Lecture Notes in Electrical Engineering

Volume 834

The book series *Lecture Notes in Electrical Engineering* (LNEE) publishes the latest developments in Electrical Engineering - quickly, informally and in high quality. While original research reported in proceedings and monographs has traditionally formed the core of LNEE, we also encourage authors to submit books devoted to supporting student education and professional training in the various fields and applications areas of electrical engineering. The series cover classical and emerging topics concerning:

- Communication Engineering, Information Theory and Networks
- Electronics Engineering and Microelectronics
- Signal, Image and Speech Processing
- Wireless and Mobile Communication
- Circuits and Systems
- Energy Systems, Power Electronics and Electrical Machines
- Electro-optical Engineering
- Instrumentation Engineering
- Avionics Engineering
- Control Systems
- Internet-of-Things and Cybersecurity
- Biomedical Devices, MEMS and NEMS

For general information about this book series, comments or suggestions, please contact leontina.dicecco@springer.com.

To submit a proposal or request further information, please contact the Publishing Editor in your country:

China

Jasmine Dou, Editor (jasmine.dou@springer.com)

India, Japan, Rest of Asia

Swati Meherishi, Editorial Director (Swati.Meherishi@springer.com)

Southeast Asia, Australia, New Zealand

Ramesh Nath Premnath, Editor (ramesh.premnath@springernature.com)

USA, Canada:

Michael Luby, Senior Editor (michael.luby@springer.com)

All other Countries:

Leontina Di Cecco, Senior Editor (leontina.dicecco@springer.com)

**** This series is indexed by EI Compendex and Scopus databases. ****

More information about this series at https://link.springer.com/bookseries/7818

Amit Kumar · Jacek M. Zurada ·
Vinit Kumar Gunjan · Raman Balasubramanian
Editors

Computational Intelligence in Machine Learning

Select Proceedings of ICCIML 2021

Springer

Editors
Amit Kumar
BioAxis DNA Research Centre
Hyderabad, India

Vinit Kumar Gunjan
Department of Computer Science
and Engineering
CMR Institute of Technology
Kandlakoya, India

Jacek M. Zurada
Department of Electrical and Computer
Engineering
University of Louisville
Louisville, KY, USA

Raman Balasubramanian
Department of Computer Science
and Engineering
Indian Institute of Technology Roorkee
Roorkee, Uttarakhand, India

ISSN 1876-1100 ISSN 1876-1119 (electronic)
Lecture Notes in Electrical Engineering
ISBN 978-981-16-8483-8 ISBN 978-981-16-8484-5 (eBook)
https://doi.org/10.1007/978-981-16-8484-5

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Preface

Information and data technologies are quickly advancing nowadays, and intelligent computers that can make decisions, analyse data, and perform other jobs that were previously done by humans are now much more efficient and competent than they used to be. The most important element of this next wave of intelligent computing is machine learning and cognitive science. Their work is underpinned by new advancements in computing power and is founded on mathematical, statistical, and data curation. And last but not least, these speedy advances are supported by the proliferation of free software, inexpensive data storage, and the massive social platform needs that have gone global. Every corporation today needs to integrate these strategies into their fabric to succeed. However, those ideas have just recently become accessible to organizations.

The aim of this book is to provide fresh input into the field of computational intelligence and machine learning, which is rapidly expanding. The editors have sought to produce a unified interpretation of the framework of this emerging research field in the novel literature by exploring current ideas in machine learning and computational intelligence and their applications. In addition to its focus on visualization, the book serves as a tool for promoting studies in machine learning and computational intelligence. Researchers and application scientists in machine learning, computational intelligence, and data technologies will find it very useful. Scholars who want to study these subjects should also use this book as a reference.

Because the material of this book is so interdisciplinary, it is light on claims regarding the reader's context. And it pulls ideas from other disciplines, when necessary, only taking up the most relevant ideas in statistics, artificial intelligence, information theory, computational intelligence, and the like. The book offers a detailed perspective on the many study directions by discussing case studies in chapters. The intent is to provide helpful information to readers seeking a comprehensive look at the application of machine learning, cognitive, and similar techniques.

This book gathers selected research papers presented at the International Conference on Computational Intelligence in Machine (ICCIML 2021), held virtually on 1–2 June 2021. The conference was jointly organized by CAIRO UTM, Malaysia; USTC, Bangladesh; and CMR College of Engineering and Technology, Hyderabad, India. This publication is organized in a total of 51 chapters, which have been structured around their contributions to the overall area of the proceedings. The book's chapters describe the central algorithms and theories that are used in technologies and applications related to facial recognition, artificial neural networks, automotive applications, automation devices, business management systems, and modern speech processing systems. Additionally, the book explores contemporary breakthroughs in medical diagnostic systems, sensor networks, and VLSI domain systems. Wherever applicable, discussions of the learning process and software modules are included in deep learning algorithms.

