DISTRIBUTED EVENT DETECTION IN WIRELESS SENSOR NETWORKS USING MACHINE LEARNING

Enrollment no - 122218

Name of Student - Aditi Kansal

Name of Supervisor - Dr. Yashwant Singh

May-2014

Submitted in Partial fulfillment of the Degree of

Master of Technology

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING,

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY,

WAKNAGHAT, DIST. SOLAN, (H.P.), INDIA

Contents

Certificate from Supervisor **iv**

Declaration **v**

Acknowledgement **vi**

Abstract **vii**

List of Figures **viii**

List of Tables **ix**

1 Introduction **1**

1.1 Event Detection 2

1.2 Distributed Event Detection 3

1.3 Machine Learning 3

1.4 Event Detection Using Machine Learning 4

1.5 Genesis of the problem 4

1.6 Problem Statement 5

1.7 Motivation 6

1.8 Objectives 6

1.9 Methodology 7

1.10 Thesis Organization 8

2 Related Work **9**

2.1 Wireless Sensor Networks 9

2.2 Event Detection 11

2.2.1 Energy-Aware Communications 12

2.2.2 Optimal Node Deployment and Localization 13

2.2.3 Resource Allocation and Task Scheduling 14

2.3 Attacks in WSN 16
2.4 Machine Learning for WSN 17
2.5 Classification of Machine Learning Techniques 18
2.5.1 Supervised machine learning 18
2.5.2 Unsupervised machine learning 18
2.5.3 Semi supervised learning 18
2.5.4 Active Learning 19
2.6 Algorithm Selection Methodologies 19
2.7 Event Detection using Machine Learning 22
2.7.1 Regression 23
2.7.2 Support Vector Machines 24
2.7.3 Decision Tree Induction 25
2.7.4 k- Nearest neighbor classifiers 26
2.7.5 Reinforcement Learning (RL) 27
2.7.6 Genetic Algorithms 30
2.7.7 Rough Set Theory 31
2.7.8 Fuzzy Logic 31
2.8 Distributed Event Detection in Wireless Sensor Networks for Forest Fires 31
2.8.1 Decision Tree 33
2.8.2 Support Vector Machine 33
2.8.3 Rough Set Theory 34
2.8.4 Fuzzy Logic 34
2.9 State of Art Method for Fire Detection 35
2.9.1 Distributed Event Detection in Wireless Sensor Network for forest fires 35
2.9.2 Forest Fire Detection in Wireless Sensor Network Using Fuzzy Logic 36
2.9.3 SWATS: Wireless Sensor Networks for Stream flood and Water flood Pipeline Monitoring 37
2.9.4 Wireless Sensors and Neural Networks for Intruders Detection and Classification 38
2.9.5 Application of Wireless Sensor Networks in Forest Fire Detection under Uncertainty (Rough Set approach) 38
2.9.6 Distributed Event Detection in Wireless Sensor Networks for Disaster Management 39
2.9.7 Circle-based Approximation to Forest Fires with Distributed Wireless Sensor Networks using clustering 40
2.9.8 Data Mining Approach to Predict Forest Fires using Meteorological Data. 41
2.10 Conclusion 44

3 **Regression and Support Vector Machines in Forest Fires.** **45**
3.1 Introduction 45
3.2 Proposed approach 46
3.2.1 Algorithm 48
3.2.2 Proposed approach and attribute selection 49
3.2.3 Division and better prediction 50
3.2.4 Regression technique for forest fire 51
3.2.5 Predictor variation in the model 52
3.2.6 Simulation Setup 53
3.3 Conclusion 53

4 **Results and Discussions** **54**
4.1 Data Collection 54
4.2 Conclusion 57

References **59**

List of Publication **65**

CERTIFICATE

This is to certify that the work titled **"DISTRIBUTED EVENT DETECTION IN WIRELESS SENSOR NETWORKS USING MACHINE LEARNING "** submitted by **" Aditi Kansal"** in partial fulfillment for the award of degree of Master of Technology in Computer Science of Jaypee University of Information Technology, Waknaghat has been carried out under my supervision. This work has not been submitted partially or wholly to any other University or Institute for the award of this or any other degree or diploma.

Signature of Supervisor:- ....

Name of Supervisor:- Dr Yashwant Singh

Designation::- Assistant Professor
Jaypee University of Information Technology

Date::- ....

DECLARATION

I declare that this thesis entitled "DISTRIBUTED EVENT DETECTION IN WIRELESS SENSOR NETWORKS USING MACHINE LEARNING",submitted by me for the award of degree of Master of Technology in Computer Science and Engineering, Jaypee University of Information Technology, Waknaghat is original and it has not submitted previously to this or any other University for any degree or diploma.

Signature of Student:- . . . .

Name of Student:- Aditi Kansal

Date::- . . . .

ACKNOWLEDGEMENT

Nothing can be accomplished without a 'Guru'.I would like to express my earnest gratitude to my Project Guide **Dr.Yashwant Singh** sir,Assistant Professor, Department of Computer Science and Engineering, Jaypee University of Information technology,Waknaghat, for his constant help and guidance. Sir helped me in formulating the idea and collection of the related material for the project proposal. His continuous monitoring to support me and my research work encouraged me a lot for doing my thesis in a smooth manner.

Last but not the least,I would like to extent my appreciation towards my parents and my friends for always being there through all phases of work,for their encouragement and patience and giving me their valuable support without which I would never be where I am today.

Signature of Student:- ....

Name of Student:- Aditi Kansal

Date::- ....

ABSTRACT

In Wireless Sensor Networks (WSN), when an usual event is noticed in the networks, an event is detected through the sensor devices placed at distributed locations. This event detection information is passed to the base station and intelligent decision is taken. We proposed an ensemble distributed machine learning approach for detecting events. This approach works in 3 steps: collection of data, defining levels of fires and division of dataset. Regression and SVM are the approaches used in proposed architecture for detection of events and prediction of forest fires. This method uses regression for calculating the detection accuracy and errors and levels of fires are defined by SVM. The predictors considered in the dataset are significant and thus help in better prediction of forest fires.

A comparison between proposed approached and other machine learning techniques has also been done which helps to prove that the proposed approach has better detection accuracy and less errors. The R- squared calculated of the R-Squared is very high as compared to other machine learning techniques. Also, the analysis time taken by the proposed approach is less as compared to other techniques. Thus, the proposed algorithm and the approach used are better in terms of detection of forest fires. The proposed algorithm helps in fast and accurate detection of forest fires.

List of Figures

1.1 Wireless Sensor Networks 2

2.1 Working of sensor network with clusters 11
2.2 Different Attacks in WSN 16
2.3 Different Attacks on protocol layers 16
2.4 Neural Networks 21
2.5 Decision Tree 25
2.6 RL –Q learning algorithm 27
2.7 Reinforcement Learning 28

3.1 Proposed Graph Based Algorithm 48
3.2 Illustration of the proposed approach 49
3.3 Illustration of the proposed approach 50
3.4 Divided dataset for the month of July 50
3.5 Divided dataset for the month of August 51
3.6 Results in minitab 52
3.7 Selecting the predictors in minitab 52
3.8 Regression Analysis 52

