Financial Time Series: Motif Discovery and Analysis Using VALMOD

Eoin Cartwright(✉), Martin Crane, and Heather J. Ruskin

Centre for Advanced Research Computing & Complex Systems Modelling (ARC-SYM), School of Computing, Dublin City University, Dublin 9, Ireland
eoin.cartwright3@mail.dcu.ie
https://www.dcu.ie/arcsym/index.shtml

Abstract. Motif discovery and analysis in time series data-sets have a wide-range of applications from genomics to finance. In consequence, development and critical evaluation of these algorithms is required with the focus not just detection but rather evaluation and interpretation of overall significance. Our focus here is the specific algorithm, *VALMOD*, but algorithms in wide use for motif discovery are summarised and briefly compared, as well as typical evaluation methods with strengths. Additionally, Taxonomy diagrams for motif discovery and evaluation techniques are constructed to illustrate the relationship between different approaches as well as inter-dependencies. Finally evaluation measures based upon results obtained from *VALMOD* analysis of a GBP-USD foreign exchange (F/X) rate data-set are presented, in illustration.

Keywords: Motif analysis · Motif evaluation · *VALMOD*

1 Introduction

Sequential data are found in many applications ranging from Healthcare [1] to Seismology [2], Machine Learning [3] and Finance [4]. Recurrent patterns (*motifs*) are common and can occur both within and between individual time series [5]. Motif identification can help pre-processing in other high level data mining tasks e.g. time series clustering and classification, rule discovery and summarisation [6]. Similarly, evaluating motif content can aid better understanding of the underlying processes generating data in a given domain.

A brief summary of state-of-the art in motif discovery and evaluation follows, with strengths and limitations indicated. Analysis of motif contributions is non-trivial, and combined analyses are usually required, depending on data features. In prior work [7], we compared two popular methods (*MrMotif* and *VALMOD*), the latter particularly suited to variable motif length analysis. Interest here is on an early evaluation of motif results from *VALMOD* for a GBP-USD F/X rate data-set.

2 Motif Discovery Techniques: Summary

Although a suite of motif discovery techniques are available (illustrated, Fig. 1), two principal approaches form the basis for many applications in the literature. These

J. M. F. Rodrigues et al. (Eds.): ICCS 2019, LNCS 11540, pp. 771–778, 2019.
https://doi.org/10.1007/978-3-030-22750-0_77

are, (i) The *CK* algorithm [8] which uses Random Projection (*RP*) to create a collision matrix and (ii) Symbolic Aggregate Approximation (*SAX*) which utilises Piecewise Aggregate Approximation (*PAA*) with breakpoints and symbolisation to discretize sequential data into a symbolic string [9].

A Taxonomy diagram of type and inter-relationships is provided, Fig. 1. Following the recent survey paper by [10], this is designed to aid analysis and interpretation. The majority of techniques apply to the time domain, and rely on *approximation* of the time-series to provide motif candidates within a suitable timeframe. In comparison, relatively few frequency-based approaches to motif discovery appear in the literature, with one notable exception, the *SIMD* [11] algorithm, which uses a *Wavelet-based* approach.

Given issues such as an initial word (or motif) target length requirement for a Brute Force (*BF*) approach and the computational expense involved, series approximation using *SAX* was considered initially. This offers significant advantages as it allows utilisation of string analysis techniques for motif detection, borrowed from the study of DNA sequences. Notable algorithms in this domain are *MrMotif* [12], an algorithm which examines *SAX* at increasing resolutions and *SEQUITUR* [13] which implements a grammar-based approach on these symbolic strings.

Initial limitations for the exact approach were overcome by use of *Brute Force* (*BF*) in combination with early abandonment, allowing motif identification in a linear timeframe (*SBF*). In consequence, the *MK* algorithm [5] has underpinned many extensions including *top-k* [14] and Variable Length Motif Discovery (*VLMD*), [15]. A twofold improvement in performance compared to *SBF* was offered by *Quick-Motif* [16] with preference shifting towards a deterministic approach to motif discovery. More recently still, performance improvements and increased scalability have been achieved through a series of algorithms based on approximation for the Matrix Profile technique: (examples include *STAMP* [24], *STOMP* [25] & *VALMOD* [26]).

A summary table of techniques and algorithms from the literature is given, Table 1. A similar split between exact and approximate methods, as noted already for motif discovery, (Fig. 1), is evident here also, with most algorithms reliant on some form of data discretisation, (notably *iSAX*).

