УДК 519.8 doi:10.15421/321603

**В. А. Громов, И. М. Воронин, В. Р. Гатыло, Е. Т. Прокопало**
*Днепропетровский национальный университет имени Олеся Гончара*

# ОЦЕНКА ГИПЕРПАРАМЕТРОВ В ЗАДАЧАХ ПРОГНОЗИРОВАНИЯ НА ОСНОВЕ КЛАСТЕРИЗАЦИИ

**Введена задача оценки прогнозной ценности кластеров, возникающих в процессе кластеризации при решении задачи прогнозирования на основе кластеризации, и предложено несколько методов решения данной задачи с использованием дополнительного валидационного множества. Отдельно исследованы составляющие ошибки прогнозирования, связанные с прогнозной субмоделью (при правильном выборе кластеров, с помощью которых осуществляется прогнозирование) и с неправильным выбором кластера, с помощью которого осуществлен прогноз, в функции объёма обучающей и валидационной выборок.**

**Уведено задачу оцінювання прогнозної цінності кластерів, що виникають у процесі кластеризації під час розв'язання задачі прогнозування на основі кластеризації. Запропоновано декілька методів розв'язання даної задачі з використанням додаткової валідаційної множини. Окремо досліджено складники помилки прогнозування, пов'язані з прогнозною субмоделлю (за правильного вибору кластерів, за допомогою яких здійснювалося прогнозування) та з неправильним вибором кластера, за допомогою якого здійснювалося прогнозування, у функції обсягу навчальної та валідаційної вибірок.**

**The paper deals with novel problem statement of estimating predictive clusters value for clusters generated in the framework of predictive clustering algorithms along with methods to solve it. We investigate separately (as functions of sizes of training and validation samples) the two addenda of an prediction error: The first one is caused by prognostic submodel (once the cluster to be used to predict is chosen correctly) and the second one is resulted from incorrect choice of this cluster.**

**Ключевые слова: прогнозирование временных рядов, хаотический временной ряд, прогнозирование на основе кластеризации, оценка прогнозной ценности кластера.**

**Введение.** Невозможность построения единой прогнозной модели для временных рядов, имеющих хаотическую природу, обусловливает необходимость построения методов прогноза, позволяющих явным или неяв-

ным образом объединять в себе множество субмоделей, отвечающих различным паттернам, наблюдаемым в ряде [22]. Среди множества различного рода подходов выделяется подход прогнозирования на основе кластеризации (predictive clustering) [3], в котором субмодели формируются на основе кластеров последовательностей наблюдений временного ряда.

Одним из возможных направлений повышения эффективности прогноза в рамках данной парадигмы является построение оценки прогнозного качества полученных кластеров. Здесь возможна оценка прогнозной ценности кластеров на основании осуществления прогноза на дополнительном валидационном множестве, отличном от обучающего (на котором осуществляется кластеризация) и тестирующего (на котором осуществляется окончательная проверка качества работы метода); в рамках данного подхода величины, характеризующие прогнозную ценность отдельных кластеров, трактуются как гиперпараметры в терминах Гудфеллоу и др. [10] или параметры алгоритма в терминах вычислительной математики. Единственным отличием от классического случая является то, что число гиперпараметров здесь совпадает с числом кластеров и потому весьма велико. В данном случае возможно использование в качестве меры прогнозной ценности кластера средней ошибки прогнозирования на валидационном множестве, индуцированной конкретным кластером, соответствующей кластеру (что позволяет использовать при прогнозе лишь кластера, соответствующие часто посещаемым областям аттрактора), другие меры.

Информация о прогнозной ценности отдельного кластера не обязательно должна представлять собой скалярную величину, но может представлять собой, например, ряд логических правил, позволяющих выделять ситуации, когда данный кластер целесообразно использовать для прогнозирования.

В любом случае ошибка прогнозирования может быть разделена на две части: первая, связанная с неправильным выбором кластера и соответствующей ему субмодели прогнозирования, вторая – с отклонением прогнозного и реального значений для правильно выбранного кластера, кластера, соответствующего области аттрактора, в окрестности которой проходит участок траектории (временного ряда), для которого осуществляется прогноз. Если вторая составляющая ошибки не может быть уменьшена в рамках выбранного алгоритма кластеризации (и рассматривается как некий теоретический минимум для данного ряда, алгоритма кластеризации, прогнозной субмодели и объёма обучающей выборки), то первая составляющая может быть уменьшена путём удаления из рассмотрения кластеров с низкой прогнозной ценностью. Совокупность процедур оценки прогнозного качества кластеров (гиперпараметров) с использованием ва-

лидационного (проверяющего) множества получила название оценки качества (quality assessment) кластеров.

В настоящей работе будет несколько подходов к оценке прогнозной ценности кластеров и их влияние на качество прогноза. В качестве критерия качества того или иного подхода используется величина первой составляющей ошибки (её доля в ошибке прогноза) и число непрогнозируемых точек тестируемой выборки [7; 8], т. е. точек, для которых алгоритм не может дать прогноз, поскольку не находит кластера, отвечающего наблюдаемой последовательности. Отметим, что, с нашей точки зрения, возможность фиксации алгоритмом непрогнозируемых точек является существенным преимуществом: гораздо лучше, когда алгоритм "честно" указывает на невозможность дать адекватный прогноз в некоторой точке, чем когда он даёт неадекватный прогноз и не указывает на опасность использования прогнозных значений.

Существенно важной особенностью алгоритмов парадигмы прогнозирования на основе кластеризации является возможность и необходимость использования в процессе создания кластеров не только временного ряда, который необходимо прогнозировать, но и совокупности подобных рядов, которые его содержат.

В качестве данных использовались ряды, полученные интегрированием системы Лоренца и зашумлённой системы Лоренца (модельные данные), а также совокупность рядов, описывающих динамику цен на электроэнергию в различных населённых пунктах Австралийского Содружества (реальные данные).

Дальнейшее изложение структурировано следующим образом. Во втором разделе представлен обзор последних результатов, связанных с парадигмой прогнозирования на основе кластеризации, в третьем рассматривается используемый метод прогнозирования, а также предложенные подходы к решению задачи оценки прогнозной ценности кластеров. Четвёртый раздел посвящён анализу полученных прогнозных результатов; в пятом разделе приведено сравнение с результатами других авторов; в завершающем, шестом, разделе сформулированы выводы.

