УДК 621.377

А.С. Корсунский

АНАЛИЗ МЕТОДОВ РЕШЕНИЯ ЗАДАЧИ РАСПОЗНАВАНИЯ НЕСАНКЦИОНИРОВАННЫХ ВОЗДЕЙСТВИЙ НАРУШИТЕЛЯ

Корсунский Андрей Сергеевич, *кандидат технических наук, окончил факультет радиосвязи Ульяновского филиала Военного университета связи, адъюнктуру Военной академии связи им. С.М. Буденного. Главный специалист ФНПЦ ОАО «НПО «Марс». Имеет статьи и изобретения в области радиоэлектронной защиты, безопасности связи и информации, а также передачи информации по беспроводным каналам связи информационно-телекоммуникационных систем. [e-mail: aksspb@mail.ru].*

Аннотация

В статье рассмотрены математические методы распознавания несанкционированных воздействий нарушителя информационной безопасности в информационно-вычислительной сети.

Ключевые слова: методы распознавания, нарушитель информационной безопасности, информационно-вычислительная сеть.

Введение

При построении эффективной системы защиты информации в информационно-вычислительной сети (ИВС) необходимо четко представлять облик нарушителя. Для адекватного описания поведения нарушителя информационной безопасности требуется провести сравнительный анализ методов решения задачи распознавания несанкционированных воздействий (НВ).

Очевидно, что нарушитель, внося НВ в информационный обмен между сегментами сети, вступит в информационное противоборство с системой защиты ИВС. Находясь в информационном противоборстве с системой защиты, он будет осуществлять НВ на процесс функционирования сети с помощью цифровых потоков (ЦП) по определенным правилам (протоколам) обмена.

Анализ математических методов распознавания несанкционированных воздействий нарушителя в ИВС

По существу решение задачи распознавания НВ заключается в проведении поверхности, разделяющей множество различных образов (классов). Именно поиску данного критерия и посвящены различные методы распознавания, направленные на решение указанной задачи.

В случае, когда эта поверхность (критерий) может быть составлена из плоскостей, метод распознавания называется кусочно-линейным. Процесс обучения считается законченным, когда проведены все разделительные плоскости между

множествами образов. Это удается выполнить при условии компактности распознаваемых образов. Наиболее сложным случаем распознавания являются многосвязанные области объектов, когда точки данного класса расположены в изолированных друг от друга областях. Для распознавания таких объектов (НВ) приходится прибегать к особым, нелинейным методам распознавания.

Первая группа методов распознавания (или, точнее, первая концепция) – геометрическая [1, 2]. Анализируя расположение точек в пространстве образов, можно предложить несколько алгоритмов обучения и решающих правил, причем обучение в этих методах сводится к линейной или нелинейной деформации (максимальному удалению множества эталонов каждого образа друг от друга) пространства признаков.

Из-за сложности вычислений геометрические методы обучения и распознавания непосредственно не находят широкого применения, а служат в основном для интерпретации других методов.

Вторая группа методов распознавания – логические методы [2, 3]. Их использование целесообразно в тех случаях, когда существенны не только количественные соотношения между величинами, характеризующими рассматриваемые процессы, но и связывающие их логические зависимости. Решение задачи распознавания в этом случае сводится к получению выражения, состоящего из изображающего числа и определенным образом связанных между собой исходных элементарных высказываний. Ограниченность использования логических методов для распознавания связана с требованием к детерминированному характеру признаков, то есть к их полной определенности, что в реальных условиях функционирования ИВС при НВ нарушителя выполнимо крайне редко. Однако отсутствие необходимости в сведениях о количественном распределении образов (НВ) в соответствующем пространстве признаков позволяет использовать логические методы для предварительной сравнительной оценки эффективности поддержки выработки решений при распознавании ситуаций.

Третья группа методов распознавания – лингвистические методы. Ориентированы на распознавание зрительных образов, что и ограничивает область их применения. При лингвистическом подходе изображение рассматривается состоящим из ряда частей, в качестве которых выступают геометрические характеристики изображения. Для их получения используется математический аппарат порождающих грамматик [1].

