Сравнительный анализ мер сходства, основанных на преобразовании скользящих аппроксимаций, в задачах классификации временных рядов

[1] *И.С. Алимова <alimovailseyar@gmail.com>*
[1] *В.Д. Соловьев <maki.solovyev@mail.ru>*
[2] *И.З. Батыршин <batyr1@gmail.com>*
[1] *Казанский федеральный университет,*
420008, Россия, г. Казань, ул. Кремлевская, д. 18
[2] *Национальный политехнический институт,*
CIC IPN, 0773, DF, Мехико, Мексика.

Аннотация. Одним из главных вопросов при решении задачи классификации временных рядов является выбор меры сходства рядов. В данной статье представлен сравнительный анализ меры сходства временных рядов, основанной на преобразовании скользящих аппроксимаций (САП трансформ), с двумя другими наиболее известными мерами: Алгоритмом Динамической Трансформации и Евклидовым расстоянием для задачи классификации. Кроме того, предложен алгоритм, улучшающий точность меры САП трансформ для временных рядов, имеющих схожие значения, но сдвинутых относительно друг друга по оси X, где координата на оси X представляет собой единицу времени.

Ключевые слова: временной ряд; классификация; мера сходства; САП трансформ; преобразование скользящих аппроксимаций.

DOI: 10.15514/ISPRAS-2016-28(6)-15

Для цитирования: Алимова И.С., Соловьев В.Д., Батыршин И.З. Сравнительный анализ мер сходства, основанных на преобразовании скользящих аппроксимаций, в задачах классификации временных рядов. Труды ИСП РАН, том 28, вып. 6, 2016 г., стр. 207-222. DOI: 10.15514/ISPRAS-2016-28(6)-15

1. Введение

В связи с непрекращающимся ростом данных, представленных в виде временных рядов, возрастает интерес к анализу временных рядов с целью извлечения новой полезной информации. Временной ряд представляет собой

упорядоченную последовательность данных, собранных в разные моменты времени. Временные ряды встречаются в области финансов, метеорологии, экономики, нефтедобыче, медицине и т.д. Примерами временного ряда являются показатели курсов валют, взятые за определенный период времени, кардиограмма человека, значения температуры воздуха окружающей среды в течении суток.

Основной задачей анализа временных рядов является выявление структуры ряда для прогнозирования его дальнейших значений. Прогноз будущих значений используется для принятия решений по дальнейшей работе системы. Например, определение дальнейшей динамики цен на акции компании может помочь в решении о покупке или продаже акций компании.

Одним из методов анализа временных рядов является метод классификации. Классификация временных рядов позволяет определить группы схожих по значению рядов с помощью различных мер сходства. В подобных задачах выбор меры сходства временных рядов влияет на точность классификации в большей степени, чем выбор метода классификации, что доказано на примере меры Алгоритма Динамической Трансформации [1]. В связи с этим, проводятся исследования по эффективности мер сходства временных для задачи классификации рядов [2, 3].

В работе [4] была представлена новая мера сходства временных рядов - САП трансформ. Данная мера применялась в области метеорологии для выявления взаимосвязи между загрязнением воздуха и метеорологическими показателями [5], а также в области нефтедобычи для задачи кластеризации нефтяных вышек [6]. Однако, не известны работы о сравнении меры САП с прочими мерами на существующих наборах временных рядов, а также не предложены способы по улучшению точности классификации меры САП трансформ. В связи с этим целью данной работы является сравнительный анализ меры САП трансформ с другими наиболее распространенными мерами сходства временных рядов, а также предложен один из способов модификации меры САП трансформ для улучшения точности.

В работе представлены результаты анализа точности трех мер сходства временных рядов на задаче классификации:

- Алгоритм динамической трансформации временной шкалы (АДТ).
- Евклидово расстояние.
- Преобразование скользящих аппроксимаций (САП трансформ) в двух вариациях: вычисленную для всех размеров скользящих окон и с модификацией, предложенной в этой работе.

В следующем разделе представлено обоснование выбора данных мер. В разделе 3 описаны формулы для расчета мер и используемый в данной работе алгоритм классификации временных рядов. Раздел 4 описывает проведенные эксперименты и наборы временных рядов на которых они проводились. В разделе 5 представлены результаты проведенного исследования.