Hyderabad, India — Amit Kumar
Louisville, USA — Jacek M. Zurada
Kandlakoya, India — Vinit Kumar Gunjan
Roorkee, India — Raman Balasubramanian

Contents

About the Editors

Amit Kumar is DNA Forensics Professional, Entrepreneur, Engineer, Bioinformatician, and an IEEE Volunteer. In 2005, he founded the first Private DNA Testing Company Bio Axis DNA Research Centre (P.) Ltd in Hyderabad, India, with a US Collaborator. He has vast experience of training 1000+ crime investigating officers and helped 750+ criminal and non-criminal cases to reach justice by offering analytical services in his laboratory. His group also works extensively on genetic predisposition risk studies of cancers and has been helping many cancer patients from 2012 to fight and win the battle against cancer. He was a member of the IEEE Strategy Development and Environmental Assessment Committee (SDEA) of IEEE MGA. He is a senior member of IEEE and has been a very active IEEE volunteer at Section, Council, Region, Technical Societies of Computational Intelligence and Engineering in Medicine and Biology and IEEE MGA levels in several capacities. He has driven several IEEE conferences, conference leadership programs, entrepreneurship development workshops, innovation, and internship-related events. Currently, he is Managing Director of BioAxis DNA Research Centre (P) Ltd and IEEE MGA Nominations and Appointments committee member.

Jacek M. Zurada is Professor of Electrical and Computer Engineering and Director of the Computational Intelligence Laboratory at the University of Louisville, Kentucky, USA, where he served as Department Chair and Distinguished University Scholar. He received his M.S. and Ph.D. degrees (with distinction) in electrical engineering from the Technical University of Gdansk, Poland. He has published over 420 journal and conference papers in neural networks, deep learning, computational intelligence, data mining, image processing, and VLSI circuits. He has authored or co-authored three books, including the pioneering text Introduction to Artificial Neural Systems, co-edited the volumes Computational Intelligence: Imitating Life, Knowledge-Based Neurocomputing, and co-edited twenty volumes in Springer Lecture Notes on Computer Science. In addition to his pioneering neural networks textbook, his most recognized achievements include an extension of complex-valued neurons to associative memories and perception networks; sensitivity concepts applied to multilayer neural networks; application of networks to clustering, biomedical image classification, and drug dosing; blind sources separation; and rule extraction as a tool for prediction of protein secondary structure.

Vinit Kumar Gunjan is an Associate Professor in the Department of Computer Science and Engineering at CMR Institute of Technology Hyderabad (affiliated to Jawaharlal Nehru Technological University, Hyderabad). An active researcher; published research papers in IEEE, Elsevier and Springer Conferences, authored several books and edited volumes of Springer series, most of which are indexed in SCOPUS database. Awarded with the prestigious Early Career Research Award in the year 2016 by Science Engineering Research Board, Department of Science and Technology Government of India. Senior Member of IEEE; an active Volunteer of IEEE Hyderabad section; volunteered in the capacity of Treasurer, Secretary and Chairman of IEEE Young Professionals Affinity Group and IEEE Computer Society. Was involved as organizer in many technical and non-technical workshops, seminars and conferences of IEEE and Springer. During his tenure, worked with top leaders of IEEE and was awarded with best IEEE Young Professional award in 2017 by IEEE Hyderabad Section.

Raman Balasubramanian is Professor in the Department of Computer Science and Engineering at the Indian Institute of Technology Roorkee. He received his Ph.D. in Mathematics (2001) from the Indian Institute of Technology Madras, India. He obtained his B.Sc. and M.Sc. in Mathematics from the University of Madras in 1994 and 1996. respectively. His research areas include computer vision, fractional transform theory, wavelet analysis, multimedia security, video skimming and summarization, medical imaging, machine learning, multilingual text recognition, EEG based pattern analysis, visualization, and volume graphics.

Machine Learning-Based Project Resource Allocation Fitment Analysis System (ML-PRAFS)

M. Rudra Kumar, Rashmi Pathak, and Vinit Kumar Gunjan

Abstract Many of the existing project automation tools can develop a project schedule with estimates of complete project summary time, budget, tracking possibilities, etc.; still one of the significant gaps in the system is about estimating how a particular resource has the necessary competence, experience, and the track record of completing the tasks more effectively. Taking cue of the gap in the system, using the fundamental approach of game theory model, a contemporary resource fitment prediction using the machine learning model is proposed. The datasets that are trained using the SVM classifiers have provided an accuracy level of around 97% which signifies the efficacy of the model. For the future research scope, if the auto-fitment of scores for the metrics are developed, it might lead to more sustainable solutions with a minimal set of human intervention to the system.

Keywords Project management · Resource allocation · Machine learning model for resource allocation · ML-PRAFS

1 Introduction

1.1 Outlook

Project management as a system is very important in ensuring timely completion of the projects, effective resource utilization, minimizing the cost overruns and toward attaining operational excellence. In the current competitive environment, it is highly important that the organizations ensure quick turnaround time in the completion of

M. Rudra Kumar (✉)
Department of CSE, GPCET, Kurnool,
Andhra Pradesh, India

R. Pathak
Department of CSE, National Institute of Technology, Silchar, India
e-mail: rashmi21_rs@cse.nits.ac.in

V. K. Gunjan
CMR Institute of Technology, Hyderabad, Telangana, India

A. Kumar et al. (eds.), *Computational Intelligence in Machine Learning*, Lecture Notes in Electrical Engineering 834, https://doi.org/10.1007/978-981-16-8484-5_1

projects, ensuring there is no lag in the project environment that might affect the desired outcome [1].

