Enabling Predictive Maintenance Strategy in Rail Sector: A Clustering Approach

JOHN VICTOR ANTONY
Supervisors' Training Centre, South Western Railway
Bangalore
INDIA
john_j_71@yahoo.com

G M NASIRA
Chikkanna Government Arts College
Bharathiar University
Tamil Nadu
INDIA
nasiragm99@yahoo.com

Abstract: - One of the imperatives of predictive maintenance of assets is to analyze, understand and act upon the failure pattern hidden in the failure data which are represented, in general, by failure code, failure description and failure instance. A maintenance plan that will be in consonance with the failure trend inferred through data mining is bound to enhance the asset reliability. The paper shows how to integrate clustering approach into the realm of asset maintenance and particularly provides a road map to implement predictive maintenance strategy in rail sector. The proposed approach has been tested on actual failure data pertaining to passenger carrying vehicles of the trains. Finally, the performance of two fundamental approaches i.e. hard and soft clustering has been investigated on the data set and a recommendation made therein.

Key-Words: - Clustering, Data Analysis, Data Description, K Means algorithm, Railways, Pattern Analysis.

1 Introduction

For operation of trains, Railways maintains two broad categories of resources in large quantity namely Rolling Stock (RS) and Permanent Way (PW). The RS consists of Coaching Stock (CS) also called as coaches, for carrying passengers, Freight Stock (FS) for goods transit and Locomotives (Locos) for hauling the trains. These RS resources are subjected to preventive maintenance at regular interval at their respective maintenance depots. The maintenance plan (MP) generally is static in nature containing predefined items to be checked and is mostly based on the Original Equipments Manufacturers' (OEM) instructions. Information Technology has made inroads in these maintenance activities and the online failures are being logged into a data base. The data base schema generally consists of, among others, a mixture of nominal and numeric attributes, invariably representing the failure categories and their occurrence in terms of distance at which the failure occurred. With the help of data mining, we attempt to study these failures, generate failure patterns and strengthen the existing static maintenance plan with the knowledge obtained. It is suggested that the priority and frequency of maintenance can be tuned based on the failure pattern exhibited by the failure history for improved and fail safe service.

2 Problem Statement

The research attempt is focused on how to apply data mining into failure history of CS for evolving a better maintenance plan for coaches. Fig.1. depicts the coach management system of railways. The CS is drawn for service from two sources such as Production Units (PU's) and Periodical Overhauling (POH) Work Shops. While Production units manufacture new vehicles, the Periodical Overhauling Shops are meant for overhauling the vehicles after they complete a specified period of life in service. The CS from these two sources is sent to different Train Care Centres (TCC's) which maintain and provide these vehicles for Train Operations. The maintenance is preventive in nature

and is done by adhering to a checklist of items to be checked at regular intervals.

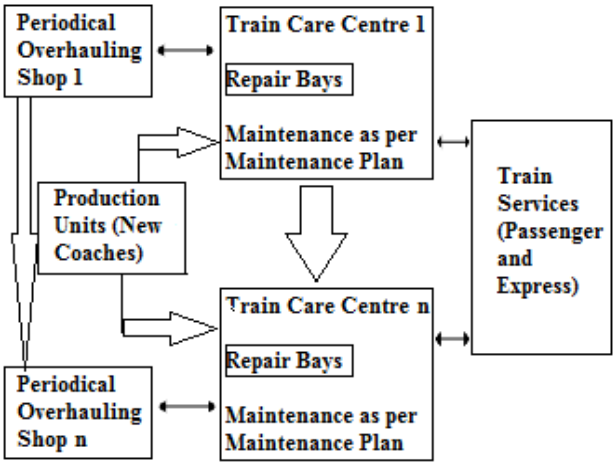

Figure 1: Coach Fleet Management

Despite the best efforts provided by TCC's, these vehicles fail prematurely leading to performance degradation and hence passenger inconvenience. Even though the failures are logged into databases, there exists a definite gap in the analysis and hence the control over failure has been a case of concern. Every now and then special maintenance instructions are issued to tackle the online failures. Any solution that provides a better understanding of the failures can be a source of intelligent maintenance plan for improved and reliable service. Is there any suitable data mining method that can bring out the failure trend of the CS? If so, how can we apply data mining in a methodological way to mine the failure history and generate the patterns? Is there a frame work to incorporate the mined knowledge into the maintenance plan? These are the pertinent questions that will be addressed in the literature. The paper addresses the problem in two parts i.e.

- Application of clustering technique to generate failure patterns of CS so that maintenance plan that is resonant with the failure can be evolved.
- Empirical study of the performance of soft and hard clustering approaches on the typical coach failure data set.

3 Clustering: The Hard and Soft Variants

Clustering is a data mining technique [1] that allows us to make groups of data from a dataset that share similarity among the data objects in certain ways. The data groups shall maintain maximum inter group dissimilarity and maximum intra group similarity. The similarity and dissimilarity are calculated using various distance measures [2], depending on the type of attribute. Clustering helps us to find out the data patterns that are not easily visible to naked eye and can not otherwise be determined using simpler means. Clustering into data, when performed in an appropriate way, brings out the hidden patterns and improves our understanding about the data at hand. The derived patterns are thus helpful in arriving at meaningful conclusions. Fundamentally, there are two clustering approaches such as hard and soft clustering.

