Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X

IJCSMC, Vol. 4, Issue. 1, January 2015, pg.01 – 18

RESEARCH ARTICLE

Mining Himachal Pradesh State Electricity Board for AT&C Losses with Data Mining Association Rules

Atul Dhiman	**Arvind Kalia**	**Mukesh Kumar**
Department of Computer Science	Department of Computer Science	Department of Computer Science
HP University, Shimla, India	HP University, Shimla, India	HP University, Shimla, India

Abstract: The leading business challenges faced by the technical world today in business field are rightful extraction of information from huge data bases and its application into future ventures to minimise the losses, maximize the profits, and multiply the dividends. In this direction data mining and its tools prove useful in extracting each and every hidden and relevant piece of information. This paper presents an empirical study on data of Himachal Pradesh State Electricity Board to find out reasons and regions of the major AT&C losses for the current year's eight months for three towns. This work is accomplished with the help of data mining association rules generated with Apriori and verified Predictive Apriori algorithm's results. The performance of these two relative algorithms on nominal datasets is also comparatively concluded in the end.

Keywords: AT&C, ARFF, CSV, HPSEB, RAPDRP

I. Introduction

Quest for knowledge and its application in daily life has been the prime activity of human race, and a result of this quest they stored information from early stages of life in forms of scripts, pages and books. Today in 21st century the form of data and knowledge storage is digital which has been increasing stupendously. The foremost important aspect of this huge amount of stored data and information is the right extraction of valuable information and its application in modern day activities like business, entertainment, research and education etc.

In last decades data accumulation in electronic format has been at an explosive rate, it has been estimated that amount of electronically stored data doubles itself in every year and half. The number of databases and their sizes are growing even faster due to different type of computing devices used by masses. Now to get a real useful piece of information from such large masses of data is like finding the needle form haystack. So, how this information could be found out or retrieved? The answer to this question is data mining.

To make understand what Data Mining is in simple words it is a detailed process of analyzing large amounts of data and picking out the hidden predictive and relevant information. It is

process of extracting or mining knowledge from large amounts of data **[12].** The data sources are databases, data warehouses, the Web, other information repositories. To extract out information we have data mining tools which predict future trends and behaviours, allowing business to make confident knowledge driven techniques. Some of the basic techniques of data mining are divided into classical and modern techniques which are:

Statistics: Classical technique which is a branch of mathematics concerning the collection and the description of data. Knowing statistics in your everyday life will help the average business person make better decisions by allowing them to figure out risk and uncertainty when all the facts either aren't known or can't be collected.

The Second classical technique of Clustering is the method by which like records are grouped together. This is done to give the end user a high level view of what is prevalent in the database. Clustering is sometimes referred to segmentation - which most marketing people will tell you are useful for coming up with a bird's eye view of the business. Cluster analysis divides data into meaningful or useful groups (clusters). **[10]**

Third classical technique of Nearest neighbour is a prediction technique that is quite similar to clustering - its essence is that in order to predict what a prediction value is in one record look for records with similar predictor values in the historical database and use the prediction value from the record that it "nearest" to the unclassified record. **[3]**

First Modern data mining technique of Decision tree is a predictive model that can be viewed as a tree where each branch of the tree is a classification question and leaves represent the partition of the data set with their classification. Another definition of Investopedia defines a Decision Tree as a schematic tree-shaped diagram used to determine a course of action or show a statistical probability **[5].** Decision trees can be viewed from the business perspective as creating a segmentation of the original data set. Thus marketing managers make use of segmentation of customers, products along with sales region for the purpose of predictive study. These predictive segments concluded come with a description of the characteristics that define the predictive segment

Second modern data mining technique of Neural Network is biological systems that detects patterns and makes predictions. The artificial ones are computer programs implementing sophisticated pattern detection and machine learning algorithms on a computer to build predictive models from large historical databases **[9].** The greatest breakthroughs in neural network in recent years are in their application to real world problems like customer response prediction, fraud detection etc. Data mining techniques such as neural networks are able to model the relationships that exist in data collections and can therefore be used for increasing business intelligence across a variety of business applications. This powerful predictive modelling technique creates very complex models that are really difficult to understand even by the experts.

Third modern data mining technique of Rule induction is one of the most weighted forms of data mining and is perhaps the most common form of knowledge discovery in unsupervised learning systems. It's perpetually the form of data mining that most closely resembles the process that masses think about when they think about data mining, namely "mining" for gold through a vast database. The rule itself is of a simple form of "if this and this and this then this". For example a rule that a supermarket might find in their data collected from

scanners would be: "if pickles are purchased then ketchup is purchased' **[13].** Or if paper plates then plastic forks etc.

This paper provides way to demonstrate combination of technique and programming for business data prediction.

II. Literature Survey

Application of data mining techniques has been effective in business field some of the research work in this field done is presented below.

Hegland Markus in his study **[4]**, in their research work has stated that Association rules are of the form "if-then rules" with two measures which quantify the support and confidence of the rule for given data set. These rules are originated from market basket He has also reviewed the tools which have the potential to deal with long item sets and which reduce the amount of item sets returned. **Ramkrishanan Srikant and Agarwal Rakesh [1],** in their study, considered the problem of discovering association rules between items in large database of sales transactions. They have presented two new algorithms for his problem that are fundamentally different from the known algorithms. Empirical evaluation showed that these algorithms outperformed the known algorithms by factors ranging from three for small problems to more than an order of magnitude for large problem. **Sharif et al**.[7], in the study, has suggested that the internet technology has brought about a significant impact in doing business. It promotes a new way of doing new business by enabling new system such as electronic commerce to the worldwide users. The paper demonstrates development of e commerce system by focusing on the use of Priori Algorithm supported feature in the algorithm. The feature is included to enhance a good customer relationship management for proposed system. **Wu Xindong et. al. [8]**, in their research work presented ten algorithms that were among most influential data mining algorithms in research community. With each algorithm, they provided a description of the algorithm, its impact and also reviewed the current and further research on algorithm. **Ramraj E. et. al.[6],** in their study suggested that association rules are discovered by identifying relationships among sets of items in transactional databases with two measures which quantify support and confidence of rule. The paper reviews the Apriori related and Éclat Algorithms with detailed discussions of various data structures. Comparison has been made on the basis of several selected datasets and analysis ends with various research issues like type of rules, execution time and space complexity of algorithms. **Divya Bansal** and **Lekha Bhambhu [2], in** their research work presented use of Apriori Algorithm of Data Mining Directed towards Tumultuous Crimes Concerning Women. The research work concluded with answering all the questions as age group of men is 20- 24 ,age group of girls who are on their target is 16-22 and mostly accused are well known by the victim.