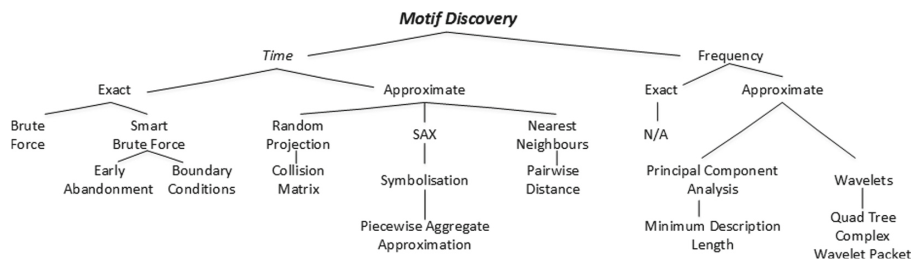

Fig. 1. Motif Discovery Technique Taxonomy: principal techniques and their interdependencies are shown here.

Table 1. Named motif discovery algorithm & model techniques: main algorithm techniques & features are illustrated. Attributes & applications are also given.

Year	Time/ Frequency	Exact Approx	Algorithm Named	Full Name	Technique Used	Algorithm Characteristics
2002	T	Approx	EMMA	**E**numeration of **M**otifs through **M**atrix **A**pproximation	Piecewise AggregateApprox (PAA) & Breakpoints to create SAX	First to address problem of repeated patterns in time series. [9]
2002	T	Approx	PERUSE	**P**attern **E**xtraction from **R**eal-valued sequences **U**sing **E**xpectation maximization	Expectation Maximization algorithm	Looks for a known pattern. [17]
2003	T	Approx	PROJECTION/CK	**C**hiu & **K**eogh Algorithm	Random Projection	Built on Projection Alg.
					Collision Matrix	Can have "Don't Care" sections for noisy data. [8]
2009	T	Exact	MK	**M**ueen & **K**eogh Algorithm	Brute Force (BF) with early abandoning	First with exact solution in reasonable time vis-à-vis BF [5]
2009	T	Exact	DAME	**D**isk **A**ware **M**otif **E**numeration	Bottom-up search on memory blocks of an "order line". While using a linear bound to prune off distance computations	"order line": created by computing distance of each point from random point & sort all such in ascending order. Applicable to large data sets.[18]
2010	T	Approx	MrMotif		Symbolic Aggregate Approximation (SAX)	Uses iSAX & Space Saving (SS) algorithm. Starting from low iSAX resolution & refined to higher. [12]
2010	T	Approx	AMG	**A**daptive **M**otif **G**eneration	Vector Suffix Tree Time Lag Matrix	Candidate motifs are generated then refined to detect desired patterns using a Time lag Matrix.[19]
2011	T	Exact	k-Motif		Generalized MK Algorithm	MK corresponds to case of finding a single motif (i.e. $k = 1$). [14]
2011	T	Approx	VLMD	**V**ariable **L**ength **M**otif **D**iscovery Algorithm	SAX	Obviates predefined window length value. Returns exact solution to variable-length motif discovery problem. [15]
					MK	
					Sliding Window	
2012	T	Approx	kBMD	**K**-**B**est **M**otif **D**iscovery	Extension of VLMD Algorithm	k-Best Motif Discovery (kBMD). Extension of VMLD above. [20]
2013	T	Exact	MOEN	**Mo**tif **En**umeration	Smart Brute Force (SBF)	Adjustment of boundary conditions. [21]
2013	T	Approx	PLMD	**P**roper **L**ength **M**otif **D**iscovery	kBMD	Continuation of kBMD above.
					VMLD	More efficient by early termination. [22]
2015	T	Approx	MDLats	**M**otif **D**iscovery method for **La**rge-scale **T**ime **S**eries	SAX	Like other discovery methods' structure but uses Hadoop for scalability. [1]
					Random Projection	
					Euclidian distance	
					DTW	
2015	F	Approx	SIMD	**S**hift **I**nvariant Feature Extraction for **M**otif **D**iscovery	Quad Tree-Complex Wavelet Packet (QT-CWP)	Applied to time & scale shifted data. [11]
2015	T	Exact	Quick-Motif		SBF	Minimum 2x faster than MK. Also scalable. [16]
					MK	
					Hilbert R-tree	
					Min bounding rectangle (MBR) pairs	
2016	T	Approx	MDM	**M**ulti-**D**imensional **M**otif	SAX	Applied to multidimensional data. [23]
					SEQUITUR	
2016	T	Approx	STAMP	**S**calable **T**ime series **A**nytime **M**atrix **P**rofile	Matrix Profile	Creates two meta time series, the matrix profile and the matrix profile index.
					Matrix Profile Index	These two data objects explicitly or implicitly contain the answers to many data mining tasks. [24]
2017	T	Approx	STOMP	**S**calable **T**ime series **O**rdered-search **M**atrix **P**rofile	STAMP	Extension of STAMP to large datasets. STAMP evaluates distance profiles in a random order while STOMP performs an ordered search. GPU unit added to improve performance [25]
2018	T	Approx	VALMOD	**V**ariable **L**ength **Mo**tif **D**iscovery	Matrix Profile	Solution returning all motifs within a given range of lengths. [26]
					Lower Bound Distance Profile	