Hard Clustering [3] also called as rigid clustering refers to partitioning method in which a scheme called exclusive cluster separation is followed i.e. each data point belongs to exactly and only one of the partitions. K means algorithm adopts such a partitioning scheme. It uses either the default Euclidean distance or the Manhattan distance measures. If the Manhattan distance is used, then centroids are computed as the component wise median rather than mean. K-means algorithm is applicable, when the mean of a set of objects is defined but it is not possible to define means of nominal variables. In such scenario, it outputs for each cluster, the frequency counts of the values of each nominal attribute. This feature can be used to analyze nominal data by treating it as k way categorical variable with k different values or states and accordingly creating data summary by either using a coding scheme [4] for all nominal attributes or otherwise.

Soft Clustering refers to partitioning method in which exclusive cluster separation is relaxed so that each data point has certain probability of belonging to each of the partitions. Soft clustering allows the objects to take part in several clusters, simultaneously, with different degrees of membership. Soft clustering is also called as fuzzy clustering [5] and probabilistic clustering. Expectation Maximization (EM) algorithm is an example of soft clustering. The differentiating point between hard and soft clustering approaches is the basis on which a data object is assigned to a cluster. A detailed account on information theoretic analysis of hard and soft assignment can be found in [6]. The current paper deals with application and suitability of these approaches in the realm of predictive maintenance strategy

4 Related Work

There have been few attempts made by researchers on topics of rail sector. These mainly belong to works on Track, Time Table, Ticket Analysis and

Train Scheduling. The methods adopted in large cases are Genetic algorithm, Artificial Neural Network; Symmetry based technique, Linear Programming, Clustering and Petri Net modeling. Use of data mining technologies on real time data exclusively on rail sector with maintenance and management as topic of interest has been rare. However, an account of few relevant literatures that carry significance to rail sector and to the present paper is provided here. Reference [7] describes pattern recognition needs and gives a solution approach in general on track failures incorporating more computer science based data analysis to reduce the testing errors. There are few similar works on track worth mentioning of which are [8], [9]. These papers talk about analysis techniques based on images provided by camera and comparisons therein to find the errors. The work of [10] is about computer trending tool for predictive monitoring of wheels and bearings using various detector outputs. Yet another work on rail track maintenance using rail profile parameters is given by [11] that predict rail profiles using linear regression model for better maintenance of the same. Reference [12] describes successful prognostics that should address several difficulties including data selection, data fusion, data labeling, model integration, and model evaluation. This paper explains these issues and presents a systematic methodology with reference to the rail and aerospace industries highlighting open problems. Reference [13] has attempted to discover temporal association using T pattern algorithm between pairs of time stamped alarms from onboard systems having positional and communicational equipments. The temporal associations between pairs of time stamped alarms, called events have been used to predict the occurrence of severe failures within a complex environment. But, the scenario visualized in the current research attempt is that the failures are more prone to the distance run than the time. The idle time after service fitness of the vehicles is quite high on account other operational requirements related to safety, traffic and road clearance. Understandably, the maintenance plan needs to be based on the distance or the impact of distance on the performance should play a significant role, especially in the maintenance of passenger coaches and wagons. Accordingly, the factors that cause failure of systems and subsystems is predominantly, the distance run among others such as the quality of materials, the maintenance procedures, type of vehicle and the forces of load acting on the vehicle as a result of speed. The present paper looks from stand point of these factors and rely on the distance patterns, there by differing from the previous work. Reference [14] has used clustering technique for the analysis of urban rail tickets. It deals with four cluster variables such as carfare frame consistency, passenger attractiveness, urban traffic coordination and city characteristic consistency, analyzes ticket type settings qualitatively and finally establishes a model in accordance with actual survey data of rail transit. Reference [15] is a work on prediction of bearing failures of machines using neural networks with vibrations signals as input. A statistical approach using condition data of railway infrastructure assets has been used in [16] wherein the data obtained from regular inspections done by a railway track measurement wagon have been analyzed to evaluate the possibility of detecting derailment, using control chart technique. An attempt to pinpoint anomalies on the status of pantograph contact of overhead power lines is found in [17] that process the data relative to voltage and current collected on high speed trains along with a set of measurements coming from photo sensors, using Support Vector Machines (SVM). Fault diagnosis of railway electrical point machines has been presented in [18] that mainly uses Wavelet Transforms, SVM and *k*-means clustering algorithm to find an appropriate parameter with drive force, current and voltage as data.

5 Data Set and Methodology

5.1 Data Set

From the data base of a TCC, a data set having around 3000 to 4000 records pertaining to 10 months of a year was formed and pre processed. All these records represent failure details due to various parts and equipments of the coaches of trains. The details of the attributes under study are provided in Table 1. Since the attributes have many labels, only select labels of attributes have been considered for study purpose and for avoiding problem of enormity. The numbers in the bracket refer to the labels available in the original dataset and the numbers without bracket indicate the ones considered for the study. For clarity, a brief account about the attributes is given here.

1. *Code:* This indicates the code attached to each vehicle, based on the purpose for which it is used such as General Compartment, AC chair car, AC Thee Tier etc. For example, WGSCN means second class three tier sleeper and GS means general second class compartment and so on. WGACCW and WGSCZ are the two

vehicles considered in this model. WGACCW means Air conditioned Two Tier coach and WGSCZ means Air conditioned Chair Car.