III. Need and Scope of Study

Apriori serves basis to most of algorithms for finding frequent item sets and generating association rules. Rules so generated serve useful for business data and market analysis as described previously. There have been considerable technical and commercial losses in electrical transmission and supply throughout the state which are needed to be minimized and data analysis and decision making procedure of HPSEB is still statistically driven and that too

also not in most of the industrial areas. Computerization of substation recently has opened opportunity for data fed to be analyzed with data mining techniques. These techniques can give accurate results with more parameters than those with simple statistical techniques. In this direction an attempt has been made to analyze and identify the regions and predict the reasons for their aggregate technical and commercial losses of Himachal Pradesh State Electricity Board (HPSEB) for period of eight months of year 2014.This work is possible help of association rules generated with Apriori and Predictive Apriori algorithms, so as to improve the electrical distribution and transmission of state and make it energy efficient state of India. So, association mining techniques can generate rules that can predict out every hidden aspect of the losses suffered by Himachal Pradesh State Electricity Board Limited in order to curb its losses and opens a scope for the work.

IV. RESEARCH METHODOLOGY

Research methodology taken up in this work includes data collection which includes description of data chosen and reason for choosing such type of data, data pre-processing which includes data set type conversion and attribute selection and attribute description are:

3.1.2 Data collection

The type of data: Power distribution and transmission is the most important process of electricity supply in any state of India. Hence this process also suffers from some technical and commercial losses in every state. So data collected of technical and commercial losses was from three major towns of Himachal Pradesh for a period of eight months from January to August for the year of 2014. Reason for selecting these three towns is given under:

Shimla: The Capital of state with largest number of electricity subdivisions.

Solan: The Town with most number of industries and high load demand town.

Dharamshala: The winter capital of state lying within largest populated district of Kangra.

All these three town of Himachal are major tourist attraction with home to Tourism Industry.

Necessity of choosing such type of data: The data provides information about aggregate technical and commercial losses of state. The concept of Aggregate Technical & Commercial losses was introduced by some state regulatory commissions in past decade. The advantage of the parameter is that it provides a realistic picture of energy & revenue loss situation in a state.

Furthermore these losses were to be monitored for towns under **R-APDRP Scheme** of India introduced for Power Sector Reforms. **Part B** of the scheme covers system strengthening, improvement and augmentation of distribution system. [**11]** This objectives shall involve:-

- Identification of high loss areas
- Preparation of investment plans for identified areas
- Implementation of plan
- Monitoring of Losses

Method of Collection

The Three Electricity Board Circles of Himachal Pradesh Shimla, Solan and Dharmshala were visited to manually collect the data. The substations at each circle had autonomous applications which automatically recorded data for each attribute of AT&C Losses. The data hence collected is stored at Storage Area Network with high storage capacity file system. The

data from this server was extracted into tabular form by initiating the batch process/commands to server and data was outputted in portable document format for each town selected along with its subdivisions and information much more.

Data Preprocessing

Data pre-processing involves Attribute selection which involves searching through all possible combinations of attributes in the data to find which subset of attributes works best for prediction. To perform this, two objects must be set up: an attribute evaluator and a search method. The evaluator determines what method is used to assign a worth to each subset of attributes. The search method determines what style of search is performed.

Step 1: Dataset Type Conversion

Name of Data Set	*AT&C Losses of HPSEB*
Type of Data Set	*Multi vibrate*
Type of Dataset Required for algorithm	*Nominal*
Filters Applied	*Unsupervised .attribute .Numeric To Nominal*
Default Dataset Format	*Portable Document Format(PDF)*
Dataset Format Required For Tool	*Arff / Comma Separated Values(CSV)*

Table 1: First Step of Data Preprocessing

Step 2: Attribute Selection

The attribute Selection feature Under Explorer tab provides in built methods to find out most important attributes among the data set. This data set AT&C Losses of HPSEB has 18 attributes which need to be ranked/selected before any conclusions are drawn upon them after rule have been generated.

Search Method Applied: *Attribute Ranking*

Attribute Evaluator: *Gain Ratio feature Evaluator*

Ranking (Close to Unity)	Ranked Attributes
0.952	*Billing Efficiency*
0.900	*Net Energy Import*
0.897	*Energy Import*
0.852	*Energy Export*
0.780	*Towns*
0.743	*Arrear Collected*
0.739	*Energy Billed*
0.735	*Collection Efficiency*
0.734	*Subdivision*
0.732	*Amount Arrear Collected*

0.728	*Amount Billed*
0.728	*Gross Amount Collected*
0.710	*Total Consumers*
0.706	*Commercial Losses*
0.705	*Feeder*
0,619	*Month*
0.540	*Technical Losses*

Table 2: Attribute Ranking

Figure next shows the result above given by attribute selection feature of weka being validated by J48 Classification algorithm tree's result. It shows billing efficiency on top as root node of tree depicting it as high information entropy attribute among all attributes

Figure 1: J48 tree based attribute ranking validation.

Hence, after ranking and bonus step of validation of results we have a list of 17 attributes from which we have selected 4 attributes including 1 attribute which hasn't being ranked i.e. AT&C losses, for our result and analysis work. Some of the brief introduction of these selected attributes is**: [11]**

Attribute Description: Most important step for highlighting causes of losses and for predictions to be made further.

Energy Export: Total Energy in kWh made input to next town feeder area after utilization at previous feeder Area**. [11]**

If it is = 0 i.e. Previous feeder area has full consumption of power input or energy hasn't been metered at all.

If it is! = 0 i.e. Input energy was surplus for previous town and energy metering has been good or the feeder area of town subdivision has low AT&C Losses.

Energy Input/Import: The amount of energy in kWh made input to a specific feeder area.

If it is = 0 i.e. All feeders do not feed the town as some of the feeders are feeding beyond boundaries or energy hasn't been metered at all**. [11]**

If it is! = 0 i.e. Energy Input has been measured by meters installed at all input points of the town.

Billing Efficiency: It is the indicator of proportion of energy that has been supplied to an area which has been billed including both metered and unmetered sales to consumers. **[11]**

$$\text{Billing Efficiency} = \frac{\text{Total Unit Sale}}{\text{Total Input}}$$

If it is = 0/infinite i.e. billing of that particular feeder town has been faulty or has been unable to bill the whole energy input due to theft ,bypassing etc.

If it is! = 0 billing percent will show the billing efficiency for that town.

AT&C Losses: Aggregate Technical and Commercial losses which are calculated as:

[1-(Billing Efficiency * Collection Efficiency)/100]

These losses have *Technical losses* which are due to: **[11]**

- Transformation losses at substations.
- Heat losses on distribution lines due to inherent resistance and poor power factor in electrical network.
- These losses vary with type of conductor used and transformation capacity of transformers.

Commercial losses are also the part which is: any illegal consumption of electrical energy which is not correctly metered billed and revenue collected**.[11]** The causes for these are:

- Meter reading discrepancy.
- Meter Tampering and Bypassing
- Theft by Direct Hooking

V. Results & Analysis

The rules generated with open source weka tool for above four attributes with 10 rules for Apriori and 100 rules for predictive Apriori gave the following results in tabular format and graphical format.