With the emergence of ICT-based project management solutions, there is a paradigm shift in the way organizations can focus on effective project management practices and leverage on the resource utilization to improve the project outcomes. However, one of the common issues that are imperative in the project management conditions could be attributed to high-cost overruns, a significant delay in the project trends, etc. But in addition to this, the other significant challenge in the case of project management is the effective utilization of resources [2].

There are many automated project management tools and solutions that are available online, which will help the organizations and the project management teams, in developing a robust project schedule. Right from auto-scheduling the tasks, to reflect on the baseline, milestones, efforts, and budget there are multiple factors that are handled by the automated tools. But one of the key areas wherein the task-related conditions are assessed for a resource utilization is not integral to many of the contemporary systems.

There is a need for a more comprehensive system that can support the project teams in understanding the composition of the resource utilization and how such task performance-based decision making can be empowered to allocate the work to right kind of resources executing the project.

In the other dimension, there are many contemporary solutions for project management that are being reviewed, wherein the AI-based solutions, predictive modeling-based WBS analysis, and other such significant models. While few such models are being more effective and is providing the necessary insights, still there is scope for improvements in terms of resource allocation for the project work break down structures [3].

1.2 Purpose of the Research

Project management as a process has three critical factors that have a significant impact on project planning.

- Task
- Time duration
- Resources required.

While these three are imperative for the process of developing a WBS, the other secondary factors that impact the WBS schedule are

- Predecessors
- Resource utilization ratio
- Cost of the resource
- Milestones
- Summary tasks.

While many of the existing project automation tools can develop a project schedule with estimates of complete project summary time, budget, tracking possibilities. However, one of the significant gaps in the system is about estimating how a particular resource has the necessary competence, experience, and the track record of completing the tasks more effectively [4].

In an illustrative scenario, it can be stated that in the case of a project "P," if for job A is one of the WBS tasks, and the team has four different resources who can be allocated to the job, the existing systems has no significance to the resources. It accepts any resource allocated to the respective task. But there are hardly any systems that can highlight as to how if resource x or y or z or m can be more effective for the specific kind of tasks and the related conditions.

If such resource advocation model can be integrated into the project management system, wherein the project team-related capability and success factor are analyzed by the system to predict the accuracy with which the schedule could be completed by the resources, it shall help in developing more comprehensive conditions [5].

Though some of the AI models are currently used in the project management toward developing a comprehensive system, still the factor of predicting the accuracy of timelines for completion of a project is relatively niche in the system planning.

Considering the abilities of the machine learning algorithms that can predict the accuracy of a system, if the similar kind of implementation plans is used for the WBS-related resource allocation classification, it can help in creating a more comprehensive and effective system, which can lead to sustainability in the project planning practices [1].

1.3 Outline of the Proposed Model

If a comprehensive system can be developed wherein the machine learning model is able to gauge the complexity score of the project task and accordingly take certain features as the competency score of the resource, impact factor score of the task and other certain parameters into account, toward estimating whether the specific resource could be able to complete the task within the timelines, it will help in choosing or allocation of right kind of resource for a project scenario.

If the machine learning model is developed for the above-mentioned condition, which can be used by the system for getting trained to identify whether an individual member is able to complete the assigned task, such model can lead to more accurate prediction of the schedule. However, the model that is proposed in this manuscript is independent of the levels of estimating whether the resource shall be able to complete the proposed task.

In furtherance, the system must be integrated into any of the existing project management tools which can support in achieving the desired input for more effective project management planning.

1.4 Organizing of the Paper

In Sect. 2 of this paper, the related work in terms of the distinct set of project management elements wherein the AI solutions are essential is discussed. Followed by in Sect. 3, the framework of the proposed model is discussed. Also, the test results of the model are discussed in Sect. 4. Followed by, in Sect. 5, the conclusion of the model is presented.

2 AI-Based PMP

Analysis, predictive modeling is paving its way into project management practices. While the role of automation in the process might increase, still the probabilities and accuracy of the project schedules can see significant improvement. Some of the key aspects that the usage of machine learning models can support in the aforementioned scenario of project management practices are explored in subsequent sections.

2.1 Effective Utilization of Resources

One of the common phenomena in the case of project success rates across the industrial verticals is about choosing the right kind of resources for a project scenario. Though many of the team members could be competent toward handling a specific task, one or few of the members among the group might be more competent for the specific contextual circumstances [6]. If the best of the resource for the specific job is chosen, the accuracy with which the job could be completed intime goes up.

2.1.1 The Analogy Drew to the Above Scenario

In the case of a sports team, all the members are equipped to handle all the tasks in the game. But one of few in the team can handle that specific process in the game more effectively than the others. The probability of estimating who among the members could handle a specific kind of circumstance or process effectively than the peer group is integral to game theory.

If the similar practice of game theory logic is extended into the machine learning analysis, it can help the project management team in choosing right kind of resources for the tasks and toward improvising the optimal performance conditions in a project environment [7].