3 Motif Evaluation Techniques Summary

In assessing importance of a given motif (or motif set) some measures are calculated exclusively based on the pattern's information-content, while others are based on how these relate to the underlying data within which they appear. Three main approaches to motif evaluation were proposed in [27]. These are: *Class-Based*(*CBM*), *Theoretical Information*(*TIM*) and *Mixed Measures*(*MM*).

CBM measures do not rely upon motif structure, but on the number of occurrences in a given category. Hence, they are applicable to any deterministic motif, whereas *TIM* measures have a probabilistic basis and *MM* a combination of both. We show a Taxonomy for motif evaluation based on examples in the literature, (Fig. 2).

To date *CBM* and *TIM* measures predominate, typically rated on the basis of *Discrimination Power* and explained variability (*F-Ratio*). Achieving a meaningful evaluation, of motif occurrence and importance, generally requires statistical inference from more than one complementary technique as well as flexible treatment of mis-(or partial) matches and identification.

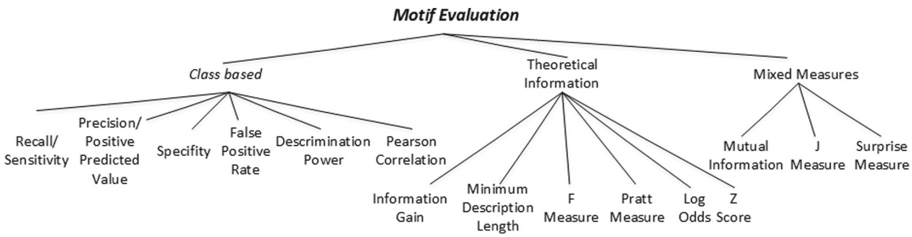

Fig. 2. Motif Evaluation Technique Taxonomy: principal techniques and their group-dependencies.

The following sub-sections outline main motif evaluation approaches with an initial application given in Sect. 4.

Class-Based Measures (*CBM*)
The ideal (or *signature*) motif [28] matches all sequences within a target family and does not overlap with any sequences outside it. As the ideal occurs rarely however the motif *quality* is illustrated by other measures: e.g. for *CBM*, these are usually *Sensitivity*, *Specificity* and *Positive Predicted Value*(*PPV*), based on comparison possibilities for a given sequence and target family, Table 2.

Sensitivity (S) is the proportion of the target family within a data-set correctly (i.e. exactly) matched by a motif. Specificity (S_p or *Recall*) indicates non-matches while Positive Predicted Value (*PPV* or *Precision*) is the percentage of data correctly matched by a motif and also belonging to the target family: formulae see [27].

Table 2. Class-based measures (*CBM*) motif comparison possibilities.

	Target family	Not target family
Matches motif	True positive (T_p)	False positive (F_p)
Does not match motif	False negative (F_n)	True negative (T_n)

A signature motif requires both *Sensitivity* and *Positive Predicted Value* = 100%. Other notable measures include the *F-Ratio* for overall quality of match and *Discrimination Power* which provides an indication of the rarity of a given pattern.

Theoretical Information Measures (*TIM*)
These measures analyse the degree and nature of information encoded in a motif. The principle of *Minimum Description Length* (*MDL*) is used to rank motifs, assuming the best is the minimum length possible (thus *reducing* the overall series length when re-encoded). *MDL* can also be used in the detection of motifs with 'wobble' (or inexact match).

