УДК 681.51

С.В. СКАКУН

МАТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ ПОВЕДЕНИЯ ПОЛЬЗОВАТЕЛЕЙ КОМПЬЮТЕРНЫХ СИСТЕМ[1]

Abstract: *In article a mathematical modelling of users' behaviour in computer systems was carried out. The dynamic of user behaviour was investigated. Statistical modelling of data that characterize user's work during his session was done.*
Key words: *user behaviour modelling, neural networks, computer systems.*

Анотація: *У статті проводилось математичне моделювання поведінки користувачів комп'ютерних систем. Вивчалася динаміка роботи користувачів за сеанс. Також здійснювалося статистичне моделювання даних, що характеризують його роботу за сеанс у цілому.*
Ключові слова: *моделювання поведінки користувача, нейронні мережі, комп'ютерні системи.*

Аннотация: *В статье проводилось математическое моделирование поведения пользователей компьютерных систем. Изучалась динамика работы пользователя во время сеанса. Также осуществлялось статистическое моделирование данных, характеризующих его работу за сеанс в целом.*
Ключевые слова: *моделирование поведения пользователей, нейронные сети, компьютерные системы.*

1. Введение

Масштабное использование компьютерных технологий практически во всех сферах человеческой деятельности приковывает все большее внимание к самому пользователю. Знание того, какие действия он выполняет (или должен выполнять), может применяться в разных областях, например, в системах безопасности [1], для создания персонализированного окружения для пользователей [2, 3] и так далее. Поэтому задача построения моделей поведения пользователей компьютерных систем является актуальной.

В работе [4] была предложена комплексная модель пользователя, состоящая из интерактивной и сеансовой частей, которые учитывают, соответственно, динамические и статистические свойства поведения пользователя. В обеих моделях для выявления отклонений от обычного или ожидаемого поведения пользователей используются нейронные сети. Так, интерактивная модель основана на прогнозировании команд[2] пользователя на основе предыдущих. Поскольку выбор архитектуры нейронной сети представляет собой нетривиальную задачу, важно знать, насколько его текущее поведение зависит от предыстории. В случае сеансовой модели возникает проблема с размером выборки, которая используется для обучения нейронной сети. Дело в том, что при небольшом размере обучающего множества нейронная сеть имеет тенденцию к локальному запоминанию образов, что нежелательно. В сеансовой модели на вход нейронной сети подаются данные, собранные за сеанс в целом. Соответственно, размер обучающего множества напрямую определяется количеством сеансов, во время которых проводилось наблюдение за деятельностью пользователя. Однако даже за

[1] Работа выполнена при содействии гранта Президента Украины для поддержки научных исследований молодых ученых № Ф8/323, "Прототип интеллектуальной мультиагентной системы компьютерной безопасности".
[2] В данной работе под прогнозированием команд будем понимать прогнозирование процессов, порожденных запуском файлов ОС Windows.

продолжительный отрезок времени этих данных будет недостаточно для качественного обучения нейронной сети. Поэтому в сеансовой модели для качественного обучения нейронной сети очень важно обеспечить более представительную выборку данных.

Вопросам, связанным с оптимизацией архитектуры нейронной сети, в частности, с размерностью входного слоя и моделированием данных, и посвящена данная статья.

2. Комплексная нейросетевая модель пользователя компьютерных систем

Комплексная модель пользователя, предложенная в работе [4], учитывает как динамические (интерактивная часть), так и статистические (сеансовая часть) свойства поведения пользователей. В основу разработанной модели положена нейронная сеть прямого распространения, которая состоит из входного, выходного и одного или нескольких скрытых слоев нейронов [5]. Выход нейрона в слое n определяется следующим отношением:

$$y_j^n = f(s_j^n), \tag{1}$$

где n – номер слоя $(n = \overline{1, p})$; p – количество слоев в нейронной сети; j – индекс нейрона в слое $(j = \overline{1, N_n})$; N_n – число нейронов в слое; f – активационная функция слоя (в нашем случае для скрытых слоев используется сигмоидная активационная функция $f(x) = \frac{1}{1+e^{-\alpha x}}$, а для выходного слоя – линейная $f(x) = \alpha x$); y_j^n – выход j-го нейрона слоя; s_j^n – постсинаптический потенциал j-го нейрона слоя, который вычисляется согласно следующим формулам:

$$s_j^n = \sum_{k=1}^{N_{n-1}} W_{jk}^n y_k^{n-1} + b_j^n;$$

$$S^n = W^n \cdot \tilde{y}^{n-1}, \tag{2}$$

где W_{jk}^n – весовой коэффициент связи k-го нейрона слоя $n-1$ с j-м нейроном слоя n; y_k^{n-1} – выход k-го нейрона слоя $n-1$; $\tilde{y}^{n-1}$ – расширенный вектор с учетом bias-нейрона; b_j^n – порог (bias-нейрон) j-го нейрона слоя n. Вход и выход нейронной сети будут определяться, соответственно, следующими соотношениями:

$$X = (x_1, x_2, ..., x_{N_1}) \equiv (y_1^1, y_2^1, ..., y_{N_1}^1); \tag{3}$$

$$Y = (y_1^p, y_2^p, ..., y_{N_p}^p). \tag{4}$$

Интерактивная модель используется для выявления аномальной деятельности во время работы пользователя. Для каждого пользователя компьютерной системы строится и обучается

нейронная сеть таким образом, чтобы прогнозировать следующую команду на основе предыдущих. При этом результат работы нейронной сети в момент времени t определяется зависимостью

$$Y_t = F(X_t),\ X_t = (c_{t-1},\ \ldots,\ c_{t-m}), \tag{5}$$

где F – нелинейное преобразование, осуществляемое нейронной сетью согласно формулам (1)–(4); c_t – t-тая команда сеанса; m – количество команд, на основе которых происходит прогнозирование следующей (глубина прогнозирования).

На основе количества команд, которые были правильно спрогнозированы нейронной сетью, делается вывод, соответствует ли текущее поведение пользователя ранее построенной модели. При этом необходимо учитывать, что пользователям свойственно изменять поведение с течением времени. Поэтому с целью обеспечения адаптации к их поведению нейронную сеть следует периодически дообучать.

Сеансовая модель предназначена для выявления нехарактерной деятельности пользователя за сеанс в целом и для этого использует статистический набор данных. Данная информация, в свою очередь, используется для построения и обучения нейронной сети, которая определяет, насколько активность пользователя соответствует ранее построенной модели. Выход нейронной сети определяется следующим соотношением:

$$Y_i = F(X_i),\ X_i = (c_i, o_i, h_i, d_i, s_i), \tag{6}$$

где i – условный номер сеанса; F – нелинейное преобразование, осуществляемое нейронной сетью согласно формулам (1)–(4); c_i – количество команд за сеанс; o_i – результаты интерактивной модели (процентное соотношение правильно спрогнозированных команд); h_i – номер компьютера; d_i – продолжительность сеанса; s_i – время начала сеанса.

При этом ожидаемый выход нейронной сети может принимать два значения: 1 – для нормального поведения пользователя и 0 – для аномального, т.е. нейронная сеть работает в качестве классификатора.

3. Описание данных

Для моделирования поведения пользователей компьютерных систем использовались реальные данные, которые были собраны в локальной сети Института космических исследований НАНУ-НКАУ за два-три месяца. В данной сети рабочие станции функционируют под управлением операционных систем (ОС) Windows 98, XP, 2000. Поскольку эти ОС необходимой информацией об активности пользователя обеспечивают не в полной мере, было разработано специальное программное приложение. Для каждого сеанса пользователя создается отдельный аудит-файл (его название однозначно определяет имя учетной записи пользователя и дату его работы), в котором сохраняется информация в следующем формате:

время запуска команды|идентификатор команды|название команды|флаг начала или завершения.

Ниже приведен пример такого аудит-файла:

...
11:28:50|620|WINWORD.EXE|STARTED
11:30:58|2276|Far.exe|STARTED
11:39:17|730|WINWORD.EXE|STARTED
11:39:17|620|WINWORD.EXE|FINISHED
...

Необходимо отметить, что идентификатор команды присваивается ОС и является уникальным (для команд с одним и тем же именем он различен, причем от сеанса к сеансу он также меняется). Поэтому при кодировании команд важно обеспечить, чтобы одинаковым командам соответствовали одни и те же значения. С этой целью для интерактивной части комплексной модели на основе собранной информации для каждого пользователя был построен алфавит команд (т.е. набор команд, которые вводились пользователем на протяжении указанного периода времени). Далее каждой команде был присвоен соответствующий десятичный номер, который впоследствии использовался при преобразовании аудит-файлов в последовательности команд. В результате для каждого пользователя имелся следующий набор данных:

$$\left\{ c_j^i \right\}_{i=1}^{N} {}_{j=1}^{N_i}, \tag{5}$$

где c_j^i – десятичный номер введенной j-ой команды i-ого сеанса; N – количество сеансов; N_i – общее количество команд в i-ом сеансе. На рис. 1 приведен пример последовательности команд, вводимых пользователем за один сеанс.