2. *POH Stn:* This attribute indicates the name of the service station which has carried out the maintenance (Periodical Overhauling) for coaches in service or the manufacturing station in case of new coach. PWP (Carriage Works, Perambur, Chennai), ICF (Integral Coach Factory, Chennai) and GOC (Golden Rock Shop, Trichy) are the codes of three stations considered here.
3. *Bogie Type:* This indicates the type of bogie used in the coach. The bogie here carries the entire coach structure on it. There are two bogies in each coach. The type of bogie considered in this model is ICF.
4. *Brake:* This indicates the type of brake system the coach is fitted with. This model considers B type (Air) brake system.
5. *Sick Desc:* This nominal attribute explains the major failure caused by the system and subsystem of the coach. Description of five failure categories appearing in this model is shown in Table 2 under the group SICKDESC.
6. *Kms Earnd*:-It is a numerical attribute indicating at what kilometer the vehicle has failed. It is the total distance run or earned by the coach at the time of occurrence of failure. For analysis, it is discretized in to ranges of every 10000 kilometers and referred to as "distance label" in ensuing literature The original dataset includes failure instance up to 700000 kilometers.

Table 1: Attributes Considered

Attribute	Type	Description	Labels Considered (Total)
Code	Nom	Coach Type	2 (51)
POH Stn	Nom	Overhauling/manufacturing station	3 (15)
Bogie Type	Nom	Bogie type or model	1 (2)
Brake	Nom	Type of brake system	1 (3)
Sick Desc	Nom	Sick/Failure Category	5 (55)
Kms Earnd	Num	Kilo meter at the time of failure	9 (70)

5.2 Proposed Methodology

The methodology is provided in graphical form in Fig.2. It shows the step by step sequence of actions followed in the methodology. The methodology can be iteratively repeated to evolve or fine tune the MP. Its generic structure makes it eligible for use in any similar situation in which previous failure data can provide meaningful direction in maintenance plan of machineries and plants.

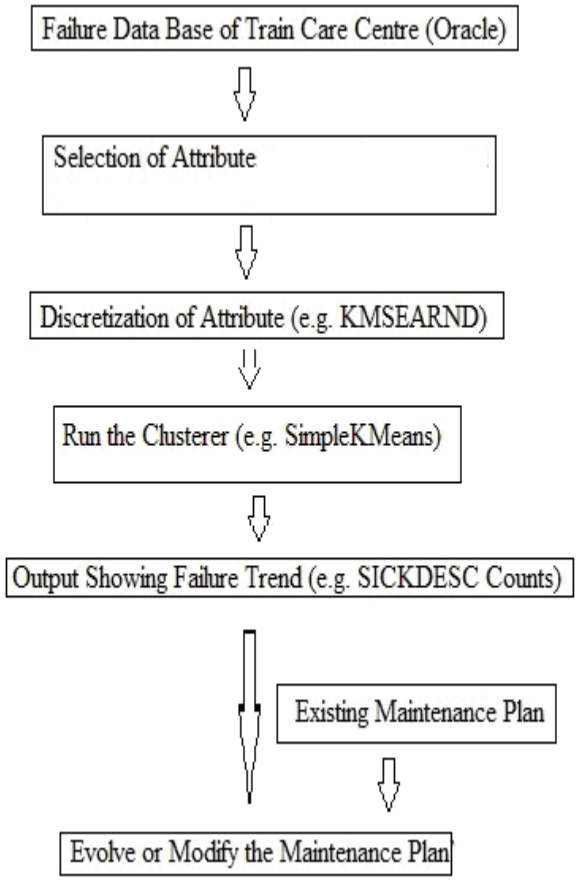

Figure 2: Methodology.

6 Experimental Setup

6.1 Data Mining Tool

Weka 3.6.7 which is an Open Source Data Mining tool is considered for use. Written in Java and developed at the University of Waikato, New Zealand, it is freely available under GNU General Public License. Weka stands for Waikato Environment for Knowledge Analysis. Weka 3.6.7 is a collection [19] of machine learning algorithms for data mining tasks. It contains tools for Data pre processing, Classification, Regression, Clustering,

Association rule mining, Feature Selection and Data visualization. Reference [20] gives an account of Weka interfaces.

6.2 Parameter Configuration

The parameter configuration deals with setting the values for the important parameters of the data mining software during experimental run. The experiment was conducted in two scenarios one using K means algorithm and the other using EM algorithm. K means and EM algorithms are representatives of hard and soft clustering respectively. The screen shot of configuration for the clusterer "SimpleKMeans" appears in Fig.3. The key parameter values are displayStdDevs = True (This setting sends the frequency counts of the nominal variables to the output which is the desired feature for the present context, otherwise it outputs standard deviations of numeric attributes), distanceFunction = Euclidean distance or Manhattan distance and number of clusters = 1 (intentionally set so). The other values are allowed to be at their default setting as found in the screen shot. The parameter configuration for EM algorithm is shown in Fig.4. The parameter values other than numClusters = 1 such as dubug, displayModelInOldFormat, maxIterations, minStdDev and seed are left at their default setting.

