УДК 004.956:519.237.8

Г.В. Карайчев

ВЫЯВЛЕНИЕ АНОМАЛЬНОЙ АКТИВНОСТИ МЕТОДОМ АДАПТИВНЫХ СЕТОК

В статье рассматривается метод выявления аномальной активности на основе адаптивного построения профиля системы в процессе самообучения высокой производительности. Первичные данные о соединениях преобразуются методом главных компонент и кластеризуются методом адаптивных. Результаты экспериментов (данные KDD CUP'99) показывают, что предлагаемый подход по эффективности не уступает другим методам аномального анализа.

Сетевая безопасность; аномальный анализ; метод главных компонент; кластеризация; адаптивные сетки; анализ ROC кривых.

G.V. Karaychev

ANOMALY DETECTION BASED ON ADAPTIVE GRID-BASED ALGORITHM

The paper provides information on high productivity unsupervised anomaly detection based on adaptive construction of system profile. Initial connection records are transformed using principal component analysis and clustered by adaptive grid-based algorithm. Evaluation (KDD CUP'99 data set) demonstrates that effectiveness of suggested approach is comparable with other anomaly analysis methods.

Network security; anomaly analysis; principal component analysis; clusterization; adaptive grid-based algorithm; ROC analysis.

Как известно, существует два основных метода обнаружения сетевых атак: определение злоупотреблений (misuse detection) и выявление аномальной активности (anomaly detection) [1] [2]. Первый метод заключается в том, что сетевые пакеты сравниваются с шаблонами, содержащими признаки атаки. Соответствие образца известной атаке называется сигнатурой. Недостаток метода выявления злоупотреблений выражается в невозможности определять новые типы атак, которые не содержатся в базе данных. Отрицательной чертой метода является также то, что, возможно, использовать только строго определенные сигнатуры, а это не допускает определения вариантов общих атак.

Указанные недостатки могут быть устранены при совместном использовании сигнатурных методов с подходом, основанным на выявлении аномальной активности. Преимущество последнего заключается в том, что он способен обнаруживать отклонения в поведении вычислительной системы и, таким образом, может выявить симптомы атак без знания их конкретных деталей. Также данный метод способен генерировать информацию, которая в дальнейшем может быть использована при сигнатурном анализе.

Существующие технологии выявления аномальной активности подразделяются на два типа: на основе управляемого построения базового профиля (*supervised*) и на основе самостоятельного построения профиля (*unsupervised*). При первом подходе эталонное состояние системы строится на основе свободного от атак трафика. Его недостаток выражается в том, что требуется дополнительная кропотливая работа по фильтрации и очистке исходного набора данных от аномалий.

Второй подход не требует очистки эталонных данных, используемых для обучения системы. Однако для применимости метода выявления аномалий на основе допустимого построения профиля требуется выполнение следующих двух условий: 1) базовый профиль основывается на наборе данных, в котором объем нормального трафика значительно превышает величину аномального; 2) набор данных, содержащий атаку, статистически отличается от нормального набора [3].

В предлагаемой работе для выявления аномальной активности предлагается использование подхода на основе самостоятельного построения базового профиля с использованием метода кластеризации данных.

В настоящее время существует множество методов кластеризации, в данной работе предлагается использовать решеточные методы (*grid-based*), разделяющие пространство объектов на конечное число клеток, формируя решеточную структуру.

Среди существующих алгоритмов кластерного анализа, сеточные методы имеют наибольшую скорость работы. Однако стоит учитывать, что их производительность напрямую зависит от размерности сетки. Тем не менее, в настоящей работе для эффективной работы предлагаемого метода, необходимости использовать сетку большой размерности нет. Преимущество сеточных методов также заключается в том, что без предварительного анализа самих данных, а только по их распределению в некоторой области, можно сгенерировать высоко достоверную модель состояния системы и выработать рекомендации по ее анализу.

Пусть $X_1,...,X_N$ – некоторая последовательность случайных событий с соответствующими числовыми характеристиками. В общем случае, пространство характеристик является многомерным, его размерность зависит от особенностей исследуемого процесса и определяется спецификой решаемой задачи. Применив метод кластеризации к множеству событий в системе, разобьем его на подмножества так, чтобы каждый кластер состоял из похожих объектов, а элементы различных кластеров имели существенные различия (внутренняя однородность и внешняя изолированность). В результате нормальные и аномальные события станут, различимы, то есть будут попадать в разные кластеры.

