International Journal of Advance Research in Science and Engineering
Volume No.06, Special Issue No.(01), December 2017
www.ijarse.com
IJARSE
ISSN: 2319-8354

Acquiring Business Intelligence through Temporal Mining of Smart Meter Data

[1]Moka Vinod, [2]Dr A V Krishna Prasad

Research Scholar, Sri Satya Sai University of Technology and Medical Sciences, (India)

ABSTRACT

More and more enterprises are switching over to Machine learning applications to improve their analyzing and predicting capabilities of their business. In this paper we propose a new outlook towards utility computing where public services can be view as a business. A public service can be better delivered by viewing it as a business model rather than a service model. The demand supply can be better analyzed and predicted by our model. This paper is about using efficient mining techniques on real time smart meter data for any utility like water, power or gas etc. The parameters that smart meters provide from time to time over a network can give us real time readings of the consumption which in itself adds enough intelligence to the service. Now by applying temporal mining techniques on this smart meter data we attempt to show how the Business intelligence can improved by data analysis and analytics. Though there is a opposition from some point of views that smart meters are hazardous to health due to its RF technology we can only improve utility computing by smarter data so that the service in efficient and effective.

Keywords — Business Intelligence, Machine Leaning, Smart Meter, Temporal Mining

1. INTRODUCTION

Digital Technology has given a new dimension to the capability of measurement. Today smart meters are available for various utilities like water, electricity, gas etc. These meters are digital and are networked. Smart meters can send event specific, time based, location based, real time parameters. This feature of smart metering opens up various point of views of storing, processing, analyzing and deducting from smart data. Though there are various ways of looking at this new set up of delivering utilities we focus on Temporal mining of information to achieve some Business intelligence.

Temporal data is data that includes time. Data mining problems can be classified into two categories: Data and Mining Operations. The main issue involved in data mining is processing data that encompasses temporal information. Temporal data mining has gained large momentum in the last decade. Various techniques of temporal data mining have been proposed. These techniques have been proved to be useful, to extract important information. In order to understand this phenomenon completely, we need to view temporal data as a sequence of events. Techniques from fields like machine learning, databases, statistics etc. are required when dealing with

temporal data mining. In this paper, we provide a brief overview of temporal data mining techniques which can been used on the smart meter data.

II TEMPORAL DATA TYPES ON SMART METER DATA.

2.1 Temporal

It is time dependent. Data and information derived from it are completely dependent on time. This might give the current real time parameters of any i\utility service. Ex: Transactional data in databases.

2.2 Time Series

This is a special case of time stamped data. It is similar to a number line. The events are uniformly separated in time variety of domains like engineering, research, medicine and finance. Smart meter data that is evenly received on regular intervals as a series.

2.3 Time Stamped

It has explicit information related to time. The volume of service delivered at any given time stamp can be read from the smart meter and the Temporal distance between data elements can be found. Inferences made can be temporal or non-temporal. Ex: data from stock exchange, inventory management.

2.4 Sequences

Sequences are ordered events with or without a concrete notion of time. These values can be event based ordered according the occurrence. Ex: customer shopping sequences, biological sequences. If an event appears before another, it means that the former event has occurred before the latter.

III TEMPORAL DATA MINING TASKS THAT YIELD USEFUL INFERENCES

Data mining has a wide range of applications. Tasks of data mining can be classified into some broad groups. In case of temporal data mining, these groups are Prediction, Classification, Clustering, Search and retrieval, Association. This categorization is not unique. Also, it is not exhaustive and does not cover all of the categories. In traditional time series analysis and pattern recognition, the first four categories have been extensively studied, understood and developed. In this section, a small overview of TDM techniques mentioned above is provided.

3.1 Clustering

Clustering groups the data on the basis of a similarity measure, like Manhattan distance, Euclidian distance. K-means, K-medics are well-known clustering techniques.

Manhattan Distance:

$$d = \sum_{i=1}^{n} |x_i - y_i|$$

Euclidian Distance:

$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Clustering of sequences, groups sequences based on their similarity measure. Clustering provides a mechanism to automatically find patterns in large data sets that would be otherwise difficult to find. In utility computing data acquisition clustering can be based on every Temporal data type like event based, time stamp based, real time volumes. This clustering proves quite useful where clusters can depict customer consumption patterns. There are various methods for clustering of sequences. We have model based sequence clustering methods like learning mixture models, a big class of model based clustering methods. For time series clustering, variations and hybrid models of ARIMA (Autoregressive Integrated Moving Average) models and Hidden Markov Models used. Another broad class in clustering of sequences uses pattern alignment-based scoring. Some techniques use mixture of both model based and alignment based methods.