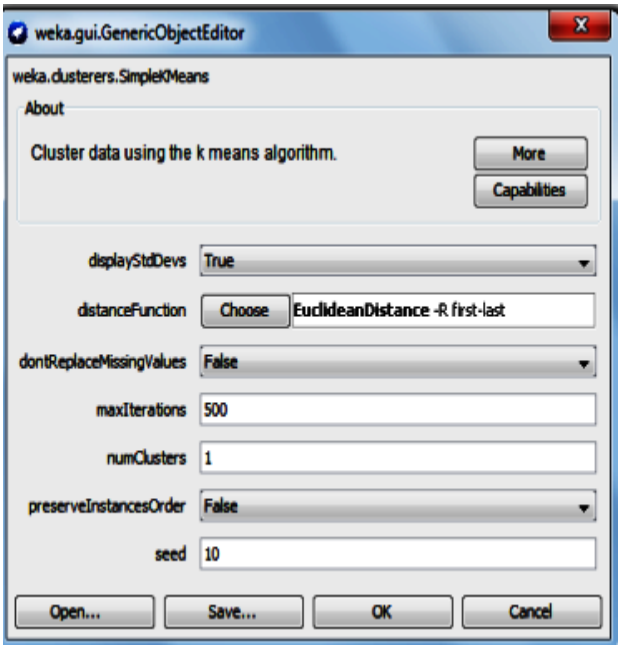

Figure 3: Configuration Interface of Simple K Means Clusterer

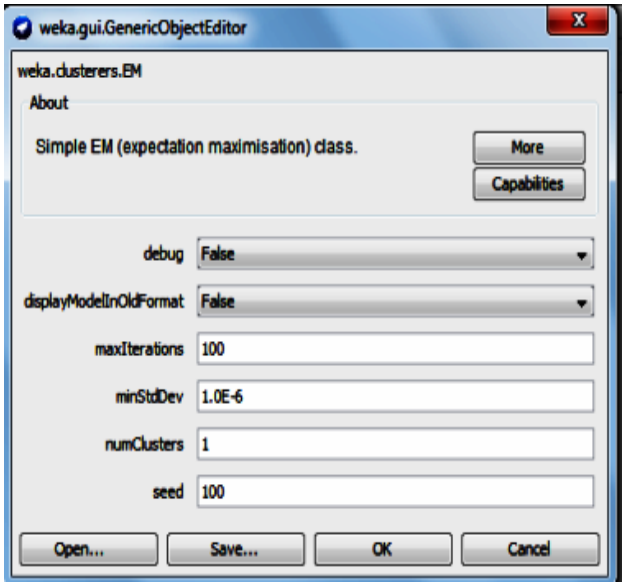

Figure 4: Configuration Interface of EM Clusterer

6.3 Data Objects for the Experiment

A dataset having 27 failure records meeting the specification provided in Table 1 was chosen for the experiment.

7 Results and Discussion

7.1 Output for Macro Level Maintenance Plan

This part of the text explains how the cluster results of the data set can effectively be used for understanding the big picture of the dataset and further effecting maintenance decisions at macro level. The macro level here means decisions addressing category wise failures, how they are distributed on different distance ranges, coach type population, and the coach distributions as owned by different overhauling / manufacturing stations etc. The output obtained from running the K means algorithm on the full dataset is provided in Table 2. Here, the frequencies of the five failure categories considered (shown as SICKDESC) and the failure frequencies of nine distance labels of the discretized attribute "KMSEARND" are considered for discussion. These two variables and their distributions can directly be invoked and correlated suitably in to maintenance decisions while others included in the dataset carry insights that are useful in managerial dimensions of higher level.

Table 2: k means Output for all Distance Ranges

Attribute	Full Data (27)	Cluster #1(27)
BRAKE		
B	27 (100%)	27
CODE		
WGACCW	19 (70%)	19
WGSCZ	8 (29%)	8
BOGIE TYPE		
ICF	27 (100%)	27
POHSTN		
PWP	23 (85%)	23
ICF	1 (3%)	1
GOC	3 (11%)	3
SICKDESC		
ALTERNATOR	4 (14%)	4
V-BELT	12 (44%)	12
PRIMARY SUSPENSION	7 (25%)	7
BOGIE MOUNTED BRAKE CYLINDER	3 (11%)	3
BUFFING GEAR	1 (3%)	1
KMSEARND		
<70000_80000)	5 (18%)	5
<80000_90000)	2 (7%)	2
<20000_30000)	3 (11%)	3
<30000_40000)	3 (11%)	3
<90000_100000)	4 (14%)	4
<10000_20000)	3 (11%)	3
<40000_50000)	5 (18%)	5
<60000_70000)	1 (3%)	1
<50000_60000)	1 (3%)	1

Here, the five failure categories depict the failure trend in varying order, as generated by various systems and subsystems of the coaches. The nine distance labels having significant frequency (counts) represent the number of failures that occurred in the respective distance range. These five failure categories and nine distance labels are representatives of global failure patterns of the coaches with respect to systems/sub systems and the distance ranges. The counts of the failure categories and their distribution with reference to the distance range can form base for macro level maintenance plan and setting the required inspection priorities. Clustering into the entire database will give a complete performance status of a TCC. This will enable the maintenance managers to frame appropriate maintenance instructions in tune with the ground reality. Given the total coach holding of a TCC, strategic maintenance decisions with respect to performance of different vehicle types, performance comparison among PU's/POH stations, failure vulnerabilities of system and subsystem, distance (D pattern) based failure distributions and the proactive actions required to be taken etc are some of the benefits to the organization.