Четвертая группа методов распознавания – статистические методы [4]. Эта группа относится к наиболее широко используемой для решения задач распознавания ситуаций, поэтому рассмотрим ее подробнее.

Самой распространенной концепцией распознавания является байесовская, взятая из теории статистических решений [1, 2, 5]. Применение метода Байеса целесообразно в случае, когда распознавание ситуации осуществляется многократно в условиях неизменного признакового пространства, стабильного описания классов и при неизменной платежной матрице. Его стратегия решений выбирается таким образом, чтобы обеспечить минимум среднего риска.

Несмотря на универсальность метода, реализовать точно байесовский алгоритм обучения практически невозможно, так как для этого требуется запоминать многомерные законы распределения вероятностей, которые в общем случае имеют бесконечные интервалы, поэтому на практике многомерные законы условных вероятностей аппроксимируются более простыми функциями, которые легко запомнить в ЭВМ (например, полиномиальными функциями). В этом случае байесовский метод превращается в дискриминантный.

При распознавании возможны такие случаи, когда априорные вероятности появления ситуаций соответствующих классов неизвестны. Минимизировать средний риск принятия решения на основе байесовской стратегии в этом случае не представляется возможным. Применительно к этой ситуации рационально использовать подход, который минимизирует максимально возможное значение среднего риска. Этот подход называют минимаксным.

При распознавании могут быть неизвестны не только априорные вероятности появления ситуаций соответствующих классов, но и платежная матрица. В подобных случаях для построения алгоритма классификации целесообразно воспользоваться критерием Неймана-Пирсона [2, 5], суть которого состоит в следующем. Исходя из того, какие решения принимаются на основании результатов распознавания неизвестных ситуаций, определяется допустимое (заданное) значение условной вероятности ошибки первого рода Q_1; затем определяется такая граница между классами, придерживаясь которой удается добиться минимума условной вероятности ошибки второго рода.

Поскольку в основе всех перечисленных статистических методов распознавания лежит байесовская стратегия в своем исходном виде, то всем им присущи ее недостатки. Далее рассмотрим метод дискриминантных функций как средство аппроксимации многомерных законов условных вероятностей, обеспечивающее тем самым условие для их запоминания в ЭВМ [1, 4].

Суть метода дискриминантных функций, используемого для распознавания, заключается в следующем: процесс обучения состоит в построении потенциальных функций, а решающее правило – в сравнении значений этих функций для распознающего значения со средним для каждого класса значением функции.

Существует три модификации этого метода: метод решающих функций; метод потенциальных функций; метод дискриминантных функций. Идея у всех этих модификаций одна и та же и заключается в следующем. Считается априори, что существуют поверхности условных плотностей распределения вероятностей $P(x/A_i)=P_{Ai}(X)$, то есть вероятностей появления значений признака x при условии, что ситуация принадлежит к классу A_i. Однако запомнить в ЭВМ эту многомерную функцию зачастую не представляется возможным, поэтому ее аппроксимируют какими-нибудь функциями, которые и называются решающими, потенциальными или дискриминантными функциями $g_i(x)$. Причем практически задача заключается не столько в аппроксимации, сколько в разработке методов построения этих функций, если задан какой-то набор эталонов или задана обучающая последовательность.

Сами функции должны выбираться так, чтобы обеспечить процесс их практического получения. Эта постановка задачи характерна для вероятностного распознавания. При детерминированном распознавании задается некоторое количество эталонов (а не функций распределения вероятностей) и требуется любую новую ситуацию отнести к определенному классу.

Дальнейшим обобщением линейных разделяющих функций в теории распознавания с помощью дискриминантных функций являются так называемые Ф-функции (по Нильсону).

Распознавание, которое использует в качестве дискриминантных функций Ф-функции, называется распознаванием с помощью Ф-функций, а машины – Ф-машинами. Очевидно, что разделяющая функция в Ф-машинах также относится к Ф-функциям, так как разность Ф-функций и функции f_i (с одинаковыми степенями свободы) тоже является Ф-функцией.