**Современное состояние проблемы.** В работах, посвящённых прогнозированию на основе кластеризации, обычно выделяют два магистральных направления [1]: в рамках первого направления путём введения различного рода метрик кластеризации подвергаются временные ряды целиком, в рамках второго – происходит выделение характерных паттернов динамического поведения, наблюдаемых во временном ряде либо же в совокуп-

ности временных рядов, (typical sequences [7 – 9], motifs [23], chunks [24], shapelets [26], pattern discovery, subsequence clustering [1]).

При описании возможных подходов к решению задачи выделения характерных паттернов динамического поведения отметим, прежде всего, работу [17], в которой акцентируется внимание на бессмысленности решения указанной задачи на одном временном ряде и на необходимости использования множества родственных временных рядов. В работах, посвящённых анализу алгоритмов данного класса, исследованию, обычно, подвергается методология формирования обучающей выборки и алгоритм кластеризации (шире – выделения родственных участков временного ряда); указанные составляющие алгоритма прогнозирования на основе кластеризации можно соотнести с концепциями data-adaptation и algorithm-adaptation [20]. В настоящей работе предложено добавить ещё одну составляющую – оценку прогнозной ценности кластеров (as it were, prediction-adaptation).

В рамках концепции data-adaptation можно выделить [1; 20] подходы, в рамках которых кластеризации подвергаются исходные данные (raw data), характеристики, извлечённые из данных, (feature-based transformation of the data) и результаты применения к исходным данным некоторой прогнозной модели (model-based transformation of the data).

Другим важным аспектом обеспечения качества прогноза в рамках выбранной парадигмы прогнозирования является выбор алгоритма кластеризации и его адаптация к решению задачи прогнозирования на основе кластеризации. Здесь было использовано значительное число работ по использованию алгоритма k-средних, c-средних (чётких и нечётких) и им подобных.

Работа [11] посвящена исследованию модификации стандартного алгоритма k-средних, удобной для выделения подобных участков во временных рядах, – TSkmeans (*Time Series k-means*). В работе [21] для прогнозирования хаотических рядов также используются k-средние; кроме того, приведены результаты прогнозирования (представленные в работах разных авторов), полученные с помощью значительного числа различных подходов к получению прогноза для временных рядов, описывающих динамику цен на электроэнергию в различных населённых пунктах Австралийского Содружества, – указанное множество рядов, похоже, становится своеобразным стандартным тестом (benchmark) для проверки эффективности алгоритмов прогнозирования хаотических временных рядов (реальные данные); расширенный вариант указанных прогнозных результатов изложен в [8].

В исследованиях [12; 13] предлагается алгоритм кластеризации пространственновременных данных, использующий модификацию евклидового расстояния, позволяющую учесть пространственные и временные закономерности. Работа [2] посвящена выделению характерных паттернов в совокупности временных рядов, порождённых энергогенерирующей системой: цель исследования – получение алгоритмов более рационального потребления электроэнергии (в рамках исследовательских задач, сформулированных в программе Европейского Союза "Горизонт 2020"); авторы используют различные модификации алгоритма k-средних.

Недостатком алгоритмов этого класса [6] является сильная зависимость структуры выделяемых классов от используемой метрики, кроме того, в большинстве случаев необходимым предусловием кластеризации является знание числа кластеров.

Указанные недостатки в известной степени не присущи методам кластеризации, опирающимся на аппарат теории графов\сложных сетей. Так, в работе [6] производится трансформация участков временного ряда в вершины графа, после чего (в рамках парадигмы теории сложных сетей) кластера выделяются с помощью алгоритмов нахождения сильносвязных подграфов (community detection). В исследовании [8] для кластеризации используется модифицированный алгоритм Уишарта [19]; автор указывает на связь между полученными кластерами (отвечающими динамическим паттернам) и областями фазового пространства с высокими значениями инвариантной меры динамической системы.

В заключение отметим, что, насколько известно авторам настоящего обзора, задача оценки прогнозной ценности кластеров в рамках парадигмы прогнозирования на основе кластеризации не ставилась, а следовательно, не рассматривались методы её решения.

**Постановки задач.** *Задача прогнозирования*. Рассматривается совокупность $S$ родственных хаотических временных рядов; совокупность наблюдений $\left\{y^{(s)}\right\} \equiv \left\{y_0^{(s)}, y_1^{(s)}, ..., y_{t_s}^{(s)}\right\}$, $s = \overline{1, S}$ (где $y_i^{(j)}$ – $i$-е наблюдение $j$-го ряда) используется для прогнозирования значений наблюдений ряда $y$ (который может принадлежать, а может и не принадлежать совокупности рядов $\left\{y^{(s)}\right\}$) – $y_{t+1}, y_{t+2}, ..., y_{t+K}$ – с требованием минимизации ошибки прогнозирования

$$I = M \sum_{p=1}^{K} \left(\hat{y}_{t+p} - y_{t+p}\right)^2 \to \min . \qquad (1)$$

Здесь $\hat{y}_{t+p}$ – прогнозные значения, по которым и осуществляется минимизация.

Предполагаем, что все переходные процессы в системе, порождающей рассматриваемый временной ряд, завершены и поведение временного ряда отражает траекторию движения системы в окрестности странного аттрактора, сколь бы сложной ни была указанная траектория. Также будем предполагать, что рассматриваемый ряд удовлетворяет условиям теоремы Такенса и, следовательно, можно проанализировать структуру аттрактора, используя наблюдения временного ряда.

Поскольку траектории системы неоднократно посещают одну и ту же область аттрактора, то среди наблюдений временного ряда можно встретить аналогичные (подобные) последовательности, связанные с движением траектории в данной области. Если можно выделить эти области, описать их и построить простейшую модель прогнозирования для каждой из них, то также можно осуществить и прогноз для рассматриваемого временного ряда на весьма значительное число шагов вперёд [7]. Рассмотренный ниже метод кластеризации используем для группировки последовательностей значений временного ряда.

Как правило, чтобы удовлетворить условиям теоремы Такенса, необходимо составить вектора из наблюдений временного ряда [15]. Интересно то, что использование векторов, состоящих из последовательных наблюдений, оказалось менее эффективными для прогнозирования, чем векторов, составленных в соответствии с некоторым шаблоном.

Следует подчеркнуть, что каждая из указанных простейших моделей это усредненное представление последовательностей наблюдений временного ряда, относящихся к одному кластеру (или, альтернативно, траекторий, проходящих по соответствующей области аттрактора). Указанное усреднение приводит к ухудшению качества прогноза из-за использования для прогноза средних значений (прогнозируемые значения получаются с использованием центров кластеров) и одновременно к его улучшению, обусловленному уменьшением экспоненциального роста ошибки прогнозирования. Используемый метод кластеризации обеспечивает компромисс между этими тенденциями.