2.2 More Effective Planning

Various elements of project planning are integral and interdependent over each of the other factors. For instance, in the project task, the time duration is one of the important factors for completion of the project. In the instance of a project requiring one week of one resource to be completed, the time frame for completion could vary based on the number of resources scheduled for the task at a time and also the budget of the task varies based on who is allocated for the project [8].

Taking such pragmatic factors into account, if the most appropriate resource for a task is selected, it is imperative that the probability of the tasks being completed within the scheduled time is more pragmatic.

2.3 Effective Cost Management

Duration of the task, human resource costs, infra requirement, and other factors have a direct impact on the cost overrun for a project. There are multiple models like activity-based costing, bottom-up costing, milestone-based costing models, etc., which have a direct impact on the project conditions. One of the critical parameters toward handling the project more effectively is about completion of the project within the estimated timelines and budget [1].

If there is significant overrun in the cost of the project, the cost–benefit viability of the model might get impacted and it might lead to more implications for the stakeholders associated with the process.

2.4 Project Dynamics

The project dynamics could be attributed to the agile environment of how the projects schedules and tasks are varying over the go. With more organizations using the path of the agile structure of project development, the onus in terms of using more competent team members toward a specific task has a high significance. But the challenge is about ensuring that the team composition is maintained in a balanced approach, where the resource cost and the resource utilization are handled below the threshold levels [2]. The ability, capacity, and the threshold values in the case of human competence toward measuring the process completion scope are essential. If such estimates are accurate, it will help in creating a comprehensive system [9].

Based on the inputs discussed above, it is evident that the model if implemented effectively, it can lead to a more integrated model which can develop sustainable features and the ones that can create significant impact. However, gaps in the system are to be addressed, in order to develop more effective project management systems that are empowered by the intelligence to assimilate the resource capabilities and deliver a comprehensive solution [4].

3 Framework for the Proposed Model

3.1 Fundamental Elements of the Model

The proposed solution is about developing an independent assessment system which can evaluate the resource mapping for a task based on the complexity ranking of the system and the competency scoring of the resource-based on various metrics that are integral to the model.

However, the fundamental principle behind the development of the system is the "game theory" model, which in parlance to the project environment stands more appropriate to the context discussed in this paper.

3.1.1 Game Theory

Game theory is a theoretical framework for conceiving social situations among competing players. In some respects, game theory is the science of strategy or at least the optimal decision making of independent and competing actors in a strategic setting [7].

In the parlance of the model to the project environment, every member of the group might be keen on working over a specific task or could have the necessary competence to execute a broad range of work. However, the optimization in terms of choosing the most appropriate to a task, among the available options, can lead to increased scope and probability of adherence to the schedule [8].

3.2 Metrics for the Models

Considering such a paradigm, some of the key aspects that are chosen in the proposed model of machine learning potential for task analysis system are.

The metrics chosen above are the key factors in the case of using the system for understanding whether the proposed model can have a significant solution. The metrics discussed in Table 1 are considered as the features for the proposed machine learning model.

Table 1 Features description

Metric	Dimension	Purpose
Task complexity score (TC)	Work break schedule task (WBS)	To understand the significance of the work complexity and the intricacies involved in the completion of the task
Task impact score (TI)	WBS	The impact score is about the rate of the effect that the specific task has on the schedule (if there are too many successors that rely on the completion of a specific task, in such condition, the score is high)
Project impact (PI)	WBS	A holistic impact of the task over the other aspects like leading to cost overruns, or risk register triggers, etc.
Resource capability score (RCS)	Resource end	The experience and efficiency of the resource toward executing the project are the score
Resource option scoring (ROS)	Resource end	The number of resources available as an option for executing the project is the weighted average score
Resource trust score (RTS)	Resource end	The trust by the team member in terms of belief about a member be able to complete the task as expected (probability score)

3.3 Data Classifiers and Training Model

In the machine learning models, there are numerous set of classifiers that are integral to the training models. It is imperative that the classifiers that are used for training the model have a significant impact on the outcome that can be envisaged in the system.

For the proposed solution, the emphasis is on multi-class classifiers that are used for training the model. The purpose of using the multi-class classifier is about understanding the probability of a resource competency defined for accuracy to a specific task.

3.3.1 Illustrated scenario

For a Task-A to be executed, the perceived impact of the task is High, and the resource to be allocated for the task too should be High or as the best alternative "Good." Any resource with lower than the "Good" ranking can be considered as In-effective allocation for the project task.

SVM is based on the fundamental that decision planes reflect the decision boundaries. The decision plane can be defined as a plane that classifies or segregate set of objects with different class membership (in the proposed project, the classes are Highly Suitable, Suitable, Moderately Suitable, In-effective) [10].

If the classification is presumed as green (Highly Suitable, Suitable) or red (Moderately Suitable, In-effective), as depicted in Fig. 1, the line defines the boundary between the green and red categories. However, in the case of machine training, the categories shall have four distinct sets.

The process followed in the above linear classifier model could be attributed as hyperplane classifiers, and the inherent reason behind the selection of the support vector machine (SVM) model is about its efficacy in handling such data classification across the training sets [11].