Наряду с методом дискриминантных функций при решении задачи распознавания ситуации могут встречаться и другие, частные методы типа корреляционного и регрессионного. Проведение системного анализа ситуаций, зависящих от большого числа признаков, их характеризующих, обусловлено большими трудностями, в частности, связанными с выявлением структуры взаимосвязей этих признаков. Проведение системного анализа до изучения взаимосвязей в многомерной совокупности требует наличия представления о связях между отдельной зависимой переменной и группой влияющих на нее показателей. Это может быть осуществлено при помощи множественного корреляционного и регрессионного анализа.

Корреляционный анализ, разработанный К. Пирсоном и Дж. Юлом, является одним из методов статистического анализа взаимозависимости нескольких признаков – компонентов случайного вектора X [4].

В настоящее время корреляционный анализ (корреляционная модель) определяется как метод, применяемый тогда, когда данные наблюдений или эксперимента можно считать случайными и выбранными из генеральной совокупности, распределенной по многомерному нормальному закону.

Основная задача корреляционного анализа состоит в оценке параметров, определяющих нормальный закон распределения k-мерного вектора X, в частности корреляционной матрицы генеральной совокупности X, по выборке. Она позволяет определить расположение множества точек в пространстве k измерений, то есть оценить природу взаимозависимости между наблюдаемыми переменными, из чего, собственно, и можно делать вывод об отнесении их к тому или иному классу.

Дополнительная задача корреляционного анализа (являющаяся основной в регрессионном анализе) состоит в оценке уравнений регрессии, где в качестве результирующего выступает признак, являющийся следствием других признаков (факторов) – причин [4, 6].

Регрессионным анализом называется метод статистического анализа зависимости случайной величины y от переменных $x_j (j = 1, 2, ..., k)$, рассматриваемых в регрессионном анализе в качестве неслучайных величин, независимо от

истинного закона распределения x_j. Следовательно, с помощью регрессионного метода строится функция $f(x_1, x_2, ..., x_к)$, описывающая зависимость условного среднего значения результативного признака y от заданных значений аргументов и называемая функцией (уравнением) регрессии. Уравнения регрессии (или их система) могут, в частном случае, служить для восстановления разделяющей функции в задачах распознавания образов (классов).

Краткий анализ корреляционного и регрессионного методов позволяет утверждать, что они в принципе могут быть использованы для решения задач распознавания ситуации, однако наиболее эффективно их использовать можно для выявления наличия статистически значимых связей между переменными и оценкой степени их тесноты (корреляционный метод), а также для математического описания конкретного вида зависимостей случайной величины (y) от переменных $x_j (j = 1, 2, ..., k)$ (регрессионный метод). Иными словами, непосредственно задача различения в них не ставится, что и позволяет говорить о них как о частных методах распознавания. К основным методам, позволяющим распознать образы с использованием ЭВМ, в статистике относят дискриминантный (различающий), который был рассмотрен выше, и кластерный.

В статистических исследованиях группировка первичных данных является основным приемом решения задачи классификации, а значит, и основой для всей дальнейшей работы с собранной информацией.

Традиционно эта задача решается следующим образом. Из множества признаков, описывающих ситуацию, отбирается один, наиболее информативный с точки зрения исследователя, и производится группировка в соответствии со значениями данного признака. Если требуется провести классификацию по нескольким признакам, ранжированным между собой по степени важности, то вначале производится классификация по первому признаку, затем каждый из полученных классов разбивается на подклассы по второму признаку и т. д. Подобным образом строится большинство комбинационных статистических группировок.

В тех случаях, когда упорядочить классификационные признаки не представляется возможным, применяется наиболее простой метод многомерной группировки – создание интегрального показателя (индекса), функционально зависящего от исходных признаков, с последующей классификацией по этому показателю.

Развитием такого подхода является вариант классификации по нескольким обобщающим показателям (главным компонентам), полученным с помощью методов факторного анализа [2, 4].

При наличии нескольких признаков (исходных или обобщенных) задача классификации может быть решена методами кластерного анализа, которые от других методов многомерной классификации отличаются отсутствием обучающих выборок, то есть априорной информации о распределении генеральной совокупности, которая представляет собой вектор X.