4.1 Graph of performance analysis 56
4.2 R-squared values comparison 57

List of Tables

2.1 Comparison of various techniques used for fire detection in wireless sensor network 43

3.1 Meaning of symbols used in the algorithm 48

4.1 Comparison of machine learning techniques 55

Chapter 1

Introduction

Wireless Sensor Networks consists of a large number of sensor nodes distributed over an area. These sensor nodes are small and of low cost. The capabilities of these are sensing and processing. Each sensor node is connected to many devices like radio transceiver, a small micro controller, a power source and multi functional sensors which can sense environment. They also help to exchange sensory data with other nodes. Then data is collected and decisions are made according to the data collected and the requirements. These help to track object and also used in monitoring objects[1]. Sensor nodes are integrated and used for low power consumption units, integrated memory, and radio transceiver and energy source. Transceiver leads to less communication and more power consumption. Information is collected from remote geographical, industrial, civil infrastructure and power plants [2, 3]. Wireless sensor networks have been used in various applications such as target monitoring, target tracking, home animation, sales tracking etc. Target tracking, vehicle tracking, forest fire detection, earthquakes etc are done in WSN but location information is not collected. So with the help of localization techniques we estimate and compute the position of sensor nodes. Nodes are confined to low power embedded system, computer memory, radio transceiver, sensors, geographical system and power source. Cost, size of hardware techniques are reduced and

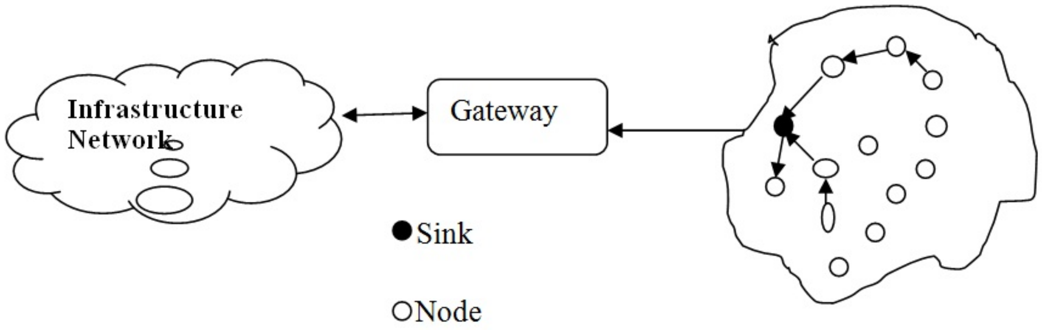


Figure 1.1: Wireless Sensor Networks

computational power has increased. Decision taking and pattern matching play an important role in event detection in WSN.

1.1 Event Detection

Event is classified as a pattern of infrequent or abnormal occurrences. But due to harsh environment there are faults in pattern matching and pattern monitoring domain. These have limited battery life. There are various real-life applications.

- Environment monitoring
- Event detection
- Habitat monitoring
- Health and medical monitoring
- Target tracking

When the data is exchanged with the other sensor nodes in their local area, then data processing is done and decisions are made about what is observed. This feature is called event detection [4, 5]. The functionality of event detection has attracted much attention in a variety of applications like safety of industries, security, environment hazards etc. Thus, in current research areas, very high performance and energy efficient sensor nodes are deployed for distributed event detection.

1.2 Distributed Event Detection

The distributed event detection is a type of event detection in which the network is heterogeneous and work is done in distributed form. The focus is on discovery of event patterns in this. The distributed event detection is divided into four phases. The first phase is called HELLO phase, in which all the nodes locate their neighboring nodes. The next phase is the CALIBRATION phase, in which sensors are evaluated by establishing the values of their neutral position and bounds are added to the background values, if needed. The next two phases do the same work of evaluation of raw data and follows classical pattern recognition algorithm [6]. These phases are named as TRAINING and RECOGNITION phases. Disaster Management provides responses whenever and wherever changes occur and helps to save lives.

1.3 Machine Learning

Machine learning is a branch of artificial intelligence that concerns the construction of the system from which learning can be done from data [7]. This is very attractive to manually construct them. Machine learning is used in Web search, spam filters, recommender systems, stock trading and many other applications. A recent report from the McKinsey Global Institute asserts that machine learning (a.k.a. data mining or predictive analytics) will be the driver of the next big wave of revolution [8]. Basically machine learning is a technique which is subfield of artificial intelligence that concerns a question that how we can improve the performance of computer programs with experience.

1.4 Event Detection Using Machine Learning

The family of machine learning algorithms is very efficient in WSN communication. In this section we will use various machine learning algorithms and techniques for efficient communication in WSN. We will use machine learning in WSN for event detection. Historical data sets are used for comparing data (sensor data) and for pattern matching. A common task in machine learning application domains involves monitoring routinely collected data from many interesting events. A number of different events like object detection, target detection and many more events can b detected. There will be no formal specification or expert knowledge needed [9]. Thus, event detection techniques for WSNs need to be light-weighted to meet limited computational capability of the sensor nodes. Also if the techniques are distributed then its computation becomes easy because a big process is divided into smaller modules and facilitates parallel processing. It also needs to detect events quickly generating timely alarms. We will be using a light-weight and accurate event detection approach for a network in decentralized event detection [10]. We will be using Decision Trees for distributed event detection and a voting method for reaching a particular conclusion among different decisions collected by the sensors. Although decision trees are simple but these are highly accurate and their simplicity fulfills WSNs requirements. On the basis of detection accuracy and time complexity the performance of the proposed approach is measured.

1.5 Genesis of the problem

Paraphrasing Mark Weiser, the late Chief Scientist of Xerox PARC and father of ubiquitous computing, “Applications are of course the whole print of sensor networking.” It is clear that wireless sensor networks hold great promise as an enabling technology for a variety of applications. Habitat monitoring is one such application that is representative of an entire class of data collection applications which have received considerable attention in literature. Fundamentally, data collection is signal reconstruction problem in which the objective is to centrally reconstruct observations of distributed phenomena. Performance metrics for such applications include accuracy and precision of signal reconstruction. Physical phenomena such as light, temperature, humidity and rain change periodically. The requirements for this type application called event detection, imply a sensing and signal processing architecture quite different from that employed for data collection problem. Fundamentally, fire detection is an event detection problem in which detection of fires should be fast and accurate. Performance metrics for such applications include probabilities of detection, false alarm, classification and misclassification, detection accuracy and system lifetime.

1.6 Problem Statement

Traditionally, distributed view of computing is not considered in event detection methods. But, with the advancement of technology, a good number of approaches have been proposed to deal with distributed event detection in WSN. Basic idea to deal with event detection in distributed environment is to collect the event data from every sensor node, then these sensor nodes sends the data to the centralized base station for further computation in order to take decision whether the event has occurred or not based on some pattern matching or prior knowledge. Different approaches are there to gather data in an effective manner. But the main problem in sensors is its detection accuracy. So, the proposed approach employs machine learning for distributed event detection. Regression is simple to use and detection accuracy problem is solved by using this. The performance of the proposed approach is determined in terms of complexity and accuracy.

1.7 Motivation

As Mark Weiser, the late Chief Scientist of Xerox PARC and father of Ubiquitous computing said that "Applications are of course the whole point of sensor networking." WSN is used in many applications. Environment monitoring is one of the applications that are representative of an entire class of data collection applications which have received considerable attention in the literature. Fundamentally, data collection is a signal reconstruction problem in which the objective is to centrally reconstruct observations of distributed phenomena with high spatial and temporal fidelity. The performance metrics for these types of applications should be accurate and precise. There should be correlation between the observed signal and the physical phenomena as well as the sensor network's life. Physical phenomena include light, temperature, pressure, humidity etc. With the help of compression and aggression, the performance of the system can be improved. The COTS platform has also helped researchers to gather a lot of data for the applications in year-wise order.

In contrast with collection of data, noise should also be observed by the applications of sensor network like intrusion detection and military surveillance. The requirements for this style of application known as event detection signify that the sensing and signaling are different from the problem of data collection. Performance metrics for such applications include probabilities of detection, false alarm, classification, detection latency, tracking accuracy and system lifetime.