SVM primarily classifies the tasks based on the constructing hyperplanes over multidimensional space that categorizes the distinct class labels. As the SVM can support in handling both regression and classification tasks and can also handle the categorical and multiple continuous variables, the model is perceived as more effective for the proposed system.

For categorical variables, the dummy variables are created with case values as 0 or 1, and the categorical dependent variable consisting of four levels as A, B, C, D is depicted for three set of variables.

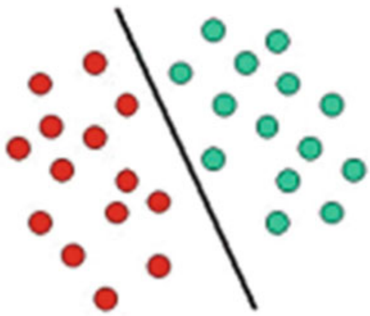
$$A : \{1\,0\,0\,0\},\ B : \{0\,1\,0\,0\},\ C : \{0\,0\,1\,0\},\ D : \{0\,0\,0\,1\}$$

In terms of developing an optimal hyperplane, SVM adapts the iterative training algorithm, which is used in the process of minimizing the error function. For the proposed training model, the C-SVM classifier is used.

The training is idealized for minimization of the error function.

$$\frac{1}{2}w^T w + C\sum_{i=1}^{N}\xi_i$$

Fig. 1 Decision plane classification



subject to the constraints:

$$y_i\left(w^T\phi(x_i)+b\right)\geq 1-\xi_i \quad \text{and} \quad \xi_i\geq 0, i=1,\ldots,N$$

In the equation above, C is the capacity constant,
Vector of coefficients is denoted by w.
b reflects the constant.

ξ_i Signifies constraints for handling non-separable data (inputs).
ξ_i The index i tags the N training cases.

It is imperative that $y\in\pm 1$ denotes the class labels and x_i indicates the independent variables. The kernel ϕ is embraced for the transformation of the data from the input state which is independent of the feature space.

Selection of C holds critical importance in the training process toward minimizing the error conditions.

3.4 Algorithm for the Proposed Model

The algorithm used for the proposed model is about.

Let T be the task for which the resource R has to be reviewed for rating.
{
T is *task selected.*
R is *resource for rating.*
{
T_p is *task criticality score*

$$T_p = \sum(T_c + T_i + P_i)$$

whereas, individually all the sub-score (T_c, T_i, P_i) are significant for the system, and the value range is

$$Tc = \geq 1 \leq 10$$
$$Ti = \geq 1 \leq 10$$
$$Pi = \geq 1 \leq 10$$

If the $T_p \geq 27$, the task is rated as “Highly Complex” denoted as HC.
If the $T_p \geq 20 \leq 26$, the task is rated as “Moderate Complex” denoted as MC.
If the $T_p \geq 15 \leq 19$, the task is rated as “Low Complex” denoted as LC.
If the $T_p \geq 1 \leq 18$, the task is rated as “Normal Task” denoted as NT.
}
{

The resource score category is considered as

RCS (resource capability score)
ROS (resource option score)
RTS (resource trust score)

For all the RCS, ROS, and RTS, the scoring pattern is at the stream range of 1–10

$$\text{RCS} = \geq 1 \leq 10$$
$$\text{ROS} = \geq 1 \leq 10$$
$$\text{RTS} = \geq 1 \leq 10$$

$$R = \sum (\text{RCS} + \text{ROS} + \text{RTS})$$

If the $R \geq 27$, *the task is rated as* W.
If the $R \geq 20 \leq 26$, *the task is rated as* X.
If the $R \geq 15 \leq 19$, *the task is rated as* Y.
If the $R \geq 1 \leq 18$, *the task is rated as* Z.
$R = either\ W\ or\ X\ or\ Y\ or\ Z$.
}
When the input "I" is tested for T-value allocation, the process followed is about
{
Eligibility score
If the TIp has the label of HC.... then only Rwmust be used for analysis.
If the TIp has the label of $MC \ldots then$ R_x *or* R_w *or* R_y *can be considered for allocation accuracy prediction.*
If the TIp has a score of LC, only R_x *or* R_y *must be considered.*
If the TIp *has a score of* NT, R_x *or* R_y *can be considered.*
}
{
If the T_p score validation is effective, then the scoring output for task allocation accuracy is estimated as

$$T = 2 * ((T_p * R)/T_p + R)$$

}
If the $T \geq 27$, *the task is rated as highly suitable.*
If the $T \geq 20 \leq 26$, *the task is rated as suitable.*
If the $T \geq 15 \leq 19$, *the task is rated as moderately suitable.*
If the $T \geq 1 \leq 18$, *the task is rated as ineffective.*
}

The inputs are evaluated in the scoring pattern, it can be comprehended that the T score indicates the potential accuracy of the system.