В данной работе мы будем использовать двумерное пространство характеристик. Для случая многомерного пространства справедливы аналогичные рассуждения. Введем в выбранном пространстве сетку размера $n_1 \times n_2$: $A = \{a_{ij}\}$, a_{ij} – количество событий в клетке (i, j). Пусть $K = \{K_{LC}, K_{RC}\}$, где K_{LC} и K_{RC} – координаты левого нижнего и правого верхнего углов сетки, соответственно: $K_{LC} = (K_{LCx}, K_{LCy})$, $K_{RC} = (K_{RCx}, K_{RCy})$, а вектор $B = \{B_1, B_2\}$, где $B_1 = (b_{1,1},...,b_{1,(n_1+1)})$, $B_2 = (b_{2,1},...,b_{2,(n_2+1)})$ содержит отношение ширины каждой границы к общей ширине сетки (рис. 1). Таким образом (A, K, B) полностью определяет нашу сетку.

При построении адаптивной сетки начнем с построения сетки с равномерным шагом $b_{ij} = \dfrac{j-1}{n_i}$, $j = \overline{1, n_i + 1}$, $i = 1, 2$.

Пусть $X = (x_1, x_2)$ – очередное событие. Рассмотрим действие функции, отображающей множество координат точек из пространства событий во множество номеров клеток сетки. Введем понятия высоты (H) и ширины (W) сетки, вычисляемые по формулам:

$$W = \left(K_{RCx} - K_{LCx}\right),$$
$$H = \left(K_{RCy} - K_{LCy}\right).$$

Тогда номер ячейки (*s*, *t*), в которой окажется новое событие *X*, будет удовлетворять следующим неравенствам:

$$\begin{aligned} W \cdot b_{1,s} < x_1 - K_{LCx} \le W \cdot b_{1,s+1}, \quad s = \overline{1,n_1}, \\ H \cdot b_{2,t} < x_2 - K_{LCy} \le H \cdot b_{2,t+1}, \quad t = \overline{1,n_2}. \end{aligned} \qquad (1)$$

Число событий в ячейке a_{st} увеличивается на единицу.

В том случае, если событие *X* попадает за пределы области сетки, изменяются границы сетки.

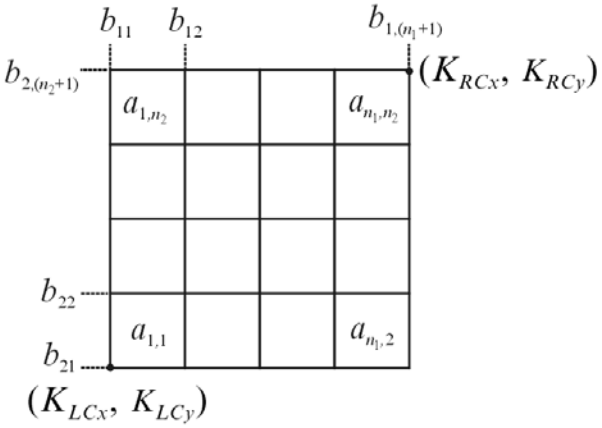

Рис. 1. Сетка (A, K, B)

Будем изменять размеры ячеек сетки таким образом, чтобы количество событий в разных клетках было примерно одинаковым. Перестраивая сетку, мы предполагаем, что события внутри каждой ячейки распределены с равномерной плотностью. Подобное предположение позволяет существенно ускорить процесс адаптации, повышая производительность всего подхода в целом.

Для реализации процесса адаптации сетки введем два множества $R_1 = \left(r_{11},...,r_{1n_1}\right)$, $R_2 = \left(r_{21},...,r_{2n_2}\right)$, где r_{ij} – сумма событий в *j*-м ряду клеток *i*-й размерности:

$$r_{1j} = \sum_{k=1}^{n_1} a_{kj}, \quad r_{2j} = \sum_{k=1}^{n_2} a_{jk}.$$

Тогда среднее число событий в ряду клеток одной размерности:

$$\bar{r}_i = \frac{1}{n_i}\sum_{j=1}^{n_i} r_{ij}, \quad i = 1,2.$$

Введем функцию плотности распределения событий в ряду по одной размерности:

$$\rho_{ij} = \frac{r_{ij}}{b_{i,j+1} - b_{i,j}}, \quad j = \overline{1,n_i}, \quad i = 1,2. \tag{2}$$

Проведя адаптацию по всем размерностям, получим новую сетку, в которой количество событий в каждой клетке примерно одинаково.

