Final Year Project Report

Automated Software to Understand Functional Relationship Between Dynamic Energy and Performance Events

Sukrat Kashyap

A thesis submitted in part fulfilment of the degree of

BSc. (Hons.) in Computer Science

Supervisor: Ravindranath Reddy Manumachu

UCD School of Computer Science

University College Dublin

December 4, 2017

Project Specification

General Information:

A energy model representing a relationship between dynamic energy consumption and performance events (PMCs) is constructed experimentally and the experimental dataset has the following format typically (k events, n records):

$E_1,\, x_{11},\, x_{12},\, x_{13} \ldots x_{1k}$
$E_1,\, x_{21},\, x_{22},\, x_{23} \ldots x_{2k}$
$\ldots$
$E_n,\, x_{n1},\, x_{n2},\, x_{n3} \ldots x_{nk}$

where E_i is the experimentally obtained dynamic energy consumption of i-th record and xij are the experimentally obtained performance events (PMCs).

Given such an experimental dataset as an input, the goal is to determine/understand the functional relationship between the dynamic energy consumption and performance events (PMCs).

Two real-life datasets will be provided to the student.
Core:

The goal is to write automated software that will detect the following:

1. Existence of records where the dynamic energy consumption is the same (within an input tolerance) but all PMCs (with the exception of one) have same values. Then the relationship between energy and the one PMC is visualized to see the nature of the functional relationship.
2. Having accomplished step (1), understand the monotonicity of the relationship between dynamic energy consumption and performance events (PMCs).
3. Existence of records where the dynamic energy consumptions are different (within an input tolerance) but all PMCs have same values (within an input tolerance) suggesting the non-existence of a functional relationship.

The software must be written using any one mainstream language but preferably one of the following: C, C++, Python

The software must be well documented and tested.

Advanced:

Given an experimental energy model dataset as an input, the goal is to write software that performs intelligent but computationally feasible simulations where combinations of inputs are generated to study the existence/non-existence of a functional relationship between dynamic energy consumption and PMCs.

The software must be written using any one mainstream language but preferably one of the following: C, C++, Python.

The software must be well documented and tested.

Abstract

With the advent of computers, data has become the new gold in the market. Most of the companies are driven by data as storage of data is no longer a expensive commodity. But new challenges have arised due to this and everyone now is focusing on analysis of data to predict, to minimize cost or to get new relations. This project tackles one aspect of data analysis that is the existence of a mathematical functional relationship of a data. It is as simple as saying true or false meaning relationship may exit or proving the non existense of functional relationship respectively.

Acknowledgments

I would like to thank my supervisor Ravindranath Reddy Manumachu and mentor Alexey Lastovetsky for the assistance and guidance of this project. I would like to thank my friends and family for supporting me in my work.

Table of Contents

Chapter 1: **Introduction**

1.1 Motivation

Modern day technology develop under incredible speed in recent decade and the computing power growth rate is truly phenomenal and lasting impact can be felt and benefit us in many ways.
For example, back in the early 90s, the power factory generate electricity at a constant speed 24 hours a day but the demand varies throughout the day. Thus, surplus electricity generally goes to waste. Although some of them are sourced from nuclear power station which produces at constant rate throughout, however, there are still large portion of electricity generated from fossil fuels, coal, petroleum, natural gas etc. control the supply of those source are practical. Thus, if we could discover the function relation between the demand of the power and the performance event (PMCs), that could enable us the ability to adjust the supply within certain network or distributed the power to some other network which power is more demanding.
Above is just only one field that possible usage of the functional relationship in data. In real world there are much more application that required such technique. Such as hybrid vehicle power management, power supply of auxiliary power units etc. There are many past work in this field but only few were focus on the relationship between dynamic energy consumption and PMCs. Hence, in this report we will try to observe the relationship between energy consumption and PMCs. is visualized to see the nature of the functional relationship. And after that we will explore the monotonicity of the relationship between dynamic energy consumption and performance events (PMCs) and suggesting the non-existence of a functional relationship as well.