K-means clustering algorithm:

Input: D = {t1, t2, …, tn} //Set of elements
K //Number of desired clusters
Output: K //Set of clusters
Algorithm:
Assign initial values for means m1, m2, …., mk;
Repeat
Assign each item ti to the cluster which has the
closest mean;
Calculate new mean for each cluster;
Until convergence criteria is met;

3.2 Classification

Classification is supervised learning. In classification, there are predefined classes into which the unknown set of attributes is classified. In temporal classification, given temporal sequence is assumed to belong to one of the predefined classes. This can be called as a training database. Using this training data, we can determine the class to which the given input belongs. Examples of sequence classification include gesture recognition, speech recognition, handwriting recognition, signature verification, etc. In smart metering based data acquisition and analysis we can find patterns of consumer consumption of any utility service. These patterns can be used for feature extraction which can yield a very analytical information about the consumer as the habit discovery.

In signature verification, the input is a sequence of pixel coordinates drawn by the user. The task here is to classify each sequence according to its pattern. In gesture recognition, trajectories of motion and other features of objects are collected from the video/frames. Feature extraction step generates a sequence of vectors, for each pattern, that is then classified. Sequence classification methods can be pattern based or model based. In pattern based methods, training database is maintained for each pattern and class. The classifier then searches entire database, for the one that is most similar to pattern of the input. Sequence aligning methods like Dynamic Time

Warping are used as similarity measures. There is another popular class of sequence recognition techniques which uses Hidden Markov Models (HMMs). It is a model based technique.

.

3.3 Association Rules

Apriori is an effective algorithm to mine association rules from a data set. In association rule mining we extract the relation among the attribute using parameters called support and confidence. Association rule mining can be applied to temporal association as well. In order to use Apriori algorithm to temporal data, some changes need to be made to the original algorithm. Support is the fraction of entities now, which has consumed the item sets in the transactions. An entity can contribute one time to the support of each item set. Large item sets are identified. The item sets with support greater than the minimum support are translated to an integer. Every sequence is changed to a new sequence having elements from large itemsets of the earlier sequence. In the next step, large sequences are found.

The algorithm first generates the candidate sequences. It then chooses the large sequences from the candidate sequences, until no candidate sequences are left. In Apriori algorithm, candidate generation is the most costly operation, as it suffers from combinatorial explosion. The general association rule X =>Y i.e. if X occurs then Y occurs, can be extended to a new rule X => tY i.e. if X occurs then Y will occur within time t. This new rule enables us to control the occurrence of an event to another event, within a time interval. The problem of discovering association rules arises from the need to unearth patterns in transactional data. Transactional data is temporal as the time of the purchase is stored in the transaction when products are purchased. This is called transaction time. In large data volumes, required for data mining purposes, there may be some information of products that did not exist throughout the data gathering period. We can find some products that have been discontinued. There may also be new products which were introduced after the beginning of the gathering process. Some of these products must be included in the associations, but may not be included in any rule because of support restrictions. Hence, these new products would not appear in interesting and potentially useful association rules. To solve this problem, we integrate time in the model of discovery of association rules. These rules are called Temporal Association Rules.

3.4 Prediction

Prediction is an important problem in data mining. Prediction problems have some specific traits that distinguish them from other problems. Prediction is a something that comes after analysis. We can predict if we perform analytics on these Temporal smart metering database. Prediction also requires the current and the historical information that can give a estimation of future consumption. This will ultimately lead to a acquiring a greater Business Intelligence through Temporal data mining. We will be able to effectively and efficiently distribute utilities to achieve better service and profitability . There has been previous work in algorithms which can be used to predict time series evolution. In prediction we forecast the future based on the data gathered in the past. In time-series prediction we predict future output of the time series using past data. An autoregressive family of models can predict a future value as a linear combination of sample values, given the condition that the time

series is stationary. Models like ARIMA, which is a linear stationary model, have been found to be useful in various industrial and economic applications, where some suitable variant of the process is assumed to be stationary. For non-stationary processes, the time series is assumed to be piece-wise stationary. This series is broken down into smaller parts called frames, within which, the stationary condition is assumed to hold and then separate models are learnt for every frame. In addition to the ARIMA family of models, there are many other nonlinear models for time series prediction like neural networks which are used for nonlinear modeling. The prediction problem is a part of Artificial. Based on various rule models such as disjunctive normal form model, periodic rule model etc. sequencegeneratingrules are found out that state some properties about the symbol which can appear next in the sequence. Prediction has huge importance in fields like medicine, finance, environment & engineering.

