УДК 004.021

В. А. Громов, А. Ю. Миняйло
Днепропетровский национальный университет им. Олеся Гончара

ПРОГНОЗИРОВАНИЕ ХАОТИЧЕСКИХ ВРЕМЕННЫХ РЯДОВ КАК ЗАДАЧА КЛАСТЕРИЗАЦИИ

Розглянуто новий метод прогнозування хаотичних часових рядів. Метод використовує алгоритм кластерізації, що спирається на застосування спеці-альним чином адаптованого генетичного алгоритму, для аналізу топологічної структури атрактора динамічної системи, що визначає поведінку часового ряду, що розглядається, та виділення характерних послідовностей, які відповідають різним частинам цього атрактора. Зазначені характерні послідовності використовуються для прогнозу значень часового ряду. Метод було застосовано для прогнозу часового ряду, побудованого за системою Лоренца, та температурного ряду для м. Львів.

Рассматривается новый метод прогнозирования хаотических временных рядов. Метод использует алгоритм кластеризации, основанный на применении специальным образом адаптированного генетического алгоритма, для анализа топологической структуры аттрактора динамической системы, определяющей поведение рассматриваемого временного ряда и выделения характерных последовательностей, отвечающим различным частям данного аттрактора. Указанные характерные последовательности используются для прогноза значений временного ряда. Метод был применён для прогноза временного ряда, построенного по системе Лоренца, и температурного ряда для г. Львов.

In this study, the novel method to predict chaotic time series is proposed. The method employs clusterization technique based on specific genetic algorithm to analyze topological structure of the attractor behind the given time series and to single out the typical sequences corresponding to the different part of the attractor. The typical sequences are used to predict the time series values. The method was applied to time series generated by the Lorenz system, and a weather time series for Lvov-city.

Ключевые слова: прогнозирование, хаотические временные ряды, кластеризация, генетический алгоритм.

Введение. Постоянный интерес к хаотическим моделям и системам со стороны исследователей самых различных научных направлений и специализаций [1; 2; 4–6] обуславливается как фундаментальной ролью, которую играют такого рода системы в описании природных и социальных процессов, так и сложностью их анализа и прогноза. Отдельной значимой и сложной задачей здесь является прогнозирование хаотических временных рядов.

В дальнейшем анализе будем предполагать, что переходные процессы в наблюдаемой динамической системе завершены, и система движется в окрестности некоторого странного аттрактора, что собственно и определяет хаотичность динамики системы. Также будем считать, что наблюда-емый временной ряд удовлетворяет условиям теоремы Такенса, т. е. ряд отражает внутреннюю динамику системы, определяемую структурой странного аттрактора.

Оценкой сверху для промежутка, на который может быть осуществлён прогноз, в задачах прогноза хаотических рядов является величина горизонта прогнозирования [4], т. е. промежутка времени необходимого для того, чтобы погрешность данных, малая в начальный момент времени, превысила некий порог (определяемый исследователем) вследствие экспоненциального расхождения близких траекторий. Эта же причина – экспоненциальное накопление погрешности – обуславливает невозможность применение классических прогнозных моделей (ARIMA-модель, DHR-модель и др.).

Другой причиной, которая обуславливает невозможность применения указанных моделей к прогнозированию хаотических рядов, как нам представляется, является то, что большинство существующих подходов к прогнозированию предполагают построение одной модели для всего наблюдаемого ряда: тем самым игнорируется внутренняя структура аттрактора. Те, модели, в которых указанная структура явным или неявным образом учитывается, оказываются успешными при прогнозировании. В качестве примера здесь можно привести нейросетевые и нейронечёткие модели [10; 22] – обзор существующих моделей, применяемых для прогнозирования хаотических временных рядов, представлен в следующем разделе.