2. Выбор мер сходства для анализа

В работе [2] представлены результаты оценки эффективности классификации 49 мер на наборе временных рядов из коллекции UCR [7]. Согласно полученным результатам, наиболее точной оказалась мера АДТ с ее модификациями. Мера АДТ [8] применяется в распознавании речи, жестов, рукописного текста, робототехнике, медицине, биоинформатике [9-12]. Преимуществом данной меры является способность распознавать схожие временные ряды даже если периоды рядов сдвинуты относительно друг друга по оси. Однако, данная мера имеет ряд недостатков. В частности, мера показывает неточные результаты для временных рядов, схожих между собой, но имеющих большую разницу максимальных и минимальных значений [13].

После меры АДТ наиболее точной мерой сравнения временных рядов оказалось Евклидово расстояние с его преобразованиями. Евклидово расстояние - это хорошо известная и наиболее простая для вычисления мера [14]. Основным ее преимуществом является маленькая вычислительная сложность, как по времени (O(n)), так и по памяти (O(1)). Однако, она уступает мере АДТ по точности классификации [2].

Мера САП трансформ основана на значении углов наклона касательных к графикам временных рядов в выбранной точке. Преимуществом данной меры является ее возможность определять как положительные зависимости, когда временные ряды одновременно увеличивают или уменьшают значения, так и отрицательные зависимости, когда значения одного временного ряда убывают, а другого возрастают и наоборот.

3. Метод классификации временных рядов

Задача классификации временных рядов связана с определением групп близких по значению временных рядов. Близость временных рядов определяется различными мерами сходства. В данном разделе представлены формулы вычисления выбранных для анализа мер сходства для временных рядов $x = (x_1, \dots x_m)$, $y = (y_1, \dots y_m)$ и алгоритм классификации.

3.1 Евклидово расстояние

Евклидово расстояние представляет из себя сумму расстояний между точками временного ряда и вычисляется по формуле:

$$d_{\text{Евкл}} = \sqrt{\sum_{i=0}^{m} |x_i - y_i|^2}$$

3.2 АДТ

В данной статье вычислялась классическая мера АДТ без модификаций. На первом этапе подсчитывалась матрица расстояний:

$$d_{ij} = |x_i - y_j|, i = 1, \dots, m, j = 1, \dots, m$$

На основе полученной матрицы вычислялась матрица деформаций:

$$D_{ij} = d_{ij} + \min(D_{i-1,j-1}, D_{i-1,j}, D_{i,j-1})$$

По матрице деформаций строился вектор трансформации W:

$$w_0 = D_{mm}$$

$$w_k = \min(D_{i-1,j-1}, D_{i-1,j}, D_{i,j-1}), k = 1, \dots, m$$

Итоговая мера АДТ вычислялась по формуле:

$$d_{\text{АДТ}} = \frac{1}{m}\sum_{i=1}^{m} w_i$$

3.3 САП

Мера САП вычислялась по формулам, описанным в [4]. Для подсчета данной меры сначала вычислялось значение локального тренда a_i:

$$a_i = \frac{6\sum_{j=0}^{k-1}(2j - k + 1)y_{i+j}}{hk(k^2 - 1)},$$

где k - размер выбранного окна, h - временной интервал (в нашем случае h=1).

После этого подсчитывалась мера ассоциаций для локальных трендов по формуле:

$$coss_k(X, Y) = \frac{\sum_{i=1}^{m} a_{yi} \cdot a_{xi}}{\sqrt{\sum_{i=1}^{m} a_{yi}^2 \cdot \sum_{j=1}^{m} a_{xj}^2}}$$

Это значение подсчитывалось для всех k = 2, ..., m. Среднее среди полученных значений является мерой САП трансформ:

$$AM(X, Y) = \frac{1}{|K|}\sum_{k=1}^{m-1} coss_k(X, Y)$$

Однако, точность базового алгоритма подсчета данной меры варьируется в зависимости от размера выбранного окна k и характеристик временных рядов. Сокращение количества и выбор наиболее оптимальных значений параметра k, а также модификации алгоритма позволили улучшить точность меры. Выбор оптимального значения параметра k и преобразование алгоритма описаны в следующем разделе.