Здесь следует подчеркнуть, что в методах парадигмы ошибка состоит из двух слагаемых: первое слагаемое связано с неправильной идентификацией наблюдаемой динамики – неправильным выбором кластера, по которому осуществляется прогнозирование; второе – с неизбежной погрешностью, вносимой используемой для прогнозирования субмоделью\субмоделями. Второе слагаемое определяется используемым алго-

ритмом кластеризации и используемыми прогнозными субмоделями и в рамках настоящей работы считается неизменным.

Задачу уменьшения значений первого слагаемого сформулируем как *задачу оценки прогнозного качества кластеров* (задача оценки качества для прогнозирования на основе кластеризации). Задачу формулируем как задачу выбора из всего множества полученных кластеров подмножество, которое бы обеспечивало (первая постановка) минимальное либо (вторая постановка) не большее чем заданное значение.

Математически задачу оценки прогнозного качества кластеров формулируем следующим образом. Пусть $\Lambda$ – множество кластеров, используемых для прогнозирования рассматриваемого временного ряда: $G:\Lambda\to R^1$ – функция оценки прогнозного качества; $\tilde{\Lambda}(G,\beta)=\{\lambda\in\Lambda: G(\lambda)\geq\beta\}$ – множество кластеров, прогнозная ценность которых не превышает заданного уровня $\beta$. Необходимо определить функцию $G^*$ и пороговое значение $\beta^*$ таким образом, чтобы в первой постановке значение ошибки прогнозирования (на тестирующем множестве) было минимальным:

$$I\left(\tilde{\Lambda}(G,\beta)\right)\to\min\,. \tag{2}$$

Во второй постановке минимизации подлежит количество кластеров, входящих в множество $\tilde{\Lambda}(G,\beta)$:

$$\left|\tilde{\Lambda}(G,\beta)\right|\to\min\,, \tag{3}$$

при ограничении

$$\left|I\left(\tilde{\Lambda}(G,\beta)\right)\right|\leq\gamma\,, \tag{4}$$

где $\gamma$ – параметр алгоритма.

В рамках первой постановки делаем упор на минимизацию ошибки прогноза, во второй – на скорость выполнения операции прогнозирования. В каждой из постановок предполагаем существенное снижения числа кластеров, т. е. уменьшение сложности общей прогнозной модели; здесь можно провести параллели с различного рода методологиями снижения сложности (числа параметров) в моделях регрессии (например: AIC, BIC, GIC [18]).

Для решения данной задачи введем дополнительное дообучающее (валидационное) множество, отличное от обучающего и тестирующего мно-

жеств; предположим, что все три множества принадлежат одной генеральной совокупности.

**Алгоритм прогнозирования.** Предлагаемый алгоритм состоит из трёх частей. Первая – это анализ временного ряда для кластеризации последовательностей, составленных из наблюдений, соответствующих предопределённым шаблонам, и последующего выявление характерных последовательностей наблюдений, таких как центры кластеров. В рамках второй части (с использованием дообучающего множества) осуществляем процедуру оценки прогнозного качества кластеров и удаления части из них с низкими значениями указанной величины; третья часть предусматривает прогнозирование с использованием полученных характерных последовательностей.

*Формирование выборки и алгоритм кластеризации.* Предполагаем, что ряд нормализован. В работе использованы два варианта процедуры нормировки: в рамках первой процедуры нормировке подвергался весь временной ряд одновременно, в рамках второй – значения вектора обучающей выборки. Указанные способы нормировки получили название глобального (G) и локального (L). В рамках локального способа нормировки мы получаем возможность кластеризировать не столько характерные амплитуды (как в случае с глобальным способом нормировки), сколько характерные профили.

Для формирования обучающей выборки использовали понятие шаблона. Под шаблоном здесь понимаем фиксированную последовательность расстояний между позициями наблюдений в последовательности, которые займут соседние позиции в формируемом векторе наблюдений.

В качестве алгоритма кластеризации использовали алгоритм Уишарта [25; 19]: метод основывается на аппарате теории графов и непараметрической оценке плотности k-ближайших соседей. Особенности применения указанного алгоритма кластеризации в задачах прогнозирования на основе кластеризации изложены в [7]. Другой используемый алгоритм – модифицированный алгоритм кластеризации FOREL (Формальный элемент) [5]. В дальнейшем при указании на применение алгоритма Уишарта мы будем ставить букву W, модифицированного алгоритма кластеризации FOREL – букву F. Указанные алгоритмы кластеризации применяли к выборкам, полученным с помощью всех возможных шаблонов фиксированной длины. Для каждой такой выборки формировали свой набор кластеров.

*Задача оценки прогнозного качества кластеров*. В работе рассмотрено два варианта решения задачи оценки прогнозного качества кластеров. В рамках первого варианта прогнозную ценность рассчитывали по формуле

$$Q_i(\beta) = \sum\nolimits_{i \in S_i} \frac{\bar{e}_i}{e_{ij}} \frac{1}{|V_i|}, \bar{e}_i = \frac{1}{|V_i|} \sum_{i \in V_i} \varepsilon_{ij} \ , \tag{5}$$

где $V_i$ – множество кластеров, которые могут прогнозировать точку $x_i$ с ошибкой, меньшей за $\beta$; $\bar{e}_i$ – среднеквадратическая ошибка по всем кластерам из $V_i$.

Другой мерой оценки прогнозной ценности кластера является частота появления последовательностей наблюдений, связанных с этим кластером, во временном ряде: чем меньше вероятность появления той или иной характерной последовательности в ряде, тем к меньшему росту ошибки прогнозирования приведёт удаление соответствующего кластера. В качестве меры оценки указанной величины было выбрано произведение значений инвариантной меры динамической системы, породившей временной ряд, для малых элементов фазового пространства, содержащих компоненты центра кластера. Для восстановления указанного распределения использовали алгоритм подсчёта ячеек [15; 17].

Наконец, третий метод решения задачи оценки прогнозного качества заключается не в определении прогнозного качества кластера с использованием одной характеристики, а в решении задачи экстракции знаний из данных о предоставляемых кластером прогнозах из наблюдений дообучающего множества.

Здесь при описании алгоритма используются следующие обозначения:

$N$ – число наблюдений дообучающего множества;

$M$ – число кластеров, оценку которых необходимо получить;

$S_{ie}(\beta)$ – множество наблюдений, которые кластер может спрогнозировать с ошибкой, меньшей $\beta$;

$S_{id}(\beta)$ – множество наблюдений, которые находятся на расстоянии, не превышающем $\beta$, от кластера $i$;

$m_i$ – количество наблюдений, для которых кластер $i$ является ближайшим в евклидовой норме;

$n_j$ – количество наблюдений, для которых использование кластера $i$ позволило получить наименьшую возможную ошибку.