Движение системы в окрестности странного аттрактора приводит к тому, что в исследуемом временном ряде наблюдаются похожие участки, которые отвечают движению системы в окрестности одной и той же облас-ти странного аттрактора. Выделение данных участков, характеризация указанных областей и построение элементарных прогнозных моделей для выделенных областей приводит к возможности прогнозирования хаотического временного ряда на промежутки времени, сравнимые с горизонтом прогнозирования [3; 11].

При этом каждая из построенных таким образом элементарных моделей представляет собой усреднённую характеристику соответствующей области, и, таким образом, качество прогноза, полученного с помощью данной модели, определяется некоторым компромиссом между потерей информации вследствие усреднения и отсутствием экспоненциального «хаотического» накопления погрешности вследствие того же

усреднения. Указанная особенность позволяет осуществлять прогнозирование по прог-нозным значениям, чем и объясняется и эффективность прогнозирования с использованием данного подхода на промежутки времени, сравнимые с горизонтом прогнозирования.

Отметим, что в [3; 11] вышеописанная методология была реализована для прогноза хаотических рядов с использованием метода «муравьиных колоний». В настоящей работе для решения вышеописанной задачи кластеризации использовался специальным образом адаптированный генетический алгоритм [14].

Обзор литературы. В последнее время было предложено значительное число моделей и методов, посвящённых восстановлению структуры странного аттрактора по временному ряду и прогнозированию данного ряда с использованием полученной при восстановлении информации. Их можно условно разделить на три группы в соответствие с теми теориями искусственного интеллекта, на которые опирается тот или иной метод при реконструкции фазового пространства [20] и анализе соответствующих временных рядов.

К первой группе можно отнести нейросетевые модели, которые являясь по природе своей универсальными адаптивными аппроксиматорами, способны выделять различные локальные тренды, присутствующие в анализируемом временном ряде, и аппроксимировать их [9; 15]. Здесь следует также упомянуть метод сингулярного спектрального разложения, использующий информацию о сингулярных значениях дисперсионно-ковариационной матрицы временного ряда для извлечения информации об указанных локальных трендах [7].

Ко второй группе следует отнести нечёткие и нейронечёткие подходы, которые используются для создания робастных и логически прозрачных прогнозных моделей [8; 10].

Наконец, третья группа включает в себя системы, основанные на подходах распределённого искусственного интеллекта, как, например, генетические алгоритмы [15], интеллект роя, метод «муравьиных колоний» [3; 11; 16; 18] и другие. Данные подходы могут использоваться как для настройки параметров нейросетевых моделей [17], так и собственно для прогнозирования. Отметим также ряд работ, в которых указанные подходы применяются для прогнозирования реальных природных или технологических процессов [12; 13; 19].

Постановка задачи. Рассматривается последовательность наблюдений хаотического временного ряда $y_t, y_{t-1}, ..., y_{t-s}$. Предполагается, что выполнены оба сформулированных выше предположения: о завершённости переходных процессов в динамической системе и о выполнении условий теоремы Такенса [4]. Необходимо построить прогнозные значения для последующих наблюдений данного временного ряда $\hat{y}_{t+1}, \hat{y}_{t+2}, ..., \hat{y}_{t+K}$, с тем, чтобы наблюдаемые значения не уклонялись от прогнозных больше чем на заданную величину:

$$\left|\hat{y}_{t+i} - y_{t+i}\right| < \varepsilon,\ i = \overline{1, K}\ .$$

Метод решения. Предложенный алгоритм прогнозирования состоит из двух частей: первая – анализ временного ряда с целью кластеризации последовательностей наблюдений временного ряда и выделения характерных последовательностей, вторая – собственно прогнозирования динамики временного ряда на основе выделенных последовательностей.

Обратимся вначале к краткому изложению генетического алгоритма, адаптированного к решению задач кластеризации. В изложении мы будем следовать [14]. Отметим, прежде всего, что указанный алгоритм самостоятельно определяет оптимальное число кластеров.