Solan Town (Jan-Aug) 2014

Sr. No.	Parameters values of rules generated for --->	Energy Export	Energy Import	Billing Efficiency	At& C Losses
1	Minimum Support	0.4	0.35	0.25	0.4
2	Minimum Confidence	0.9	0.9	0.9	0.9
3	Number of Cycles	12	13	15	12
4	Number of large item sets	3	5	4	3
5	Time Taken (in seconds)	1	1	1	1
6	Confidence Achieved	100	100	100	100
7	Instances with minimum support	29	26	18	29

Table 3: Parametric values obtained with Apriori on Solan Dataset

Sr. No	Parameters values of rules generated for --->	Energy Export	Energy Import	Billing Efficiency	At& C Losses
1	Best Accuracy Achieved	98.286	96.955	96.775	97.116
2	Time Taken(in seconds)	4	9	6	4
3	Total Number of attributes	18	17	14	18
4	Number of Rules generated	55	42	19	49

Table 4: Parametric values obtained with Predictive Apriori on Solan Dataset

Energy Export: For this attributes best rules were generated at 40% minimum support and it achieved confidence of full 100% and Predictive Apriori gave 98.286 % as best accuracy for its rules generated for in both antecedent and consequent positions in rule.

Rules generated have shown that 58 out of 72 sub towns do not export energy to next town i.e. they have heavy consumption. AT&C losses have been infinite in 33 sub towns where Energy Export is zero means there is lot of theft and illegal consumption in these sub towns. Calculation of net energy import has been incorrect due 0 energy import and export of 27 towns i.e. faulty metering. Billing efficiency for 27 towns has been incorrect due to incorrect metering and incorrect billing of energy input and heavy consumption at 27 sub towns. The subdivision which has shown significant AT&C and poor billing efficiency as these subdivision needs more correct metering and loss prevention schemes has also been identified. **[11]**

Energy Import:

For this attributes best rules were generated at 35% minimum support, it gave confidence of full 100% and Predictive Apriori gave 96.95 % as best accuracy for its rules generated for in both antecedent and consequent positions in rule.

Rules generated predict that 31 sub towns input feeder meters do not meter the correct energy input. Billing efficiency if improved for 28 towns could give the total energy input reducing losses hence it needs correct metering at town feeder installed at boundaries, to reduce AT&C losses. Moreover total energy supplied to energy billed has been not good for 27 sub towns i.e. they have faulty billing process .Billing efficiency is related with AT&C Losses as these losses are result of power theft ,transformation losses and low quality material used **[13].**

Billing Efficiency:

For this attributes best rules were generated at 25% minimum support, it achieved confidence of full 100% and Predictive Apriori gave 96.775 % as best accuracy for its rules generated for in both antecedent and consequent positions in rule

28 Sub towns have total energy supplied to energy billing method poor leading to high commercial losses and technical losses within these sub towns as revealed by rules. Sub division 11213 has highlighted 19 sub towns for poor billing efficiency and high losses. Subdivision 11211has been identified with 10 sub towns for 0% efficiency and gives causes for it to be no arrear collection.

Aggregate Technical and Commercial Losses

For this attributes best rules were generated at 40% minimum support and it achieved confidence of 100% and Predictive Apriori gave 97.116 % as best accuracy for its rules generated for in both antecedent and consequent positions in rule.

33 sub towns of Solan have high AT&C Losses due heavy consumption possibly due to meter tampering, hooking and meter bypassing **[11].** 28 sub towns has undergone poor billing practices which has lead to incorrect energy estimation and high losses and 27 have shown reason of incorrect energy evaluation as incorrect metering at feeder inputs i.e. meter do not account for energy at all or are metering it beyond feeder area**[11].** Sub division which has shown highest percentage of AT&C losses these sub division need to undertaken under strict energy monitoring and transmission methods, has been revealed.

Dharamshala Town (Jan-Aug) 2014

Sr. No.	Parameters values of rules generated for --->	Energy Export	Energy Import	Billing Efficiency	At& C Losses
1	Minimum Support	0.5	0.45	0.45	0.45
2	Metric Confidence	0.9	0.9	0.9	0.9
3	Number of Cycles	10	11	11	11
4	Number of large item sets	3	5	3	3
5	Time Taken (in seconds)	1	1	1	1
6	Confidence Achieved	100	100	100	100
7	Instances with minimum support	28	25	25	25

Table 5: Parametric values obtained with Apriori on Dharamshala Dataset

Sr. No.	Parameters values of rules generated for --->	Energy Export	Energy Import	Billing Efficiency	At&C Losses
1	Best Accuracy Achieved	98.235	96.419	96.549	96.783
2	Time Taken(in seconds)	8	5	4	4
3	Total Number of attributes	18	17	14	14
4	Number of Rules	57	67	28	27

Table 6: Parametric values obtained with Predictive Apriori on Dharamshala Dataset

Energy Export

For this attributes best rules were generated at 50% minimum support and it achieved confidence of full 100% and Predictive Apriori gave 98.235 % as best accuracy for its rules generated for in both antecedent and consequent positions in rule.

56 out of 57 towns of Dharamshala have 0 energy export means all theses sub towns have heavy load requirements or illegal consumption of energy as revealed by rules. Sub division 12221 has 48 sub towns which do not export energy hence identifying the regions. Another rule generated implies that technical losses such as heat losses and energy transformation losses for 28 sub towns have main reason behind energy exported from these towns to be zero and as same implied by rule these technical losses accumulated with commercial losses account for 0 Energy export. Poor/no metering of energy for 26 towns has accounted for zero energy export and non estimation of net energy input for the sub towns; hence correct metering at input points could reduce these errors **[11].** Reasons for losses in 12221 sub division with town name have also been identified.

Energy Import

For this attributes best rules were generated at 45% minimum support and it achieved confidence of full 100% and Predictive Apriori gave 96.419 % as best accuracy for its rules generated for in both antecedent and consequent positions in rule.

For energy input system 26 town's feeder input meters failed to account for energy input to the area and is the reason behind non estimation of net energy as implied by rule generated. Zero metering or inaccurate metering for these 26 sub towns has accounted for inefficient billing and incorrect energy estimation. Faulty energy input has been accounted for AT&C Losses and identify the subdivision 12221 as one with major losses. Faulty energy measurements for 12221 sub division's 22 sub towns to be main reasons behind inefficient billing. Sidhpur area has been identified for inefficient billing and AT&C Losses.

Billing Efficiency

For this attributes best rules were generated at 45% minimum support, it achieved confidence of full 100% and Predictive Apriori gave 96.459 % as best accuracy for its rules generated for in both antecedent and consequent positions in rule

Rules generated has identified 26 sub towns where total energy input to energy billed ratio has been erroneous and has led to major AT&C Losses . Dharamshala feeder has 8 sub towns which has been billing inefficient and due this they are prone to major AT&C Losses.

AT&C Losses

For this attributes best rules were generated at 45% minimum support and it achieved confidence of full 100%. Predictive Apriori gave 96.783 % as best accuracy for its rules generated for in both antecedent and consequent positions in rule.