3.4 САП с динамической трансформацией локальных трендов

В базовом варианте САП мера считается по всем возможным размерам окна, однако, сокращение их количества и выбор оптимальных размеров позволили улучшить точность меры. Кроме того, мера САП не приспособлена для выявления сходства между временными рядами, значения которых имеют схожую динамику, однако, сдвинуты относительно друг друга по оси X. С целью улучшения точности на подобных временных рядах были произведены модификации базовой меры САП трансформ, описанные в этом разделе.

Для подбора значений k временные ряды в обучающем файле были разделены на два подмножества случайным образом с одинаковым количеством рядов в каждом из них. Одно подмножество принималось за контрольную выборку, другое - за обучающее. Далее для каждого значения окна k <= 30 мы проводили классификацию обучающего множества по алгоритму, описанному ниже. Размер окна k, для которого точность имела максимальное значение, использовался при построении матрицы косинусов для классификации основного тестового множества временных рядов.

Для полученного на предыдущем этапе значения k вычислялись вектора локального тренда a_x и a_y. Для вычисленных векторов строилась матрица косинусов:

$$coss_{p,q}(a_x, a_y) = \frac{\sum_{i=1}^{m-\max\{p,q\}+1} a_{y_{i+q}} \cdot a_{x_{i+p}}}{\sqrt{\sum_{i=1}^{m-q+1} a_{y_{i+q}} \cdot \sum_{j=1}^{m-p+1} a_{x_{i+p}}}}$$

По матрице косинусов аналогично мере АДТ вычислялся вектор трансформации:

$$w_0 = coss_{mm}$$

$$w_k = w_{k-1} + \min\left(coss_{i-1,j-1}, coss_{i-1,j}, coss_{i,j-1}\right), k = 1, \dots, m$$

Итоговое значение меры вычислялось по формуле:

$$d_{\text{САП+ДТ}} = \sum_{i=1}^{m} w_i$$

Описанную в данном разделе меру в дальнейшем будем называть САП с динамической трансформацией локальных трендов (САП+ДТ).

3.5 Алгоритм классификации

Входные данные: T — тестовый набор временных рядов, S — обучающий набор данных, M — мера сходства
Результат: k — количество верно классифицированных рядов
k ← 0
Для каждого t ∈ T:
Находится s ∈ S такой, что $M(t,s)$ — минимальна
Если номера классов s и t совпадают:
k ← $k + 1$

Алгоритм 1. Алгоритм подсчета количества верно классифицированных рядов для заданной меры M.

Algorithm 1. The algorithm for counting the number of correctly classified time series for a given measure M.

Классификация временных рядов проходила по методу ближайшего соседа. Класс временного ряда выбирался следующим образом: для каждого временного ряда из тестового файла подбирался ряд из обучающего файла такой, что используемая мера для выбранных временных рядов была минимальна. Если номера классов выбранных рядов совпадали, считалось, что ряд классифицирован верно.

4. Эксперименты

Для оценки точности мер сходства временных рядов была проведена классификация временных рядов, где в качестве критерия для сравнения рядов применялись выбранные меры сходства. В данном разделе описаны наборы данных и критерии по которым проводилась оценка точности мер сходства временных рядов.

4.1 Набор данных

Мы использовали временные ряды из 43 наборов данных из коллекции UCR [7]. Коллекция содержит в себе как реальные данные, полученные в результате измерений, так и синтетические. Например, временные ряды в наборе данных GUN_POINT являются результатом измерения траектории движения центра ладони руки человека, когда он достает пистолет из кобуры на бедре, целится в мишень и кладет пистолет обратно. Временной ряд Synthetic Control получен генерированием различного вида графиков (циклических, возрастающих, убывающих и т.д.).

В табл. 1 представлена статистика по наборам, которые использованы в дальнейшем для количественного анализа мер. Подробная статистика по длине временных рядов, количеству классов и временных рядов в 43 наборах

представлена в [7]. Каждый набор данных содержит обучающий и тестовый файлы. В каждом из файлов записаны временные ряды с номером класса к которому они принадлежат.

Табл. 1. Статистика по некоторым наборам временных рядов из коллекции UCR.

Table 1. Statistics for some sets of time series of UCR collection.