3.5 Search and Retrieval

In searching, we aim to locate subsequences in large database of sequences in a single sequence, efficiently. To locate exact matches of substrings is a trivial problem, however to handle efficiency when looking for approximate matches is a difficult task. In data mining, we are more interested in approximate matching rather than exact string matching. When a query is given by the user, similar results must be given because the user might not be looking for exact results. We define similarity measures by considering distances between two corresponding sequences. Similarity measures like Euclidian distance or Manhattan distance can be used. Similarity measures based on DFT (Discrete Fourier Transform) and DWT (Discrete Wavelet Transform) have been discovered as well. Choice of similarity or dissimilarity measures is just one part of the sequence matching problem. When we are determining similarity between two sequences, the sequences can be of different size. So it is not possible to calculate distances between corresponding sequences. Hence we use sequence alignment. We insert gaps in the two sequences or decide which should be corresponding elements in the given pair of sequences. For sequence classification and matching, time warping methods have been used. In speech applications, Dynamic Time Warping (DTW) is an efficient method that uses dynamic programming to identify correspondence among the vectors of two sequences to determine similarity between them. Symbolic sequence matching problems find applications in biological sequences
such as proteins, genes, etc.

IV TEMPORAL DATA MINING ALGORITHMS

The goal of temporal data mining is to find hidden relations between given sequence of events. An efficient approach to mining such relations is sequence mining. It involves three steps:

1) Transformation: converting given data into suitable form.

2) Similarity Measure: defining the similarity measure to be used.

3) Mining Operation: applying mining operation to get desired results.

Some of the Sequence Mining algorithms are :

4.1 Generalized Sequential Pattern(GSP) Algorithm

GSP is used for sequence mining. It is based on the Apriori algorithm. We first discover all the frequent items level-wise by counting the occurrences of all singleton elements in the data set. The transactions are then filtered. Non frequent items are removed. After this step, each transaction consists of only the frequent elements. This is the input to the algorithm. GSP Algorithm makes multiple passes. In the 1st pass, all single items are counted. A set of candidate 2-sequences are formed from the frequent items, and one more pass is made to find out their frequency. Candidate 3-sequences are generated from frequent 2-sequences. This process is repeated until no more frequent sequences are found.

Two main steps in the algorithm are:

Candidate Generation: The candidates for the next pass are generated by joining F(k-1) with itself. Pruning is done in order to eliminate any sequence at least one of whose subsequences is not frequent.

Support Counting: A hash-tree based search is used for counting support efficiently. Non-maximal frequent sequences are removed.

GSP Algorithm:

F1 = the set of frequent 1-sequence
k=2,
do while F(k-1)!= Null;
Generate candidate sets Ck (set of candidate ksequences);
For all input sequences s in the database D
do
Increment count of all a in Ck if s supports a
Fk = {a Є Ck such that its frequency exceeds the threshold}
k= k+1;
Result = Set of all frequent sequences is the union of all Fks
End do
End do.

4.2 Sequential Pattern Discovery using Equivalence Classes (SPADE)

SPADE is based GSP. SPADE uses a vertically structured database. SPADE initiates from the bottom-most element of the lattice and works in a bottom-up fashion to generate all frequent sequences. It maintains the vertical structure as it proceeds from the less elements to more elements.

Algorithm: SPADE (min_sup, D):

F1 = {frequent items or 1-sequences};
F2 = {frequent 2 sequences};
E = {Equivalence Class [X]θ1};
for all [X] Є E do Enumerate-Frequent-Seq([X]);

Enumerate-Frequent-Seq(S):
for all atoms Ai Є S do
Ti = φ ;
for all atoms Aj Є S with j>=i do
R = Ai v Aj;
if (Prune(R) == FALSE) then
L(R) =L(Ai) ∩ L(Aj);
if σ(R) >= min_sup then
Ti = Ti U {R}; F|R| = F|R| U {R};
end
if(Depth-First-Search) then Enumerate-Frequent-Seq(Ti);
end
if(Breadth-First-Search) then
for all Ti != φ do Enumerate-Frequent-Seq(Ti);
Prune (β):
for all (k-1)-subsequences, α < β do
if([α1] has been processed, and α not Є F(k-1) then
return true;
return false;

V COMPARISON BETWEEN GSP AND SPADE

	GSP	**SPADE**
Purpose	It is used for extracting frequently occurring sequences	It is used for fast discovery of sequential pattern
Approach	Apriori Based	Apriori Based
Candidate Sequence	Candidate Sequence Are required to be generated.	Candidate Sequences are required to be generated
Database Format	Uses Horizontal Format Database	Uses Vertical Format Database
Performance	1. Iterative algorithm 2. Makes multiple passes over the database depending on the length of the longest frequent sequences in database. 3. I/O cost is high if database has	1. Outperforms the GSP. 2. Excellent Scale up properties w.r.t parameters like number of input sequences, event size, size of potential maximal frequent sequences etc.