В этих обозначениях алгоритм поиска характеристик сформулируем следующим образом.

1. Инициализация: для любого $j$:

$S_{jd} \neq \varnothing$, $S_{je} \neq \varnothing$, $m_j = 0$ , $n_j = 0$ .

2. Положим $i = 0$ .
3. Положим $j = 0$ .
4. Для точки $x_i$ находим расстояние $d_{ij}$ от неё до кластера $c_j$ .

   Если $d_{ij} < \beta$, то $S_{jd} = S_{jd} \cup x_i$ .

5. Для точки $x_i$ находим ошибку прогнозирования $e_{ij}$ с использованием кластера $c_j$ .

   Если $e_{ij} < \beta$, то $S_{je} = S_{je} \cup x_i$ .

6. $d_{i\min} = d_{ik} = \min_j \{d_{ij}\}$ . $m_k = m_k + 1$.
7. $e_{i\min} = e_{ip} = \min_j \{e_{ij}\}$ и расстояние $d_{ip}$ . $n_k = n_k + 1$.
8. $j = j + 1$. Если $j < M$ , то переходим к шагу 3.
9. $i = i + 1$. Если $i < N$ , то переходим к шагу 2.

*Процедура получения прогнозных значений*. Для прогнозирования временного ряда использовались центры полученных кластеров (характерные последовательности), рассчитанные для всех используемых шаблонов: для позиции, для которой требуется получить прогноз, из предыдущих наблюдений временного ряда составлялись векторы в соответствии с каждым из использованных шаблонов таким образом, чтобы последняя позиция в шаблоне совпадала с позицией, для которой требуется получить прогноз. Размерность полученных таким образом векторов и центров кластеров уменьшали на единицу и рассчитывали эвклидово расстояние между усечённым вектором наблюдений и центром кластера. Среди всех кластеров и всех шаблонов отыскивали кластер, для которого данное расстояние было минимальным. Если указанное расстояние было меньше, чем некоторое пороговое значение, то последний элемент центра данного кластера использовали в качестве прогнозного значения. В противном случае – динамику считали неидентифицированной, а наблюдение добавляли к множеству непрогнозируемых точек.

**Численные результаты.** Метод, представленный в предыдущем разделе, был применён для прогнозирования ряда Лоренца, совокупности зашумлённых рядов Лоренца, совокупности рядов, описывающих цены на электроэнергию в различных населённых пунктах Австралийского Содружества. Прогнозирование осуществлялось на один шаг вперёд. Для всех рядов использовались 4-точечные шаблоны с максимальным рас-

стоянием между наблюдениями, равным 10; таким образом, максимальное число шаблонов равно 1000.

Информацию о каждом из исследованных рядов представим в табличном виде. В качестве меры ошибки прогнозирования использовали три величины: среднеквадратичное уклонение прогнозных значений от наблюдаемых (*RMSE*), относительную погрешность прогнозирования (*MAE*) и число непрогнозируемых точек. Все три величины вычисляли для тестирующего множества (множество не используется ни для обучения, ни для дообучения). Первые две величины определяли формулами

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1,n}(y_i - \hat{y}_i)^2}\ , \tag{6}$$

$$MAE = \frac{1}{n}\sum_{i=1,n}|\ y_i - \hat{y}_i\ |. \tag{7}$$

Далее в таблице приведены значения ошибки прогнозирования и доли непрогнозируемых точек для различных вариантов метода. В первых двух колонках указан объём обучающего и дообучающего (валидационного) множеств, в трёх следующих – информация об используемом методе: способ нормализации (третья колонка), алгоритм кластеризации (четвёртая колонка) и алгоритм оценки прогнозной ценности кластеров (пятая колонка). Здесь использованы следующие обозначения: G – глобальный способ нормировки, L – локальный; W – алгоритм Уишарта, F – модифицированный алгоритм кластеризации FOREL; числа 1, 2 отвечают, соответственно, алгоритму оценки прогнозного качества кластеров с помощью формулы (5) и алгоритму смены прогнозного кластера.

В следующих трёх колонках приведены значения среднеквадратичного уклонения, значения относительной погрешности и процент непрогнозируемых точек; в последней колонке приведены значения величины среднеквадратичного уклонения для случая, когда для прогнозирования выбирается наилучший кластер (т. е. первая составляющая ошибки прогноза равна нулю): указанная величина служит своеобразным “теоретическим минимумом”, с которым сравниваются значения ошибок прогнозирования фиксируемые для того или иного метода.

*Таблица 1*

**Результаты, полученные для ряда Лоренца**

| Count | Train | Normalize | Algo | PE | RMSE % | MAE | Non % | Min |
|---|---|---|---|---|---|---|---|---|
| $10^5$ | $10^4$ | G | W | 1 | 1.82 | 0.012 | 0.73 | 0.00531 |
| $10^5$ | $10^5$ | G | W | 1 | 1.023 | 0.008 | 0.61 | 0.00389 |
| $10^5$ | $10^6$ | G | W | 1 | 0.89 | 0.008 | 0.59 | 0.00379 |
| $10^5$ | $10^7$ | G | W | 1 | 0.83 | 0.008 | 0.52 | 0.00362 |
| $10^5$ | $10^4$ | G | W | 2 | 1.45 | 0.01 | 0.74 | 0.0053 |
| $10^5$ | $10^5$ | G | W | 2 | 1.027 | 0.008 | 0.64 | 0.00392 |
| $10^5$ | $10^6$ | G | W | 2 | 0.87 | 0.007 | 0.6 | 0.00377 |
| $10^5$ | $10^7$ | G | W | 2 | 0.78 | 0.007 | 0.57 | 0.00358 |
| $10^5$ | $10^4$ | L | F | 1 | 0.781 | 0.005 | 0.07 | 0.00213 |
| $10^5$ | $10^5$ | L | F | 1 | 0.774 | 0.005 | 0.06 | 0.00213 |
| $10^5$ | $10^6$ | L | F | 1 | 0.774 | 0.005 | 0.05 | 0.00213 |
| $10^5$ | $10^7$ | L | F | 1 | 0.773 | 0.005 | 0.03 | 0.00213 |
| $10^5$ | $10^4$ | L | F | 2 | 0.762 | 0.005 | 0.03 | 0.00213 |
| $10^5$ | $10^5$ | L | F | 2 | 0.733 | 0.004 | 0.03 | 0.00213 |
| $10^5$ | $10^6$ | L | F | 2 | 0.714 | 0.004 | 0.03 | 0.00213 |
| $10^5$ | $10^7$ | L | F | 2 | 0.702 | 0.004 | 0.03 | 0.00213 |

Count – размер выборки обучения; Train – размер выборки дообучения; Normalize – способ нормализации; Algo – алгоритм кластеризации; PE (Post-education) – алгоритм оценки прогнозной ценности кластеров; RMSE – среднеквадратическая отклонение; MAE – относительная погрешность; Non – процент не прогнозируемых точек; Min – «теоретический минимум».