1.8 Objectives

The aim of this thesis is increase the accuracy of the detection of fires. Also, to introduce a self-healing event detection framework for distributed wireless sensor networks. We present working of event detection that includes approaches like

Support Vector Machine and Regression for fast and accurate detection of forest fires. We focus on failures caused by deterioration of a component's quality, rather than on systematic attacks of malicious participants. Moreover, we attempt to minimize the Root Mean Squared Error for accurate detection of fires. Initially, we present the overall working of the system. In addition, we look into more detail of how the Regression works and how the events are detected accurately. Root Mean Square Error is also calculated and the detection accuracy. The algorithm called Maximum Sensor Accuracy is used and compared to prove its accuracy and better performance.

1.9 Methodology

In this research work, I have applied Maximum Sensor Accuracy algorithm for maximizing the event detection accuracy of forest fires. This algorithm works very efficiently as it takes efficiency issues into consideration. It is based on division of dataset of forest fires. A forest fire dataset is taken into consideration and the dataset is divided on the basis of months for fast detection.

The forest fire dataset is divided into levels according to the level of fires. This forest fire dataset has 8 parameters out of which all are significant. The response on the basis on which accuracy is detected is taken in form of level i.e. level is taken as response variable. Afterwards, Regression technique is applied on the divided dataset. A regression equation is calculated. After that taking all the parameters as significant parameters and level as response variable R-squared and Root Mean Squared Error is calculated. Also, a comparative analysis of some machine learning techniques has been done.

1.10 Thesis Organization

This thesis is organized into four chapters. The first chapter presents the problem statement, motivation followed by different objectives to be achieved. The remainder is organized as follows: In Chapter 2, related work on event detection algorithms using machine learning techniques and how those techniques can be used for fire detection is reviewed and discussed. In Chapter 3, a proposed algorithm for accurate detection of forest fires is discussed. In Chapter 4, results are discussed and a comparative analysis of proposed technique with other machine learning techniques has been done.

Chapter 2

Related Work

In this chapter we give an idea of related work that is important for the investigations carried out in this research. The sections are ordered accordingly. In distributed event detection in wireless sensor network a heterogeneous area is considered where multifunctional sensors are deployed and event has to be detected. Event can be in form of storm, fire, earthquake etc. An alarm is generated when the sensors detect the values within the level defined in a particular month. Hence, with the help of various machine learning techniques defined as follows various events can be detected according to suitability and requirements. Firstly, we will discuss the work related to WSNs.

2.1 Wireless Sensor Networks

Wireless sensor networks is an active area of research. In WSN, there are various networking issues and application issues. In the network associated issue, different machine learning algorithms are applied in WSNs to improve network performance. In application associated issue, machine learning methods which are being used for information processing in WSNs have been summarized. Basically, the distributed wireless sensor architecture is two- tier architecture. This is a hierarchical cluster topology. Since multiple nodes can report a cluster head nearby its location very easily, we can use this topology for deployment of nodes. The local region in which the region nearby the node is cluster and the gathering node is the cluster head. By using this topology the network deployment becomes attractive in heterogeneous settings and the cluster nodes become more powerful on the basis on communication and computation. This approach fragments a large network into separate zone fragments within which the data processing is carried out[10]. There are two types of sensor nodes:

(A). Forwarding nodes or simple nodes.

(B). Cluster head.

The forwarding nodes sense the activity or the event and forward the data to the base station. The cluster head, also known as simple data gathering point node collects all the data from all the nodes of the sensors. In this figure we have four clusters. Each cluster selects a cluster head which is responsible for collection of data from all the sensor nodes and sends that to the base station (BS). It is like all the other sensor nodes. A base station controlled dynamic clustering protocol is a clustering based protocol which uses high energy base station and do different tasks [11].

So we will use various algorithms based on machine learning techniques and with the help of those techniques we will detect events in a distributed system. According to the various classifications of the machine learning algorithms we will implement various events.

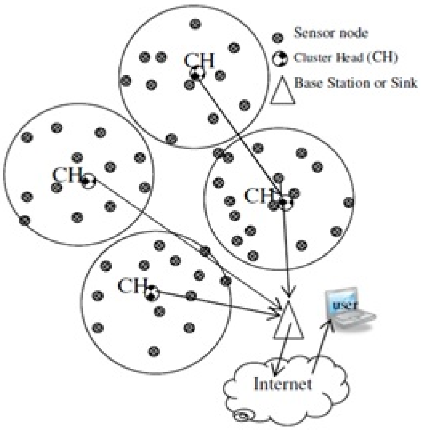


Figure 2.1: Working of sensor network with clusters

2.2 Event Detection

The event detection problem can be tackled from different perspectives like defining some threshold values and when the sensor readings are lower or higher or equal to the required or predefined threshold then alarm is generated. Since events cannot be detected by simple pattern matching, distributed pattern matching fulfills almost all the requirements and makes machine learning technique known as decision tree to be used in an efficient way. Therefore, involvement of more sensors makes failures easy to detect, leads to prevention of failures and makes the system robust. Thus, it leads to reduction of communication overhead also. The four phases defined above are HELLO phase, CALIBRATION phase, TRAINING phase and RECOGNITION phase. The TRAINING phase is used to learn new reference patterns for future runs of the RECOGNITION phase. During the supervised training, the sensors of the node are added temporary values for a specified class of events. From these temporary values the features of each class are extracted and sent to the neighboring nodes. The energy consumption should be as low as possible and the nodes must operate at high volumetric densities. The components must be cheap and the individual nodes must be dispensable [12].

2.2.1 Energy-Aware Communications

Improvement in the performance of the network and optimization is the major objective of WSN. Its main motive is energy conservation. So machine learning techniques are used to fulfill this motive. Most of the WSN applications depend on fast, efficient and reliable communication of data. But the communication links of WSN were inherently unreliable because these are unbounded in nature. Since communication protocols employ situation-aware adaptation to identify healthy and energy-efficient routes, we will use supervised leaning approach for routing optimization and optimize communications. Machine learning techniques help to discover the correlations between input features and output. The input features include node level and network level metrics such as buffer length, occupancy, node residual energy. The output features include link quality and optimal route. There are four steps in which machine learning is carried out:

1) Feature selection and output labeling.

2) Sample collection

3) Offline training

4) Online classification[13].

The feature selection chooses the most appropriate vector that best represents the problem in hand. It includes fast fading, slow fading, traffic pattern, signal strength etc. The output labeling is used to classify outputs using domain knowledge. Sample collection is the process which is used to collect data for the purpose of training. Background server software is also used in the system as a data collector and processor. Therefore, in terms of architecture this is known as centralized learning. However, in many real-life applications, deploying a background server in WSNs is not efficient or even impossible. Thus, a decentralized/distributed learning architecture is sometimes preferred. Decision Tree and Rule- based classifiers are the classifiers to be evaluated. Since this approaches a centralised solution, therefore, are not considered as best solutions for WSN. Learning overhead is considered only in online classification phase because there is background server software used in offline training.

2.2.2 Optimal Node Deployment and Localization

Sensor node deployment and localization are two interrelated issues in WSNs. With different node deployment strategies, the localization algorithms can be completely different. For example, sensor nodes can be deployed manually can be localized if we will use a walking GPS method. However, the GPS method can be too costly and time-consuming if sensor nodes will be deployed randomly in large scale. In this case, parameters such as signal strengths, message delivering speeds, relative orientation are used to estimate the sensor location using machine learning techniques.

Location information is an important parameter in both networking and application domains of WSN. Accurate location estimation is a basic prerequisite for energy-aware routing and sensor event localization and reporting. There are generally two approaches to tackle the localization problem in WSN, namely hardware-based and probabilistic estimation based approaches.