Common statistical techniques such as the *Z-Score*, (based on Gaussian assumptions), can be used to identify functionally important regions within a data set and as an initial pruning mechanism before other significance measures are calculated. In determining incidence of unexpected motifs, the *Log-Odds* calculates the probability of occurrence in relation to a given distribution, e.g. *Binomial*, *Uniform* or other. Commonly, either *Bernoulli* or *Markov* models are used for motif symbol counts, depending on whether those symbols occurring within a sequence are independent or conditional.

Another useful *TIM* is the *Pratt* measure, used to rank motifs when 'flexible gaps' are permitted in symbol content. A two-step approach applies, whereby information is first encoded by the motif, then a penalty factor is introduced when gaps occur.

Hybrid (or Mixed) Measures (*MM*)
For *MM*, *Class-Based* and *Theoretical Information* measures are combined to gain a better appreciation of a motif's functional significance within a given data-set. Numerous occurrences within a data-set of a given motif do not necessarily imply importance, while a functionally significant motif may occur infrequently but still contain valuable information. *MM* techniques include *Mutual Information*, the *J-Measure* and the *Surprise (or S-)measure* [27].

4 Motif Evaluation Examples

We briefly illustrate points from Sects. 2 and 3 with reference to Financial data from [29]. A GBP vs USD daily F/X series provides input for *Mr Motif*, *SBF*, *Mueen Keogh* and *VALMOD* algorithms with motif data set location for the same motif length criteria the objective (Table 3). Similar motif locations are returned (even for small sample size) so that algorithm features best suited to the application can guide tool choice.

Table 3. *MrMotif*, *Quick Motif*, *Smart Brute Force*, *Mueen Keogh* & *VALMOD* Compared

Motif length		100		150		200	
FX series	Algorithm	Execution time (s)	Dataset location	Execution time (s)	Dataset location	Execution time (s)	Dataset location
GBP V USD	MrMotif	0.123	41, 1921, 2161, 2461, 2601, 4181 2461, 2601, 4181	0.125	1081, 1591, 2251	0.122	161, 201, 601
	Quick Motif	0.111	1751, 2584	0.093	3039, 1701	0.05	1045, 244
	Smart Brute Force	0.247	1751, 2584	0.243	3039, 1701	0.269	244, 1045
	Mueen Keogh	0.171	1751, 2584	0.163	1701, 3039	0.154	244, 1045
	VALMOD (single length)	0.484	2585, 1752	0.468	3040, 1702	0.437	1046, 245

The *VALMOD* algorithm was chosen for further tests due to its ability to parse a *user-provided* range of lengths. *VALMOD* source code was amended to return a complete set of candidate motifs for given length, serving as input for a bespoke application written in C#. The original series can be displayed, *VALMOD* criteria chosen and motif evaluation measures applied to the discovered motif set, as shown, (Fig. 3).

Fig. 3. Sample *VALMOD* motif results analysis of GBP vs USD FX dataset (Motif Length 100, Z-Score technique)

The user can choose a target family, based on an area of interest in the series, allowing T_p, F_p, T_n & F_n values with corresponding formulae to be calculated. Similarly motif locations can be shown within the data set and ordered by *Z-Score*, (Fig. 3).

5 Conclusions and Future Work

The growing importance of identifying repeated sub-sections or motifs in sequential data is outlined. Taxonomy diagrams illustrating motif discovery and evaluation techniques are provided while state of the art discovery algorithms are listed and characterised. The *VALMOD* algorithm is found to provide a sound basis for evaluation of motif occurrence in a financial data-set and examples are provided, indicative of its potential as an investigative tool in this context.

Clearly desirable for the future however, is an implementation of *SAX*, permitting refinement of the MDL principle applied to motif ranking and analysis, (with 'wobble' or less precise matching), as well as to discovery of common motifs over multiple series.