На первых двух рисунках результаты, представленые в таблице, визуализированы: здесь представлены зависимости среднеквадратичного уклонения и относительной погрешности от объёма дообучающего множества в логарифмическом масштабе (при фиксированном значении объёмов обучающей и тестирующей выборок).

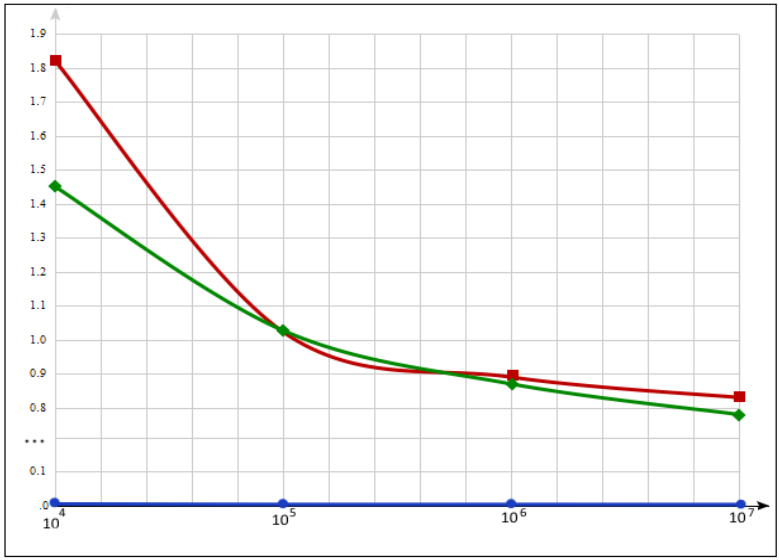


**Рис. 1. Ряд Лоренца. Алгоритм кластеризации Уишарта. Линии с кружочками (соответствующими расчётным значениям) – «теоретический минимум», с квадратиками – первый метод оценки прогнозного качества кластеров, с ромбиками – второй метод**

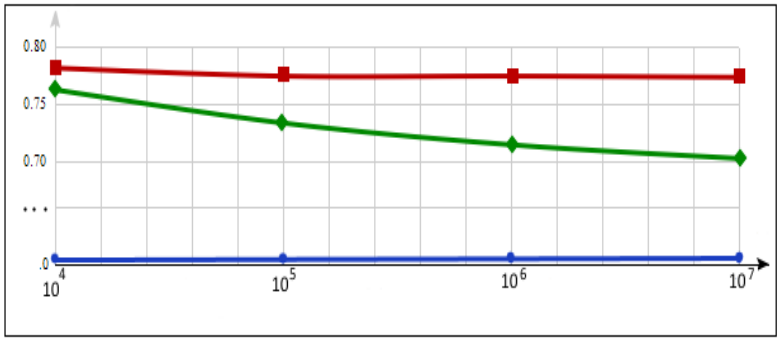


**Рис. 2. Ряд Лоренца. Алгоритм кластеризации FOREL. Линии с кружочками (соответствующими расчётным значениям) – «теоретический минимум», с квадратиками – первый метод оценки прогнозного качества кластеров, с ромбиками – второй метод**

Наконец, последний из серии рисунков, отвечающих одному ряду (группе родственных рядов), описывает изменение качества прогноза в зависимости от использования метода прогнозирования.

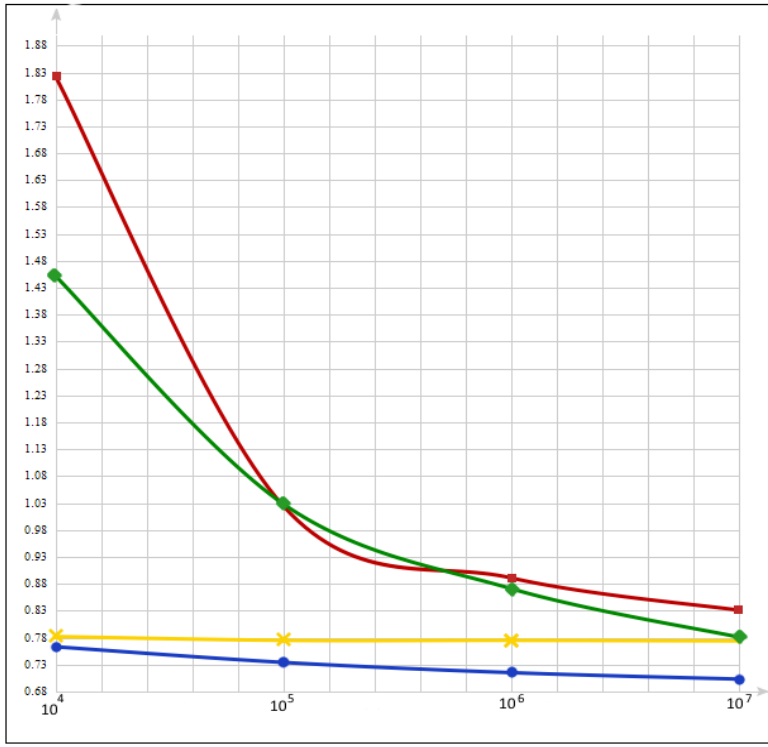


**Рис. 3. Ряд Лоренца. Сравнение результатов, полученных разными методами. По оси абсцисс отложен размер выборки дообучения; линии с квадратиками – кластеризация методом Уишарта и дообучение первым методом, с ромбиками – кластеризация методом Уишарта и дообучение вторым методом, с крестиками – кластеризация модифицированным методом FOREL и дообучение первым методом, с кружочками – кластеризация модифицированным методом FOREL и дообучение вторым методом**

В начале указанный алгоритм был применён к временному ряду, полученному интегрированием системы Лоренца [14]. Указанный ряд, с одной стороны, является типичным примером хаотического временного ряда – вычисленное нами значение старшего показателя Ляпунова равнялось 0.92

(использовался метод аналога [15; 17]), что хорошо согласуется с известными из литературы данными [17], – а, с другой стороны, традиционно используется при проверке эффективности того или иного метода прогнозирования хаотических временных рядов.