1.2 Approach

Our main goal here is to find non existence of functional relationship. In other words, it means proving that there cannot be a functional relationship.
Our first approach to do so is to find two tuples such that n number of performance events are equal to each other within some tolerance, but their dynamic energy consumption is the same. Existence of such a tuple in database will lead us to prove the non-existence of a functional relationship.
Second approach is to group n-1 performance events and find dataset where the group size is greater than equal to three. In this approach we first assume here that the dataset provided has dynamic energy consumption equal to a linear combination of its performance events. Then using arithmetic we compute the constant relating to the nth event. This constant will then be used to find the dynamic energy consumption of the third tuple in the group. If this test fails then we would be able to conclude the non-existence linear functional relation.
The third approach which follows the second is to operate on the existing records and make new records by simple addition, subtraction with each other or multiplication of records with the constant. And then on these new records, the first approach is run which means finding same dynamic energy consumption and validating the equality of the performance events associated with it. The above three approaches help us to find the non existence of linear relation and also the functional relation. Our approaches are based on finding one irregularity which helps us defy the assumptions that the dataset is linear or have functional relationship. This methodology does

not prove or find the relationship. It rather defies the possibility of having one.

1.3 Structure of report

In this report, you must have already seen the motivation behind the project and brief description of the approaches that will be taken to reach the objective.
This Introductory chapter is followed by Background research, work done and work plan and the timeline.
Background research introduces you to the application of our project in various fields like power supply management, load balancing etc. It talks about Big data as the data for which the existence of relationship has to be found can be very large. It discusses different softwares available in the market which can be used to do the computation explaining the pros and cons of each. It explains the approaches mentioned above with proofs. So that one can be sure about the correctness of the program and understand the result. Performance of these approaches is also been discussed as performance is one of the key when the data is really big.
Following Background research is the work done. Work done contains simple observation that one found and analysis which will reflect the work plan. Work plan contains the timeline for the project.

Chapter 2: **Background Research**

2.1 Proof

The following proofs are for the different approaches that will be imployed to understand and find any descrepancy in the dataset to negate our hypothesis.

First proof:

Prove: We need to prove that finding atleast 2 equal dynamic energys with different performance events ensures that there exists no functional relationship in the dataset:

Proof:
We know that the data will be in the following format:

Let k be the number of parameters for the energy and n be the number of tuples in the dataset

$E_1, x_{11}, x_{12}, \dots x_{1k}$
$\dots$
$E_n, x_{n1}, x_{n2}, \dots x_{nk}$
where E_n is the dynamic energy for the nth tuple and x_{nk} corresponds to the kth performance event for nth tuple.

Let us assume that there exists a functional relation such that:
$f(x_{n1}, x_{n2}, \dots x_{nk}) = E_n$
where f is the functional relations for the dataset.

Our task is to find $f(x_{i1}, x_{i2}, \dots x_{ik}) = E_i$ and $f(x_{j1}, x_{j2}, \dots x_{jk}) = E_j$ where $i \neq j$ and $E_i = E_j$.

If we are able to find one in the dataset. Then we compare the parameters $x_{i1}, x_{i2}, \dots x_{ik}$ and $x_{j1}, x_{j2}, \dots x_{jk}$ of E_i and E_j respectively.

If the parameter tuples are not equal to each other meaning:
$f(x_{i1}, x_{i2}, \dots x_{ik}) = f(x_{j1}, x_{j2}, \dots x_{jk}) = E_i = E_j$ with the parameter tuple not equal that is $x_{i1}, x_{i2}, \dots x_{ik} \neq x_{j1}, x_{j2}, \dots x_{jk}$

From this we can conclude that the function f is not a function by using the definition of a function as this assumed function has two images.

Which contradicts from our hypothesis stated above. Hence by proof of contradiction we could say that f is not a function on the dataset.

Restating the above we can say dataset does not contain a functional relation.

Second proof:

Prove: Assuming that the dataset given has linear relationship and if we are able to find the constant for any one of the events. And if it does not apply to the other tuples of data that means linear relationship doesnot exist between the dataset.

Proof:
Lets first visualize and assume our dataset: Let k be the number of parameters for the energy and n be the number of tuples in the dataset

$E_1, x_{11}, x_{12}, \ldots x_{1k}$
$\ldots$
$E_n, x_{n1}, x_{n2}, \ldots x_{nk}$
where E_n is the dynamic energy for the nth tuple and x_{nk} corresponds to the kth performance event for nth tuple.