Рассматривается множество из N векторов, подлежащих кластеризации. В этом случае каждая хромосома, используемого генетического алгоритма, будет представлять собой $N+1$-мерный целочисленный вектор. Последний, $N+1$-й, компонент вектора кодирует число кластеров, соответствующих данному решению. Целое число, содержащееся в произвольной, i-й $i \le N$, компоненте хромосомы отвечает номеру кластера, которому в этом решении принадлежит i-й вектор из множества кластеризируемых данных.

Функция приспособленности (фитнесса) хромосомы в рамках данного алгоритма базируется на понятии силуэта:

$$s(i) = \frac{b(i) - a(i)}{\max\{b(i), a(i)\}},$$

где $a(i)$ - среднее расхождение между i-м вектором и всеми остальными векторами, отнесёнными к тому же кластеру A, что и i-й вектор, согласно решению, закодированному в данной хромосоме. $b(i) = \min_{B \neq A} a(i)$, т. е. под $b(i)$ понимается минимальное расхождение между i-м вектором и всеми остальными кластерами, кроме кластера A. Легко видеть, что $-1 \le s(i) \le 1$. Чем больше величина $s(i)$,

тем выше уровень принадлежности i-го вектора к соответствующему классу. Если кластер состоит из одного элемента, то величина $s(i)$ не определена, и самый разумный вариант положить $s(i) = 0$. Таким образом, функция приспособленности определяется как среднее значение:

$$f = \frac{1}{N}\sum_{i=1}^{N} s(i).$$

Необходимость соблюдения соответствия между числом классов, хранящимся в последней компоненте хромосомы, и числом классов, определяемых из решения, хранящегося в первых N компонентах хромосомы, накладывает определённые ограничения на операторы мутации и скрещивания.

Оператор скрещивания работает следующим образом. Сначала выбираются 2 генотипа (G1 и G2). Далее, считая, что G1 включает в себя k1 кластеров, алгоритм случайным образом выбирает $c \in \{1,2,...,k1\}$ кластеров и копирует их в G2. Неизмененные кластеры G2 остаются, а те, что изменились, перемещаются в ближайший из кластеров (ближайший относительно центров кластеров). Таким образом, потомок G3 получен. Такая же процедура используется для получения потомка G4, но теперь изменённые кластеры G2 копируются в G1. Для иллюстрации этой процедуры, рассмотрим следующие два генотипа:

G1 - 11232451254325334**2**4;
G2 - 12123232124423221321.

Например, предположим, что кластеры 2 и 3 в генотипе G1 были случайным образом выбраны (отмечены жирным шрифтом ниже):

G1 - 11**232**451**2**54**32**5**33**4**2**4.

Когда эти кластеры копируются в G2, они изменят кластеры {1, 2, 3} в G2, в то время как кластер 4 не меняется:

G2 - 12**232**3212**2**44**32**2**33**321.

Подчёркнутые позиции, которые соответствуют генам, на которые косвенно влияют кластеры 2 и 3 с G1, теперь изменены на 0 (ноль):

G3 - 00232000244320330**2**0.

Гены равные нулю затем будут помещены соответственно в ближайшие из кластеров (согласно их центров тяжести). Такая же процедура используется для получения потомка G4, за исключением того, что выбранные кластеры будут копироваться в G1.

Следует обратить внимание, что такая процедура скрещивания может привести как к увеличению числа кластеров, так и их уменьшения.

В генетическом алгоритме используются два оператора для мутации. Первый оператор работает только для генотипов, которые кодируют более чем два кластера. Он устраняет один из случайно выбранных кластеров путём размещения его в ближайший из оставшихся кластеров (в соответствии с их центрами тяжести). Второй оператор делит случайно выбранный кластер на два новых. Первый кластер формируется из последовательностей, которые находились ближе к исходному центру тяжести. Второй кластер – из тех последовательностей, которые находились ближе к самой дальней последовательности исходного кластера.

