, and A. Hassanien, “Hybrid

intelligent intrusion detection scheme,” in Soft Computing in Industrial

Applications, vol. 96. Berlin, Germany: Springer-Verlag, 2011,

ser. Advances in Intelligent and Soft Computing, pp. 293–303.

[76] S. Sen and J. A. Clark, “Evolutionary computation techniques for intrusion

detection in mobile ad hoc networks,” Comput. Netw., vol. 55,

no. 15, pp. 3441–3457, Oct. 2011.  
90 IEEE COMMUNICATION SURVEYS & TUTORIALS, VOL. 17, NO. 1, FIRST QUARTER 2015

[77] K. Shafi and H. Abbass, “Evaluation of an adaptive genetic-based

signature extraction system for network intrusion detection,” Pattern

Anal. Appl., vol. 16, no. 4, pp. 549–566, Nov. 2013.

[78] J. Song, H. Takakura, Y. Okabe, and K. Nakao, “Toward a more practical

unsupervised anomaly detection system,” Inf. Sci., vol. 231, pp. 4–14,

May 2011.

[79] R. Tarannum and M. Lamble, “Hybrid approach: Detection of intrusion

in MANET,” IJCA Proc. Innov. Conf. Embedded Syst., Mobile Commun.

Comput., vol. ICEMC2, no. 1, pp. 24–28, Sep. 2011.

[80] T. Tuncer and Y. Tatar, “Detection DoS attack on FPGA using fuzzy

association rules,” in Proc. IEEE 10th Int. Conf. Trust, Security Privacy

Comput. Commun., Nov. 2011, pp. 1271–1276.

[81] R. Vijayasarathy, S. Raghavan, and B. Ravindran, “A system approach to

network modeling for DDoS detection using a naïve Bayesian classifier,”

in Proc. 3rd Int. Conf. Commun. Syst. Netw., Jan. 2011, pp. 1–10.

[82] J. Visumathi and K. Shunmuganathan, “A computational intelligence for

evaluation of intrusion detection system,” Indian J. Sci. Technol., vol. 4,

no. 1, pp. 40–45, Jan. 2011.

[83] W. Yu, C. Xiaohui, and W. Sheng, “Anomaly network detection model

based on mobile agent,” in Proc. 3rd Int. Conf. Meas. Technol. Mechatron.

Autom., Jan. 2011, vol. 1, pp. 504–507.

[84] B. Zhang, “A heuristic genetic neural network for intrusion detection,”

in Proc. Int. Conf. Internet Comput. Inf. Serv., Sep. 2011, pp. 510–513.

[85] S. Abdulla, N. Al-Dabagh, and O. Zakaria, “Identify features and parameters

to devise an accurate intrusion detection system using artifi-

cial neural network,” World Acad. Sci., Eng. Technol., vol. 46, no. 70,

pp. 627–631, Oct. 2010.

[86] V. Das et al., “Network intrusion detection system based on machine

learning algorithms,” Int. J. Comput. Sci. Inf. Technol., vol. 2, no. 6,

pp. 138–151, Dec. 2010.

[87] M. Faizal et al., “Time based intrusion detection on fast attack for

network intrusion detection system,” in Proc. 2nd Int. Conf. Netw. Appl.

Protocols Serv., Sep. 2010, pp. 148–152.

[88] R. Fanelli, “Further experimentation with hybrid immune inspired network

intrusion detection,” in Artificial Immune Systems, vol. 6209.

Berlin, Germany: Springer-Verlag, Jul. 2010, ser. Lecture Notes in

Computer Science, pp. 264–275.

[89] D. M. Farid and M. Z. Rahman, “Attribute weighting with adaptive

NBTree for reducing false positives in intrusion detection,” Int. J.

Comput. Sci. Inf. Security, vol. 8, no. 1, pp. 19–26, 2010.

[90] G. Folino, C. Pizzuti, and G. Spezzano, “An ensemble-based evolutionary

framework for coping with distributed intrusion detection,”

Genetic Programm. Evol. Mach., vol. 11, no. 2, pp. 131–146,

Jun. 2010.

[91] P. Gogoi, B. Borah, and D. Bhattacharyya, “Anomaly detection

analysis of intrusion data using supervised & unsupervised approach,”

J. Convergence Inf. Technol., vol. 5, no. 1, pp. 95–110,

Feb. 2010.

[92] H. Sarvari and M. Keikha, “Improving the accuracy of intrusion

detection systems by using the combination of machine learning

approaches,” in Proc. Int. Conf. Soft Compu. Pattern Recog., Dec. 2010,

pp. 334–337.

[93] N. Seliya and T. Khoshgoftaar, “Active learning with neural networks

for intrusion detection,” in Proc. IEEE Int. Conf. Inf. Reuse Integr.,

Aug. 2010, pp. 49–54.

[94] M. Sheikhan, Z. Jadidi, and M. Beheshti, “Effects of feature reduction

on the performance of attack recognition by static and dynamic neural

networks,” World Appl. Sci. J., vol. 8, no. 3, pp. 302–308, 2010.