References

1. Liu, B., Li, J., Chen, C., Tan, W., Chen, Q., Zhou, M.: Efficient motif discovery for large-scale time series in healthcare. IEEE Trans. Ind. Inform. **11**, 583–590 (2015)
2. Cassisi, C., et al.: Motif discovery on seismic amplitude time series: the case study of Mt Etna 2011 eruptive activity. Pure Appl. Geophys. **170**, 529–545 (2012)
3. Lecun, Y., Bengio, Y., Hinton, G.: Deep learning. Nature **521**, 436–444 (2015)
4. Guan, Q., An, H., Liu, N., An, F., Jiang, M.: Information connections among multiple investors: evolutionary local patterns revealed by motifs. Sci. Rep. **7**, 14034 (2017)
5. Mueen, A., Keogh, E., Zhu, Q., Cash, S., Westover, B.: Exact discovery of time series motifs. In: Proceedings of 2009 SIAM International Conference on Data Mining (2009)
6. Son, N., Anh, D.: Discovery of time series k-motifs based on multidimensional index. Knowl. Inf. Syst. **46**, 59–86 (2015)
7. Cartwright, E., Crane, M., Ruskin, H.J.: Abstract: Motif Discovery & Evaluation Focus on Finance. https://sites.google.com/view/econophysics-colloquium-2018
8. Chiu, B., Keogh, E., Lonardi, S.: Probabilistic discovery of time series motifs. In: Proceedings of 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (2003)
9. Lin, J., Keogh, E., Lonardi, S., Patel, P.: Finding motifs in time series. In: Proceedings of 2nd Workshop on Temporal Data Mining, pp. 53–68 (2002)
10. Torkamani, S., Lohweg, V.: Survey on time series motif discovery. Wiley Interdiscip. Rev.: Data Min. Knowl. Disc. **7**, e1199 (2017)
11. Torkamani, S., Lohweg, V.: Shift-Invariant Feature Extraction for Time-Series Motif Discovery (2015)
12. Castro, N., Azevedo, P.: Multiresolution motif discovery in time series. In: Proceedings of 2010 SIAM International Conference on Data Mining (2010)
13. Nevill-Manning, C., Witten, I.: Identifying hierarchical structure in sequences: a linear-time algorithm. J. Artif. Intell. Resol. **7**, 67–82 (1997)
14. Lam, H., Pham, N., Calders, T.: Online discovery of top-k similar motifs in time series data. In: Proceedings of 2011 SIAM International Conference on Data Mining (2011)

15. Nunthanid, P., Niennattrakul, V., Ratanamahatana, C.: Discovery of variable length time series motif. In: The 8th (ECTI) Association of Thailand (2011)
16. Li, Y., U, L., Yiu, M., Gong, Z.: Quick-motif: an efficient and scalable framework for exact motif discovery. In: 2015 IEEE Proceedings of International Conference on Data (2015)
17. Oates, T.: PERUSE: an unsupervised algorithm for finding recurring patterns in time series. In: 2002 IEEE Data Mining, Proceedings (2002)
18. Mueen, A., Keogh, E., Bigdely-Shamlo, N.: Finding time series motifs in disk-resident data. In: 2009 IEEE Data Mining (2009)
19. Wang, L., Chng, E., Li, H.: A tree-construction search approach for multivariate time series motifs discovery. Pattern Recogn. Lett. **31**, 869–875 (2010)
20. Nunthanid, P., Niennattrakul, V., Ratanamahatana, C.: Parameter-free motif discovery for time series data. In: 9th (ECTI) Association of Thailand (2012)
21. Mueen, A., Chavoshi, N.: Enumeration of time series motifs of all lengths. Knowl. Inf. Syst. **45**, 105–132 (2014)
22. Yingchareonthawornchai, S., Sivaraks, H., Rakthanmanon, T., Ratanamahatana, C.: Efficient proper length time series motif discovery. In: IEEE Data Mining (2013)
23. Balasubramanian, A., Wang, J., Prabhakaran, B.: Discovering multidimensional motifs in physiological signals for personalized healthcare. IEEE J-STSP **10**(5), 832–841 (2016)
24. Yeh, C., et al.: Time series joins, motifs, discords and shapelets: a unifying view that exploits the matrix profile. Data Min. Knowl. Disc. **32**, 83–123 (2017)
25. Zhu, Y., et al.: Exploiting a novel algorithm and GPUs to break the ten quadrillion pairwise comparisons barrier for time series motifs and joins. Knowl. Inf. Syst. **54**, 203–236 (2017)
26. Linardi, M., Zhu, Y., Palpanas, T., Keogh, E.: Matrix profile X. In: Proceedings of 2018 International Conference on Management of Data - SIGMOD 2018 (2018)
27. Ferreira, P., Azevedo, P.: Evaluating deterministic motif significance measures in protein databases. Algorithms Mol. Biol. **2**, 1–20 (2007)
28. Jonassen, I., Collins, J., Higgins, D.: Finding flexible patterns in unaligned protein sequences. Protein Sci. **4**, 1587–1595 (1995)
29. Time Series Data Library - Data provider – DataMarket. https://datamarket.com/data/list/?q=provider:tsdl