Для применения вышеописанного алгоритма кластеризации к прогнозированию временных рядов вводится понятие шаблона $(k_1, k_2, ..., k_l)$ длины $l+1$ как набора элементов временного ряда вида $y_t, y_{t+k_1}, ..., y_{t+k_l}$. Множество векторов, составленных из элементов временного ряда, номера которых удовлетворяют некоторому шаблону, составляют множество кластеризируемых данных.

Центры кластеров, полученных при кластеризации векторов, отвечающих всем возможным шаблонам всех возможных длин, и образуют множество характерных последовательностей.

Для каждой из полученных таким образом характерных последовательностей вычисляется её прогностическая ценность как величина обратная средней ошибке прогноза, полученного с помощью данной характерных последовательности. Для вычисления прогностической ценности характерных последовательностей использовалась та часть рассматриваемого временного ряда, которая не участвовала в формировании данных для алгоритма кластеризации. При прогнозе используются лишь те характерные последовательности, прогностическая ценность которых высока.

Отметим, что существенно важным параметром данного алгоритма является максимально допустимое число кластеров M. Увеличение данного числа до некоторого предела (равного теоретически возможному числу для регулярных рядов) приводит к улучшению работы алгоритма, но, естественно, замедляет скорость его работы.

Анализ полученных результатов. Работа приведенного выше алгоритма вначале была протестирована на зашумлённых регулярных рядах. В частности рассматривался временной ряд вида $y_t = \cos t + \alpha\varepsilon_t$, где случайные составляющих ε_t удовлетворяют условиям Гаусса-Маркова, а параметр α, определяющий уровень шума, менялся. Размер ряда составлял 1000 наблюдений, шаг дискретизации был равен $\Delta t = 0.01$.

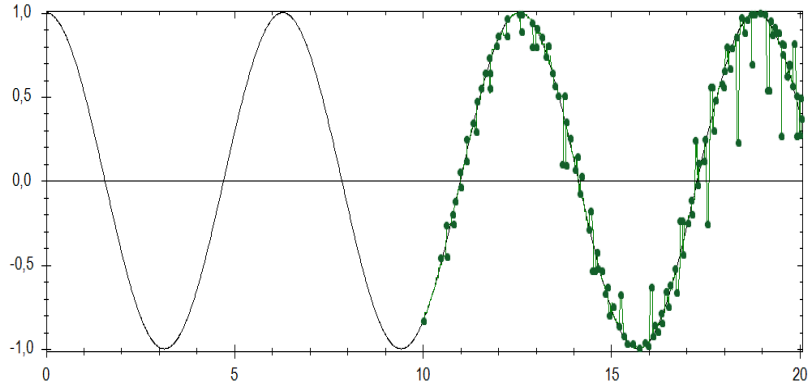


Рис. 1. Результаты применения алгоритма прогнозирования к регулярному временному ряду

На рис. 1 представлен результат прогноза для случая $\alpha = 0.01$. Максимально допустимое число кластеров составляло 50% (420 кластеров) от теоретически возможного. Сплошной линией представлен график $y_t = \cos t$, точками отмечены прогнозные значения. На рис. 2а для этих же значений параметров представлена зависимость средней ошибки прогноза от числа раз, которое последовательно применялся данный алгоритм к уже спрогнозированным с его помощью значениям. На рис 2б приведена аналогичная зависимость для случая $M = 100\%$ (840 кластеров). Представленные зависимости позволяют сделать вывод, что