Название	Количество классов	Кол-во рядов в обучающем файле	Кол-во рядов в тестовом файле	Длина рядов
50Words	50	450	455	270
Car	4	60	60	577
ECGFiveDays	2	23	861	136
FaceFour	4	24	88	350
FacesUCR	14	200	2050	131
Gun_Point	2	50	150	150
MoteStrain	2	20	1252	84
SonyAIBORobotSurface	2	20	601	70
SonyAIBORobotSurfaceII	2	27	953	65
Symbols	6	25	995	398
Trace	4	100	100	275
TwoLeadECG	2	23	1139	82
uWaveGestureLibrary_X	8	896	3582	315
WordsSynonyms	25	267	638	270

4.2 Качественный анализ классификации

Точность (Acc) классификации меры для набора данных подсчитывалась по формуле из [13]:

$$Acc = \frac{\mathrm{k}}{|\mathrm{T}|},$$

где k - количество верно классифицированных рядов, |T| - количество рядов в тестовом наборе данных.

После этого подсчитывалось среднее значение точности по всем наборам данных. Полученное число считалось точностью меры.

Для статистической оценки эффективности проведенных модификаций для меры САП трансформ был использован критерий Вилкоксона.

5. Результаты

5.1 Оценка точности мер сходства

В табл. 2 представлены макро усредненные оценки точности выбранных мер ассоциаций временных рядов, подсчитанные в данной работе на 43 наборах коллекции UCR. Наилучшие результаты согласно табл. 2 показали мера АДТ и Евклидово расстояние, следом идут САП + ДТ и САП.

Несмотря на то, что в целом меры САП трансформ уступают по точности Евклидову расстоянию и АДТ мере, предложенные модификации для меры САП позволили улучшить точность данной меры на 14 наборах данных из 43-х на которых проводились тесты. Кроме того, на 4 наборах данных из этих 14-ти мера САП+ДТ показывает большую точность, чем мера АДТ и Евклидово расстояние. Результаты для наборов данных представлены в табл. 3. Жирным шрифтом в табл. 3 выделены максимальные показатели точности для каждого набора данных.

Табл.2. Точность рассматриваемых мер по всем наборам коллекции, полученное макро усреднением.

Table 2. The average accuracies of the measures for all sets of the collection.

Мера	**Точность**
АДТ	0,91
Евклидово расстояние	0,9
САП + ДТ	0,88

САП	0,86

Для анализа вклада предложенной модификации, меры САП+ДТ и САП были подсчитаны улучшения точности меры САП+ДТ по отношению к САП в процентах. В табл. 3 полученные значения указаны в скобках в столбце для меры САП+ДТ. Согласно полученным значениям, проведенные модификации для меры САП улучшили ее точность в среднем на 9% для 14 наборов данных, на которых мера САП+ДТ превышает по точности меру САП.

Табл. 3. Значения точности рассматриваемых мер для наборов данных для временных рядов из коллекции UCR.

Table 3. The values of accuracy of the described measures for the data sets, on time series from UCR repository.

Набор данных	Евклидово расстояние	АДТ	САП	САП + ДТ
50Words	**0,8**	0,71	0,65	0,67 (+3%)
Car	**0,8**	0,73	0,65	0,67(+2%)
ECGFiveDays	0,81	0,77	0,68	**0,97** (+29%)
FaceFour	0,78	**0,84**	0,82	**0,84** (+2%)
FacesUCR	0,85	**0,94**	0,78	0,87 (+9%)
Gun_Point	0,93	0,87	0,84	**0,94** (+10%)
MoteStrain	0,88	**0,89**	0,83	0,86 (+3%)
SonyAIBORobotSurface	0,71	0,72	0,72	**0,78** (+6%)
SonyAIBORobotSurfaceII	0,87	0,85	0,85	**0,88** (+3%)
Symbols	0,9	**0,95**	0,88	0,94 (+6%)
Trace	0,8	**0,98**	0,71	0,95 (+24%)
TwoLeadECG	0,84	**0,97**	0,76	0,92 (+16%)
uWaveGestureLibrary_X	**0,79**	0,72	0,66	0,71(+5%)
WordsSynonyms	**0,72**	0,68	0,59	0,63(+4%)
Beef	1	1	1	1(0%)
Coffee	1	1	1	1(0%)
Cricket_X	1	1	1	1(0%)
Cricket_Y	1	1	1	1(0%)
Cricket_Z	1	1	1	1(0%)
ECG200	1	1	1	1(0%)