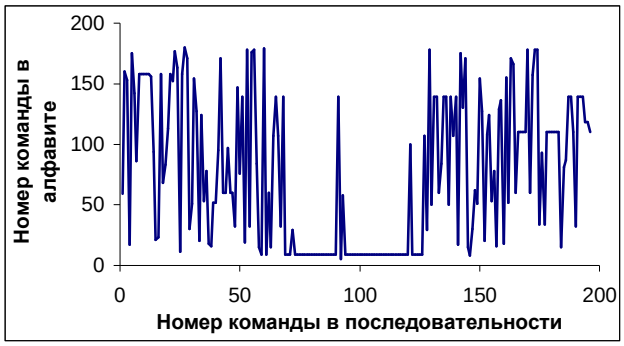


Рис. 1. Пример последовательности команд, вводимых пользователем за один сеанс

В свою очередь, при использовании информации из аудит-файлов для сеансовой части комплексной модели был получен следующий набор данных:

$$\{c_i, h_i, d_i, s_i\}_{i=1}^{N},$$

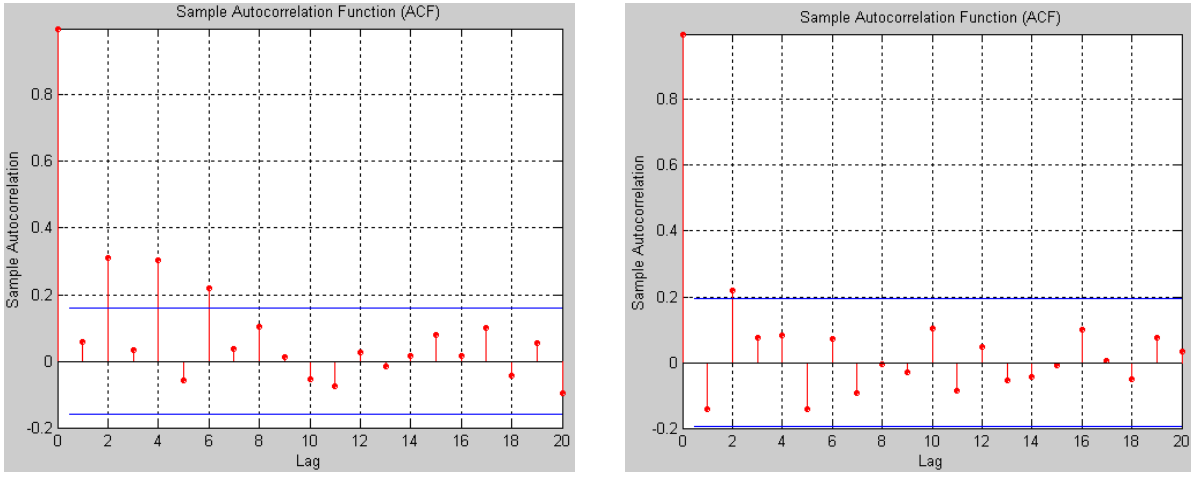
где i – условный номер сеанса; N – количество сеансов; параметры c_i, h_i, d_i, s_i определены в соотношении (4).

4. Изучение динамики поведения пользователя во время сеанса

Поскольку интерактивная модель основана на прогнозировании нейронной сетью команд пользователя, важно знать, на сколько его поведение в данный момент времени зависит от предыдущего. Для этого для каждого пользователя были построены автокорреляционные кривые, определяемые соотношениями следующего вида:

$$\rho^i(n) = corr(\vec{\xi}_n^i, \vec{\eta}_n^i), \quad \vec{\xi}_n^i = \left(c_1^i, \; c_2^i, \; \ldots, \; c_{N_i-n}^i\right);$$
$$\vec{\eta}_n^i = \left(c_{n+1}^i, \; c_{n+2}^i, \; \ldots, \; c_{N_i}^i\right).$$

где $\rho^i(n)$ – коэффициент корреляции для последовательности команд, введенных в i-ом сеансе; n – значение лага (временное смещение между элементами последовательности). На рис. 2 приведены примеры автокорреляционных кривых для разных пользователей.