28 sub towns of Dharamshala have been under major losses due theft incorrect metering bypassing hooking and various other factors. Another rule has identified billing inefficiency as a major factor behind these losses in 26 sub towns. 8 sub towns of feeder named Dharamshala have AT&C Losses combined with billing inefficiency. Another rule reveals that only 26 sub towns out of 28 are causing AT&C losses due to billing inefficiency hence need more strict billing procedures. Last rule identifies 12221 sub division has suffered major AT&C losses and billing inefficiency for its 7 sub towns in last eight months **[11].**

Shimla Town (Jan-Aug) 2014

Sr. No.	Parameters values of rules generated for --->	Energy Export	Energy Import	Billing Efficiency	At& C Losses
1	Minimum Support	0.9	0.75	0.3	0.3
2	Metric Confidence	0.9	0.9	0.9	0.9
3	Number of Cycles	2	5	14	14
4	Number of large item sets	3	4	3	3
5	Time Taken (in seconds)	1	1	1	1
6	Confidence Achieved	100	100	100	100
7	Instances with minimum support	406	338	135	137

Table 7: Parametric values obtained with Apriori on Shimla Dataset

Sr. No.	Parameters values of rules generated for --->	Energy Export	Energy Import	Billing Efficiency	At& C Losses
1	Best Accuracy Achieved	98.82	98.85	98.265	98.265
2	Time Taken(in seconds)	129	145	151	152
3	Total Number of attributes	16	16	13	13
4	Number of Rules	35	56	47	50

Table 8: Parametric values obtained with Predictive Apriori on Shimla Dataset

Energy Export

For this attributes best rules were generated at 90% minimum support and it achieved confidence of full 100%. Predictive Apriori gave 98.82 % as best accuracy for its rules generated for in both antecedent and consequent positions in rule.

88 Sub towns of Shimla 11122 sub division aren't exporting energy where we can predict that they are under heavy electricity consumption or are incorrectly metering the energy exported. Two months out of eight months January and April had most consumption of electricity in 57 sub town reason might be harsh climatic conditions and faulty metering as revealed by rule 3,during April month there have been most technical losses due to heat losses inefficient power transformation etc. as revealed by rules generated. Sub division 11114 and 11121 have been billing inefficient and suffering with AT&C losses revealed. Another Rule reveals that 421 sub towns have been more prone to technical losses out of 424 sub towns and can reduce AT&C losses if taken care of.

Energy Import

For this attributes best rules were generated at 75% minimum support and it achieved confidence of full 100%. Predictive Apriori gave 98.85 % as best accuracy for its rules generated for in both antecedent and consequent positions in rule.

Faulty metering at input feeders and no arrear collection has accounted for Shimla's 91 towns' billing inefficiency and AT&C Losses if these are improved they will reduce the losses. Technical losses i.e. heat loss on transmission lines and inefficient transformation is the major reasons behind billing inefficiency and AT&C losses of Shimla's 85 towns. Identified sub division 11122 has been suffering from AT&C and billing inefficiency in 55 sub towns due no accounting of energy input. Identifies another sub division 11126 has been suffering with losses due to zero energy import at 54 sub towns. Another Rule direct relationship of incorrect metering at input feeders of 345 sub towns for billing inefficiency and AT&C losses, hence correct accounting of energy input can improve the billing efficiency and reduce losses for HPSEB.

Billing Efficiency

For this attributes best rules were generated at 30% minimum support and it achieved confidence of full 100% and predictive Apriori gave 98.26 % as best accuracy for its rules generated for in both antecedent and consequent positions in rule

Sub division 11122 and 11126 has been identified by the association rules in the dataset with most billing inefficient sub divisions accounting for major AT&C Losses as implied by rule generated. The month of August and another subdivision 11114 where billing efficiency has been negatively infinite for 54 sub towns of Shimla leading to AT&C losses have been predicted. Zero amount collection with billing inefficiency has been due to no amount collected excluding arrears leading to Commercial losses for 26 sub towns, these sub towns

need to improve collection efficiency. The total energy billed was zero leading to AT&C losses and resulting in billing inefficiency and commercial losses for 15 and 12 sub towns respectively, the causes here can be predicted are that no energy was billed due heavy theft or faulty metering practices which have shown zero commercial or AT&C losses but has lead to billing inefficiency or commercial losses as revealed in rules. Another rule shows that 6 sub towns have billing efficient.

AT&C Losses

For this attributes best rules were generated at 30% minimum support and it achieved confidence of full 100%. Predictive Apriori gave 98.26 % as best accuracy for its rules generated for in both antecedent and consequent positions in rule.

Rules identify sub division 11122 and month August with negatively infinite billing efficiency as a major cause of AT&C losses. Identified sub division 11121 and month of January are having major AT&C losses due to billing inefficiency. Another Rule reveals 11112 with 40 sub towns has suffered with technical and commercial losses, hence this sub division need improvement all around. Erroneous amount billed and energy billed with no amount collected except arrears have been major causes of AT&C losses for 19 sub towns. 6 towns have been identified with most billing efficient towns i.e. 345 towns have been contributing majorly for AT&C losses **[11].**

Graphical Analysis

The important aspect of result analysis is graphical analysis which here has given algorithmic comparison results along with comparison among above three towns in context with four attributes selected. Some of the key points to understand before the graph and rule analysis are that:

I. Support given for rules is given by weka itself while generating the rules it starts from 100% and then it is decremented by a decrementing factor set in weka for association algorithms.
II. Reaching upto a best support value, the best rules are generated for a dataset.
III. Minimum Confidence is set to 90% by default, as we aspire to have best rules which are 90% above confident.
IV. Rules with high value of support and confidence are the most valuable mined nuggets.

Apriori Vs Predictive Apriori (Average Execution Time)

The line graph below depicts that average execution time of Apriori is less than predictive Apriori i.e. Apriori is faster algorithm than Predictive Apriori for generation of Association rules. The average time in case all three dataset for Apriori was less than 1or 1 second which was equal in all three datasets. In case of predictive Apriori is varied from 4 seconds to 200s in case of Shimla dataset which has more than 400 instances. The Apriori algorithm is faster than Predictive Apriori has been also proven in previous research works **[2].**

The graphs for four attributes: energy export, energy import, billing efficiency and AT&C losses have followed after the graph below.