На рис. 2 представлен пример построения адаптивной сетки, совмещенной с пространством случайных событий. По оси *X* и *Y* рассматриваются главные компоненты, соответствующие середине интервала собственных значений. Более подробно о методе главных компонент будет сказано ниже. Темным цветом отмечены клетки с высокой плотностью событий, светлым — с низкой. Кружками отмечены отдельные события.

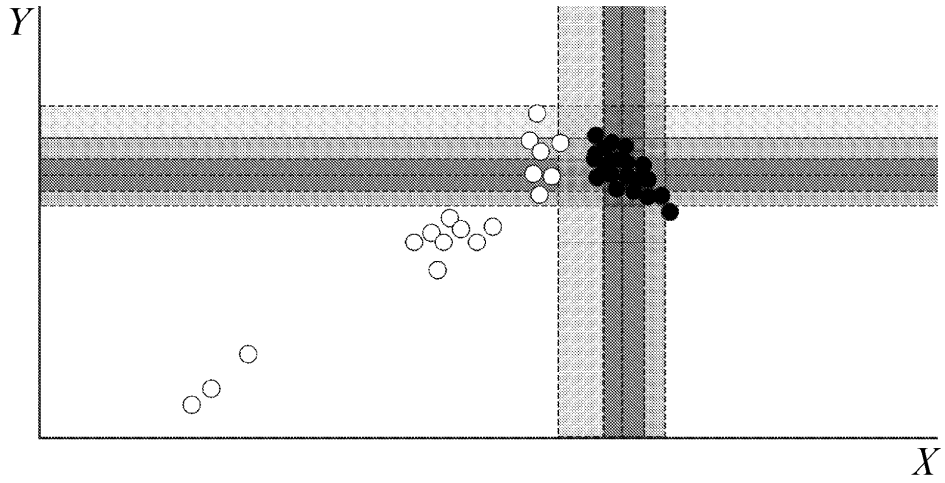

Рис. 2. Результат построения адаптивной сетки: черные кружки – нормальные события, белые кружки – атаки

Размерность пространства характеристик может быть много больше двух [8] и предлагаемый метод легко может быть обобщен на многомерный случай.

Обратимся к одному из основных способов уменьшения размерности данных, теряющему наименьшее количество значимой информации — методу главных компонент (преобразованию Кархунена-Лоэва). Данный метод успешно применяется в различных областях науки и техники. Вычисление главных компонент сводится к вычислению собственных векторов и собственных значений ковариационной матрицы исходных данных.

В задаче анализа главных компонент можно выделить несколько основных направлений: аппроксимация данных линейными многообразиями меньшей размерности; поиск подпространства меньшей размерности, в ортогональной проекции на которые разброс данных максимален; поиск подпространства меньшей

размерности, в ортогональной проекции на которые среднеквадратичное расстояние между точками максимально и др.

Задача метода поиска ортогональных проекций с наибольшим рассеянием сводится к поиску такого ортогонального преобразования в новую систему координат, для которого были бы верны следующие условия: 1) выборочная дисперсия данных вдоль первой координаты максимальна (эту координату называют первой главной компонентой); 2) выборочная дисперсия данных вдоль второй координаты максимальна при условии ортогональности первой координате (вторая главная компонента); 3) выборочная дисперсия данных вдоль значений k-й координаты максимальна при условии ортогональности первым k–1 координатам.

Следуя методу поиска ортогональных проекций с наибольшим рассеянием [4] [5], рассмотрим корреляционную матрицу M размера h, вычисленную по N случайным событиям $X = (X_1,...,X_N)$, где $X_k = (x_{k1},...,x_{kh})$:

$$M = \left\{ \sum_{k=1}^{N} v_{ki} \cdot v_{kj} \right\}, \quad v_{ki} = \frac{x_{ki} - \overline{x_i}}{\sigma_i},$$

где $\overline{x}_i$ – математическое ожидание i-й характеристики, σ_i – ее дисперсия.

Для пар собственных значений и собственных векторов матрицы M $(\lambda_1, e_1), (\lambda_2, e_2), ..., (\lambda_h, e_h)$ выполняется свойство: $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq ... \geq \lambda_h \geq 0$. А i-я выборочная главная компонента для k-го события будет иметь следующий вид:

$$w_i = e_{i1} v_{k1} + e_{i2} v_{k2} + ... + e_{ih} v_{kh},$$

где e_{ij} – собственный вектор.