Пользователь № 1 Пользователь № 2

Рис. 2. Примеры автокорреляционных функций для разных пользователей

Анализ построенных кривых показывает, что с ростом числа лагов автокорреляционные функции убывают. При этом экстремумы наблюдаются при значениях лагов от 1 до 8. Таким образом, при прогнозе команд пользователя следует использовать именно данное количество команд. При этом необходимо учитывать следующее: использование слишком большого количества команд приведет к тому, что на протяжении этого периода времени будет невозможно осуществлять прогноз команд, что снизит возможности по выявлению аномальной деятельности пользователей.

5. Моделирование данных

В общем случае функционирование нейронной сети значительно зависит от качества обучающей выборки. Дело в том, что при небольшом размере обучающего множества нейронная сеть имеет тенденцию к жесткому запоминанию образов, что приводит к уменьшению ее способности к обобщению. Так, при построении интерактивной модели пользователя в нашем случае проблема с представительной выборкой данных не возникала, поскольку даже за непродолжительный период времени работы пользователя может быть собрано достаточное количество образов (представительных) для обучения нейронной сети. (Например, с учетом того, что пользователь в среднем вводит от 80 до 150 команд за сеанс, то за десять сеансов обучающая выборка может насчитывать до 1000 образов.)

В случае сеансовой модели размер обучающего множества напрямую определяется количеством сеансов, во время которых проводилось наблюдение за работой пользователя. Так, за три месяца количество таковых сеансов может достигать 100, что в нашем случае было недостаточно для качественного обучения нейронной сети и оптимизации ее архитектуры. Для решения этой проблемы можно использовать два подхода. Первый из них состоит в значительном увеличении отрезка времени, в рамках которого происходит наблюдение за поведением пользователя (скажем, до 8-10 месяцев). Однако в данном случае велика вероятность того, что за этот промежуток времени оно изменится и, таким образом, обучающееся множество будет содержать противоречивые образы. Второй подход заключается в статистическом моделировании данных на основе имеющейся выборки. Он и будет использован в данной работе.

Поскольку в сеансовой модели для обучения нейронной сети используется информация о количестве вводимых команд за сеанс, номере компьютера, продолжительности и времени начала сеанса, для каждого пользователя проводилось моделирование именно этого набора данных. Для этого проверялось соответствие эмпирического распределения набору теоретических (нормальному, логарифмически нормальному, равномерному и т.д.). В качестве критерия согласия использовался критерий χ^2. Рассмотрим полученные зависимости более подробно.

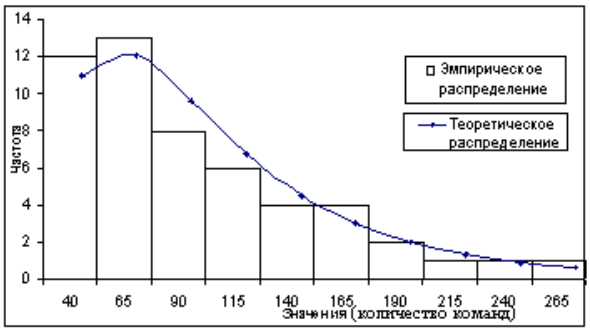


Рис. 3. Эмпирическое и теоретическое распределения для количества вводимых команд

Количество вводимых команд. Был проведен анализ различных распределений, но наилучшее значение χ^2 получено для логарифмического нормального распределения, что обеспечивало 97%-ое соответствие гипотезы реальным данным. На рис. 3 приведено эмпирическое и построенное теоретическое распределение для этого параметра.

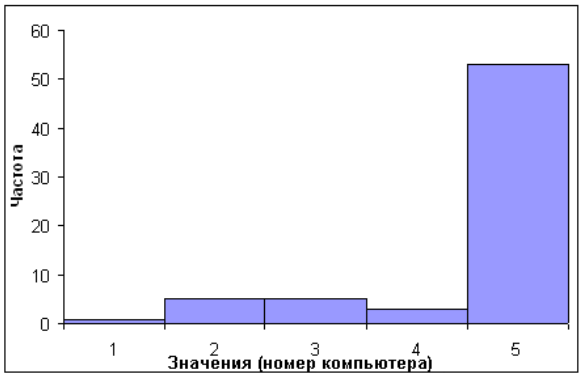


Рис. 4. Распределение по номерам компьютеров

Номер компьютера. При анализе значений этого параметра необходимо учитывать следующее: пользователь, как правило, имеет свое основное место за компьютером и очень редко работает за другими. Поэтому всегда существует значение, вероятность которого наибольшая и составляет 0,8…0,95 (рис. 4). События же, связанные с работой пользователя за другими рабочими станциями, можно считать равновероятными.