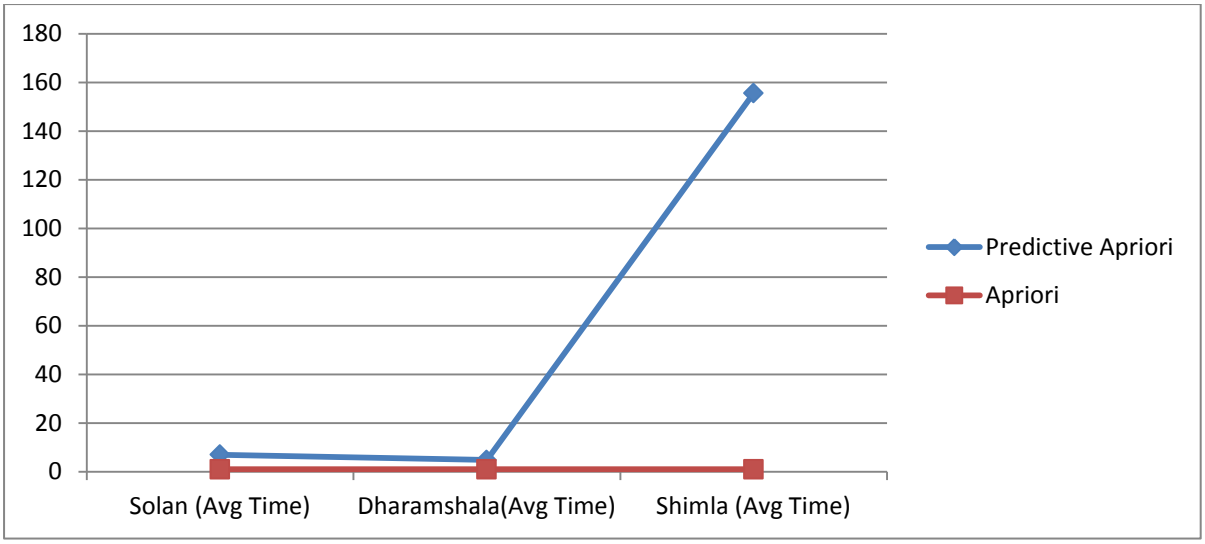

Figure 2: Average Execution Time Line Graph

Energy Export

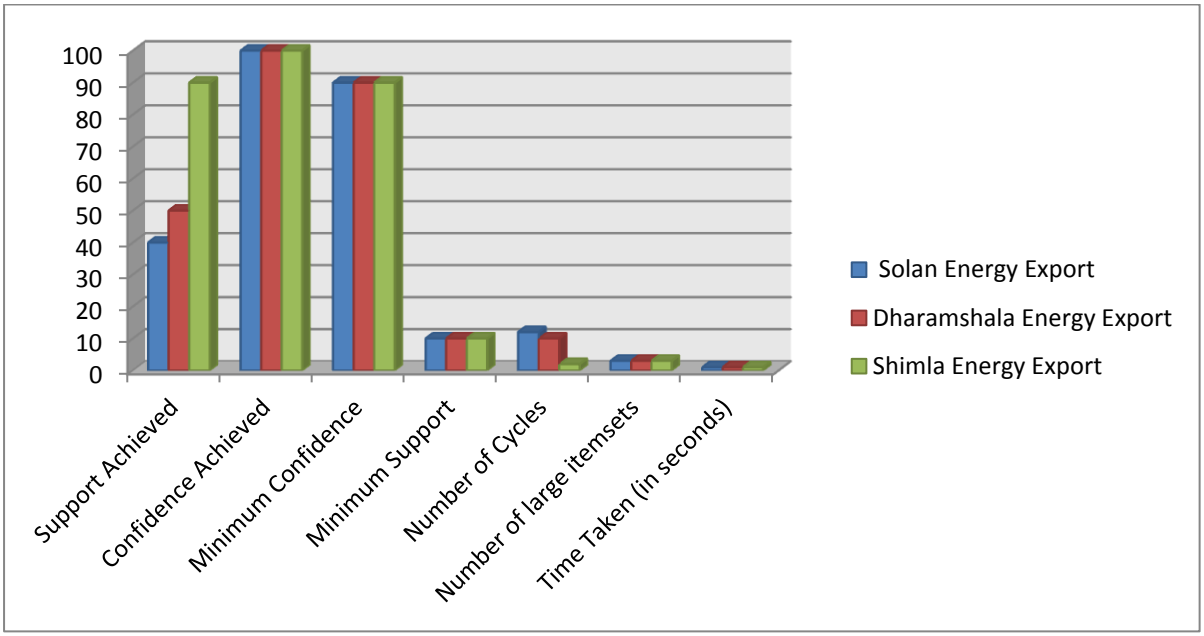

Figure 3: Energy Export Graph with Apriori Algorithm

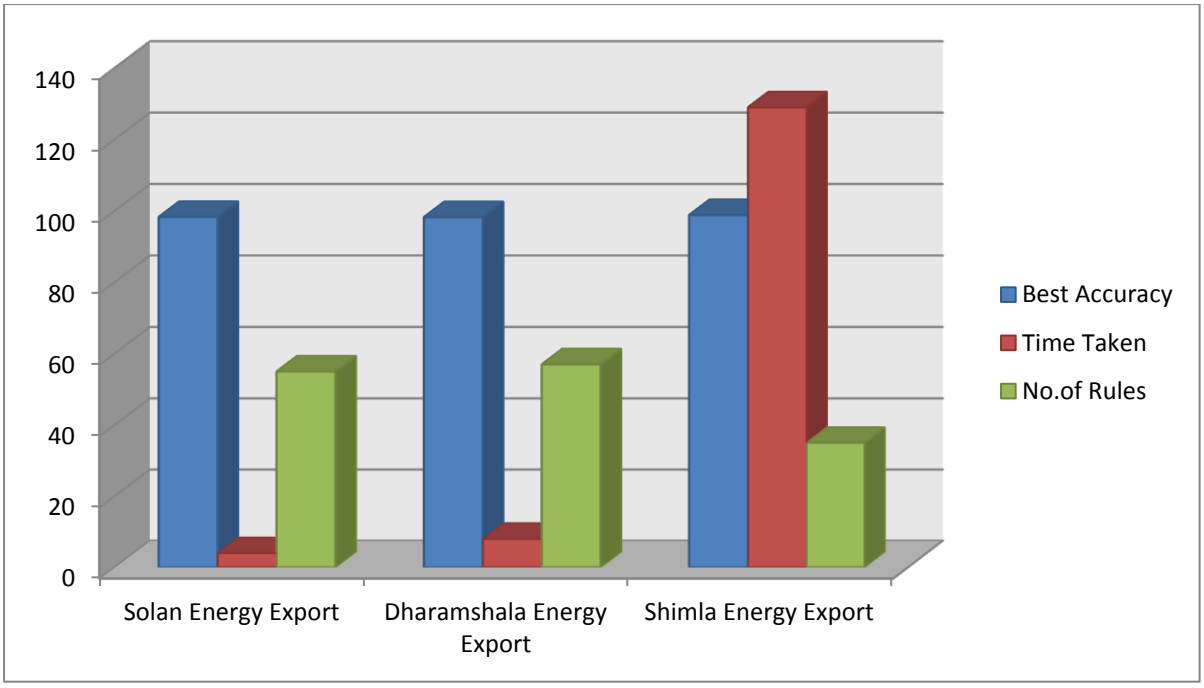

Figure 4: Energy Export Graph with Predictive Apriori Algorithm

The best rules with maximum support for this attribute were generated for Shimla dataset with best accuracy, least number of cycles and average number of large item sets. Predictive Apriori took maximum execution time for rule generation with maximum number of rules generated for Dharamshala dataset followed by Solan and Shimla dataset. Accuracy of rules generated is approximately equal for all three datasets.