FISH	1	1	1	1(0%)
OliveOil	1	1	1	1(0%)
OSULeaf	1	1	1	1(0%)
Plane	1	1	1	1(0%)
SwedishLeaf	1	1	1	1(0%)
Herring	1	1	1	1(0%)
InsectWingbeatSound	1	1	1	1(0%)
BeetleFly	1	1	1	1(0%)
BirdChicken	1	1	1	1(0%)
PhalangesOutlineCorrect	1	1	1	1(0%)
ShapeletSim	1	1	1	1(0%)
ToeSegmentation1	1	1	1	1(0%)
ToeSegmentation2	1	1	1	1(0%)
Lighting2	1	1	1	1(0%)
wafer	1	1	1	1(0%)
FaceAll	**0,78**	0,77	0,74	0,74(0%)
ItalyPowerDemand	**0,99**	0,93	0,95	0,95(0%)
Adiac	**0,99**	0,59	0,45	0,31(-14%)
CBF	0,88	**1**	0,94	0,89(-5%)
DiatomSizeReduction	0,96	0,96	**0,97**	0,86(-11%)
Lighting7	0,6	**0,8**	0,53	0,49(-4%)
synthetic_control	0,94	**0,98**	0,88	0,87(-1%)
Two_Patterns	0,92	**1**	0,96	0,9(-6%)

Для анализа наборов данных, на которых мера САП улучшила точность, для каждого из них были построены графики. На каждом из них отображались данные двух временных рядов, принадлежащих одному классу из обучающего и тестового наборов данных.

На рис 1, 2 и 3 представлены графики временных рядов, принадлежащих одному классу, из обучающего и тестового наборов данных Gun_Point, ECGFiveDays и Symbols соответственно.

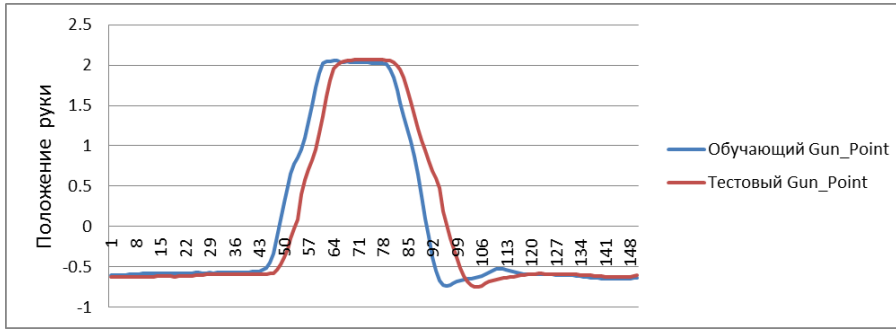

Рис. 1. График обучающего и тестового временных рядов из набора данных Gun_Point, описывающего положение руки при выстреле.

Fig. 1. The charts of the training and test sets of time series of GunPoint dataset describing the hand position in shooting.

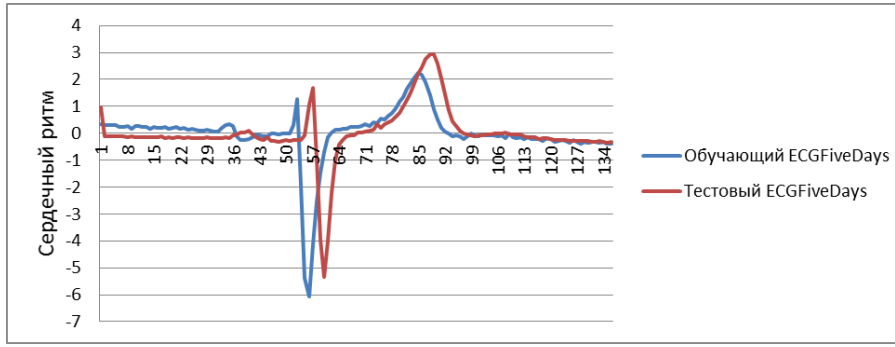

Рис. 2. График обучающего и тестового временных рядов из набора данных ECGFiveDays, describing сердечный ритм.

Fig. 2. The charts of the training and test sets of time series of ECGFiveDays dataset describing heart rate.