Заметим, что w_i имеет выборочную дисперсию λ_i и выполняется свойство $\lambda_1 + \lambda_2 + \lambda_3 + ... + \lambda_h = h$. Это означает, что все случайные величины исходной выборки учтены в главных компонентах [6] [7].

Главные компоненты, соответствующие большим собственным значениям, в основном, связаны с регулярными зависимостями характеристик друг от друга; соответствующие малым собственным значениям – связаны со случайными несущественными отклонениями. Поэтому наиболее эффективно использование компонент, соответствующих середине интервала собственных значений.

Из приведенного рис. 2 видно, что атаки попадают в области с низкой плотностью, а большинство нормальных событий – в области с высокой плотностью. Поэтому для разделения нормальных и аномальных событий введем пороговое значение ρ_0 плотности клетки в пространстве характеристик. Тогда ячейки, величина плотности (2) которых выше ρ_0 ($\rho_{st} \geq \rho_0$), считаются нормальными, остальные ($\rho_{st} < \rho_0$) признаются аномальными. И если очередное событие по (1) помещается в такую ячейку, оно классифицируется как потенциальное вторжение.

Для определения эффективности предлагаемого метода обнаружения вторжений используется набор данных 1999 KDD CUP, который был впервые представлен на V Международной конференции «Knowledge Discovery and Data Mining» (KDD'99) и использовался в конкурсе по созданию обучаемых систем обнаружения вторжения [8]. Так Lincoln Labs создала инфраструктуру, условно имитирую-

щую ЛВС ВВС США и в течение 9 недель записывала необработанные данные. Они пользовались сетью, как если бы это и вправду была инфраструктурная сеть ВВС США, подверженная большому количеству атак.

Полученные данные, объемом более 4Гб, были преобразованы примерно в пять миллионов записей о соединениях. Каждое соединение отмечалось либо как нормальное, либо как содержащее атаку конкретного типа. Для каждого соединения вычислялось порядка 30 различных характеристик трех основных типов: основные характеристики; характеристики соединений, рассчитанные по информации о доменах; характеристик, рассчитанные с использованием двухсекундного окна. В итоговом наборе данных содержится 24 вида атак четырех категорий: DOS (отказ в обслуживании, например, syn flood); R2l (несанкционированный доступ с удаленного компьютера, например, подбор пароля); U2R (несанкционированный доступ к правам локального администратора, например, различные атаки переполнения буфера); зондирующего (наблюдение и другие виды сбора информации, например, сканирования портов).

Для оценки эффективности предлагаемого подхода обнаружения аномалий используем метод ROC характеристик (Receiver operating characteristic) [9]. Основой этого метода оценок является анализ графика ROC кривой, при построении которого используется относительная частота истинного (TPR) и ложного (FPR) срабатываний системы. Формулы для вычисления кривой ROC приведены ниже:

$$TPR = \frac{TP}{TP + FN}, \qquad FPR = \frac{FP}{FP + TN}, \tag{3}$$

где TP – число выявленных аномалий, FN – число пропущенных аномалий, FP – число ложных срабатываний, TN – число верно идентифицированных событий как нормальные.

Графики поведения ROC кривых для различных типов атак и методов обнаружения вторжения имеют различный вид. Пример построения таких графиков приведен на рис. 3. Чем площадь фигуры под ROC кривой больше, тем соответствующий ей алгоритм эффективнее.

При построении ROC кривой в качестве варьируемого параметра используется значение, определяющее пороговую плотность событий в клетке:

$$\rho_0 = k \cdot \max\left(a_{ij}\right),$$

где $\max(a_{ij})$ – показатель плотности ячейки с максимальной величиной, k – коэффициент плотности распределения событий в клетке.