Some of the authors have proposed fuzzy logic to do this by using strict probabilistic rules and set up heuristic fuzzy rules. The algorithm is based on a grid approach in which location of a node is represented by its confidence level that it stands at a certain point in this grid. The confidence level computed is based on the fuzzy logic system and the input variables of the fuzzy system which are sensor measurements (signal strength, time difference of arrival and etc). Some authors have adopted an evolutionary approach called a micro genetic algorithm. The extension of this algorithm has been used to enhance the accuracy of the present localization methods. This is a post-optimizer for any other localization algorithms. It makes use of two genetic operators; the mutation and crossover operators, whose aim is to decrease the objective values obtained to mutate out the location of the estimated current node, or for the localisation by crossover of points for any pair of existing chromosomes. Simulation results indicate that this post optimizer improves the accuracy of various localization algorithms from 11.0 percent to 18.6 percent on average.

2.2.3 Resource Allocation and Task Scheduling

The major research challenges in the field of WSNs in terms of system wise interaction are resource allocation and task scheduling. Unlike energy-aware communication or optimal deployment and localization that did the work of optimizing aim of optimization problems formulated under these two scenarios were from a more global point of view. Problem was how to manage a group of sensor nodes and how to schedule them to achieve some system objectives, such as a swapping the network lifetime and the information gain[**?**].

Some of the authors have explored three machine learning algorithms for task scheduling in radar sensor networks and have compared their simulation results. The algorithms used were fuzzy Lyapunov synthesis, genetic algorithms and neural networks. The simulation results showed that genetic algorithms were over fitted as compared to other algorithms[14]. This study was initially carried out to address the radar scheduling problem, however, the result could be applied to WSN due to similar system architecture setting (WSNs and radar sensor network). Many authors have proposed an adaptive distributed resource allocation scheme which specifies relatively simple local action to be performed by individual nodes in a WSN for management of system modes. Each node adapts its operation over time in response to the status and feedback of its neighbouring nodes. The adaptive operations were defined locally and the optimal global behaviour results were calculated from these local interactions. The scheme was being studied in two separate application scenarios, namely an acoustic WSN for field surveillance and camera network for traffic monitoring. Simulation results showed that it provided good tradeoffs between performance objectives like accuracy of the target being tracked, coverage area of the network and network lifetime[15].

Also, some have introduced a novel fuzzy approach for cluster head election in WSN. Strictly speaking, cluster head assigning is a step in hierarchical routing in WSN. However, it was being viewed as a resource allocation scheme in WSN as assigning the role of cluster head to nodes is functionally equivalent to allocating resources to the node. The fuzzy system in this research takes in three parameters – node energy, node concentration and centrality with respect to the entire cluster. The node which becomes the cluster head is taken as the output. The scheme allowed the energy consumption across the network and also helped to improve the overall network lifetime[16].

Wireless Sensor Networks have been considered as key-enabler for distributed applications with respect to home security, health care, monitoring environment because these are reliable. In all of the cases like target tracking, monitoring vehicle etc, a sensor network has been used to detect events which are based on the raw data being sampled by the various sensors nodes deployed in the network.

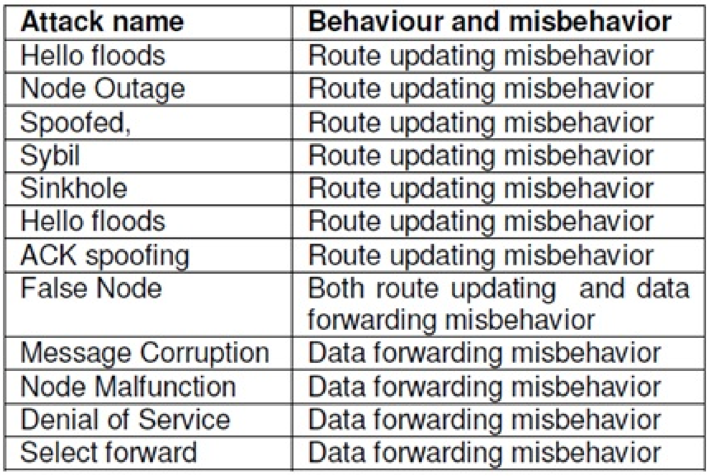
There are two key points to solve the problem of event detection in WSNs:

1) By transmitting the raw data sampled by the sensors as it is to the base station for centralized evaluation. But the drawback of this is that there will be fast energy consumption due to continuous channelling of data. Thus, it will shorten the lifetime of the WSN.

2) By evaluating the raw data on each sensor node, reporting the information to the base station, and examining all the reports. In this case, the overall detection accuracy suffers from the fact that the initial classification is based on the data of merely a single sensor node.

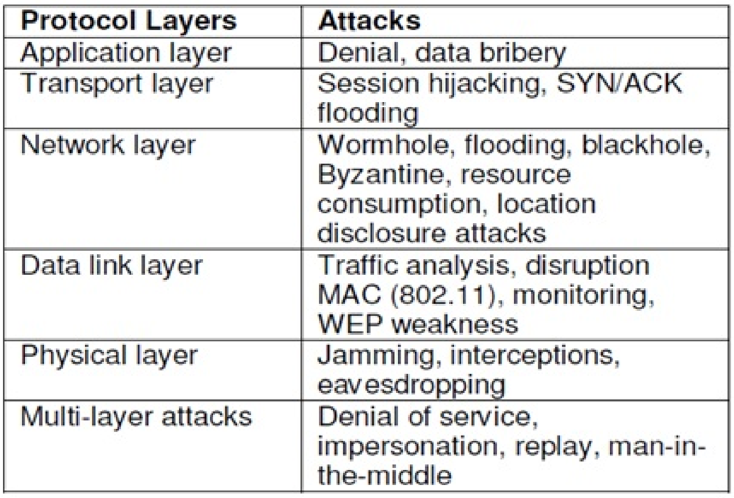
2.3 Attacks in WSN

There are two types of attacks in WSN leading to interruption in network. These are active attack and passive attack. In passive attack the attacker is out from the communication network and plans to attack by eaves dropping the link between the client and the server. But in case of active attack the attacker sends false data to both the client and the server and can modify the whole information. There are different attacks with respect to their behavior [11].



Attack name	Behaviour and misbehavior
Hello floods	Route updating misbehavior
Node Outage	Route updating misbehavior
Spoofed,	Route updating misbehavior
Sybil	Route updating misbehavior
Sinkhole	Route updating misbehavior
Hello floods	Route updating misbehavior
ACK spoofing	Route updating misbehavior
False Node	Both route updating and data forwarding misbehavior
Message Corruption	Data forwarding misbehavior
Node Malfunction	Data forwarding misbehavior
Denial of Service	Data forwarding misbehavior
Select forward	Data forwarding misbehavior

Figure 2.2: Different Attacks in WSN [11]



Protocol Layers	Attacks
Application layer	Denial, data bribery
Transport layer	Session hijacking, SYN/ACK flooding
Network layer	Wormhole, flooding, blackhole, Byzantine, resource consumption, location disclosure attacks
Data link layer	Traffic analysis, disruption MAC (802.11), monitoring, WEP weakness
Physical layer	Jamming, interceptions, eavesdropping
Multi-layer attacks	Denial of service, impersonation, replay, man-in-the-middle

Figure 2.3: Different Attacks on protocol layers [11]

There are various differences between WSN and MANET (Mobile Ad-hoc networks). The WSN contains more number of nodes as compared to MANET.

1)The capacity of nodes in WSN is more.

2)Because of the deployment circumstances, there are more chances of failure of sensors in WSN.