Table 2 Datasets

Particulars	Dataset records	Remarks
Total data used for analysis	1700	
Three sets used for training	1200	Every set constituting 400 records for training purpose
Positive records in the training set	950	The records that can be classified for highly suitable or moderate suitable or suitable conditions
Records of ineffective, in the training set	250	Records that constitute the ineffective training records
Total records used for testing the system	500	Test data
Positive records in the test data	376	Records that are to be classified for positive category
Negative records in the test data	124	Records to be traced by the system as negative or ineffective for task

3.5 Datasets

To understand the efficacy of the proposed system, a synthesized dataset is generated with random score value for all the six-feature metrics, three of task complexity, and three of the resource abilities. Table 2 indicates the cumulative set of records that are integral to the testing process. Also, in terms of classification, the test is evaluated based on the metrics of efficacy which is defined for a model as precision, recall, $F1$ score, specificity, and the sensitivity for the project conditions.

For the purpose of accuracy estimations from the datasets, the ones that are categorized to the high scores constituting as Highly Suitable, Suitable, or Moderate Suitable as positive records, and the ones that are to be classified as to be Ineffective are chosen as negative records for the inference purpose.

Table 2 indicates the dataset used for the analysis.

4 Test Result Analysis

The tests are carried out for the data over the system application executed for the algorithm discussed in the earlier section. The test data results from the process generated are classified for the evaluation purpose (Table 3).

Based on the inputs that are garnered for the respective record conditions, the estimation of the accuracy levels for the system is evaluated. From the contextual conditions that are evaluated for the process, it can be stated that the scoring pattern of the model are (Table 4).

Table 3 Result classification

Metrics	Detection from test records
True positive	364
False positive	12
False negative	14
True negative	110

Table 4 Performance metrics

Metrics	Performance
Precision	96.81
Recall	96.30
Accuracy	94.80
$F1$ score	96.55
Specificity	90.16

When the metrics are reviewed, the precision rate of 96.81% is imperative in the case of the data tested. Though the data garnered for the evaluation parameters are synthesized data, still the rate of accuracy could be more pragmatic, considering the correlation estimated for the records chosen for the analysis (Fig. 2).

In terms of the recall rate, it depicts the cost of a false negative high. In the instances of false negative being high, it reflects the risks in embracing the model. However, the recall rate indicates that for the tested model, the rate of false negatives is well within the controllable levels of around 3%, which indicates that the model can be used in integration to many of the existing project management models.

The $F1$ score of the proposed model reflects the rate of 96% which indicates that the model has the overall accuracy, taking into account the precision and recall values in totality. With a good $F1$ score attained, the model is imperative for a low rate of false positives and low set of false negatives. Also, it can be inferred that the model

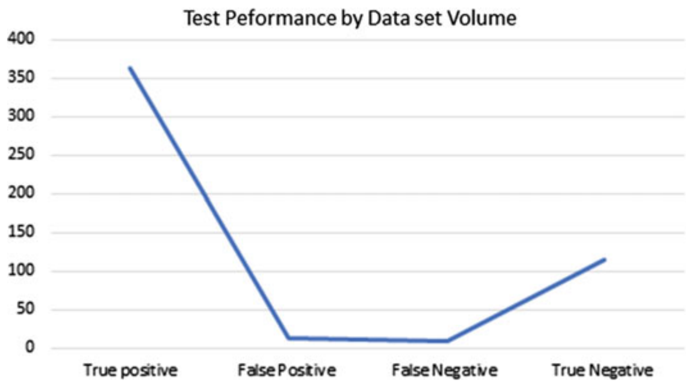


Fig. 2 Test performance by records

Table 5 Records classification for classes

Classification	Actual segregation	Positive tracing	Accuracy %
Highly suitable	176	169	96.0
Suitable	127	124	97.6
Moderately suitable	73	71	97.3
Ineffective	124	115	92.7

does not have much of real threats and do not be intrigued by certain levels of false alarms [12].

The other parameter in which the study is evaluated is to focus on the metrics of how accurately the system has identified the respective classes in the test records. The comprehensive Table 5 indicates the quantum of records that are considered for different classes in the testing sets and the accuracy of prediction for the respective class category (Fig. 3).

The model has performed ineffective ways of identifying most of the classes close to accurate parameters. This set of inputs can be inferred as the effectiveness of the SVM classifier that is used for the training of dataset patterns [12]. However, the model has certain level issues in the case of detecting the ineffective set, which can be focused upon by working on reinforced learning model of the training sets.