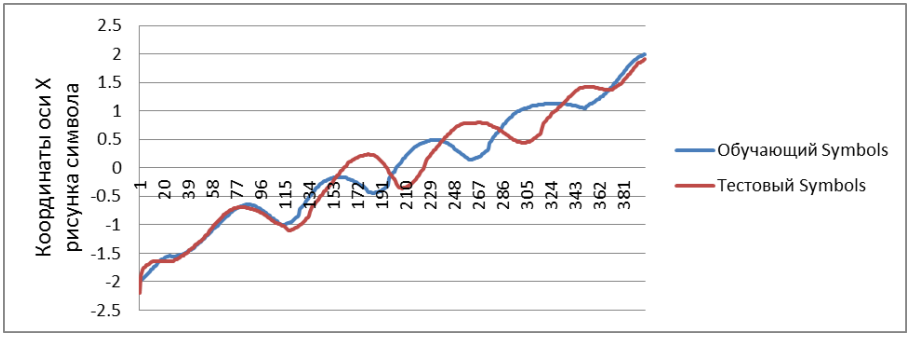

Рис. 3. График обучающего и тестового временных рядов из набора данных Symbols, описывающего координату X для рисунка символа.

Fig. 3. The charts of the training and test sets of time series of ECGFiveDays dataset describing the X-Axis motion in drawing the shape.

На рис. 1, 2 и 3 видно, что графики тестового и обучающего временных рядов имеют схожую форму, однако, сдвинуты относительно друг друга по оси X. Похожие закономерности наблюдаются для остальных 10-ти наборов данных из 14-ти, в табл. 3 они выделены курсивом. Для проверки гипотезы о том, что мера САП+ДТ улучшает точность, были выбраны дополнительно 10 рядов для которых наблюдается сдвиг по оси X относительно друг друга. В табл. 3 они так же выделены курсивом. Лишь на одном из выбранных наборов значение меры ухудшилось. Таким образом, мера САП+ДТ может быть применима при классификации рядов, для которых заранее известно, что принадлежащие к одному классу ряды имеют сдвиг относительно друг друга по оси X.

5.2 Статистическая оценка эффективности меры САП+ДТ

Для проверки гипотезы о том, что мера САП+ДТ имеет более высокую точность классификации временных рядов, была посчитана T статистика критерия Вилкоксона для значений точности мер САП и САП+ДТ. Наборы данных, для которых разность значений точности равна 0, были заранее исключены из рассмотрения. Таким образом, статистика считалась для 20 оставшихся наборов данных. В результате вычислений, было получено $T_{\text{эксп}} = 35$, при $T_{\text{крит}} = 43$ для $p = 0.01$. Поскольку $T_{\text{эксп}} < T_{\text{крит}}$, гипотеза о том, что мера САП+ДТ показывает более высокую точность, является достоверной. Исходя из полученных результатов, можно сделать вывод, что для задачи классификации временных рядов лучше подходит мера САП+ДТ, чем мера САП.

5.3 Оценка производительности алгоритмов

В табл. 4 приведены значения вычислительной сложности алгоритмов для вычисления мер сходства, рассматриваемых в данной статье. Наиболее быстрой для вычисления является Евклидово расстояние. Несмотря на то, что остальные меры имеют одинаковые оценки сложности все же мера САП вычисляется дольше, поскольку на каждом шаге алгоритма необходим проход по подпоследовательности длины k. Наиболее долгая по времени для вычислений мера САП+ДТ, так как для нее необходимы дополнительные вычисления для подбора оптимального значения длины окна k.

Табл.4.Оценки производительности алгоритмов для вычисления рассматриваемых мер сходства для временного ряда длины m.

Table 4. The estimates of productivity of the algorithms for calculating similarity measures for the time series of length m.

Мера	**Сложность**
Евклидово расстояние	$O(m)$
АДТ	$O(m^2)$
САП	$O(m^2)$
САП + ДТ	$O(m^2)$ (для фиксир. значения окна k)

6. Заключение

В данной статье проведен сравнительный анализ точности меры САП трансформ с мерами АДТ и Евклидовой на задаче классификации временных рядов. Помимо этого, предложен один из способов по улучшению точности меры САП для классификации рядов схожих по значению, однако смещенных относительно друг друга по оси X, где координата на оси X представляет собой единицу времени. Результаты исследования показали, что меры САП и САП+ДТ уступают по точности Евклидовой и АДТ мерам на наборах временных рядов различных предметных областей. Однако, для наборов временных рядов, у которых явно прослеживается сдвиг значений относительно оси X мера САП+ДТ превзошла значения точности меры САП, а для 4-х наборов данных значения точности меры САП+ДТ превосходит значения точности всех рассматриваемых в статье мер. Исходя из этого, можно сделать вывод, что для классификации временных рядов, обладающих подобным свойством, в качестве альтернативы можно рассматривать меру САП+ДТ.