[95] P. Srinivasulu, J. R. Rao, and I. R. Babu, “Network intrusion detection

using FP tree rules,” J. Adv. Netw. Appl., vol. 1, no. 1, pp. 30–39, 2009.

[96] C. Torrano-Giménez, A. Pérez-Villegas, and G. Álvarez Marañón, “An

anomaly-based approach for intrusion detection in web traffic,” J. Inf.

Assur. Security, vol. 5, no. 4, pp. 446–454, 2010.

[97] J. Wang, T. Li, and R. Ren, “A real time IDSs based on artificial bee

colony-support vector machine algorithm,” in Proc. 3rd Int. Workshop

Adv. Comput. Intell., Aug. 2010, pp. 91–96.

[98] Y. Wang, D. Gu, M. Wen, J. Xu, and H. Li, “Denial of service detection

with hybrid fuzzy set based feed forward neural network,” in Advances

in Neural Networks—ISNN 2010, vol. 6064. Berlin, Germany:

Springer-Verlag, Jun. 2010, ser. Lecture Notes in Computer Science,

pp. 576–585.

[99] J. Xiao-Pei and W. Hou-Xiang, “A new immunity intrusion detection

model based on genetic algorithm and vaccine mechanism,” Int. J.

Comput. Netw. Inf. Security, vol. 2, no. 2, pp. 33–39, Dec. 2010.

[100] X. Ye, J. Li, and Y. Li, “An anomaly detection system based on hide

Markov model for MANET,” in Proc. 6th Int. Conf. Wireless Commun.

Netw. Mobile Comput., Sep. 2010, pp. 1–4.

[101] B. Zeng, L. Yao, and Z. Chen, “A network intrusion detection system

with the snooping agents,” in Proc. Int. Conf. Comput. Appl. Syst.

Modeling, Oct. 2010, vol. 3, pp. V3-232–V3-236.

[102] F. Zhu, J. Long, W. Zhao, and Z. Cai, “A misleading attack against

semi-supervised learning for intrusion detection,” in Modeling Decisions

for Artificial Intelligence, vol. 6408, V. Torra, Y. Narukawa, and

M. Daumas, Eds. Berlin, Germany: Springer-Verlag, 2010, ser. Lecture

Notes in Computer Science, pp. 287–298.

[103] I. Ahmad, A. B. Abdullah, and A. S. Alghamdi, “Application of artificial

neural network in detection of DOS attacks,” in Proc. 2nd Int. Conf.

Security Inf. Netw., Oct. 2009, pp. 229–234.

[104] W. Al-Sharafat and R. Naoum, “Significant of features selection for detecting

network intrusions,” in Proc. Int. Conf. Internet Technol. Secured

Trans., Nov. 2009, pp. 1–6.

[105] F. Barika, N. El Kadhi, and K. Ghedira, “MA IDS: Mobile agents for

intrusion detection system,” in Proc. IEEE Int. Adv. Comput. Conf.,

Mar. 2009, pp. 900–910.

[106] F. A. Barika, N. E. Kadhi, and K. Ghédira, “Agent IDS based on misuse

approach,” J. Softw., vol. 4, no. 6, pp. 495–507, Aug. 2009.

[107] Y. Changguo et al., “Improvement of association rules mining algorithm

in wireless network intrusion detection,” in Proc. Int. Conf. Comput.

Intell. Natural Comput., Jun. 2009, vol. 2, pp. 413–416.

[108] D. Md. Farid, J. Darmont, N. Harbi, H. H. Nguyen, and M. Z. Rahman,

“Adaptive network intrusion detection learning: Attribute selection and

classification,” in Proc. Int. Conf. Comput. Syst. Eng., Bangkok,

Thailand, Jul. 2009, pp. 154–158.

[109] J. Gao, W. Hu, X. Zhang, and X. Li, “Adaptive distributed intrusion

detection using parametric model,” in Proc. IEEE/WIC/ACM

Int. Joint Conf. Web Intell. Intell. Agent Technol., Sep. 2009, vol. 1,

pp. 675–678.

[110] A. Orfila, J. Estevez-Tapiador, and A. Ribagorda, “Evolving high-speed,

easy-to-understand network intrusion detection rules with genetic programming,”

in Applications of Evolutionary Computing, vol. 5484.

Berlin, Germany: Springer-Verlag, 2009, ser. Lecture Notes in

Computer Science, pp. 93–98.

[111] M. Panda and M. R. Patra, “Ensemble of classifiers for detecting network

intrusion,” in Proc. Int. Conf. Adv. Comput., Commun. Control,

Jan. 2009, pp. 510–515.

[112] B. Shanmugam and N. Idris, “Improved intrusion detection system using

fuzzy logic for detecting anamoly and misuse type of attacks,” in Proc.

Int. Conf. Soft Comput. Pattern Recog., Dec. 2009, pp. 212–217.