Energy Import

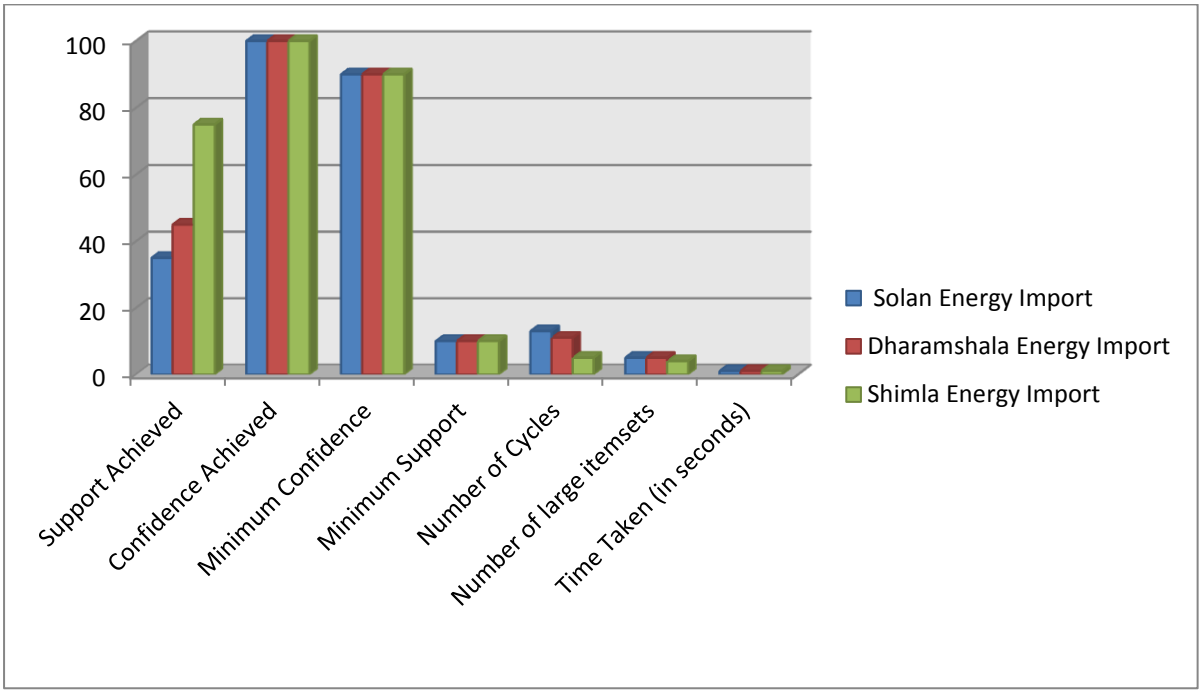

Figure 5: Energy Import Graph with Apriori Algorithm

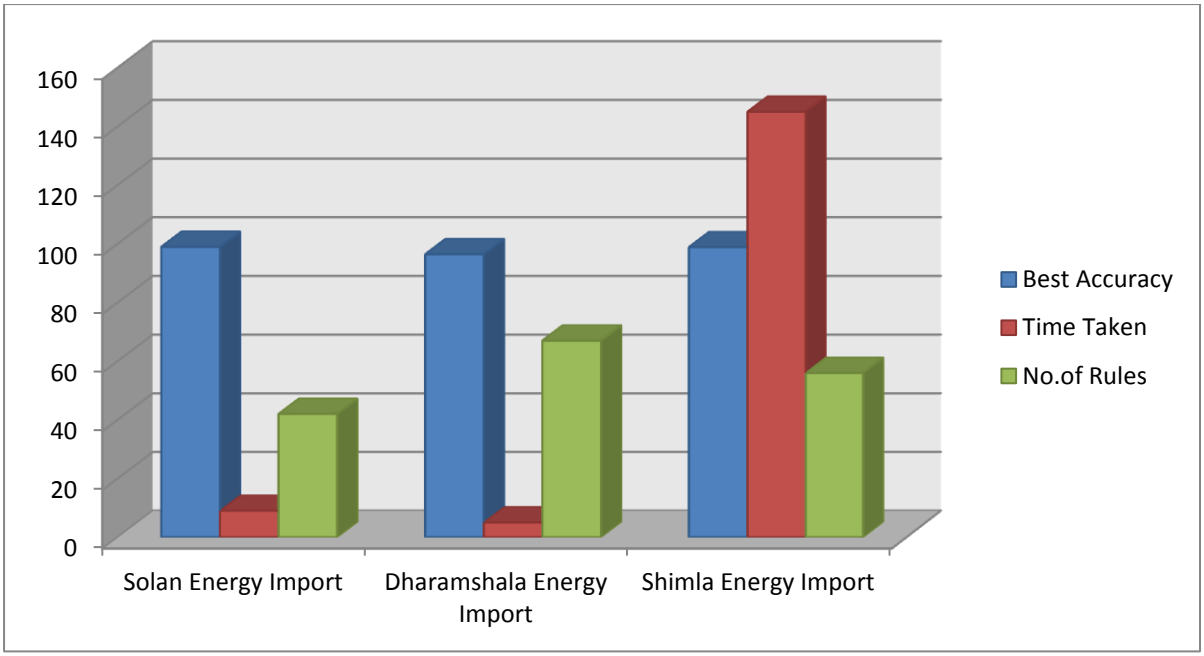

Figure 6: Energy Import Graph with Predictive Apriori Algorithm

The rules were generated with full 100% confidence for all three towns with Apriori being the fastest form predictive Apriori. Maximum Support was obtained by Shimla dataset in case this case with best accuracy near to 100. Least number of cycles or scan into dataset was made in case of Shimla dataset .The most number of rules in case of Predictive Apriori was generated for Dharamshala dataset. Performance of Solan dataset was minimal among all three.