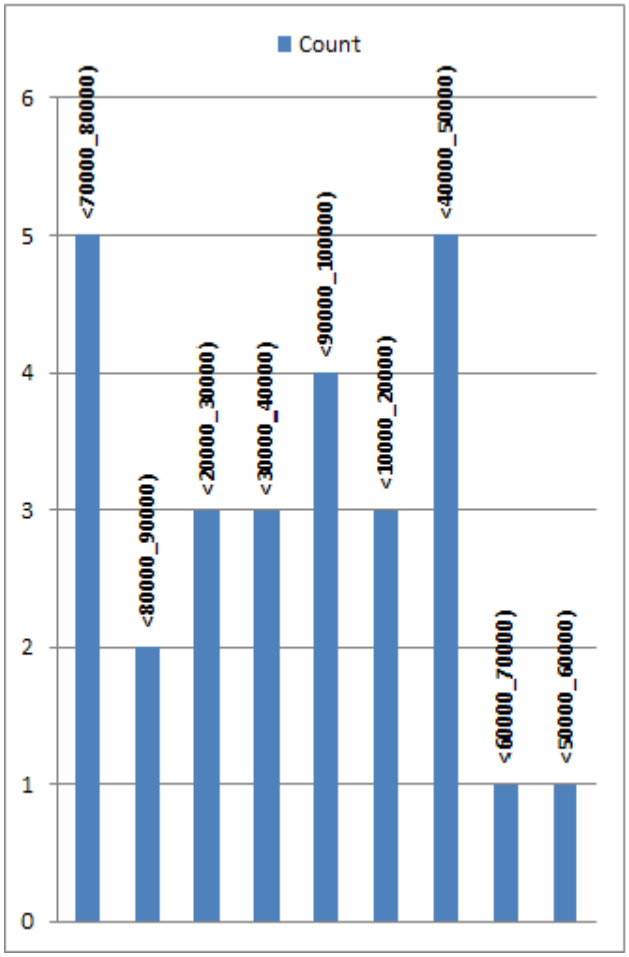

Figure 5: Distance wise Failure Distribution

Fig.5. graphically represents the failure frequency distributions with respect to all the distance ranges

available in the dataset and Fig.6 represents the category wise failure distributions with respect to all the distance ranges.

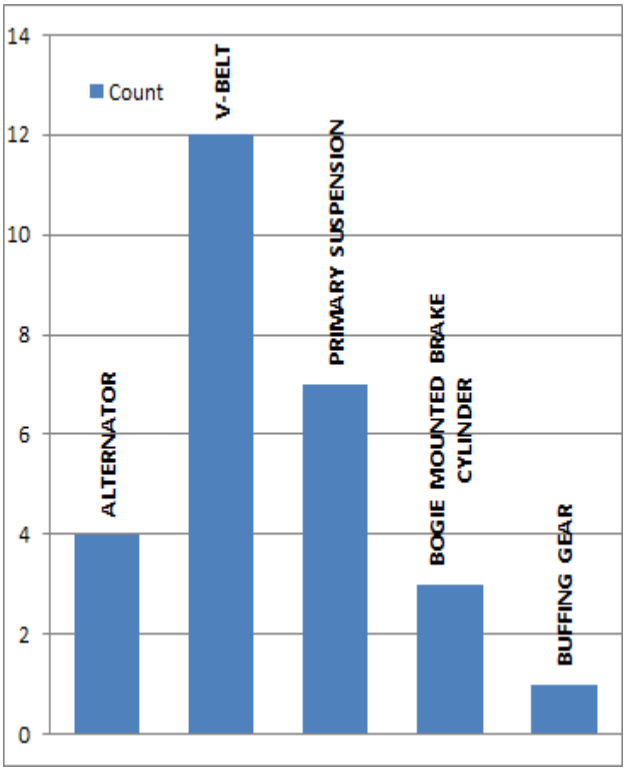

Figure 6: Category wise Failure Distribution

7.2 Output for Micro Level Maintenance Plan

The coaches return to TCC for maintenance after having run some distance which shall fall within a maximum of 5000 to 10000 kms. Hence, for micro level understanding of the failures patterns, a distance (D pattern) based failure trend would be a desired requirement for fine tuning the maintenance plan with respect to the distance range concerned. Determining the range of distance label can be left to discretion of the managers involved in maintenance. The distance label with maximum failure counts signifies the priority and gravity of attention required to be paid during the maintenance for coaches whose distance covered falls within that specified distance range. For example, D pattern was generated by filtering other data instances and running the algorithm on those data instances pertaining to distance range <70000_80000) with same parameter setting outlined earlier.

Table 3: k Means Output for Distance Range <70000_80000)

Attribute	Full Data (5)	Cluster #1 (5)
BRAKE		
B	5 (100%)	5
CODE		
WGACCW	2 (40%)	2
WGSCZ	3 (60%)	3
BOGIE TYPE		
ICF	5 (100%)	5
POHSTN		
PWP	4 (80%)	4
GOC	1 (20%)	1
SICKDESC		
ALTERNATOR	2 (40%)	2
V-BELT	1 (20%)	1
PRIMARY SUSPENSION	2 (40%)	2
KMSEARND		
<70000_80000)	5 (100%)	5

The output obtained is provided in Table 3 which is self explanatory and will enable meaningful inferences by the maintenance department. Fig.7. graphically represents the failure distribution with respect to the distance range <70000_80000). This has been compared with the ground truth available in dataset and found tallying with the macro level realities. The failure distributions of different distance labels can suitably be correlated by the maintenance managers to further sensitize the entities involved in maintenance.

