Advanced Automated Pattern Search in Industrial Sensor Data

MASTERARBEIT

ZUR ERLANGUNG DES AKADEMISCHEN GRADES
MASTER OF SCIENCE IN ENGINEERING

DER
FACHHOCHSCHULE FH CAMPUS WIEN
MASTER-STUDIENGANG BIOINFORMATIK

Vorgelegt von:
Dr. Gerhard Bilek
Personenkennzeichen: 1810542015

FH-Hauptbetreuer*in:
Ass.Prof.in Dipl.-Ing.in Dr.in Alexandra Posekany

Abstract

German Abstract

Industrieprozesse nutzen zeitlich geordnete Daten um die Produktionsleistung und Produktqualität zu überwachen. Diese Daten werden als Zeitserien dargestellt und ermöglichen somit eine Nachverfolgung qualitätsrelevanter Signalmuster über weite Zeiträume. Hohe Qualitätsstandards erfordern eine hohe Datenpunktdichte und eine lückenlose Prozessüberwachung. Es ist daher eine enorme Herausforderung, alle relevanten Kriterien in Zeitserien zu identifizieren und zu extrahieren. Einige Muster ändern ihre Form je nach Produktionsstufe oder verwendetem Aufzeichnungsinstrument. Trotzdem müssen sie alle identifiziert werden.

Zur Bewältigung dieser Data-Mining Aufgabe, wurden zwei vielversprechende Similarity-Search Algorithmen mit Hilfe einer State-of-the-Art Analyse identfiziert und anschließend im Hinblick auf die Erweiterung des aktuellen Suchprogramms, das vom *VRVis Zentrum für Virtual Reality und Visualisierung Forschungs-GmbH* entwickelt wurde, getestet. Diese beiden Algorithmen verkörpern zwei komplett unterschiedliche Prinzipien. MASS (Mueen's ultra-fast Algorithm for Similarity Search) ist ein profilbasierter Ansatz und SAX (Symbolic Aggregate approXimation) nützt eine symbolische Repräsentation.

Diese Arbeit konnte zeigen, dass beide Methoden unterschiedlichste Data-Mining Aufgaben erfolgreich lösen können. Auch Muster mit unterschiedlichen Ausprägungsformen konnten mit hoher Sensitivität detektiert werden. Lediglich die Art der Datensätze und die Parameterwahl mussten für den Erfolg berücksichtigt werden. MASS arbeitet fast parameterfrei und liefert exakte Lösungen. Daher wurde dieser Algorithmus als die bessere Wahl betrachtet, um die bestehende Suchsoftware zu erweitern.

English Abstract

Industrial processes use temporally ordered data to monitor production performance and product quality. This data is visualized in time series in order to trace quality-relevant signal patterns over long time ranges.

Higher quality standards require a high data-point density and a consistent process monitoring. Therefore, it become more challenging to identify and extract all quality relevant information from different time series. Target patterns might even change their shape depending on process step or sensor type. Yet, the goal is the detection of all relevant appearances.

For this data-mining challenge, two similarity search algorithms were identified in a state-of-the-art research and subsequently examined with the goal to enhance an existing similarity search tool of the *VRVis Zentrum für Virtual Reality und Visualisierung Forschungs-GmbH*. The selected algorithms cover two different approaches. MASS (Mueen's ultra-fast Algorithm for Similarity Search) represents a profile-based approach, whereas SAX (Symbolic Aggregate approXimation) is a symbolic representation.

This work demonstrated that both search methods can be successfully applied for various data mining tasks. They have the potential to sensitively detect patterns of different occurrence. However, the detection quality depends highly on the nature of data and appropriately adjusted parameters. Since MASS requires fewer parameters and returns exact solutions, it was found to be the better choice to extend the existing tool.

Contents

1. Introduction **8**
1.1. Problem statement . . . 8
1.2. VRVis . . . 9
1.3. Motivation to improve similarity search technology . . . 10
1.4. Aim . . . 12

2. State of the Art **15**
2.1. Similarity measures . . . 15
2.2. Data mining approaches in time series . . . 19
2.3. The profile-based approach . . . 21
2.4. Symbolic representations . . . 25

3. Materials and Methods **28**
3.1. Visplore . . . 28
3.2. Python prototyping . . . 29
3.2.1. Used data sets . . . 31
3.2.2. MASS: Basics of the algorithm . . . 32
3.2.3. SAX: Basics of the algorithm . . . 32
3.3. The evaluation system . . . 33