Система Лоренца (со стандартными «хаотическими» параметрами $\sigma = 10, b = \frac{8}{3}, r = 28$) интегрировалась с помощью метода Рунге-Кутта 4-го порядка (с шагом интегрирования 0.05). Для полученного ряда – ряд Лоренца – первые 3000 наблюдений отбрасываются, чтобы гарантировать, что движение траектории происходит в окрестности соответствующего странного аттрактора. Тестирующее множество состоит из 100000 наблюдений, тогда как размер обучающего и дообучающего множества изменяется и является существенным параметром рассматриваемого метода.

На рис. 1 представлены результаты прогнозирования на 1 шаг вперёд для ряда Лоренца. Размер обучающей выборки – 100000 наблюдений, дообучающей – $10^7$. Число непрогнозируемых наблюдений составило 570, среднеквадратическая ошибка прогнозирования для остальных точек равна 0.0072, относительная погрешность 0.78%. В таблице 1 представлены результаты прогнозирования для различных вариантов алгоритма и при различных объёмах обучающего и дообучающего множеств.

Видно, что метод кластеризации FOREL с локальной нормировкой в сочетании с вторым методом дообучения показал наилучшие результаты, что, однако, компенсируется и наибольшим временем работы в данном случае. Примечательным фактом также является стабилизация (в первом методе) процента отбрасываемых кластеров, соответствующего минимальному значению ошибки прогнозирования, при росте объёма дообучающей выборки на уровне $10^6$ или выше.

Далее рассмотренный алгоритм был применён к совокупности родственных рядов, полученных зашумлением ряда Лоренца: таким образом моделировалось возможность переноса результатов кластеризации на родственные временные ряды. Здесь в качестве базового ряда использовали участок нормализированного ряда Лоренца длиной 100000, к которому добавлялся белый шум с последующей повторной нормализацией. Дисперсия шума была различной для различных рядов и была равна реализации нормально распределённой случайной величины со средним 0.7 и дисперсией 0.3. Всего было сгенерировано 4 ряда, которые использовались для формирования дообучающей выборки. Таким образом, объём обучающей выборки здесь составил $10^7$, дообучающей – вариировался от $10^4$ до $10^7$, тестирующей – $10^5$. Старший показатель Ляпунова для этих

рядов меняется в пределах от 0.98 до 1.23, при этом его среднее значение составило 1.14. Полученные результаты сведены в таблицу 2 и проиллюстрированы рисунками: на рис. 4 представлены зависимости среднеквадратичного уклонения и относительной погрешности в функции от объёма дообучающей выборки в логарифмическом масштабе (при фиксированном значении объёмов обучающей и тестирующей выборок).

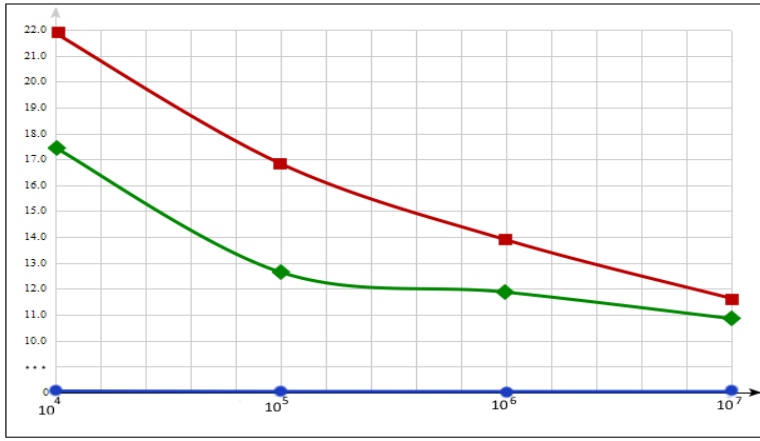


**Рис. 4. Зашумлённый ряд Лоренца. Кластеризация методом Уишарта. Линии с кружочками (соответствующими расчётным значениям) – «теоретический минимум», с квадратиками – первый метод оценки прогнозного качества кластеров, с ромбиками – второй метод**

*Таблица 2*

**Результаты для зашумлённого ряда Лоренца**

| *Count* | *Train* | *Normalize* | *Algo* | *PE* | *RMSE %* | *MAE* | *Non %* | *Min* |
|---|---|---|---|---|---|---|---|---|
| $10^5$ | $10^4$ | G | W | 1 | 21.87 | 0.1637 | 18.69 | 0.0551 |
| $10^5$ | $10^5$ | G | W | 1 | 16.83 | 0.1438 | 19.51 | 0.0405 |
| $10^5$ | $10^6$ | G | W | 1 | 13.89 | 0.0932 | 15.13 | 0.0374 |
| $10^5$ | $10^7$ | G | W | 1 | 11.63 | 0.0874 | 14.69 | 0.0359 |
| $10^5$ | $10^4$ | G | W | 2 | 17.45 | 0.1529 | 16.54 | 0.0413 |
| $10^5$ | $10^5$ | G | W | 2 | 12.64 | 0.1326 | 15.34 | 0.0387 |
| $10^5$ | $10^6$ | G | W | 2 | 11.87 | 0.0874 | 14.97 | 0.0368 |
| $10^5$ | $10^7$ | G | W | 2 | 10.85 | 0.0687 | 13.78 | 0.0354 |

Count – размер выборки обучения; Train – размер выборки дообучения; Normalize – способ нормализации; Algo – алгоритм кластеризации; PE (Post-education) – алгоритм оценки прогнозной ценности кластеров; RMSE – среднеквадратическая отклонение; MAE – относительная погрешность; Non – процент не прогнозируемых точек; Min – «теоретический минимум».

Здесь, лучшим показал себя вариант с алгоритмом Уишарта и заменой активного кластера. Стабилизация (для первого метода) процента отбрасываемых кластеров здесь не наблюдалась, что связано, вероятно, с разной природой обучающего и тестирующего множеств (незашумлённые и зашумлённые ряды Лоренца, соответственно).

Наконец, предложенный метод был применён к анализу и прогнозированию цен на электроэнергию в различных населённых пунктах Австралийского Содружества. Здесь на рис. 5 показано зависимость относительной погрешности от размера выборки дообучения и выбора метода дообучения.