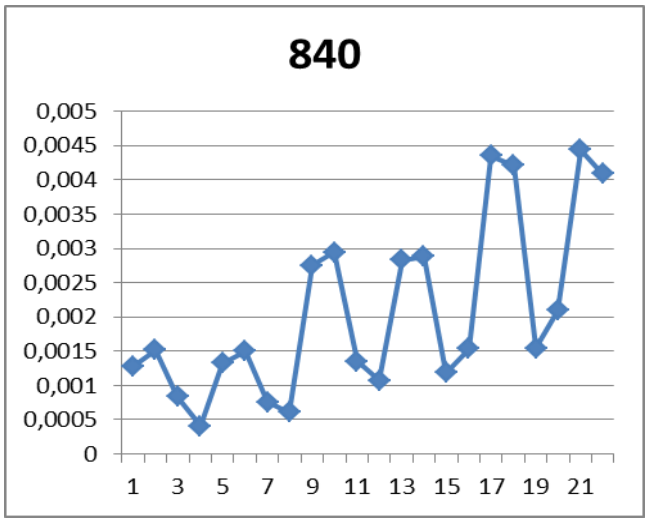


Рис. 2. Зависимость средней ошибки прогноза от числа раз, которое последовательно применялся данный алгоритм к уже спрогнозированным с его помощью значениям

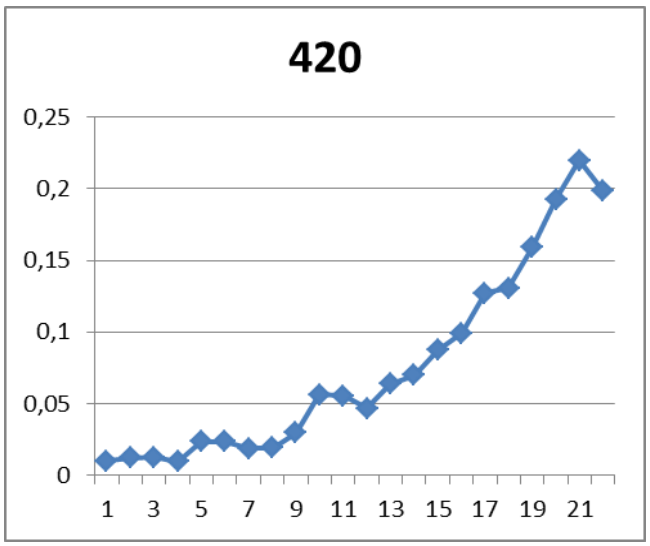
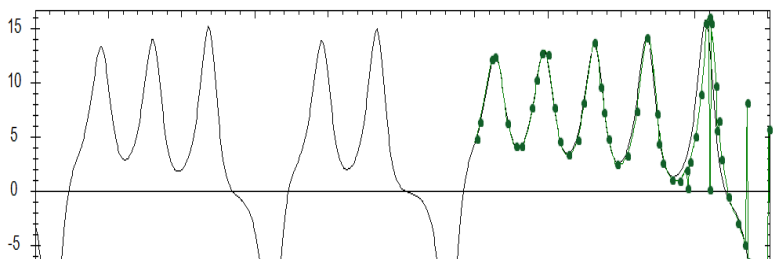


Рис. 2а. Зависимость средней ошибки прогноза от числа раз, которое последовательно применялся данный алгоритм к уже спрогнозированным с его помощью значениям. $M = 50\ \%$



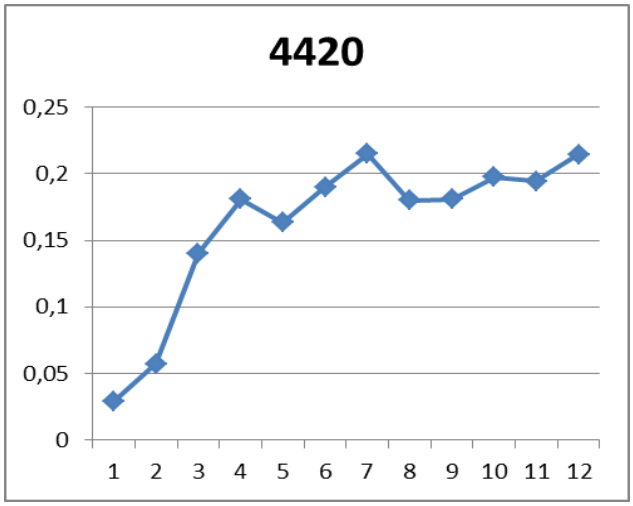


Рис. 4. Ряд Лоренца. Зависимость средней ошибки прогноза от числа раз, которое последовательно применялся алгоритм

алгоритм позволяет осуществлять прогноз на значительное число шагов вперёд, при сохранении приемлемой точности прогноза.