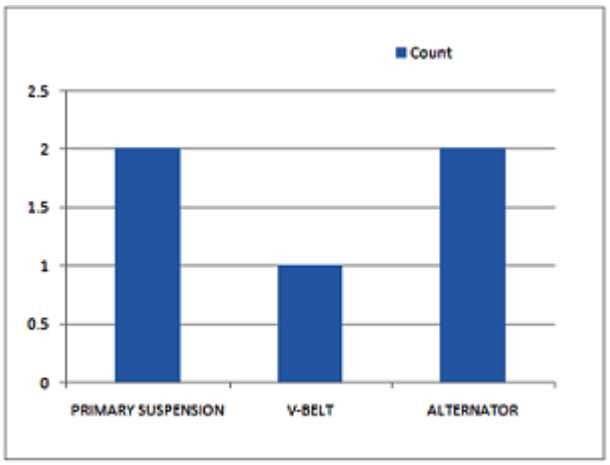

Figure 7: Failure Distribution for Distance Range <70000_80000)

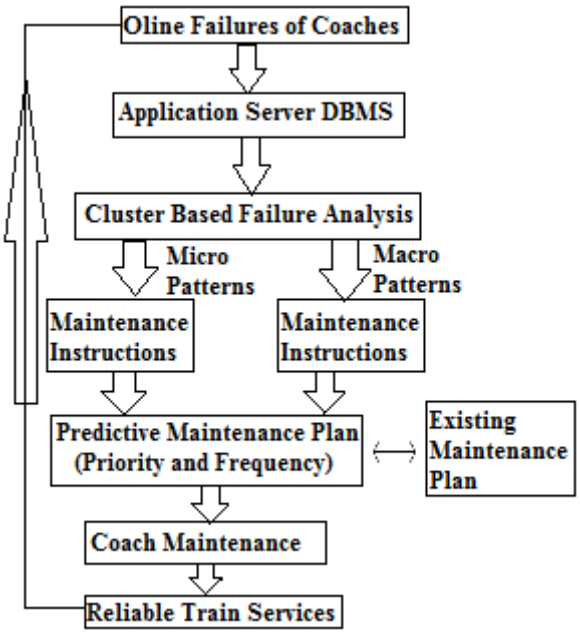

Figure 8: Coach Maintenance System.

The coach maintenance system with our methodology integrated into it is shown in Fig.8. It is pointed out that the MP derived by incorporating the predictive behaviour pattern of the failure history in collaboration with the existing MP will undoubtedly lead to improved and reliable train services. However, for successful and sustained results, the predictive maintenance has to go hand in hand with the preventive maintenance, complementing each other in a massive safety oriented organization like railways that considers safety as paramount factor. This work is to be taken as a frame work to convert failure data into insights and transform the existing MP from being static into information driven. The MP that is information driven is bound to enhance the reliability and bring down the overall cost of maintenance. The routine inspection of certain items that do not generate any failure pattern over a span of time can either be avoided or judiciously postponed. The items that fall into the failure pattern mined can be given more preference in terms of inspection priority and frequency. It is thus emphasized that clustering based failure analysis can help us frame predictive maintenance plan with failure information infused into it, for improved coach maintenance system.

7.3 Comparing Soft and Hard Clustering

With an intention of studying the performance of two fundamental approaches i.e. hard and soft clustering, the experiment was conducted using soft clustering algorithm also, namely Expectation Maximization. The output obtained from running the two clustering algorithms i.e. SimpleKMeans (Hard Clustrer) and EM (Soft Clusterer) is provided in Table 4 for the purpose of comparison.