Let us assume that the energy consumption is a linear combination of the performance events.
$f(x_{i1}, x_{i2}, \ldots x_{ik}) = E_i$
$f(x_{i1}, x_{i2}, \ldots x_{ik}) = \alpha_1 \times x_{i1} + \alpha_2 \times x_{i2} + \cdots + \alpha_k \times x_{ik} + \alpha_{k+1}$
where α_i are the constants in the linear combination of performance events and $i \in [1 \ldots k+1]$

Lets find α_i where $i \in [1 \ldots k+1]$
To find this we will have to find atleast 3 records which have their parameter events equal $x_1, x_2, \ldots x_k$ except x_i where this x values are value belonging to a row in the dataset.

Suppose we found three records:
$E_m, x_{m1}, x_{m2}, \ldots x_{mk}$
$E_n, x_{n1}, x_{n2}, \ldots x_{nk}$
$E_o, x_{o1}, x_{o2}, \ldots x_{ok}$
where the tuples $(x_{m1}, x_{m2}, \ldots x_{mk})$, $(x_{n1}, x_{n2}, \ldots x_{nk})$ and $(x_{o1}, x_{o2}, \ldots x_{ok})$ are equal to $(x_1, x_2, \ldots x_k)$ except x_i for some $i \in [1 \ldots k+1]$ where $m, n, o \in [1, n]$ and m, n, o are not equal to each other.

Then
$E_m - E_n$
$= f(x_{m1}, x_{m2}, \ldots x_{mk}) - f(x_{n1}, x_{n2}, \ldots x_{nk})$
{ By our assumption that E is a linear combination of its parameters }
$= (\alpha_1 \times x_{m1} + \alpha_2 \times x_{m2} + \cdots + \alpha_k \times x_{mk} + \alpha_{k+1}) - (\alpha_1 \times x_{n1} + \alpha_2 \times x_{n2} + \cdots + \alpha_k \times x_{nk} + \alpha_{k+1})$
{ Gathering terms }
$= \alpha_1 \times (x_{m1} - x_{n1}) + \ldots + \alpha_k \times (x_{mk} - x_{nk}) + (\alpha_{k+1} - \alpha_{k+1})$
{ Since we know except x_{mi} and x_{ni} all are equal }
$= \alpha_i \times (x_{mi} - x_{ni})$

From the above we get:
$\alpha_i = (E_m - E_n)/(x_{mi} - x_{ni})$
where $(x_{mi} - x_{ni}) \neq 0$ as $x_{mi} \neq x_{ni}$ by above during our finding phase.

Then we know that using the α_i and applying to result to this equation $E_m - E_o = \alpha_i \times (x_{mi} - x_{oi})$ must be true as well.

If this is false then α_i is not a constant which contradicts our assumption that our E is linear combination of its parameter is false.

Hence using proof by contradiction we can say that the dataset is not linear combination of its parameters.

Third proof:

Prove: Keeping our assumption that the dataset is linear. Creating Pseudo records by simple addition or subtraction with other records if results into a record whose energy consumption lies in the dataset but with different performance events shows that the dataset is not linear.

Proof:
Let us assume that the energy consumption is a linear combination of the performance events.

$f(x_{i1}, x_{i2}, \dots x_{ik}) = E_i$
$f(x_{i1}, x_{i2}, \dots x_{ik}) = \alpha_1 \times x_{i1} + \alpha_2 \times x_{i2} + \dots + \alpha_k \times x_{ik} + \alpha_{k+1}$
where α_i are the constants in the linear combination of performance events and $i \in [1 \dots k+1]$

We know that if f is linear combination of its parameters they can be added and subtracted with each other to generate other data points.
$f(x_{i1}, x_{i2}, \dots x_{ik}) = E_i$
$f(x_{j1}, x_{j2}, \dots x_{jk}) = E_j$

Adding above to equations we get:
$f(x_{i1} + x_{j1}, x_{i2} + x_{j2}, \dots x_{ik} + x_{jk}) = E_j + E_i$

we will not prove this as it is a know property of linear function f. Lets rename it as:
$f(x_1, x_2, \dots x_k) = E$

Now if we are able to find data record with $E = E_m$ where $m \in [1, n]$ in the dataset provided.

Then if after comparing the parameters we find that $(x_1, x_2, \dots x_k) \neq (x_{m1}, x_{m2}, \dots x_{mk})$.

Then by the proof provided above, f has two images. So, above dataset is not linear combination.

Chapter 3: **Work Done**

Chapter 4: **Work plan**

Bibliography

[1] Robert Zembowicz and Jan M. Zytkov. *Testing the existence of functional relationship in data.* In Proceedings of the AAAI-93 Workshop on Knowledge Discovery in Databases, 1993.