Предложенный алгоритм был применён для прогнозирования временного ряда, полученного интегрированием системы Лоренца.

Система Лоренца является классическим примером для проверки работоспособности прогнозных алгоритмов. Для получения временного ряда система интегрировалась методом Рунге-Кутта 4-го порядка, размер полученного таким образом ряда составил 1000 наблюдений.

На рис. 3 представлен временной ряд, построенный по системе Лоренца, (сплошная линия) и результаты применения алгоритма (точки). Представленный случай соответствует случаю - 50 % от теоретически возможного числа кластеров.

На рис. 4 для указанных значений параметров представлена зависимость средней ошибки прогноза от числа раз, которое последовательно применялся данный алгоритм к уже спрогнозированным с его помощью значениям.

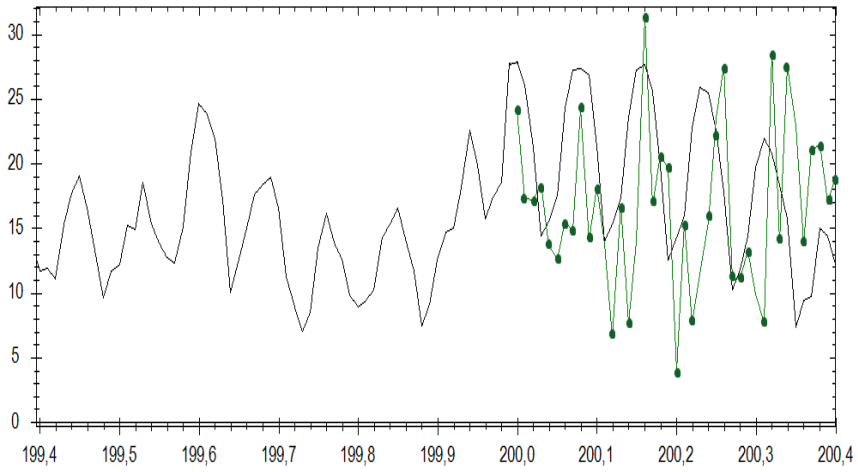


Рис. 5. Результаты применения алгоритма прогнозирования к температурному временному ряду

Наконец, обратимся к исследованию качества прогноза, полученного с помощью данного алгоритма для температурных временных рядов. В качестве примера был выбран временной ряд, описывающий температурный режим г. Львов за последние семь лет; таким образом, общее число наблюдений составило 20000. На рис. 5 представлены наблюдаемые значения (сплошная линия) и прогнозные (точки) для данного ряда.

Выводы.

1. В настоящей работе предлагается новый метод прогнозирования хаотических временных рядов. Метод базируется на выделении характерных последовательностей в рассматриваемом временном ряде с помощью клас-теризации на множестве всех возможных последовательностей, наблюдаемых в данном временном ряде. При этом прогноз осуществляется по центрам полученных кластеров, что позволяет избежать накопления погрешности, естественного для хаотических рядов. Кластеризация осуществляется с помощью специальным образом модифицированного генетического алгоритма.
2. В работе анализируется качество работы метода, определяется набор оптимальных параметров.
3. Метод был протестирован на слабозашумлённых регулярных рядах и применён к прогнозированию хаотических рядов. В частности, были рассмотрены «классический» хаотический ряд – ряд Лоренца – и температурный ряд для г. Львов. Метод продемонстрировал хорошие результаты прогноза для зашумлённых регулярных рядов и удовлетворительные – для хаотических.