3)The need for mobility causes dynamic change of WSN topology.

4)There are large resource constraints in WSN in terms of power, storage, communication and processing capability [17].

2.4 Machine Learning for WSN

Machine learning is a branch of artificial intelligence that concerns the construction of the system from which learning can be done from data. This is very attractive to manually construct them. Machine learning is used in Web search, spam filters, recommender systems, stock trading and many other applications. The family of machine learning algorithms is very efficient in WSN communication. In this section we will use various machine learning algorithms and techniques for efficient communication in WSN. We will use machine learning in WSN for event detection. Machine learning techniques are Support Vector Machines(SVM), Artificial Neural Networks(ANN), Decision Trees(DT), Regression(Reg), Clustering[18]. In this thesis, various machine learning techniques for fire detection have been discussed.

2.5 Classification of Machine Learning Techniques

2.5.1 Supervised machine learning

It is very useful. Learning algorithms are now used in many domains and different performance metrics are appropriate for each domain. For example Precision/Recall measures are used in information retrieval; medicine prefers ROC area; Lift is appropriate for some marketing tasks, etc. We can evaluate the performance of various algorithms used. We evaluate the performance of SVMs, neural nets, logistic regression, naive Bayes, memory-based learning, random forests, decision trees, bagged trees, boosted trees, and boosted stumps on eleven binary classification problems using a variety of performance metrics: accuracy, F-score, Lift, ROC Area, average precision, precision/recall break-even point, squared error, and cross-entropy [19].

2.5.2 Unsupervised machine learning

It has a problem that hidden structures need to be found in unlabeled data. Since there are unlabeled examples given to the learner, there is no error or reward signal to evaluate a potential solution. The unsupervised learning is closely related to density estimation problem. Many methods in unsupervised learning are based on data mining methods used to pre-process the data.

2.5.3 Semi supervised learning

This is basically combination of supervised and unsupervised learning as it uses both labeled and unlabeled data for training, typically a small amount of labeled data with large amount of unlabelled data. Many machine-learning researchers have found that unlabeled data, when used in conjunction with a small amount of labeled data, can produce considerable improvement in learning accuracy. The acquisition of labeled data for a learning problem often requires a skilled human agent (e.g. to transcribe an audio segment) or a physical experiment (e.g. determining the 3D structure of a protein or determining whether there is oil at a particular location). The cost associated with the labeling process thus may render a fully labeled training set infeasible, whereas acquisition of unlabeled data is relatively inexpensive. In such situations, semi-supervised learning can be of great practical value. Semi-supervised learning is also of theoretical interest in machine learning and as a model for human learning [10].

2.5.4 Active Learning

It is an umbrella term that refers to several models of instruction that focus the responsibility of the learner. This includes class discussion, think-pair share, learning cell, short written exercise, collaborative learning group, student debate, class game etc. Domain expert is also involved during learning. Goal is to optimize the model quality by actively acquiring knowledge from human users. Given a constraint on how many examples they can ask to label [20].

2.6 Algorithm Selection Methodologies

In machine learning, **support vector machines** are supervised learning models with associated learning algorithms that analyze data and recognize patterns, used for classification and regression analysis [9]. The basic SVM takes a set of input data and predicts, for each given input, which of two possible classes forms the output, making it a non-probabilistic binary linear classifier. Given a set of training examples, each marked as belonging to one of two categories, an SVM training algorithm builds a model that assigns new examples into one category or the other. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as wide as possible. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

The **Artificial neural networks** are kind of learning based algorithms. These basically work on principle of neuron. The first model of neuron contained two inputs and one output. Both the inputs should be active for correct output. The weights for both the inputs were equal and output was binary. The mathematical model functions of these are $F : X$. The data flows from input nodes to the output nodes through the whole network. Ability to learn is important property of artificial neural networks and to adjust input/output weights to reflect the exactly learned function. Thus, for training an artificial neural network, a set of data for training are needed in which inputs are already mapped to get the possible output. For example classification of different numbers, the pictures can be considered as inputs and the numbers can be considered as outputs. In contrast to decision trees, inputs cannot be described as attribute pairs. The neural network is also known as supervised offline learning algorithm. This consists of a training set, which has already been classified. Offline is how much necessary is the training set which will be used for classification.

There are also unsupervised and online learning neural networks. For example a network which is used for learning the data model for sensor readings. Neural networks are very suitable for such problems where features or attribute-values pairs are not available. However, they have large memory and processing requirements like decision tree learning. There are also some techniques which are applicable in WSN for static classification problems such as data models or link quality estimation. They can be efficiently implemented even on standard sensor nodes because of their low requirements.

We will present a system for distributed event detection in WSNs that allows several sensor nodes to integrate in order to recognise application-specific events. Our

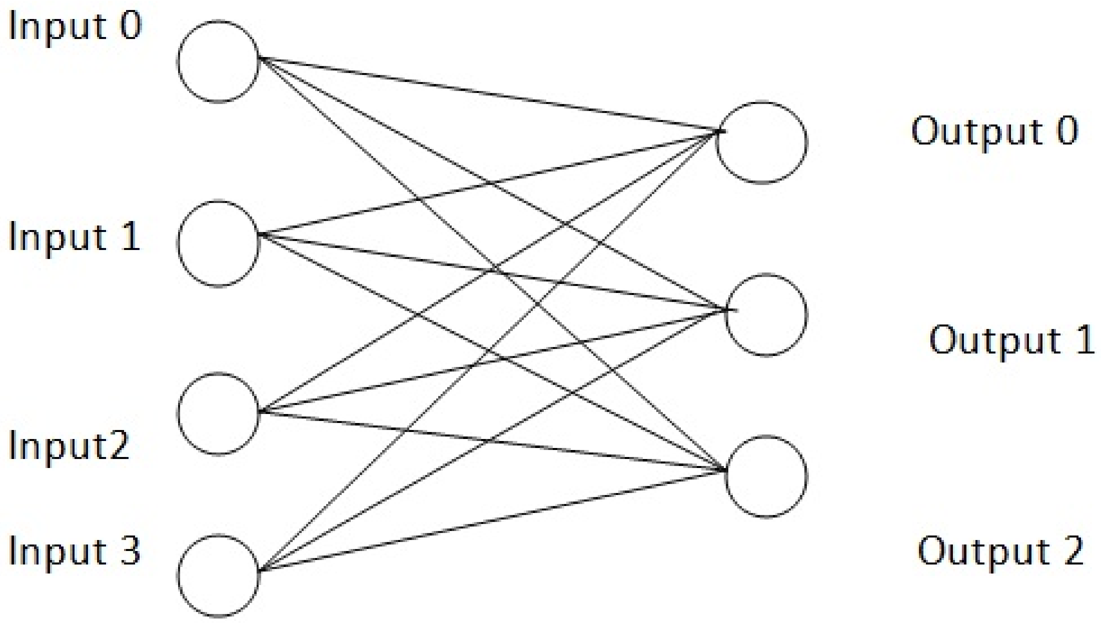


Figure 2.4: Neural Networks

system is capable to correctly identify various classes of events, which can be freely trained on the system being deployed. There is no need of formal event specification or expert knowledge about the attributes of the event. Communication cost will be reduced to a minimum as raw data is assessed directly on the sensor nodes. Only small feature vectors, i.e., condensed but pictorial sampled data, are locally exchanged between nodes; only the information about which event was detected is being reported back to the base station. Our general approach is to adapt algorithms from the field of pattern recognition to WSNs and train the deployed sensor network to recognize new events. The algorithms presented consist of the feature vector which uses an automatically generated, application-specific selection of attributes across multiple sensor nodes. The extraction of features from the data samples has been performed locally on each sensor. Hence, our approach combines the energy savings of local data processing with a distributed evaluation to yield high detection accuracy. Further, our system is not confined to any particular application scenario because the pattern recognition algorithms are not specific to the type of sensor used or the characteristics of the deployment area.