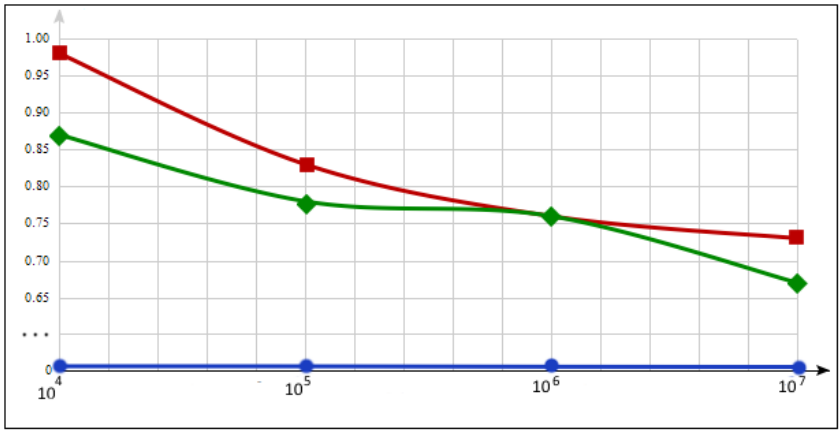


**Рис. 5. Энергетические ряды. RMSE в функции размера выборки дообучения. Линии с кружочками (соответствующими расчётным значениям) – «теоретический минимум», с квадратиками – первый метод оценки прогнозного качества кластеров, с ромбиками – второй метод**

**Сравнение с результатами других авторов**. В данном разделе представлено сравнение с результатами других авторов. В таблицах 4 и 5 представлено сравнение результатов, полученных различными методами: часть таблицы заимстовована из работы [21], см. также [8].

*Таблица 3*

**Результаты для энергетических рядов**

| *Count* | *Train* | *Norm alize* | *Algo* | *PE* | *RMSE %* | *MAE* | *Non %* | *Min* |
|---|---|---|---|---|---|---|---|---|
| $10^5$ | $10^4$ | G | W | 1 | 0.98 | 0.00701 | 0.35 | 0.00549 |
| $10^5$ | $10^5$ | G | W | 1 | 0.83 | 0.00662 | 0.29 | 0.00487 |
| $10^5$ | $10^6$ | G | W | 1 | 0.76 | 0.00627 | 0.25 | 0.00449 |
| $10^5$ | $10^7$ | G | W | 1 | 0.73 | 0.00617 | 0.17 | 0.00447 |
| $10^5$ | $10^4$ | G | W | 2 | 0.87 | 0.00674 | 0.37 | 0.00543 |
| $10^5$ | $10^5$ | G | W | 2 | 0.78 | 0.00631 | 0.34 | 0.00452 |
| $10^5$ | $10^6$ | G | W | 2 | 0.74 | 0.00623 | 0.29 | 0.00451 |
| $10^5$ | $10^7$ | G | W | 2 | 0.67 | 0.00608 | 0.23 | 0.00449 |

Count – размер выборки обучения; Train – размер выборки дообучения; Normalize – способ нормализации; Algo – алгоритм кластеризации; PE (Post-education) – алгоритм оценки прогнозной ценности кластеров; RMSE – среднеквадратическая отклонение; MAE – относительная погрешность; Non – процент не прогнозируемых точек; Min – «теоретический минимум».

Отметим, что уровень ошибки прогноза рассматриваемых в работе алгоритмов ниже соответствующего уровня для остальных методов, если исключить из рассмотрения точки, определяемые алгоритмом как непрогнозируемые, – а их число мало, и сравнима с указанными величинами, если заставить алгоритм давать прогноз в этих точках в принудительном порядке.

*Таблица 4*

**Прогнозирование для определенных дней 2004 года (Австралийские национальные рыночные цены на электроэнергию)**

| День | 5 Июня | 17 Июня | 20 Июня | 21 Июня | Среднее |
|---|---|---|---|---|---|
| ARIMA (%) | 32,31 | 29,09 | 33,73 | 24,18 | 29,82 |
| SVM(%) | 18,09 | 13,31 | 17,11 | 19,2 | 16,93 |
| PSF(%) | 16,72 | 8,31 | 14,23 | 18,93 | 14,55 |
| PCW(%) | 1,94 | 1,72 | 1,32 | 1,94 | 1,73 |
| PCW(1)(%) | 0,87 | 0,78 | 0,64 | 0,84 | 0,74 |
| PCW(2)(%) | 0,76 | 0,72 | 0,58 | 0,83 | 0,69 |

ARIMA – авторегрессивная интегрированная скользящая средняя; SVM – метод опорных векторов; PSF – модель на основе последовательности прогнозирования; PCW – прогнозирование на основе кластеризации с помощью алгоритма кластеризации Уишарта; PCW(1) – прогнозирование на основе кластеризации с помощью алгоритма кластеризации Уишарта с дообучением методом оценки прогнозного качества кластеров; PCW(2) – прогнозирование на основе кластеризации с помощью алгоритма кластеризации Уишарта с дообучением методом замены активного кластера.

*Таблица 5*

**Прогнозирование для определенных недель 2004 года (Австралийские национальные рыночные цены на электроэнергию)**

| Неделя | Вторая Января | Первая Июля | Первая Августа | Третья Декабря | Среднее |
|---|---|---|---|---|---|
| DWT(%) | 12,94 | 12,23 | 16,17 | 10,01 | 12,84 |
| SVM(%) | 23,37 | 15,03 | 36,18 | 33,74 | 27,08 |
| PSF(%) | 15,62 | 9,12 | 13,98 | 10,23 | 12,23 |
| PCW (%) | 1,33 | 1,47 | 1,28 | 1,11 | 1,30 |
| PCW(1) (%) | 0,96 | 0,78 | 0,83 | 0,62 | 0,76 |
| PCW(2) (%) | 0,89 | 0,81 | 0,74 | 0,59 | 0,72 |

DWT – дискретное волновое преобразование; SVM – метод опорных векторов; PSF – модель на основе последовательности прогнозирования; PCW – прогнозирование на основе кластеризации с помощью алгоритма кластеризации Уишарта; PCW(1) – прогнозирование на основе кластеризации с помощью алгоритма кластеризации Уишарта с дообучением методом оценки прогнозного качества кластеров; PCW(2) – прогнозирование на основе кластеризации с помощью алгоритма кластеризации Уишарта с дообучением методом замены активного кластера.

**Выводы.**

1. Применение подхода оценки прогнозного качества кластеров и удаления кластеров с низкой прогнозной ценностью позволяет существенно повысить качество прогноза как для модельных, так и для реальных данных в рамках парадигмы прогнозирования на основе кластеризации.