Библиографические ссылки

1. **Анищенко В.С.** Сложные колебания в простых системах. – М., 2009.
2. **Глазунова О.И.** Синергетика творчества. – М., 2012.
3. **Громов В.А.** Применение метода “муравьиных колоний” для прогнозирования хаотических временных рядов / В.А. Громов, А.Н. Шульга // Питання прикладної математики і математичного моделювання: зб. наукових праць. – 2011. – С. 74-84.
4. **Малинецкий Г.Г.** Современные проблемы нелинейной динамики / Г.Г. Малинецкий, А.Б. Потапов. – М., 2000.
5. **Милованов В.П.** Синергетика и самоорганизация. Общая и социальная психология. – М., 2010.
6. **Мюррей Дж.** Математическая биология. Введение. – Ижевск, 2009.
7. **Elsner J.B.** Singular spectrum analysis: A new tool in time series analysis (1st ed.) / J.B. Elsner, A.A. Tsonis. – N.Y, 1996.
8. **Fu Y.Y.** ARFNNs with SVR for prediction of chaotic time series with outliers / Y.Y. Fu, C.Y. Wub, J.T. Jeng, C.N. Ko // Expert Systems with Applications. – 2010. – №37. – P. 4441–4451.
9. **Gan M.** A locally linear RBF network-based state-dependent AR model for nonlinear time series modeling / M. Gan, H. Peng, X. Peng, X. Chen, G. Inoussa // Information Sciences. – 2010. – №180. – P. 4370–4383.
10. **Gu H.** Fuzzy prediction of chaotic time series based on singular value decomposition / H. Gu, H. Wang // Applied Mathematics and Computation. – 2007. – №185. – P. 1171–1185.
11. **Gromov V.A.** Chaotic time series prediction with employment of ant colony optimization / V.A. Gromov, A.N. Shulga // Expert Systems with Applications. – 2012. – №39. – P. 8474-8478.
12. **Hong W.C.** Application of chaotic ant swarm optimization in electric load forecasting. Energy Policy. – 2010. – №38. – P. 5830–5839.
13. **Hong W.C.** Forecasting urban traffic flow by SVR with continuous ACO / W.C. Hong, Y. Dong, F. Zheng, C.Y. Lai // Applied Mathematical Modeling. – 2011. – №35. – P. 1282–1291.
14. **Hrushka E.R.** Extracting rules from multilayer perceptrons in classification problems: A clustering-based approach / E.R. Hrushka, N.F.F. Ebecken // Neurocomputing. – 2006. – №70. – P. 384-397.
15. **Mirzaee H.** Linear combination rule in genetic algorithm for optimization of finite impulse response neural network to predict natural chaotic time series // Chaos, Solitons and Fractals. – 2009. – №41. – P. 2681–2689.
16. **Niu D.** Power load forecasting using support vector machine and ant colony optimization / D. Niu, Y. Wang, D.D. Wu // Expert Systems with Applications. – 2010. – №37. – P. 2531–2539.
17. **Pan Y.** Predicting the net heat of combustion of organic compounds from molecular structures based on ant colony optimization / Y. Pan, J.C. Jiang, R. Wang, J.J. Jiang // Journal of Loss Prevention in the Process Industries. – 2011. – №24. – P. 85–89.
18. **Toskari M.D.** Estimating the net electricity energy generation and demand using the ant colony optimization approach. Energy Policy. – 2009. – №37. – P. 1181–1187.
19. **Unler A.** Improvement of energy demand forecasts using swarm intelligence: The case of Turkey with projections to 2025. Energy Policy. – 2008. – №36. – P. 1937–1944.
20. **Wang J.** Chaotic time series method combined with particle swarm optimization and trend adjustment for electricity demand forecasting / J. Wang, D. Chi, J. Wu, H. Lu // Expert Systems with Applications. – 2011. – №38, – P. 8419–8429.

Надійшла до редколегії 18.04.2012