	very long frequent Sequences Candidate	
Speed	It is slower than SPADE	It is faster than GSP

VI. CONCLUSION AND FUTURE SCOPE

We have attempted to give a new dimension to utility computing where a service of an essential commodity is viewed as a profit model an create a win-win situation for both consumer and service provider. We have referred about the different TDM techniques that can be used over Smart meter data to improve Business Intelligence. Mining, analysis and analytics on the temporal data can make the system to acquire more machine intelligence so that efficient and effective utility management can be achieved. The system will ultimately be capable of give real-time decision parameters.

There is a lot of scope in this area using TDM on smart metering. Analysis and analytics can be applied on specific utilities with different parameters like water, gas, power etc. Based on each service the prediction may differ.

REFERENCES

Journal Article

[1] "Verified Computational Differential Privacy with Applications to Smart Metering" Gilles Barthe, George Danezis, Benjamin Gregoire, Cesar Kunz, Santiago Zanella-Beguelin. IMDEA Software Institute, Spain , INRIA Sophia Antipolis – M´editerran´ee, France, Microsoft Research, UK.

[2] "Smart Meter Driven Segmentation: What Your Consumption Says About You" Adrian Albert, Electrical Engineering and Management Science. Ram Rajagopal Civil and Environmental Engineering Department, Stanford University, Stanford USA, Manuscript draft May 30, 2013

[3]"You Are How You Consume: Mining Structure in Smart Meter Data" Adrian Albert Electrical Engineering and Management Science. Ram Rajagopal Civil and Environmental Engineering Department, Stanford University, Stanford USA, Manuscript draft May 30, 2013.

[4] "Cluster Analysis of SmartMetering Data An Implementation in Practice", Dipl.-Wi.-Ing. Christoph Flath Dipl.-Wi.-Ing. David Nicolay, Dr. Tobias Conte, PD Dr. Clemens van Dinther Dr. Lilia Filipova-Neumann, Research Center for Information Technology, Karlsruhe Germany, BISE Research paper Published online: 2012-01-12, doi: 10. 1007/s11576-011-0309-8.

[5] "Towards Automatic Classification of Private Households Using Electricity Consumption Data" Christian Beckel , Institute for Pervasive Computing, Zurich, Switzerland, Leyna Sadamori, Silvia Santini, Wireless Sensor Networks Group Darmstadt, Germany, Buildsys'12, November 6, 2012, Toronto, ON, Canada.

Conferences and Worshops

[1] "Application of a Data Mining Framework to Energy Usage Profiling in Domestic Residences Using UK Data ", Ian Dent, Uwe Aickelin, Tom Rodden, 1,2Intelligent Modelling and Analysis Group, University of Nottingham, Conference on "Buildings Don't Use Energy, People Do?" – Domestic Energy Use and CO2 Emissions in Existing Dwellings 28 June 2011, Bath, UK.

[2] "Computing Electricity Consumption Profiles from Household Smart Meter Data" Omid rdakanian, Rayman Preet Singh, Luk asz Golab, S. Keshav University of Waterloo, Canada Workshop Proceedings of the EDBT/ICDT 2014 Joint Conference (March 28, 2014, Athens, Greece) on CEUR-WS.org (ISSN 1613-0073).

[3]Symbolic Representation of Smart Meter Data, TriKurniawan Wijaya, Julien Eberle, Karl berer, School of Computer and Communication Sciences, Switzerland, EDBT/ICDT '13 , March 18 – 22, 2013, Genoa, Italy.

[4] A Habit Discovery Algorithm for Mining Temporal Recurrence Patterns in Metered Consumption Data, Rachel Cardell-Oliver CRC for Water Sensitive Cities and University of Western Australia, 1st International Workshop on ML for Urban Sensor Data (SenseML)15Sept2014.

Books, Articles and Reports

[1] " Big Data, Smart Energy, and Predictive Analytics, Time Series Prediction of Smart Energy Data", Rosaria Silipo, Phil Winters , A report by KNIME.

[2] "What are Smart Meters?" Fact Sheet, An Article by Australian Water Association.

[3] " Data mining: Concepts and techniques" (Han J, Kamber (2001), San Fransisco, CA: Kauffmann)

[4] "A survey of temporal data mining" Srivatsan Laxman and P S Sastry (2006), Sadhna , Vol. 31, Part 2, pp. 173–198

[5] "Data Mining Techniques" A.K. Pujari (2007), , University Press ISBN 8173713804.