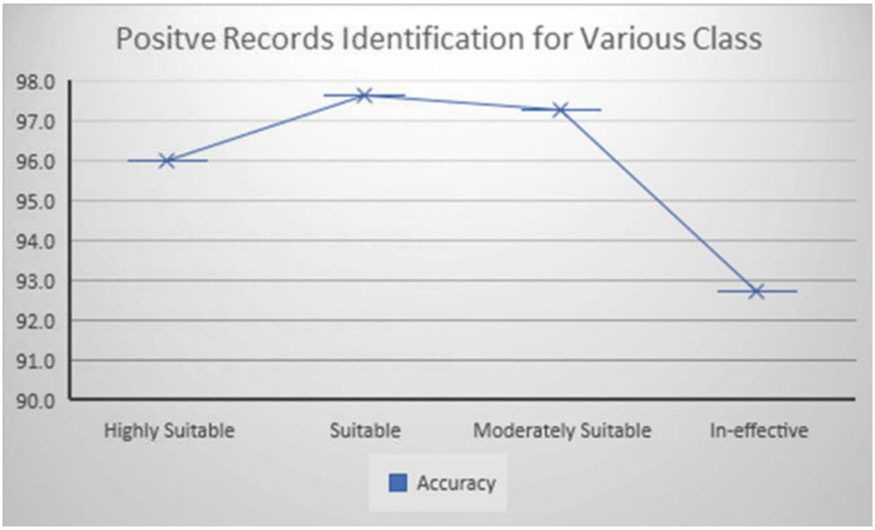


Fig. 3 Class performance

5 Conclusion

While many of the existing project automation tools can develop a project schedule with estimates of complete project summary time, budget, tracking possibilities. However, one of the significant gaps in the system is about choosing the resource allocation with a high probability of fitment for the task. Though some of the AI models are currently used in the project management toward developing a comprehensive system, still the factor of predicting the accuracy of timelines for completion of a project is relatively niche in the system planning. Comprehending on the game theory model wherein the scores of individual members capability play a vital role can lead to improving the accuracy of planning. The proposed machine learning model when trained with the datasets using SVM classifier provides considerable performance as discussed in the test results. For the future solutions, the model of focusing on autoscoring for the complexity of the task score and the resources could be used as the backward integrated solution, which might lead to more optimization of the services

References

1. J. Thomas, T. Mengel, Preparing project managers to deal with complexity—advanced project management education. Int. J. Proj. Manage. (2008)
2. C.P. Killen, K. Jugdev, N. Drouin, Y. Petit, Advancing project and portfolio management research: applying strategic management theories. Int. J. Proj. Manage. (2012)
3. J.K. Pinto, Project management, governance, and the normalization of deviance. Int. J. Proj. Manage. (2014)
4. S. Pemsel, A. Wiewiora, R. Müller, M. Aubry, K. Brown, A conceptualization of knowledge governance in project-based organizations. Int. J. Proj. Manage. (2014)
5. L. Project, *Rights-Based Approaches* (2009)
6. A. Perini, A. Susi, P. Avesani, A machine learning approach to software requirements prioritization. IEEE Trans. Softw. Eng. (2013)
7. R. Breen, Game theory, in *The Oxford Handbook of Analytical Sociology* (2017)
8. E. Kalai, Game theory: analysis of conflict. *Games Econ. Behav.* (1991)
9. K. Bredin, People capability of project-based organisations: a conceptual framework. Int. J. Proj. Manage. (2008)
10. K. Bennett, E. Bredensteiner, Duality and geometry in SVM classifiers. ICML (2000)
11. C.C. Chang, C.J. Lin, Training ν-support vector classifiers: theory and algorithms. Neural Comput. (2001)
12. D.N. Carvajal, P.C. Rowe, Sensitivity, specificity, predictive values, and likelihood ratios. Pediatr. Rev. **31**(12), 511–513 (2010)

Electric Theft Detection Using Unsupervised Machine Learning-Based Matrix Profile and *K*-Means Clustering Technique

Saddam Hussain, Mohd. Wazir Mustafa, Steve Ernest James, and Shadi Khan Baloch

Abstract Electric theft is the major issue faced by utility companies in different countries as it causes significant revenue losses and affects the power grid reliability. This paper presents a novel electric theft detection framework based on an unsupervised machine learning technique employing matrix profile and *K*-means clustering algorithm. The proposed framework is based on three stages to identify the fraudster consumers in a conventional electric consumption meter dataset acquired from Pakistan's power distribution company. Initially, the missing and inconsistent observations are filtered out from the acquired dataset. After that, the matrix profile from each consumer's consumption profile is computed to identify the irregular and sudden changes present in them. Later, the *K*-means clustering algorithm is used on the datasets divided based on their computed matrix profile values in order to label each consumer into "Healthy" and Theft." The developed framework is compared against the latest state of art machine learning algorithms and statistical-based outlier detection methods. The proposed technique achieved an accuracy of 93% and a detection rate of 91%, which is greater than all compared models.

Keywords Electric power distribution · Electric theft · Matrix profile · *K*-means clustering algorithm

S. Hussain (✉) · Mohd. W. Mustafa
School of Electrical Engineering, University Technology Malaysia, 81310 Skudai, Johor Bahru, Malaysia
e-mail: hussainsaddam@graduate.utm.my

S. E. James
Department of Actuarial Science and Risk Management, Institute of Business Management, Korangi Creek, Karachi, Sindh 75190, Pakistan

S. K. Baloch
Department of Mechatronics Engineering, Mehran University of Engineering and Technology, Jamshoro, Sindh 76062, Pakistan

A. Kumar et al. (eds.), *Computational Intelligence in Machine Learning*, Lecture Notes in Electrical Engineering 834, https://doi.org/10.1007/978-981-16-8484-5_2

1 Introduction

Restructuring of electrical distribution companies to meet standards of the modern grid has already begun in most of the advanced countries [1, 2], among numerous modernization in the old grid, especially smart meters has altered various outdated applications [3–6]. However, Pakistan still has an old conventional system of metering [7]. The biggest problem faced by power distribution companies is power theft [8], which worsens the economic state of the electric utilities [9, 10], restrains new funds for capacity expansion [11], which ultimately leads to electricity scarcity [12]. It happens due to bad governance [13], poor accountability, and an elevated level of corruption [14]; comprehensive studies on various electricity crisis in Pakistan and proposed solutions for those can be found here [15–17].