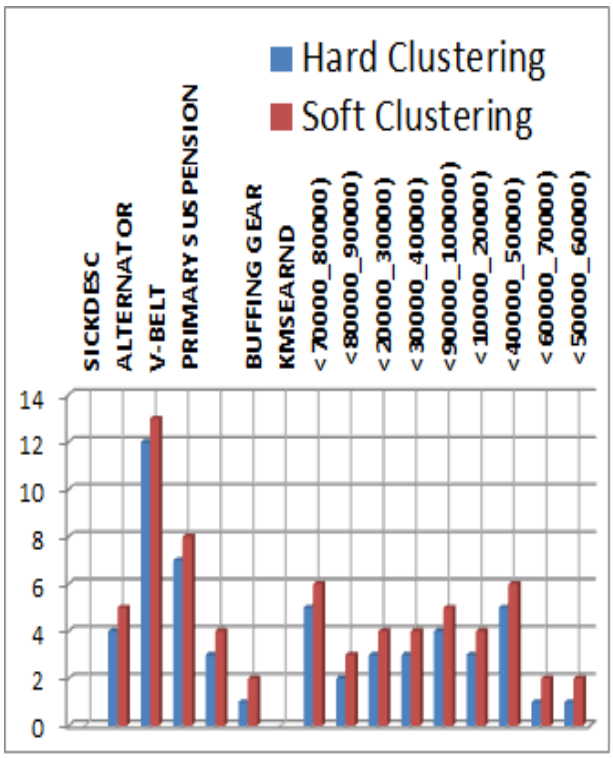

Figure 9: Comparing the outputs of K means and EM Clusterers

Table 4: Output of K means and EM Clusterers

Clustering Type		Hard (K means)	Soft (EM)
Attribute	Full Data (27)	Cluster #1 (27)	Cluster #2 (27)
BRAKE			
B	27 100%)	27	28
CODE			
WGACCW	19 (70%)	19	20
WGSCZ	8 (29%)	8	9
BOGIE TYPE			
ICF	27 100%)	27	28
POHSTN			
PWP	23 (85%)	23	24
ICF	1 (3%)	1	2
GOC	3 (11%)	3	4
SICKDESC			
ALTERNATOR	4 (14%)	4	5
V-BELT	12 44%)	12	13
PRIMARY SUSPENSION	7 (25%)	7	8
BOGIE MOUNTED BRAKE CYLINDER	3 (11%)	3	4
BUFFING GEAR	1 (3%)	1	2
KMSEARND			
<70000_80000)	5 (18%)	5	6
<80000_90000)	2 (7%)	2	3
<20000_30000)	3 (11%)	3	4
<30000_40000)	3 (11%)	3	4
<90000_100000)	4 (14%)	4	5
<10000_20000)	3 (11%)	3	4
<40000_50000)	5 (18%)	5	6
<60000_70000)	1 (3%)	1	2
<50000_60000)	1 (3%)	1	2

The frequency count provided by hard clustering algorithm matches the ground truth of the attribute values. But, the frequency count provided by soft clustering algorithm deviates consistently upward from the ground truth by one for every attribute value within the attribute domain. It thus gives a distorted insight about the data under study. The experiment was repeated with different data sets also and the deviation was observed. The comparison is shown graphically in Fig.9 for two variables namely SICKDESC and KMSEARND. Detailed failure summarization against each distance label (range) was obtained by conducting the same experiment on records pertaining to that distance label. Here again, the deviation in frequency count among the soft and hard clusterer was evident. The comparison highlights the fact that the hard clustering approach is better and suitable than the soft counter part, in the present context involving our typical data set.

7.4 Cluster Evaluation Criteria

The two significant cluster evaluation criteria such as number of clusters to be created and measurement of clustering quality are discussed here. Since the value of frequency count of failures shall determine the order and priority of inspection, it is purposely ensured that the number of clusters required to be created is one. As regards the cluster quality, extrinsic and intrinsic methods are the two ways of measuring cluster quality. The extrinsic method is applicable, when the ground truth of the data is available and the latter can be used in case of non availability of ground truth. The frequency count (of attributes) in each cluster obtained using hard clustering approach was compared with ground truth of the data and found to be correct, while the result obtained out of soft clustering approach was not matching with the ground truth.

8 Applicability: Another Example

Apart from the context described above that attempt to strengthen a MP of CS, the usage of the methodology can be extended to another live scenario as explained in the following text. The above context essentially provides frequency distribution of failures categories and relates the failure frequency with distance patterns. As an extension, each failure can further be associated to

the respective components with an idea of ensuring quality control over the manufacturers/suppliers. The components/parts could come from outside manufacturers or could have been from different railway units acting as production centers. The parts that go into a coach can range from those belonging to suspension, draw gear, bogies, brake system, electrical power system, safety and passenger amenities to name a few. The following are the data fields.

- Part Number
- Part Description
- Supplier Code (both outside and from within railways)
- Part Fitted Date
- Coach Code (with which the part is fitted)
- Date Failed
- Failed Kilo Meter (the distance at which the part failed)
- Stock Type (whether the part is regularly stocked or none stocked, indicating the mode of purchase Railways generally follows on the basis of cost, lead time etc)
- Down Time (of the coach due to the failure)

Clustering into the above data schema can provide valuable insights with respect to which part has led to more downtime, manufacturer wise part failure frequency distribution, any relationship between failure and distance etc. A good understanding of material failures could pay way for better quality materials being allowed into the system and hence better train services.

9 Conclusion

The paper has attempted to show a methodology that uses clustering technique for bringing out the failure trend or patterns hidden in the failure history of passenger carrying vehicles of trains. It has also been shown that the information and knowledge thus mined can be used to significantly tune the Maintenance Plan. The Maintenance Plan founded upon the knowledge of failure pattern brought out by clustering technique will surely be an effort towards predictive maintenance and management of assets envisioned in Vision 2020 statement of Railways [21]. The key idea behind the predictive approach is to effectively analyze and use the data captured so that further failures can be avoided before their occurrence, by fine tuning the maintenance plan. This idea has been adequately explained in this paper, by providing a coach maintenance system that incorporates clustering based failure analysis to infuse the failure information into the maintenance plan.

Further, the study has conducted trials to test the methodology on two different clustering approaches (using K means and EM algorithms available in Weka 3.6.7). While the hard clustering approach is able to count the attribute values exactly, the soft clustering approach suffers from this ability, especially in a scenario, as has been outlined here, in which the absolute count of attribute values can serve to develop a true predictive asset maintenance plan, precisely embedded on the failure trend hidden in the data. Again the fundamental idea behind the predictive approach is to effectively analyze the data and derive meaningful insights from the data in order to use the same constructively and effectively in every facet of decision making in an organization. In this endeavor, it has been highlighted that hard clustering performs and suits well than soft clustering; when the absolute counts of attribute values are required. This literature has adequately demonstrated this aspect and recommends hard clustering approach as an enabler to implement predictive maintenance strategy for the fleet of rail coaches in rail sector.