Billing Efficiency

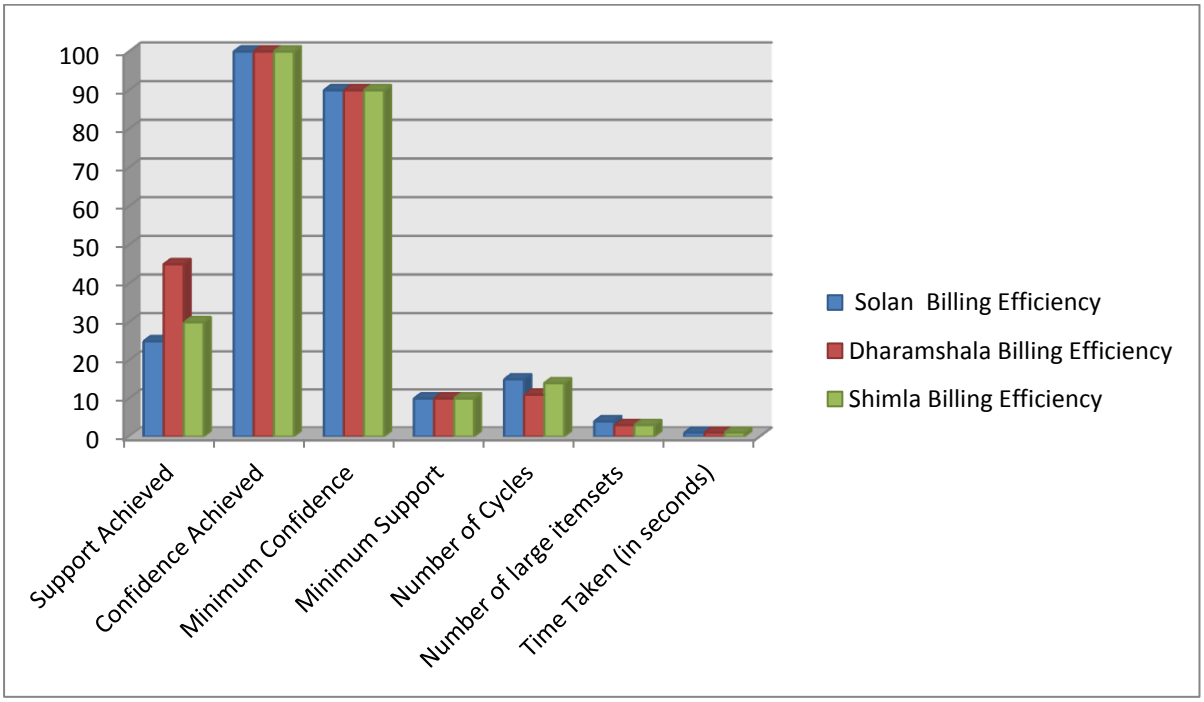

Figure 7: Billing Efficiency Graph with Apriori Algorithm

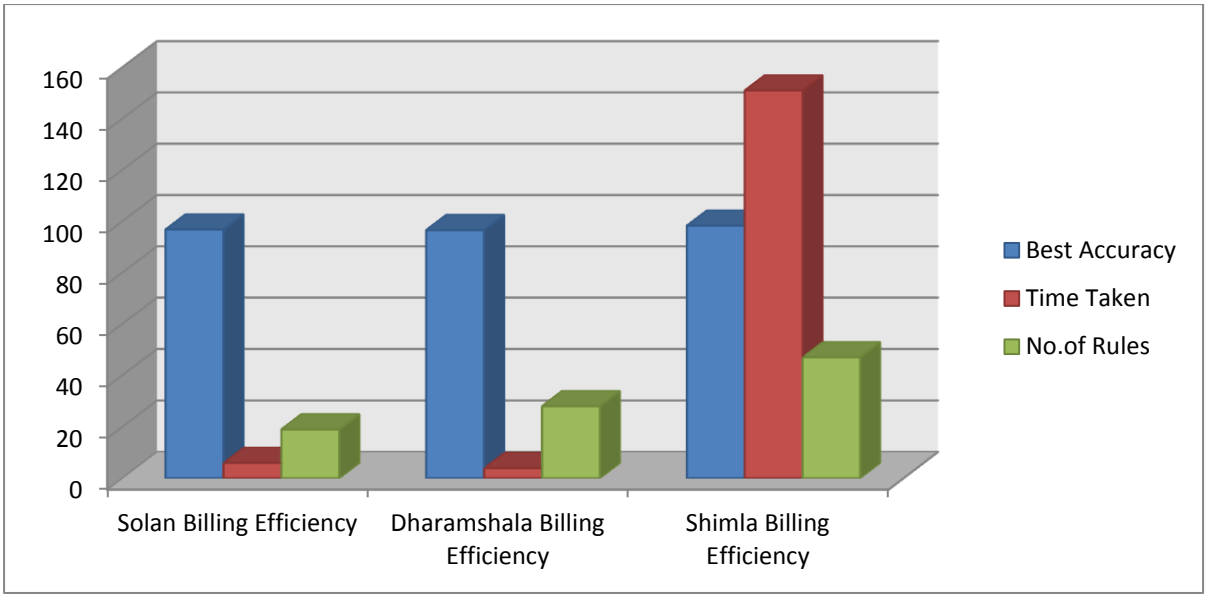

Figure 8: Billing Efficiency Graph with Predictive Apriori Algorithm

Best rules for billing efficiency as shown in graph with Apriori algorithm were generated for Dharamshala dataset with least number of cycles /scan into dataset and with average number of frequent item sets. Best accurate rules were generated for Shimla dataset followed by Dharamshala and then for Solan. Most number of predictive accurate rules was generated for Shimla among all three datasets. Predictive Apriori was slowest in execution in case of Shimla dataset and fastest for Dharamshala Dataset.