We will evaluate our system on the example of a wireless alarm system which consists of some sensor nodes that are attached to the fence surrounding a real-world construction site. The function of the WSN is to detect incidents which are applicable to security. This is done by identification of four previously trained patterns in the side fluctuation of the fence elements.

We will propose a system for event detection in WSNs which is based on distributed pattern recognition algorithms that can work on application-specific events by the use of supervised machine learning techniques. We will evaluate our system on the basis of quantity which is being applied to a real-world use case in a major WSN deployment. We will conclude our results by considering previous lab experiments and compare our work qualitatively to similar approaches.

2.7 Event Detection using Machine Learning

The basic aim of event detection can be fulfilled by defining some threshold values within a particular range. The alarm will be generated when the sensor reads the value which is below or above the defined threshold range. The events are can be highly developed and cannot be detected by defining the simple threshold values; therefore, we will use pattern matching and machine learning techniques [21]. On the basis of the network scale, requirements of the application and attributes, local base station in the sensor nodes is considered as pattern matching. Among all these, distributed pattern matching completes the basic needs of WSN. Since we are using more than one sensor nodes and if one of the nodes fails then it does not affect the whole event detection process and the system becomes more robust. Also, energy consumption and overhead needs to be reduced.

Naive Bayes feed forward neural networks, support vector machines (SVM) and Regression were proposed to detect an event locally on single individual sensor nodes. Distributed in network studies engulf the merging of nodes and exchange of data. But the range from techniques based on distributed fuzzy engine, map the base pattern matching, feed forward neural networks and the naive Bayes classifiers.

2.7.1 Regression

Linear regression is widely used approach amongst the wealth of machine learning techniques .This approach is used for prediction and forecasting and uses massive datasets having high dimensionality. The results obtained are beneficial for both the customer and the organization for better decision making. Various models have been proposed for selecting the best regression model i.e. is the model having high accuracy and less dimensionality. These models select the attributes based only on their significance without considering the linear relationship between the attributes. In simple linear regression, there is a dataset in which there are variable which are independent. Only those variables can be changed. These variables are called independent variables/regression/ control variables. It is the change of these variables which bring change in the value of some dependent/response variable. Simple Linear regression is a model which has one explanatory variable(X) which has a linear relationship with the response variable (Y). A Multiple Linear regression model is:-

$$Y_i = \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i$$

This equation can be written in matrix form as follows :- where Y is the vector of observation

X is the Regression variable

β is the vector of parameter

ε_i is the vector of error

β_1 is the Slope

Least square method is used to estimate the coefficients in the model and $\widehat{\beta}$ which is the unbiased estimator of β is calculated as:-

$$\widehat{\beta} = (X'X)-1)X'Y$$

The datasets that occur in real time have large dimensionality and instances which have to be reduced achieving the highest possible accuracy and lowest redundancy of variables. When the numbers of response variables is more than one and we estimate the relationship by a linear fit it is termed as multiple linear regressions. It has n-k degrees of freedom unlike simple linear regression which has n-2 degrees of freedom where n is the total number of attributes and k is the number of constraints. Least R- square method is used for the estimation of the coefficient of the model. Linear relationship among some of the predictors involved in the model (multi-co linearity) violates the linear regression assumptions. So, because some decisions like the stock exchange returns, fire detection or climate forecasting may be of utmost importance to the user and any prediction too far from the true one may cause fatal damage. To make a dataset for regression is also a cumbersome task which is prepared based on certain experiments and collection from various different places. We need to analyze the variables values and keep only the variables that are independent, thus keeping the dimensionality of the dataset less.

2.7.2 Support Vector Machines

In machine learning, support vector machines are supervised learning models with associated learning algorithms that analyze data and match patterns . The basic SVM takes a set of input data and estimates, which of two possible classes forms the output for each input value, making it a non-probabilistic binary linear classifier. Consider example of a training value set, which are marked as belonging to one of

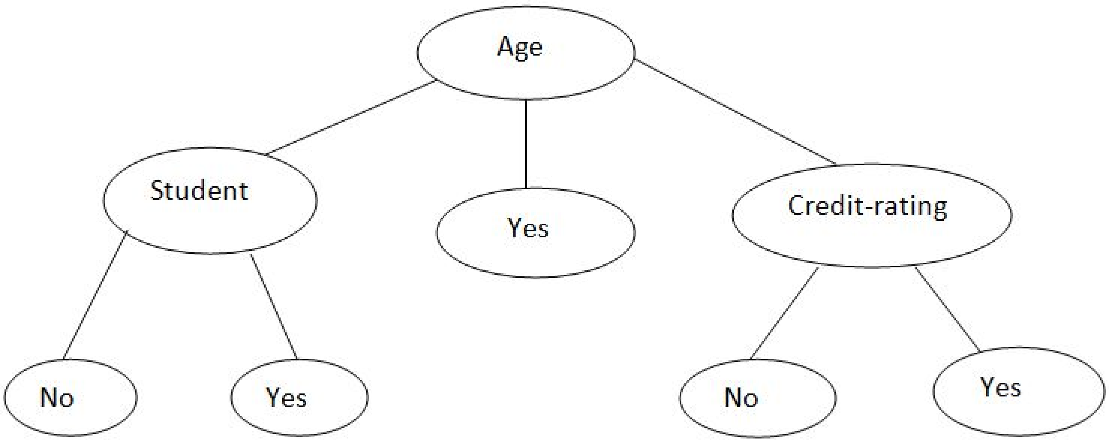


Figure 2.5: Decision Tree [22]

two categories, an SVM training algorithm which builds a model that assigns new examples into one category or the other. An SVM model can be described as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as large as it can be. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

2.7.3 Decision Tree Induction

An attribute selection measure is a heuristic for selecting the splitting criteria that best separates the given data partitioning criteria, D of the class labeled training tuples into individual classes.

The Decision Tree classifiers are usually successful in many diverse areas such as radar signal detection, remote sensing, medical diagnosis, expert systems, and speech recognition. It has capability to break down complex decision making process into simpler one, thus providing proper solutions which are easier to interpret. They can be easily represented in form of graphs or in form of rules [22] This example basically tells whether a person is student or not. This tree shows the relation of a person as a student, his/her age and the credit rating as fair in the left and excellent on the right side. Most popular algorithms are chi- square automatic induction (CHAID), cart. Decision trees are upside side down. They are built from root at the top to leaves at the bottom [4]. A decision tree is a flow chart like tree structure where each internal node represents a test on an attribute. Each branch represents an outcome of the test and leaf nodes represent class or class distributions. We use goal attributes to form a decision tree. **Chi square (C4.5) and ID-3** are the most widely used decision tree algorithms. The C4.5 computes the information gain and finds which feature splits clusters best. C4.5 takes feature of highest information gain and puts that information in the root node of the tree and then recursively computes the information gain for the subclasses until all the samples from training set are classified. In wireless sensor networks various problem can arise that how to classify links as good or bad based on the data such as signal strength or delivery rate.