4. Results **37**
4.1. Parameter overview . . . 37
4.1.1. Identification of MASS key parameters . . . 38
4.1.2. Identification of SAX key parameters . . . 41
4.2. Comparison: Visplore, MASS and SAX . . . 43
4.3. Permissive search: Tolerance for temporal distortions . . . 48
4.4. Domain search: A query with gap . . . 53

5. Discussion **58**
5.1. Comparison of the result quality between profile-based and symbolic methods . . . 58
5.1.1. Permissive Search: Benefits for MASS and SAX . . . 60
5.1.2. Domain Search: Distance profiles enable new perspectives . . . 63
5.2. Recommendation: User experience and customer requirements . . . 64

6. Conclusion **66**

7. Glossary **68**

A. Data import **72**

B. MASS code **73**
B.1. MASS: Core functions . . . 73
B.2. Permissive MASS: Peak widening function . . . 73
B.3. Domain MASS: Simultaneous MASS . . . 76
B.4. Domain MASS: Key functions of sequential MASS . . . 77

C. SAX code 80
C.1. SAX: Core functions 80
C.2. Jupyter Notebook: SAX basics 80

List of Figures

1. Example of a current similarity search result 11
2. Idea behind permissive and domain search 13
3. Principle of the Euclidean distance in time series analysis 15
4. Illustration of Euclidean distance and dynamic time warping matching . 17
5. Relation of time series and distance profile 22
6. Visual explanation of matches and trivial matches 23
7. Scheme of the matrix profile calculation 24
8. PAA dimensionality reduction 25
9. Illustration of SAX 26
10. Visplore dashboard 28
11. Workflow of the Visplore/Python prototype 31
12. Example line plot 34
13. Stacked bar chart group 36
14. Synthetic time series 39
15. Effect of k-motifs 40
16. Effect of the exclusion zone 41
17. Effect of trimmed query 42
18. Effect of fPAA and alphabet size on result quality 44
19. Scenario 1 - Differentiate two groups 45
20. Scenario 2 - Find the uppercut 46
21. Scenario 3 - Detect gradual changing patterns 48
22. Retrieval sequence of gradual changing patterns 48
23. Permissive search in accelerometer data 49
24. Scores of permissive MASS search in accelerometer data 50
25. Score of permissive SAX search in accelerometer data 51
26. Permissive search in industrial data 52
27. Domain Search in complex accelerometer data 54
28. Overlay plot of two patterns witch dynamic region. 56
29. Domain search scores of MASS, simultaneous MASS and sequential MASS 57
30. Overview of used approaches and derivates 58
31. Workflow of permissive MASS and permissive SAX 60
32. Sensitivity summary of Visplore, permissive MASS and permissive SAX 62
33. Workflow of sequential and simultaneous MASS 63

List of Tables

1. Overview of MASS parameters 37
2. Overview of SAX parameters 38

Listings

1. Import data from CSV into a Pandas dataframe 72
2. MASS procedure to retrieve motifs from a distance profile 73
3. Permissive search using simultaneous MASS 76
4. Sequential MASS function definitions 77
5. SAX procedure to create a symbolic representation 80

1. Introduction

1.1. Problem statement

Industrial companies have a great pool of data in their possession. However, it is difficult to gain full advantage from this data. As the amount of data rises, so rises the need for automated procedures to handle it. Yet, most companies are stuck with outdated solutions, and thus require great effort to gain meaningful information from their data. To solve slightly more complex problems, companies have to launch costly data science projects.These projects are very time consuming and include phases such as data preparation, exploration, analysis, visualization, reporting and archiving. Most of the tasks that are associated with this kind of projects are performed manually by experts of a technical service department. Therefore, this inefficient process consumes many resources, slows down production and leads to challenges such as incremental updates, backups, reproducibility and a reasonable documentation.

As suggested above, an automated approach would be desirable to analyse large amounts of data efficiently. Depending on the respective field, this implementation can vary in its degree of difficulty. Time series are chronological observations, and thus have to be considered as a whole. From this point of view, fields that rely on data mining from time series data are the supreme discipline because of the difficulties involved. When it comes to data mining, three inherent characteristics of time series make it difficult to handle them: data size, dimensionality, and the need for continuous updates. So, it has become an interesting research field in data mining. The main study categories are representation & indexing, similarity measure, segmentation, visualization and mining. These fields are described in a detailed review by Fu (2011).