With regard to further work and future direction, it is prudent to approach the problem possibly with models employing data mining concepts such as classification, association rule mining etc. and verify its applicability. It is also suggested to extend such model to other critical areas such as Locomotives, Freight Stock and Signaling. Finally, it is emphasized that applying DM techniques to data rich organization like Railways is expected to pay good dividends in terms of better decision support towards maximizing capacity utilization, reliability of assets, safety etc.

Acknowledgement

We acknowledge the contribution on data collection made by Mr Ravi, Senior Section Engineer, Southern Railways, Chennai, India.

References:

[1] M. Kantardzic, *Data Mining: Concepts, Models, Methods and Algorithms*, 2nd ed., Wiley-IEEE Press, 2011.

[2] J. Han, Data Mining Concepts and Techniques, 3rd ed., USA: MK Publishers, 2012.

[3] David Poole, Alan Mackworth, *Artificial Intelligence: foundations of computational agents.* Cambridge University Press, 2010

[4] Pedhazur, Elazar J., and Liora Pedhazur Schmelkin. *Measurement, design, and analysis: An integrated approach*. Psychology Press, 2013.

[5] Zimmermann, Hans Jürgen. *Fuzzy set theory-and its applications*. Springer, 2001.

[6] Kearns, Michael, Yishay Mansour, and Andrew Y. Ng. "An information-theoretic analysis of hard and soft assignment methods for clustering." *Learning in graphical models*. Springer Netherlands, pp. 495-520, 1998.

[7] Sholl, H., Ammar, R., Greenshields, I., Pagano, D., "Application of Computing Analysis to Real-Time Railroad Track Inspection," *Automation Congress, 2006. WAC '06. World IEEE* , 2006, pp. 1-6.doi: 10.1109/WAC.2006.376027

[8] Trinh, H., Haas, N., Ying Li, Otto, C., Pankanti, S., "Enhanced rail component detection and consolidation for rail track inspection," *Applications of Computer Vision (WACV), 2012 IEEE Workshop on* , 2012, pp. 289-295 doi: 10.1109/WACV.2012.6163021

[9] Velten, J., Kummert, A., Maiwald, D., "Image processing algorithms for video-based real-time railroad track inspection," *Circuits and Systems, 1999. 42nd Midwest Symposium on* , vol.1, 1999, pp.530-533. doi: 10.1109/MWSCAS.1999.867321

[10] Bladon, Keith, et al. "Predictive condition monitoring of railway rolling stock."*Proc., Conference on Railway Engineering*. 2004.

[11] Faiz, R. B., Singh, S., "Predictive Maintenance Management of Rail Profile in UK Rail," *Computing, Engineering and Information, 2009. ICC '09. International Conference on* , 2009, pp. 370-375 doi: 10.1109/ICC.2009.69

[12] Létourneau, Sylvain, et al. "A domain independent data mining methodology for prognostics." Essential technologies for successful prognostics: *proceedings of the 59th Meeting of the Society for Machinery Failure Prevention Technology, Virginia Beach, Virginia, April 18-21*, 2005.

[13] Sammouri, Wissam, et al. "Mining floating train data sequences for temporal association rules within a predictive maintenance framework." *Advances in Data Mining. Applications and Theoretical Aspects*. Springer Berlin Heidelberg, pp. 112-126, 2013.

[14] Wang Zhansheng; Ding Ling; Yang Liqiang; Zhang Ning, "Cluster Analysis on Urban Rail Transit Ticket Types," *Intelligent Computation Technology and Automation (ICICTA), 2010 International Conference on*, vol.1, 2010, pp. 950-953, doi: 10.1109/ICICTA.2010.829

[15] Gebraeel, N., Lawley, M., Liu, R., Parmeshwaran, V., "Residual life predictions from vibration-based degradation signals: a neural network approach," *Industrial Electronics, IEEE Transactions on* , vol.51, 2004, no.3, pp. 694-700 doi: 10.1109/TIE.2004.824875

[16] Bergquist, B., Söderholm, P. "Data Analysis for Condition-Based Railway Infrastructure Maintenance," *Quality and Reliability Engineering Interantional*, 2014, doi: 10.1002/qre.1634

[17] Barmada, S., Raugi, M., Tucci, M., Romano, F. "Arc detection in pantograph-catenary systems by the use of support vector machines-based classification", *IET Electrical Systems in Transportation*, 2013, doi: 10.1049/iet-est.2013.0003

[18] Asada, T., Roberts, C., Koseki, T. "An algorithm for improved performance of railway condition monitoring equipment: Alternating-current point machine case study", *Transportation Research Part C: Emerging Technologies*, vol.30, May 2013, pp. 81-92, ISSN 0968-090X

[19] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and W. Ian H, *The WEKA Data Mining Software: An Update*, vol. XI, 2009.

[20] Witten, Ian H., and Eibe Frank. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, 2005.

[21] "Indian Railway Vision 2020," 2009. [Online]. Available: http://www.iritm.indianrailways.gov.in/uploads/files/1365142127276-2%20VISION%202020%20Eng%20SUBMITTED%20TO%20PARLIAMENT.pdf