2.7.4 k- Nearest neighbor classifiers

They are basically how we learn by analogy. The training samples can be defined by n-dimensional numeric attributes. Each sample shows a point in an n-dimensional space. This is the way of storing all of the training samples in an n-dimensional pattern space. When given an unknown sample, a k-nearest neighbor classifier searches the pattern space for the k training samples that are closest to the unknown samples. These k training samples are the k “nearest neighbors” of those unknown samples. “Closeness” is defined in terms of Euclidean distance, where the Euclidean distance between two points, $X = (x_1, x_2...x_n)andY = (y_1, y_2, ..., y_n)is$

$$d(X,Y) = \sqrt{\sum (x_i–y_i i)^2}$$

The unknown sample is assigned the most common class among its k nearest neighbors. When k=1, the unknown sample is assigned the class of the training sample that is closest to it in pattern space. Nearest neighbor classifiers are instance based or lazy learners in that they store all of the training samples and do not build a classifier until a new (unlabeled) sample needs to be classified. This contrasts with eager learning methods, such as decision tree induction and back propagation, which construct a generalization model before receiving new samples to classify. Lazy learner can incur inexpensive computational costs when the number of potential neighbors with which to compare a given unlabeled sample is great. Therefore, they require efficient indexing techniques. As expected, lazy learning methods are faster at training than eager methods, but slower at classification since all computation is delayed to that time. Unlike decision tree induction and back propagation, nearest neighbor classifiers assign equal weight to each attribute

2.7.5 Reinforcement Learning (RL)

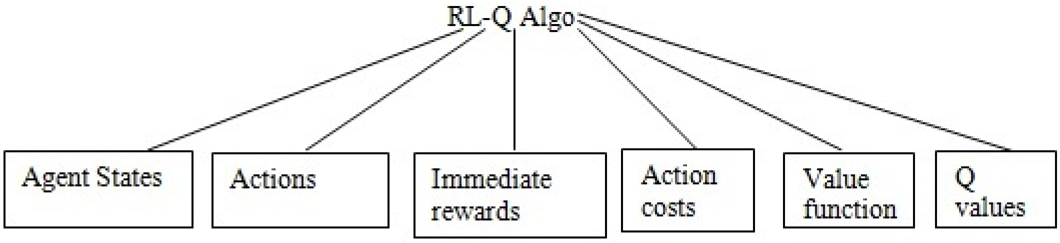


Figure 2.6: RL –Q learning algorithm [23]

Reinforcement learning (RL) is biologically considered as one of the machine learning technique. In this technique, learning agents gain its knowledge by directly interacting with its environment. This can be explained by example.

The RL –Q learning algorithm consists of the following:

Reinforcement learning (RL) is biologically considered as one of the machine learning technique. In this technique, learning agents gain its knowledge by directly interacting with its environment. This can be explained by example. A mouse trying to find cheese in a maze and it has to select a direction to move [23]. This is a reinforcement learning technique in which agents select actions and receive rewards from the environment.

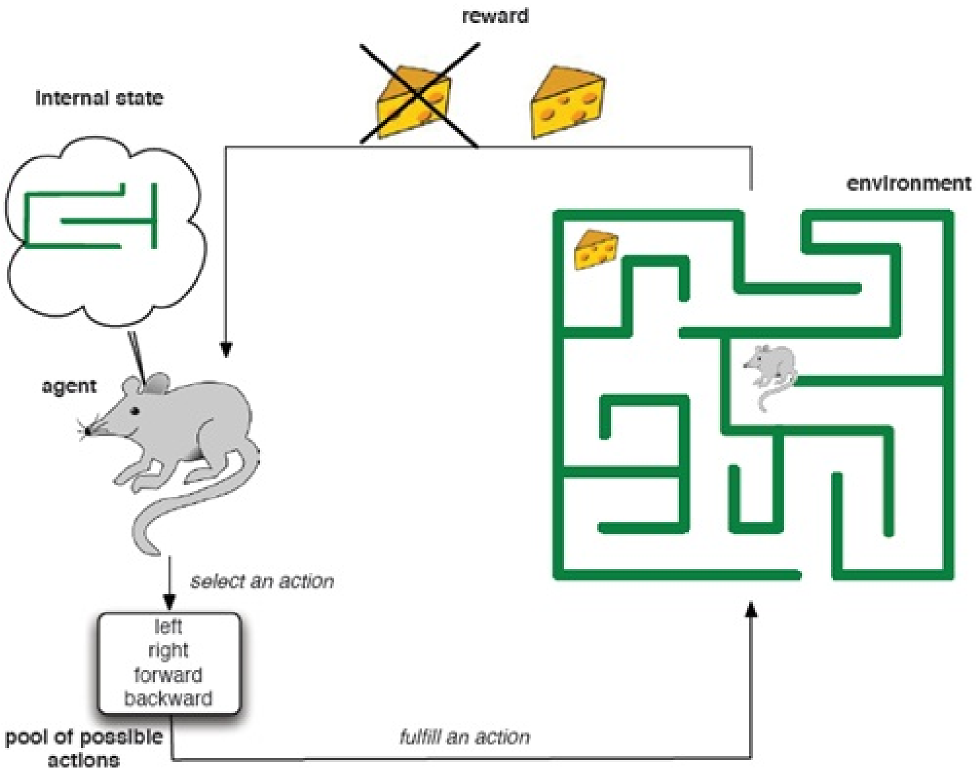


FIGURE 2.7: Reinforcement Learning [23].

Agent states- A set of learning agents was considered which consist of finite set of possible states T and c to be considered as the coordinate representing the state of the agent at time step t. The current state of the mouse is in maze. Actions-Q-Learning basically associates a different set of actions and to each of the states in T. Let the actions of the mouse are left, right, backward, forward.

Immediate rewards. There is an associate with each example and in our example, all of the state transitions that do not lead to the goal state have immediate rewards of 0(no cheese) and the ones leading to the goal state have an immediate reward of 1(cheese reached). The agent can see only the actions with their associated rewards from its current state. The global knowledge about the environment is not there but only the states information and the rewards.

Action costs- There are also costs in addition to rewards which are associated with each action in each state. This is a scalar value which represents the cost of the action. In this example the cost is one unit of energy of one cheese bite for the movement of the mouse. The negative reward costs are directly subtracted from immediate reward. Value function- The value function represents the expected total accumulated reward, in contrast to immediate rewards which are associated to each action in each state. These are easily observable; the value function represents the expected total accumulated reward. The goal of the agent is to learn a sequence of actions with a maximum value function, such that the reward on the taken path is maximized.

Q-Values-To represent the currently expected total future reward at any state, a Q-Value is associated to each action and state T (c, a). The Q-Value represents the memory of the learning agent in terms of the quality of the action in this particular state. In the beginning Q-Values are usually initialized with zeros, representing the fact that the agent knows nothing. Through trial and experience the agent learns how good some action was. The Q-Values of the actions change through learning and finally represent the absolute value function. After convergence, taking the actions with the greatest Q-Values in each state guarantees taking the optimal decision (path). Updating a Q-Value: A simple rule exists to update a Q-Value after each step of the agent: $Q(c+1,T) = Q(c,T)+(R(c,T)Q(c,T))$. The new Q-Value of the pair $c+1, T$in state$c+1$ after taking action at in state C is computed as the sum of the old Q-Value and a correction term, which includes the received reward and the old Q-Value. is the learning constant. It prevents the Q-Values from changing too fast and thus oscillating. The total computation of the received reward is as follows:

$R(c,T) = r(c,T) + c(c,T)(2); wherer(c,T)$is the immediate reward as defined above and c(c, T) is the cost of taking the action at in state c. Exploration strategy (action selection policy): Learning is performed in episodes, e.g., the mouse takes actions in its environment and updates the associated Q-Values until reaching the cheese. After completion, a new episode begins, repeating until the Q-Values no longer change. The question is how to select the next action. Always taking the actions with maximum Q-Value (greedy policy) will result in finding locally minimal solutions. On the other hand, selecting always random (random policy) will mean ignoring prior experience and spending too much energy to learn the complete environment[24].