Исследование предлагаемого в работе метода проводилось для заранее подготовленного набора данных, содержащего аномалии произвольных типов, затрагивающие протокол передачи данных http. Для построения профиля системы по этому набору данных было использовано 10 характеристик соединений: 1) число байт, переданных с адреса источника на адрес назначения; 2) число байт, переданных с адреса назначения на адрес источника; 3) продолжительность связи (в секундах); 4) число «сбойных» фрагментов; 5) количество срочных пакетов; 6) коли-

чество соединений с тем же хостом, что и в данном соединении, в течение последних двух секунд; 7) % соединений с одной и той же сетевой службой (тот же хост); 8) % подключений к разным сетевым службам (тот же хост); 9) количество соединений с той же сетевой службой, что и в данном соединении, в течение последних двух секунд; 10) % подключений к различным хостам (та же служба). Всего было проанализировано 494 000 записей о соединениях, содержащих более 10% атак. Сетка была выбрана размерности 11×11 клеток. Частота адаптации – каждые 5000 соединений.

Результаты вычисления по формуле (3), в зависимости от коэффициента плотности распределения событий в клетке, представлены в табл. 1.

Таблица 1

Результаты вычисления (3) для набора данных KDD CUP'99

k	0,002	0,01	0,03	0,1
TPR	63,1840%	77,6119%	94,5273%	97,5124%
FPR	0,8899%	2,0975%	7,5117%	16,8097%

Теперь выясним эффективность (по аналогии с площадью под ROC кривой) исследуемого метода в выбранных узловых точках (в нашем случае при фиксированном значении коэффициента плотности распределения событий в клетке). Для чего проведем вычисления по формуле

$$E = \frac{TPR}{FPR}. \tag{4}$$

Результаты вычислений (4) для набора данных KDD CUP'99), в зависимости от коэффициента плотности распределения событий в клетке, представлены в табл. 2. Как видно, эффективность предлагаемого метода снижается с ростом показателя *TPR*, таким образом, становится очевидным необходимость взвешенного подхода к выбору коэффициента плотности распределения событий в клетке.

Таблица 2

Результаты вычисления (4) для набора данных KDD CUP'99

k	0,002	0,01	0,03	0,1
E	71	37	12,58	5,8

Сопоставим результативность выявления аномальной активности методом адаптивных сеток с другими методами обнаружения вторжений. Результаты сравнения представлены в табл. 3. На рис. 3 изображены кривые, иллюстрирующие эффективность обнаружения атак предлагаемым методом в сравнении с методами, разработанными победителями KDD CUP'99 [10] [11] [12].

В заключение следует отметить, что предлагаемый метод адаптивных сеток устойчиво работает даже при достаточно зашумленном аномалиями исходном обучающем трафике, давая высокий показатель обнаружения атак уже при небольшой величине ложных срабатываний.

Таблица 3

Сравнение результативности выявления аномалий различными методами обнаружения вторжений для набора данных KDD CUP'99

TPR FPR	Адаптивная сетка	Canberra	NN	KDD	LOF
1%	64,51%	4,12%	58,25%	0,6%	0,03%
4%	83,56%	6,13%	81,3%	73,74%	98,7%
8%	94,67%	26,2%	92,78%	87,12%	99,04%
10%	95,32%	28,11%	93,96%	88,99%	99,13%

К предлагаемому в работе подходу применим метод весовых функций [13], позволяющий существенно увеличить его производительность в тех случаях, когда анализируемый интервал времени является большим. Изучение данного вопроса является предметом дальнейших исследований по теме.

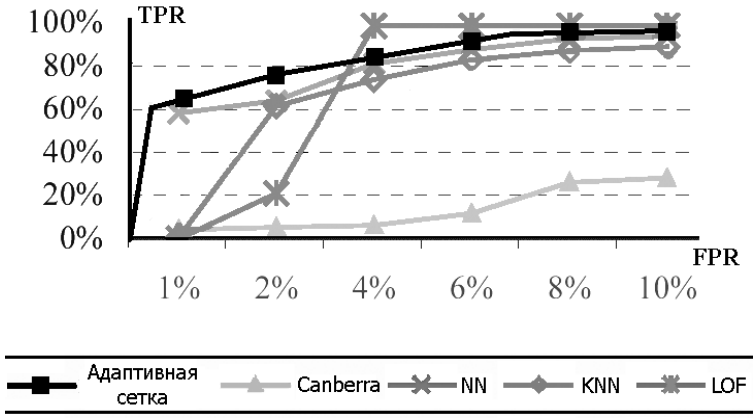

Рис. 3. График ROC кривых для различных методов выявления вторжений

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. *Denning D.E.* An intrusion detection model. IEEE Transactions on Software Engineering, SE-13, 1987. – P. 222–232.

2. *Javitz H.S., Valdes A.* The NIDES statistical component description and justification. Technical report, Computer Science Laboratory, SRI International, Menlo Park, California, March 1994.