2. Проведенный широкомасштабный вычислительный эксперимент позволяет предположить, что с ростом объёма обучающей выборки составляющая ошибки прогнозирования, связанная с прогнозной моделью,

(при правильном выборе кластеров, с помощью которых осуществляется прогнозирование) стремится к нулю.

3. Аналогично, наблюдалось убывание с ростом дообучающего множества составляющей ошибки, связанной с неправильным выбором кластера, с помощью которого осуществляется прогноз и соответствующей ему субмодели прогнозирования.

4. Уровень ошибки прогноза рассматриваемых в работе алгоритмов ниже соответствующего уровня для методов машинного обучения, с которыми производилось сравнение, если исключить из рассмотрения точки, определяемые алгоритмом как непрогнозируемые, а их число мало, и остаётся одного порядка с указанными методами, если заставить алгоритм давать прогноз в этих точках в принудительном порядке.

**Библиографичиские ссылки**

1. **Aghabozorgi, S.** Wah Time-series clustering–A decade review [Text] / S. Aghabozorgi, A. S. Shirkhorshidi, T. Y. Wah // Information Systems. – 2015. – Vol. 53. – P. 16–38.

2. **Benítez, I.** Dynamic clustering of residential electricity consumption time series data based on Hausdorff distance Electric Power Systems Research [Text] / I. Benítez, J. L. Díezb, A. Quijanoa, I. Delgado // Online First. – 2016.

3. **Blockeel, H.** Top-down induction of clustering trees [Text] / H. Blockeel, L. De Raedt, J. Ramon // 15th international conference on machine learning. – 1998. – P. 55–63.

4. **D'Urso, P.** GARCH-based robust clustering of time series [Text] / P. D'Urso, L. De Giovanni, R. Massari // Fuzzy Sets and Systems.

5. **Elkin, E.A.** The Possibility of Application of Taxonomy Methods in Paleontology [Text] / E.A. Elkin, V.N. Elkina, N.G. Zagoruiko // Geology and Geophisics. – 1967. – Vol.9.

6. **Ferreira, L. N.** Time series clustering via community detection in networks [Text] / L.N. Ferreira, L. Zhao // Information Sciences. – 2016. – Vol. 326. – P. 227–242.

7. **Gromov, V. A.** Chaotic time series prediction with employment of ant colony optimization [Text] / V. A. Gromov, A. N. Shulga // Expert Systems with Applications. – 2012. – Vol. 39, № 9. – P. 8474–8478.

8. **Gromov, V. A.** Chaotic time series prediction and clustering methods [Text] / V. A. Gromov, E. A. Borisenko // Neural Computing and Applications. – 2015. – №. 2. – P. 307–315.

9. **Gromov, V. A.** Precocious identification of popular topics on Twitter with the employment of predictive clustering [Text] / V. A. Gromov, A. S. Konev // Neural Computing and Applications. – 2016. – Online First.

10. **Goodfellow, I.** Deep Learning [Text] / I. Goodfellow, Y. Bengio, A. Courville // MIT Press. – 2015. – P. 643.

11. **Huang, X.** Time Series k-Means: A New k-Means Type Smooth Subspace Clustering for Time Series Data [Text] / X. Huang, Y. Ye, L. Xiong, R.Y.K. Lau, N. Jiang, S. Wang // Information Sciences. – 2016.

12. **Izakian, H.** Agreement-based fuzzy c-means for clustering data with blocks of features [Text] / H. Izakian, W. Pedrycz // Neurocomputing. – 2014. – Vol. 127. – P. 266–280.

13. **Izakian, H.** Clustering spatiotemporal data: an augmented fuzzy c-means [Text] / H. Izakian, W. Pedrycz, I. Jamal // IEEE Trans. Fuzzy Syst. – 2013. – Vol. 21 (5). – P. 855–868.

14. **Jackson, E. A.** The Lorenz System: I. The Global Structure of its Stable Manifolds [Text] / E. A. Jackson. – Phys. Scr. – 1985. – Vol. 32. – P. 469–475.

15. **Kantz, H.** Nonlinear Time Series Analysis [Text] / H. Kantz, T. Schneider // Cambridge University Press. – 2004. – P. 388.

16. **Kattan, A.** Time-series event-based prediction: An unsupervised learning framework based on genetic programming [Text] / A. Kattan, S. Fatima, M. Arif // Information Sciences. – 2015. – Vol. 301. – P. 99–123.

17. **Keogh, E.** Clustering of time-series subsequences is mean-ingless: implications for previous and future research [Text] / E. Keogh, J. Lin // Knowledge and information systems. – 2005. – Vol. 8 (2). – P. 154–177.

18. **Konishi, S.** Information Criteria and Statistical Modeling [Text] / S. Konishi, G. Kitagava // Springer. – 2008. – P. 280.

19. **Lapko, A. V.** Nonparametric information processing systems [Text] / A.V. Lapko, S.V. Chentsov // Nauka. – 2000. – P. 350.

20. **Liao, T.W.** Clustering of time series data-a survey [Text] / T. W. Liao. – Pattern Recogn. – 2005. – Vol. 38 (11). – P. 1857–1874.

21. **Martınez-Alvarez, F.** Energy time series forecasting based on pattern sequence similarity [Text] / F. Martınez-Alvarez, A. Troncoso, J.C. Riquelme, J.M. Riquelme // IEEE Trans Knowl Data. – 2011. – Vol. 23 (8). – P. 1230–1243.

22. **Palit, A.K.** Computational intelligence in time series forecasting. Theory and engineering applications [Text] / A.K. Palit, D. Popovich // Springer. – 2005.

23. **Phu, L.** Motif-based method for initialization the K-means clustering for time series data [Text] / L. Phu, D.T. Anh // Springer. – 2011. – Vol. 7106. – P. 11–20.

24. **Widiputra, H.** A novel evolving clustering algorithm with polynomial regression for chaotic time-series prediction [Text] / H. Widiputra, H. Kho, R. Pears, N. Kasabov // Neural Inf Process. – 2009. – Vol. 5864. – P. 114–121.

25. **Wishart, D.** A numerical classification methods for deriving natural classes [Text] / D. Wishart. – Nature. – 1969. – Vol. 221, – P. 97–98.

26. **Zakaria, J.** Clustering time series using unsupervised-shapelets [Text] / J. Zakaria, A. Mueen, E. Keogh // 12th International Conference on Data Mining: IEEE Computer Society. – 2012. – P. 785–794.

*Надійшла до редколегії 29.07.2016*