7. Благодарности

Работа выполнена при финансовой поддержке проекта РФФИ 15-01-06456 и за счет средств субсидии, выделенной в рамках государственной поддержки Казанского (Приволжского) федерального университета в целях повышения его конкурентоспособности среди ведущих мировых научно-образовательных центров.

Авторы выражают благодарность Тутубалиной Елене Викторовне за помощь при подготовке статьи.

Список литературы

[1]. Weiss S. M. A novel approximation to dynamic time warping allows anytime clustering of massive time series datasets. Proceedings of the 2012 SIAM International Conference on Data Mining, 2012, pp. 999-1010. DOI: 10.1137/1.9781611972825.86.

[2]. Giusti R., Batista G. E. An empirical comparison of dissimilarity measures for time series classification. Intelligent Systems (BRACIS), 2013 Brazilian Conference on. – IEEE, 2013, pp. 82-88.

[3]. Ding H. et al. Querying and mining of time series data: experimental comparison of representations and distance measures. Proceedings of the VLDB Endowment, 2015, vol. 1, issue 2, pp. 1542-1552.

[4]. Batyrshin, I., Herrera-Avelar, R., Sheremetov, L., & Suarez, R. Moving approximations in time series data mining. Proc. Int. Conf. Fuzzy Sets and Soft Computing in Economics and Finance FSSCEF, 2004, pp. 62-72.

[5]. Almanza V., Batyrshin I. On trend association analysis of time series of atmospheric pollutants and meteorological variables in Mexico City Metropolitan Area. Mexican Conference on Pattern Recognition. Springer Berlin Heidelberg, 2011, pp. 95-102.

[6]. Батыршин И.З., Кошульски А., Шереметов Л.Б., Климова А.С., Панов А.М. Анализ взаимодействия нефтяных скважин на основе гибридной кластеризации временных рядов продуктивности скважин. Нечеткие системы и мягкие вычисления. Тверской государственный университет, том 2, вып. 4, 2007 г., стр. 63-73.

[7]. E. Keogh, X. Xi, L. Wei, and C. A. Ratanamahatana. (2006) The UCR time series classification/clustering homepage. Доступно по ссылке: http://www. cs. ucr. edu/~ eamonn/time series data.

[8]. M. Muller. Dynamic time warping. Inf. Retr. Music Motion. Information retrieval for music and motion. Springer, Berlin, 2007, pp. 69–84.

[9]. Lu G. et al. A novel framework of change-point detection for machine monitoring. Mechanical Systems and Signal Processing, 2017, vol. 83, pp. 533-548.

[10]. Rath T. M., Manmatha R. Word image matching using dynamic time warping. Computer Vision and Pattern Recognition. Proceedings IEEE Computer Society Conference on, 2003, vol. 2, pp. 521-527.

[11]. Muda L., Begam M., Elamvazuthi I. Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. Journal of Computing, 2010, vol. 2, issue 3, pp. 138-143.

[12]. Vakanski A. et al. Trajectory learning for robot programming by demonstration using hidden Markov model and dynamic time warping. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2012, vol. 42, issue 4, pp. 1039-1052.

[13]. Keogh E. J., Pazzani M. J. Derivative Dynamic Time Warping .Sdm, 2001, vol. 1, pp. 5-7.

[14]. Faloutsos C., Ranganathan M., Manolopoulos Y. Fast subsequence matching in time-series databases. Proceedings of the 1994 ACM SIGMOD international Conference on Management of Data, 1994, vol. 23, issue 2, pp. 419-429.

Comparative analysis of the similarity measures based on the moving approximation transformation in problems of time series classification

[1] *I.S. Alimova <alimovailseyar@gmail.com>*
[1] *V.D. Solovyev <maki.solovyev@mail.ru>*
[2] *I.Z. Batyrshin <batyr1@gmail.com>*
[1] *Kazan Federal University,*
18, Kremlyovskaya st., Kazan, 420008, Russia.
[2] *Instituto Politecnico Nacional,*
CIC IPN, 0773, DF, Mexico.