Data mining tools of the last decades had to make certain compromises in order to analyse time series as a whole efficiently. In other words, precision had to be traded for efficiency. For that purpose, a variety of approximation methods had been developed with the goal to reduce the dimensionality of time series(Lin, E. Keogh, Lonardi, and Chiu 2003). In the end, the time series data should be small enough to fit into the memory as a whole, and thus enable efficient computation methods. Only in recent years exact methods have become possible due to constant development in this area (Mueen, Yan Zhu, et al. 2017; C.-C. M. Yeh et al. 2016). Approaches using approximated data and exact values, will be discussed in more detail in the following sections.

So, we can agree that the foundation for quality control is the meaningful handling of data. A producing company therefore has the responsibility to find the best possible methods to remain competitive. Nevertheless, no one has to reinvent the wheel. For that purpose, service companies have emerged, which have great knowledge in data science, and thus can offer efficient tools to handle big data. Here the goal is to offer the customer an easy access to programs that address data science questions. Over the last 20 years, VRVis Zentrum für Virtual Reality und Visualisierung Forschungs-GmbH (VRVis)[1] has developed and gained a lot of experience in this field.

1.2. VRVis

"A picture tells a thousand words". This saying could also be a short description of the company strategy of VRVis. In case of VRVis, "the picture" stands for high-quality and interactive visual representations of multi-dimensional data pools. In other words, VRVis provides customers with software solutions that transform data into visualizations. It is important for their customers to receive a tool to make profound data-based decisions. These tools are universally applicable and are not limited to a certain industrial sector. Therefore, the company has a wide range of research projects, leading to a great knowledge base concerning data types and visualization applications. With this approach, VRVis has been bridging the gap between research and industry for 20 years now.

The *Visual Analytics Group* has a clear focus on time-dependent multi-dimensional data. These topics are a great foundation for the development of decision supporting tools. The group creates new technology, empowering their customers for insights in their data structures. This goal should be achieved with dynamic systems that allow a user to easily interact with data (Arbesser et al. 2017). A profound knowledge of the own data leads to an increased confidence in decision making. The quick response of the analysis system should lead to an almost dialogue-like process between user-input and the resulting visual representation. Such an advanced user interface makes multi-dimensional analyses efficient and new perspectives on data are provided. In this regard, a vivid discussion on time series visualizations is given by Aigner et al. (2007). In this article, the authors stresses the importance of task-orientation when designing visualization methods in order to improve the user experience. Overall, this challenging field is a great opportunity for VRVis to support customers with existing solutions or to develop new ones.

[1] https://www.vrvis.at/

1.3. Motivation to improve similarity search technology

Similarity searches in time series are a good example for a complex data mining task that can easily overwhelm a user. These kind of tasks are rather time-consuming and require multiple parameters as input. Since these tasks are difficult to initiate and to control, normally they are only performed by experts.

The current analysis tools by VRVis deal with time series on a regular basis. Therefore, they are already equipped with a similarity search method. However, the current method operates on a very basic level. Here, the user selects a pattern of interest in a time series. Also large time series can be processed in a reasonable time consumption, since the search algorithm has been implemented very efficiently in C++. There is almost no need for parameters. Only the initial pattern width must be defined in order to generate a query for the program.

As a result from the similarity search, similar patterns are highlighted in the time series. These labels give the user a quick estimation about found matches. In a next step, the user dynamically defines the error rate up to which matches are accepted. Again, the final matches are labeled within a representation of the original time series. The intuitive user interface is already a big advantage for this solution.

Yet, there is room for improvement. Some patterns possess challenging properties that make it difficult to find all correct matches. Two properties in particular can reduce the amount of meaningful matches, if only basic search algorithms are used. First, time distortion of patterns might decrease the success rate drastically. Although the human eye is trained to tolerate distortions in time fairly well, it is quite a challenge for automatic similarity searches. Secondly, complexity of patterns must be addressed. The shape of complex patterns can be built up by several domains. This circumstance rises the need for a weighted query input in order to address different features within one pattern. Also, the more subtle the search pattern becomes, the more accurate and efficient the algorithm must be. In other words, even the smallest deviation in a process can provide quality-relevant signals, which can be a valuable control tool.

Figure 1 shall give an impression of the current search result quality and the lost potential. Actually, the selected query was supposed to find all signals shown in the figure. However, only a minor fraction of signals had been found due to slight temporal distortion as can be seen by the variance on the horizontal axis.