3. *Leung K., Leckie C.* Unsupervised Anomaly Detection in Network Intrusion Detection Using Clusters. In Proceedings of Twenty-Eighth Australasian Computer Science Conference (ACSC2005), Newcastle, Australia, 1–3 February 2005. – P. 333–342.

4. *Shyu M.-L., Chen S.-C., Sarinnapakorn K., Chang L.* A novel anomaly detection scheme based on principal component classifier. // Proceedings of the IEEE Foundations and New Directions of Data Mining Workshop, Melbourne, FL, USA, 2003. – P. 172–179.

5. *Wang W., Battiti R.* Identifying Intrusions in Computer Networks with Principal Component Analysis. // Proceedings of the First International Conference on Availability, Reliability and Security (ARES'06). – P. 270–279, April 20–22, 2006.

6. *Liao W.-k., Liu Y., Choudhary A.* A Grid-based Clustering Algorithm using Adaptive Mesh Refinement. // 7th Workshop on Mining Scientific and Engineering Datasets in conjunction with SIAM International Conference on Data Mining (SDM), pp. 61–69, April 2004, Lake Buena Vista, Florida, USA.

7. *Kwitt R., Hofmann U.* Robust Methods for Unsupervised PCA-based Anomaly Detection. IEEE/IST Workshop on Monitoring, Attack Detection and Mitigation, Tuebingen, Germany, September 28–29, 2006.

8. *Lincoln labs*. KDDCup'99. http://kdd.ics.uci.edu/databases/kddcup99/kdd-cup99.html, 2003.

9. *Gu G., Fogla P., Dagon D., Lee W., Skoric B.* Measuring Intrusion Detection Capability: An Information-Theoretic Approach. ASIACCS'06, March 21–24, 2006 Taipei, Taiwan.

10. *Levin I.* KDD-99 Classifier Learning Contest: LLSoft's Results Overview. ACM SIGKDD Explorations 2000, pp. 67–75, January 2000.

11. *Pfahringer B.* Winning the KDD99 Classification Cup: Bagged Boosting. ACM SIGKDD Explorations 2000, pp. 65–66, January 2000.

12. *Miheev V., Vopilov A., Shabalin I.* The MP13 Approach to the KDD'99 Classifier Learning Contes». ACM SIGKDD Explorations 2000. – P. 76–77, January 2000.

13. *Карайчев Г.В., Нестеренко В.А.* Применение весовых функций для определения локальных статистических характеристик потока пакетов в сети // Известия высших учебных заведений. Северо-Кавказский регион. Естественные науки. – Ростов н/Д, 2008. № 1. – С. 10–14.

Карайчев Глеб Викторович
Южный федеральный университет.
E-mail: kgv_rostov@mail.ru.
344091, г. Ростов-на-Дону, пр. Стачки, 235/1, кв. 47.
Тел.: +7 (928) 1252607.
Факультет математики, механики и компьютерных наук; кафедра информатики и вычислительного эксперимента; ассистент.

Karaychev Gleb Viktorovich
South Federal University.
E-mail: kgv_rostov@mail.ru.
App. 47, 235/1, prosp. Stachki, Rostov-on-Don, 344091, Russia.
Phone: +7 (928) 1252607.
Faculty of mathematics, mechanics and computer science; Department of informatics and computing experiment; junior member of teaching staff.

УДК 510.6:656.001

Е.А. Пакулова

МОДЕЛЬ СОВРЕМЕННОЙ СИСТЕМЫ МОНИТОРИНГА ПОДВИЖНЫХ ОБЪЕКТОВ С ГАРАНТИРОВАННОЙ ДОСТАВКОЙ СООБЩЕНИЙ В ГЕТЕРОГЕННОЙ БЕСПРОВОДНОЙ СЕТИ

Основной целью данной статьи являлось построение модели системы мониторинга транспортных средств (ТС) с использованием нескольких беспроводных технологий связи. В связи с этим были решены следующие задачи: выбраны методы исследования, основанные на теории множеств и теории конечных автоматов, построена модель системы мониторинга ТС, определены события и команды в модели, а также связи между компонентами модели. В заключение статьи выделены дальнейшие планы работы, направленные на разработку и реализацию методов рационального управления технологиями беспроводной связи.

Моделирование; теория множеств; беспроводные технологии связи; система мониторинга транспортных средств.