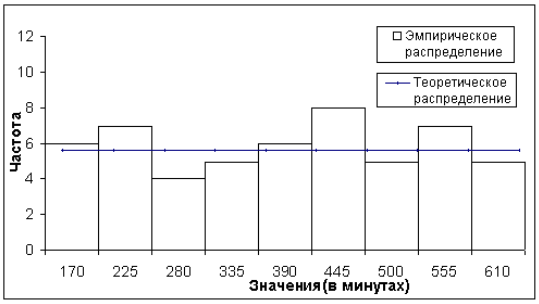


Рис. 5. Эмпирическое и теоретическое распределения для продолжительности сеанса

Продолжительность сеанса. Анализ значений этого параметра показывает, что они распределены равновероятно (рис. 5). Значение критерия χ^2 обеспечивает 95%-ое соответствие гипотезы реальным данным.

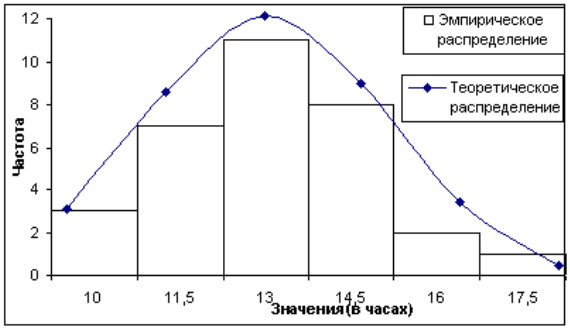


Рис. 6. Эмпирическое и теоретическое распределения для времени начала сеанса

Время начала сеанса. Наилучшее значение критерия χ^2 для этого параметра было получено для нормального распределения, что обеспечивало 90%-ое соответствие гипотезы реальным данным. На рис. 6 приведено эмпирическое и построенное теоретическое распределение.

На основе полученных теоретических зависимостей была разработана программа, моделирующая работу пользователя за сеанс. Сформированные с ее помощью данные позволили оптимизировать архитектуру нейронной сети и улучшить качество ее функционирования.

6. Заключение

В данной работе проводилось математическое моделирование поведения пользователей компьютерных систем. Для этого использовалась комплексная модель, предложенная в работе [4].

Изучалась динамика работы пользователя во время сеанса. Поскольку определение оптимальной архитектуры нейронной сети является нетривиальной задачей, в интерактивной модели важно знать, сколько команд следует использовать при обучении для прогноза следующей. На основе построенных автокорреляционных кривых было выявлено, что для этого следует учитывать до восьми команд.

Также было проведено статистическое моделирование данных, используемых для обучения нейронной сети в сеансовой модели. Эмпирические распределения были аппроксимированы теоретическими и на их основе сгенерирован необходимый набор данных, что дало возможность оптимизировать архитектуру нейронной сети и улучшить ее функционирование.

СПИСОК ЛИТЕРАТУРЫ

1. Куссуль Н., Соколов А. Адаптивное обнаружение аномалий в поведении пользователей компьютерных систем с помощью марковских цепей переменного порядка. Ч. 2: Методы обнаружения аномалий и результаты экспериментов // Проблемы управления и информатики. – 2003. – № 4. – С. 83 – 88.
2. Manavoglu E., Pavlov D., Lee Giles C. Probabilistic User Behavior Models // Proc. of the 3rd IEEE International Conf. on Data Mining (ICDM 2003). – Melbourne, Florida (USA). – 2003. – P. 203 – 210.
3. Davison B. D., Hirsh H. Probabilistic Online Action Prediction // Working Notes of the AAAI Spring Symposium on Intelligent Environments. – 1998. – P. 148 –154.
4. Скакун С.В., Куссуль Н.Н. Нейросетевая модель пользователей компьютерных систем // Кибернетика и вычислительная техника. – 2004. – Вып. 143. – С. 55 – 68.
5. Haykin S. Neural Networks: a comprehensive foundation. – Upper Saddle River, New Jersey: Prentice Hall, 1999. – 842 p.