Особенность распознавания без обучения (адаптации) заключается в том, что для его реализации требуется объем первоначальной априорной информации больший, чем при случае с обучением; а в последнем случае, в свою очередь, тре-

буется исходной априорной информации больше, чем для случая распознавания с самообучением. Указанный факт важен тем, что позволяет обоснованно выбирать необходимые системы распознавания в различных условиях с учетом имеющейся априорной информации, а также возможных затрат на получение дополнительных исходных данных.

На практике иногда сталкиваются с необходимостью распознавания, когда провести классификацию ситуаций либо невозможно, либо по тем или иным соображениям нецелесообразно. В данном случае число классов заранее не известно, поэтому информация о принадлежности каких-либо ситуаций к тем или другим классам отсутствует и единственный путь формирования системы распознавания – применение методов самообучения. К самообучению приходится прибегать и тогда, когда хотя заранее и известно число классов, однако обучающая выборка не задана, а имеется лишь некоторая совокупность ситуаций.

Последней (по порядку, но не по значимости) в группе статистических методов распознавания рассмотрим процедуру последовательных решений.

Ранее предполагалось, что решение о принадлежности распознаваемого объекта (НВ) W соответствующему классу Ω_i, $i = 1, ..., m$ принимается после измерения всей совокупности признаков этой ситуации $x_1, ..., x_m$. Однако возможен и другой подход к решению этой задачи: после измерения каждого очередного признака включается алгоритм распознавания и решается задача распознавания на основе данных об измеренных к определенному моменту времени признаках неизвестной ситуации. При этом, в зависимости от результатов сравнения полученного решения с некоторыми установленными заранее границами, либо измеряется очередной признак ситуации W, либо прекращается дальнейшее накопление информации об этой ситуации. Такая процедура решения задачи распознавания, называемая последовательной, обязана своим возникновением одному из разделов статистики – последовательному анализу [1, 3, 5].

Последовательное многократное решение задачи распознавания с использованием на каждом шаге все возрастающего числа измеренных признаков особенно целесообразно в случаях, когда:

– определение признаков сопряжено с затратами на проведение анализа;

– их процесс накопления недопустим из-за ограниченного количества времени, обусловленного особенностями предметной области;

– ситуации ряда классов из их общей совокупности надежно распознаются по ограниченному количеству признаков и т. п.

Наиболее удобной формой представления значений выделенных признаков (данных) в динамических системах являются временные ряды. Обоснование данного утверждения можно найти в ряде исследований [7, 8]. Заметим, что под динамическими системами будем понимать такие системы, в которых в реальном масштабе времени (или с темпом поступления информации на их вход) происходят изменения хранящихся в них данных.

Традиционным при анализе временных рядов является предположение о том, что статистические свойства наблюдаемого ряда или свойства порождающего его

механизма сохраняют определенное постоянство во времени или медленно изменяются. Вместе с тем многие практические задачи, например текущего контроля производства, технической диагностики, геофизики и т. п., сводятся к последовательному (то есть в темпе с поступлением очередного наблюдения) обнаружению скачкообразного изменения свойств наблюдаемого временного ряда, происходящего в неизвестный момент времени [6, 9, 10].

Анализ статистических методов, применяемых для распознавания ситуаций, позволяет сделать вывод о том, что наиболее перспективными из них для применения на практике являются дискриминантный (в случае с адаптацией) и кластерный (без адаптации), а при решении задачи распознавания, связанной с отнесением ситуации к одному из двух классов, – последовательный А. Вальда в виде алгоритма кумулятивных сумм.

Пятую группу методов распознавания назовем в силу их разнородности (по способам представления) «Другие методы» и отнесем к ним модели многослойных распознающих машин, методы нейросетей и метод группового учета аргумента (МГУА) [11–14].

Сравнительный анализ методов распознавания пятой группы позволяет сделать следующие выводы: во-первых, модель персептрона (многослойной распознающей машины), построенная на линейных функциях активации, не является универсальным распознающим устройством, так как не в состоянии решить целый ряд самых простых задач; во-вторых, главным достоинством нейросетей является то, что они позволяют использовать при распознавании некий универсальный нелинейный элемент, характеристики которого могут изменяться и настраиваться в широком диапазоне, однако их конкретные конфигурации в виде однослойной сети Хопфилда и многослойной сети со слоями «скрытых» нейронов обладают рядом недостатков, ограничивающих их практическое применение; в-третьих, методом, сохраняющим возможности получения нелинейных разделяющих функций, обеспечивающим «прозрачность» и точность вычислений на каждой итерации, является МГУА и его аналог – метод Ф-функций.