AT&C Losses

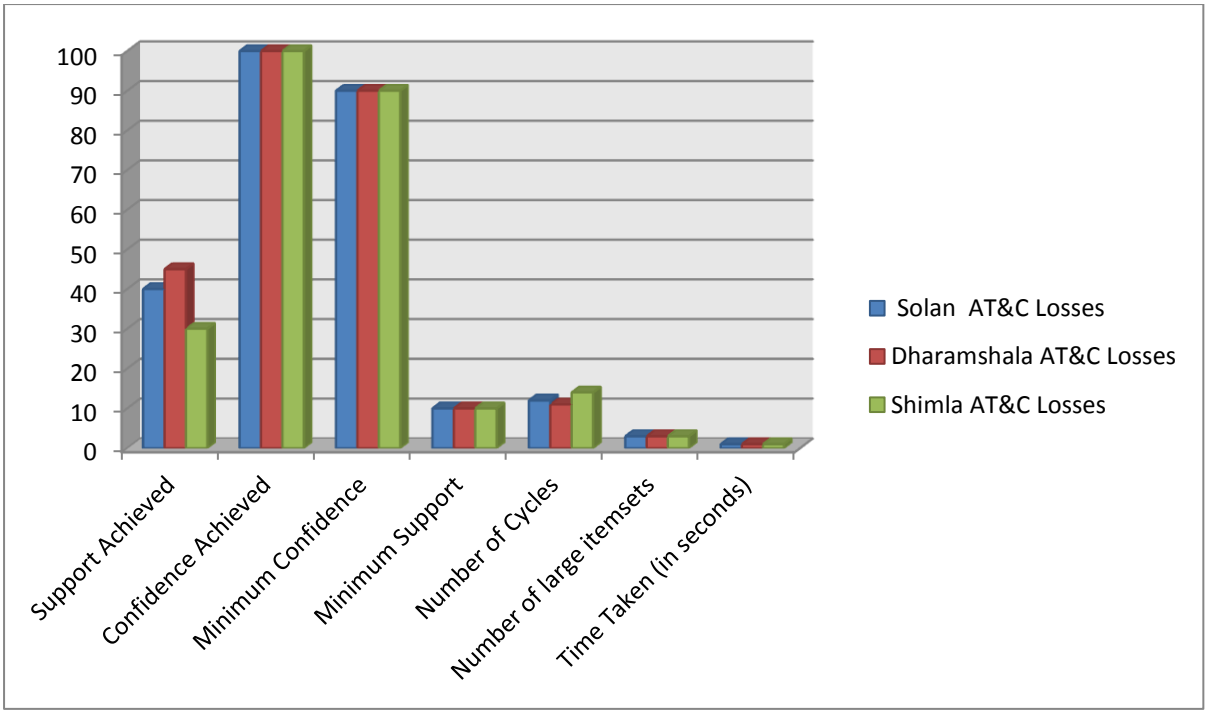

Figure 9: AT&C Losses Graph with Apriori Algorithm

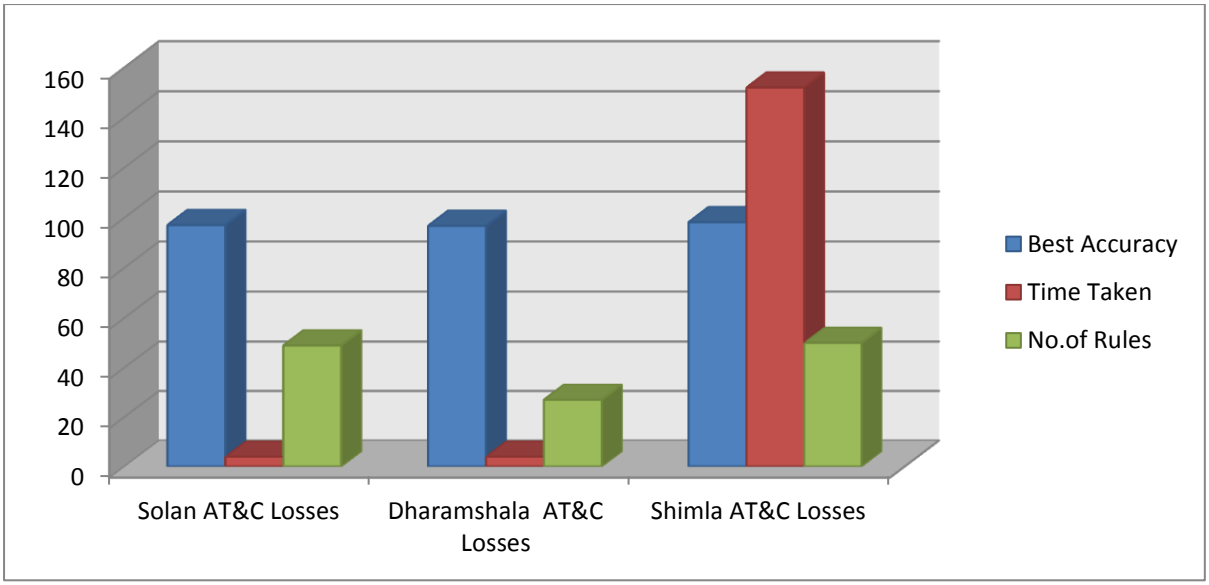

Figure 10: AT&C Losses Graph with Predictive Apriori Algorithm

The best value of support for rules generated in case of Aggregate Technical and commercial losses was for Dharamshala dataset with least number of cycles and frequent large item sets. Best accurate rules were generated for Shimla dataset with large execution time in case of predictive Apriori algorithm. Most number of rules for this attribute was generated equally for solan and Shimla dataset. Rules generated for all three datasets obtained full 100% confidence with. With best accurate and maximum supported rule good inferences can be made.

VI. Conclusion & Future Scope

The Apriori algorithm given by Aggrawal and Srikant **[1]** serves a major basis for this work. It can be concluded that general association rules give varying support accuracy and confidence values for rules generated hence leaving a large horizon for prediction/analysis to be made. It is also proven in case of Shimla dataset that least number of cycles or scan give accurate rules with maximum support. Rules were considered in both positions of antecedent and consequent with their confidence and accuracy achieved hence provided versatile analysis and causes behind them. AT&C Losses predicted in each case of every dataset have left with several causes behind them and reason and prevention for these causing factors have been well explained in previous chapters. Hence, for Himachal Pradesh State Electricity Board it can prove useful to bring the value of AT&C losses down to low as in some cases sub division, months and number of towns with major losses have also been identified with possible causes. These data mining algorithms and open source weka has proven to be good combination of technique and programming for business data prediction. For future work such type business datasets can be subjected to clustering and other algorithms in association rules to seek improvement the prediction/analysis made which inherently support numeric type attributes as Apriori algorithm have also given quite good results.

References

[1] Agrawal, Rakesh, and Ramakrishnan Srikant. "Fast algorithms for mining association rules." *Proc. 20th int. conf. very large data bases, VLDB*. Vol. 1215. 1994.

[2] Bansal, Divya, and Lekha Bhambhu. "Execution of APRIORI Algorithm of Data Mining Directed Towards Tumultuous Crimes Concerning Women." *International Journal Of Advanced Research in Computer Science and Software Engineering, ISSN* 2277 (2013).

[3] Berson, Alex, Stephen Smith, and Kurt Thearling. *Building data mining applications for CRM*. New York: McGraw-Hill, 2000.

[4] Hegland, Markus. "Algorithms for association rules." *Advanced lectures on machine learning*. Springer Berlin Heidelberg, 2003. 226-234.

[5] Madhun V. Joseph, Lipsa Sadeth and Vanaja Rajan ,"Data mining : A Comparative study on various Techniques and Methods", Published in IJARCSSE Journal Volume 3rd, Issue 2nd , February 2013.

[6] Ramaraj, E. "An efficient pattern mining analysis in health care database." (2009).

[7] Sharif, Ahmad, et al. "Using a priori algorithm for supporting an e-commerce system." *Journal of Information Technology Impact* 5.3 (2005): 129-138.

[8] Wu, Xindong, et al. "Top 10 algorithms in data mining." *Knowledge and Information Systems* 14.1 (2008): 1-37.

[9] www.cs.ccsu.edu/~markov/ccsu_courses/DataMining-6.html accessed on 15 October at 12:00 PM.

[10] www.hevellum.files.wordpress.com/2012/03/datamining-paper-03152012-v01-morecomplete-2chapters.pdf retrieved on 16 October at2:00 PM.

[11] www.hpseb.com/rapdrp.htm accessed on 14 October 2014 at 11: 00 A.M.

[12] www.investopedia.com/terms/d/datamining.asp accessed on 15 October at 12:00 PM.

[13] www-users.cs.umn.edu/~kumar/dmbook/ch6.pdf retrieved on 3 October at 11:30 AM.