Abstract. One of the major issues dealing with time-series classification problem is the choice of similarity measure. This article presents a comparative analysis of the similarity measure for time series based on moving approximations transform (MAP transforms) with other two most useful measures: Algorithm Dynamic Transformation and Euclidean distance for classification task. In addition, algorithm, that improves the precision of the measure for time series, that have similar values, but shifted relative to each other on the axis X, where coordinate on the X axis represents the time unit, is proposed.

Key words: time series; classification; similarity measure; MAP transform; Moving Approximation Transform.

DOI: 10.15514/ISPRAS-2016-28(6)-15

For citation: Alimova I.S., Solovyev V.D., Batyrshin I.Z. Comparative analysis of the similarity measures based on the moving approximation transformation in problems of time series classification. Trudy ISP RAN/Proc. ISP RAS, vol. 28, issue 6, 2016, pp. 207-222 (in Russian). DOI: 10.15514/ISPRAS-2016-28(6)-15

References

[1]. Weiss S. M. A novel approximation to dynamic time warping allows anytime clustering of massive time series datasets. Proceedings of the 2012 SIAM International Conference on Data Mining, 2012, pp. 999-1010. DOI: 10.1137/1.9781611972825.86.

[2]. Giusti R., Batista G. E. An empirical comparison of dissimilarity measures for time series classification. Intelligent Systems (BRACIS), 2013 Brazilian Conference on. – IEEE, 2013, pp. 82-88.

[3]. Ding H. et al. Querying and mining of time series data: experimental comparison of representations and distance measures. Proceedings of the VLDB Endowment, 2015, vol. 1, issue 2, pp. 1542-1552.

[4]. Batyrshin, I., Herrera-Avelar, R., Sheremetov, L., & Suarez, R. Moving approximations in time series data mining. Proc. Int. Conf. Fuzzy Sets and Soft Computing in Economics and Finance FSSCEF, 2004, pp. 62-72.

[5]. Almanza V., Batyrshin I. On trend association analysis of time series of atmospheric pollutants and meteorological variables in Mexico City Metropolitan Area. Mexican Conference on Pattern Recognition. Springer Berlin Heidelberg, 2011, pp. 95-102.

[6]. Batyrshin I.Z., Koshulski A.1, Sheremetov L.B.2, Klimova A.S.3, Panova A.M.4. Oil wells interaction analysis based on hybrid clustering of wells productivity time series. Nechetkie sistemy i mjagkie vychislenija [Fuzzy Systems and Soft Computations]. Tverskoj gosudarstvennyj universitet [Tver State University], 2007, vol. 2, issue 4, pp. 63-73 (in Russian).

[7]. E. Keogh, X. Xi, L. Wei, and C. A. Ratanamahatana. (2006) The UCR time series classification/clustering homepage. Available at http://www.cs.ucr.edu/~eamonn/time_series_data.

[8]. M. Muller. Dynamic time warping. Inf. Retr. Music Motion. Information retrieval for music and motion. Springer, Berlin, 2007, pp. 69–84.

[9]. Lu G. et al. A novel framework of change-point detection for machine monitoring. Mechanical Systems and Signal Processing, 2017, vol. 83, pp. 533-548.

[10]. Rath T. M., Manmatha R. Word image matching using dynamic time warping. Computer Vision and Pattern Recognition. Proceedings IEEE Computer Society Conference on, 2003, vol. 2, pp. 521-527.

[11]. Muda L., Begam M., Elamvazuthi I. Voice recognition algorithms using mel frequency cepstral coefficient (MFCC) and dynamic time warping (DTW) techniques. Journal of Computing, 2010, vol. 2, issue 3, pp. 138-143.

[12]. Vakanski A. et al. Trajectory learning for robot programming by demonstration using hidden Markov model and dynamic time warping. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), 2012, vol. 42, issue 4, pp. 1039-1052.

[13]. Keogh E. J., Pazzani M. J. Derivative Dynamic Time Warping .Sdm, 2001, vol. 1, pp. 5-7.

[14]. Faloutsos C., Ranganathan M., Manolopoulos Y. Fast subsequence matching in time-series databases. Proceedings of the 1994 ACM SIGMOD international Conference on Management of Data, 1994, vol. 23, issue 2, pp. 419-429.