Также необходимо отметить, что методы обучения распознаванию традиционно классифицируются на параметрические и непараметрические. В их основе лежат те же математические методы, которые были рассмотрены выше.

Заключение

Таким образом, сравнительный анализ методов решения задачи распознавания ситуации позволяет сделать следующие выводы:

1. Наиболее распространенными на практике являются статистические методы, причем для решения задач анализа и синтеза системы защиты ИВС наилучшими свойствами обладают формальные грамматики теории распознавания образов, позволяющие, во-первых, наиболее полно учитывать структурные свойства и искажения, моделируемые ЦП, во-вторых, избежать неоднозначности их описания. В то же время эти методы относительно просты в реализации и эффективно применяются при обнаружении НВ нарушителя.
2. Для случая с априорным обучением наиболее эффективным считается дис-

криминантный метод распознавания, для случая без обучения – кластерный, а если решается задача обнаружения разладки (при классификации по двум образам) – последовательный метод А. Вальда в виде алгоритма кумулятивных сумм.

3. Альтернативой статистическим методам является метод группового учета аргумента, предназначенный для восстановления разделяющей функции по небольшому числу экспериментальных данных и основанный на принципе самоорганизации, позволяющем ему моделировать определенные функции должностных лиц, выполняемые им на этапе классификации.

4. В случае отсутствия сведений о каких-либо количественных распределениях классов (образов) в соответствующем пространстве признаков, а также предварительной сравнительной оценки эффективности поддержки выработки решений для распознавания ситуации при ручной и автоматической ее реализации в ИВС необходимо использовать логические методы.

СПИСОК ЛИТЕРАТУРЫ

1. Кузин Л.Т. Основы кибернетики // Основы кибернетических моделей: учеб. пособие для вузов. – М. : Энергия, 1979. – 584 с.

2. Математический энциклопедический словарь / Гл. ред. Ю.В. Прохоров. – М. : Сов. энциклопедия, 1988. – 847 с.

3. Горелик А.Л., Скрипник В.А. Методы распознавания: учеб. пособие для вузов. – М. : Высш. школа, 1977. – 222 с.

4. Дубров А.М., Мхитарян В.С., Трошин Л.И. Многомерные статистические методы: учебник. – М. : Финансы и статистика, 1998. – 352 с.

5. Мушик Э., Мюллер П. Методы принятия технических решений: пер. с нем. – М. : Мир, 1990. – 208 с.

6. Бухалев В.А. Распознавание, оценивание и управление в системах со случайной скачкообразной структурой. – М. : Наука. Физматлит, 1996. – 288 с.

7. Бокс Дж., Дженкинс Г. Анализ временных рядов. Прогноз и управление : пер. с англ. А.Л. Левшина. Выпуски 1 и 2. – М. : Мир, 1974. – 406 с. и 198 с. соответственно.

8. Гуляев А.И. Временные ряды в динамических базах данных. – М. : Радио и связь, 1989. – 128 с.

9. Никифоров И.В. Последовательное обнаружение изменения свойств временных рядов. – М. : Наука, 1983. – 200 с.

10. Жиглявский А.А., Красковский А.Е. Обнаружение разладки случайных процессов в задачах радиотехники. – Л. : Изд-во Ленинградского университета, 1988. – 224 с.

11. Дуда Р., Харт П. Распознавание образов и анализ сцен. – М. : Мир, 1976. – 512 с.

12. Ивахненко А.Г., Лапа В.Г. Предсказание случайных процессов. – Киев : Наукова думка, 1971. – 416 с.

13. Ивахненко А.Г. Персептроны. – Киев : Наукова думка, 1974. – 397 с.

14. Ивахненко А.Г. Персептрон – система распознавания образов. – Киев : Наукова думка, 1975. – 431 с.