[113] M. Sheikhan and A. Sha’bani, “Fast neural intrusion detection system

based on hidden weight optimization algorithm and feature selection,”

World Appl. Sci. J.—(Special Issue Comput. IT), vol. 7, pp. 45–53,

2009.

[114] S. Singh and S. Silakari, “Generalized discriminant analysis algorithm

for feature reduction in cyber attack detection system,” Int. J. Comput.

Sci. Inf. Security, vol. 6, no. 1, pp. 173–180, Oct. 2009, abs/0911.0787.

[115] J. Tian and M. Gao, “Network intrusion detection method based on high

speed and precise genetic algorithm neural network,” in Proc. Int. Conf.

Netw. Security, Wireless Commun. Trusted Comput., Apr. 2009, vol. 2,

pp. 619–622.

[116] Q. Xu, Z. Bai, and L. Yang, “An improved perceptron tree learning model

based intrusion detection approach,” in Proc. Int. Conf. Artif. Intell.

Comput. Intell., Nov. 2009, vol. 4, pp. 307–311.

[117] A. Zainal, M. Maarof, and S. Shamsuddin, “Ensemble classifiers for

network intrusion detection system,” J. Inf. Assur. Security, vol. 4,

pp. 217–225, Jul. 2009.

[118] S. Zaman and F. Karray, “Fuzzy ESVDF approach for intrusion

detection systems,” in Proc. Int. Conf. Adv. Inf. Netw. Appl., May 2009,

pp. 539–545.

[119] M. Zamani, M. Movahedi, M. Ebadzadeh, and H. Pedram, “A DDoSaware

IDS model based on danger theory and mobile agents,” in Proc.

Int. Conf. Comput. Intell. Security, Dec. 2009, vol. 1, pp. 516–520.

[120] G. Zargar and P. Kabiri, “Identification of effective network features

to detect Smurf attacks,” in Proc. IEEE Student Conf. Res. Develop.,

Nov. 2009, pp. 49–52.

[121] M. Zhenying, “Reason for hierarchical self organized map-based intrusion

detection system incapable of increasing detection rate,” in Proc.

Int. Symp. Inf. Eng. Electron. Commerce, May 2009, pp. 150–154.

[122] R. Fanelli, “A hybrid model for immune inspired network intrusion

detection,” in Artificial Immune Systems, vol. 5132. Berlin, Germany:

Springer-Verlag, 2008, ser. Lecture Notes in Computer Science,

pp. 107–118.

[123] W. Hu, W. Hu, and S. Maybank, “AdaBoost-based algorithm for network

intrusion detection,” IEEE Trans. Syst., Man, Cybern. B, Cybern.,

vol. 38, no. 2, pp. 577–583, Apr. 2008.  
WELLER-FAHY et al.: SURVEY OF DISTANCE AND SIMILARITY MEASURES USED WITHIN NIAD 91

[124] M. Rehak, M. Pechoucek, P. Celeda, J. Novotny, and P. Minarik, “CAMNEP:

Agent-based network intrusion detection system,” in Proc. 7th

Int. Joint Conf. Auton. Agents Multiagent Syst., Ind. Track, 2008,

pp. 133–136.

[125] H. Zhengbing, L. Zhitang, and W. Junqi, “A novel network intrusion

detection system (NIDS) based on signatures search of data mining,”

in Proc. 1st Int. Workshop Knowl. Discov. Data Mining, Jan. 2008,

pp. 10–16.

David J. Weller-Fahy (S’11–M’12) received the

B.S. degree in computer science from the University

of Illinois at Springfield, Springfield, IL, USA,

in 2010 and the M.S. degree in cyber operations

from the Air Force Institute of Technology, WrightPatterson

Air Force Base, Dayton, OH, USA, in

2013. He is a Senior Master Sergeant with the

Air Force. His research interests include network

intrusion detection, machine learning, artificial

intelligence, network characterization, and data

visualization.

Brett J. Borghetti received the B.S. degree in electrical

engineering from the Worcester Polytechnic

Institute, Worcester, MA, USA, in 1992; the M.S.

degree in computer systems from the Air Force Institute

of Technology (AFIT), Wright-Patterson Air

Force Base, Dayton, OH, USA, in 1996; and the

Ph.D. degree in computer science from the University

of Minnesota, Twin Cities, MN, USA, in 2008.

He is an Assistant Professor of computer science

at AFIT. His research interests focus on anomaly

detection, artificial intelligence, multi-agent systems,

human–machine teams, game theory, and machine learning.

Angela A. Sodemann (M’12) received the M.S. degree

in mechanical engineering from the University

of Wisconsin–Milwaukee, Milwaukee, WI, USA, in

2006 and the Ph.D. degree in mechanical engineering

from the Georgia Institute of Technology, Atlanta,

GA, USA, in 2009. She is currently an Assistant

Professor of mechanical engineering with the Department

of Engineering, College of Technology

and Innovation, Arizona State University, Mesa, AZ,

USA. Her current research interests include applications

of artificial intelligence and machine learning

in mechatronics, robotics, and manufacturing.