Pakistan electric distribution sector is based on public and private distribution companies [18], majority of electric power is distributed through the public sector [17], few statistics are shown in Fig. 1, and details can be seen in annual reports published by NEPRA Pakistan [19]. The existing process practiced by power distribution companies (PDC) for computing and registering electric consumption data through the electromechanical meter is illustrated in Fig. 2.

As from Fig. 2, it can be observed that the entire process is manual and subjected to numerous issues to falsify meter reading. For example, the bribe is given to meter reader employees by the consumer for registering wrong meter reading; the metering process is reversed or made a hole to slow down its counter by consumers, bogus meter installation, bypassing meter registering process, etc. The dedicated study and practical demonstration of various power theft approach in Pakistan can be found here [20]. Similarly, based on an on-site inspection conducted by the largest PDC,

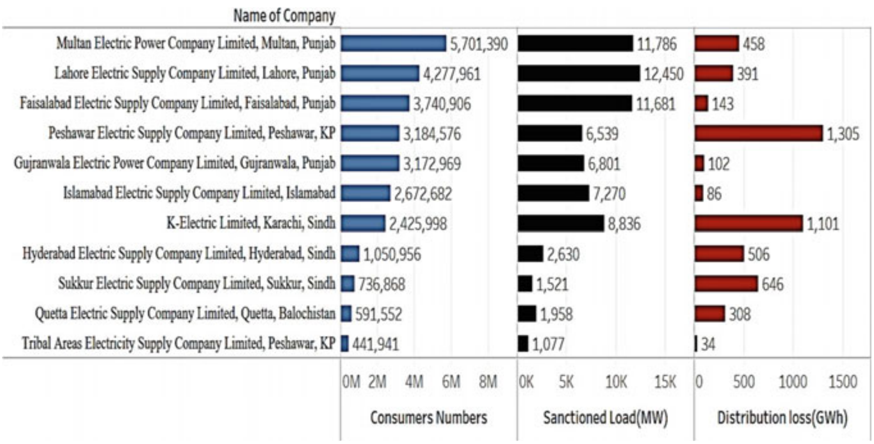


Fig. 1 Pakistan electricity public distribution companies, number of consumers, sanctioned load, and distribution loss

Fig. 2 Current process of data collection, processing, and registering in power distribution companies of Pakistan

Multan electric power supply company (MEPCO), in 2016, most occurring theft methods were computed and illustrated in Fig. 3.

Electricity theft is accomplished by manipulating consumption data in order to reduce electricity charges [21]. It is claimed that power theft conducted through electromechanical meters cannot be eliminated entirely. However, it can be controlled by increasing theft detection hit rate accuracy. In [22], author has asserted to increase accuracy from the labor-intensive procedure of theft detection (of 3% accuracy) to utilizing a combination of data mining approaches and SVM algorithm (achieving 60% theft detection hit rate accuracy) by choosing TNB Malaysia power utility company consumer's historical power consumption of 25 months as a dataset; the proposed method claimed to benefit also in reducing on-site inspection time by efficiently going only toward suspected labeled consumers rather than the manual approach of checking each and every consumer possible. In [23], author has taken Brazil Light SA company historical monthly consumption dataset in order to distinguish irregular consumption patterns from the pool of normal consumption patterns

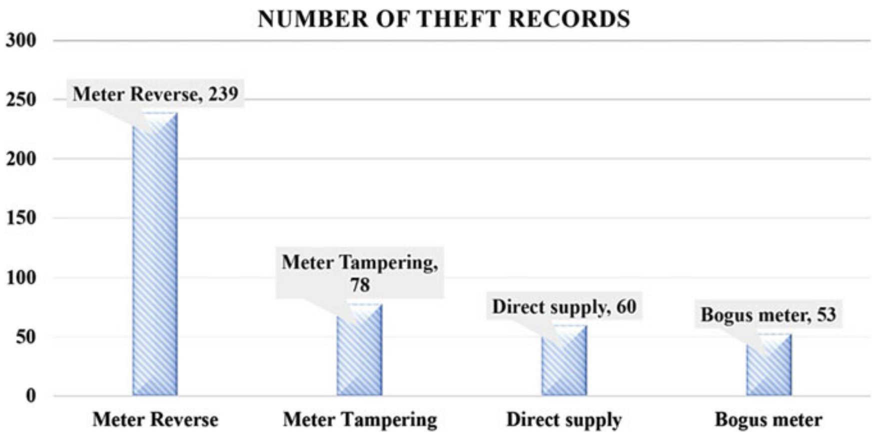


Fig. 3 Most occurring theft methods in MEPCO, PDC of Pakistan