2.7.6 Genetic Algorithms

Genetic algorithms attempt to incorporate ideas of natural evolution. In general, genetic learning starts as follows. An initial population is created consisting of randomly generated rules. Each rule can be represented by a string of bits. As a simple example, suppose that samples in a given training set are described by two Boolean attributes, A1 and A2, and that there are two classes, C1 and C2. The rule "IF A1 AND NOT A2 THEN C2" can be encoded as the bit string"100", where the two leftmost bits represent attributes A1 and A2, respectively, and the rightmost bit represents the class. Similarly, the rule "IF NOT A1 AND NOT A2 THEN C1" can be encoded as "001". If an attribute has k values, where$k > 2$, then k bits may be used to encode the attribute values. Classes can be encoded in a similar fashion. Based on the notion of the survival of the fittest, a new population is formed to consist of the fittest rules in the current population, as well as offspring of these rules. Genetic algorithms are easily parallelizable and have been used for classification as well as other optimization problems. In machine learning, they may be used to evaluate the fitness of other algorithms.

2.7.7 Rough Set Theory

Rough set theory can be used for classification to discover structural relationships within imprecise or noisy data. It implies to discrete-valued attributes. Continuous-valued attributes must therefore be discrete prior to its use. Rough set theory is based on the establishment of the equivalence classes within the training data. All of the data samples forming an equivalence class are indiscernible and the samples are identical with respect to attributes describing the data. Rough sets can also be used for feature reduction and relevance analysis. The problem of finding the minimal subsets of attributes that can describe all off the concepts in the given data set is NP-hard. However, algorithms to reduce the computational intensity have been proposed. In one method, for example, a matrix is used that stores the differences between attribute values for each pair of data samples. Rather than searching on the entire training set, the matrix instead searched to detect redundant attributes.

2.7.8 Fuzzy Logic

Fuzzy logic is useful for machine learning as it provides advantage of working at high level of abstraction. In general, the use of fuzzy logic in rule based systems involves the following: Attribute values are converted to fuzzy values. Fuzzy logic comes in when conventional logic fails. It can deal with virtually any proposition expressed in natural language. The meaning of propositions like this can be determined.

2.8 Distributed Event Detection in Wireless Sensor Networks for Forest Fires

In distributed event detection for forest fires, there are various machine learning techniques that can be applied. The following describes the working of various machine learning techniques (Artificial Neural Networks, Decision Trees, Support Vector Machines, Fuzzy Logic, Rough Sets, Regression and Clustering) and how these can be applied for event detection. Artificial Neural Networks The Artificial neural networks are kind of learning based algorithms. These basically work on principle of neuron. The first model of neuron contained two inputs and one output. Both the inputs should be active for correct output. The weights for both the inputs were equal and output was binary. The mathematical model functions of these are $F : X \to Y$. The data flows from input nodes to the output nodes through the whole network. Ability to learn is important property of artificial neural networks and to adjust input/output weights to reflect the exactly learned function. Thus, for training an artificial neural network, a set of data for training are needed in which inputs are already mapped to get the possible output. For example classification of different numbers, the pictures can be considered as inputs and the numbers can be considered as outputs. In contrast to decision trees, inputs cannot be described as attribute pairs.

The neural network is also known as supervised offline learning algorithm. This consists of a training set, which has already been classified. Offline is how much necessary is the training set which will be used for classification. There are also unsupervised and online learning neural networks. For example a network which is used for learning the data model for sensor readings. Neural networks are very suitable for such problems where features or attribute-values pairs are not available. However, they have large memory and processing requirements like decision tree learning. There are also some techniques which are applicable in WSN for static classification problems such as data models or link quality estimation. They can be efficiently implemented even on standard sensor nodes because of their low requirements. An attribute selection measure is a heuristic for selecting the splitting criteria that best separates the given data partitioning criteria, D of the class labeled training tuples into individual classes [25].

2.8.1 Decision Tree

The Decision Tree classifiers are usually successful in many diverse areas such as radar signal detection, remote sensing, medical diagnosis, expert systems, and speech recognition. It has capability to break down complex decision making process into simpler one, thus providing proper solutions which are easier to interpret. They can be easily represented in form of graphs or in form of rules. Chi square (C4.5) and ID-3 are the most widely used decision tree algorithms. The C4.5 computes the information gain and finds which feature splits clusters best. C4.5 takes feature of highest information gain and puts that information in the root node of the tree and then recursively computes the information gain for the subclasses until all the samples from training set are classified. In wireless sensor networks various problem can arise that how to classify links as good or bad based on the data such as signal strength or delivery rate [26].

2.8.2 Support Vector Machine

In machine learning, support vector machines are supervised learning models with associated learning algorithms that analyze data and match patterns [27]. The basic SVM takes a set of input data and estimates, which of two possible classes forms the output for each input value, making it a non-probabilistic binary linear classifier. Consider example of a training value set, which are marked as belonging to one of two categories, an SVM training algorithm which builds a model that assigns new examples into one category or the other. An SVM model is a representation of the examples as points in space, mapped so that the examples of the separate categories are divided by a clear gap that is as large as it can be. New examples are then mapped into that same space and predicted to belong to a category based on which side of the gap they fall on.

2.8.3 Rough Set Theory

Rough set theory can be used for classification to discover structural relationships within imprecise or noisy data. It implies to discrete-valued attributes. Continuous-valued attributes must therefore be discretized prior to its use. Rough set theory is based on the establishment of the equivalence classes within the training data. All of the data samples forming an equivalence class are indiscernible and the samples are identical with respect to attributes describing the data. Rough sets can also be used for feature reduction and relevance analysis. The problem of finding the minimal subsets of attributes that can describe all off the concepts in the given data set is NP-hard. However, algorithms to reduce the computational intensity have been proposed. In one method, for example, a discernibility matrix is used that stores the differences between attribute values for each pair of data samples. Rather than searching on the entire training set, the matrix instead searched to detect redundant attributes[24].

2.8.4 Fuzzy Logic

Fuzzy logic is useful for machine learning as it provides advantage of working at high level of abstraction. In general, the use of fuzzy logic in rule based systems involves the following: Attribute values are converted to fuzzy values. Fuzzy logic comes in when conventional logic fails. Fuzzy logic can deal with virtually any proposition expressed in natural language. For example, the proposition, “It is very unlikely that the price of gold will significantly increase in near future,” which beyond the classical first-order predicate logic, is perfectly manageable by fuzzy logic. The meaning of propositions like this can be determined. An important concept in fuzzy logic lies in the concept of linguistic variables whose values are words or sentences in natural language. In general, any relation between two linguistic variables can be expressed in terms of fuzzy if-then rules [4].

2.9 State of Art Method for Fire Detection

2.9.1 Distributed Event Detection in Wireless Sensor Network for forest fires

The approach progresses by dividing the training observations into constant size groups of sample vectors. These techniques work in incremental manner and hence considered useful and use only the fractional processed data at each step. In DFP-SVM or distributed fixed partitioning, we find the estimates of the hyper planes by series of incremental steps that occur at each cluster. Only the current estimation value is send and not all the past data as is there in the previous approaches .By following such a technique we achieve reduction in energy consumption followed by other advantages, some of which may include-reduction in complexity and increased efficiency of the system.

The setup involves deploying sensors which have the ability to sense carbon dioxide, carbon monoxide, temperature and other parameters that can predict fire. Each sensor will detect the requisite parameter and send it. This data will a large stream data. The clusters are created dynamically and this large stream of data is then passed and clusters are made which contain data that needs large storage value. The data is send from the cluster head to the base station. We use clustream algorithm which is basically used for clustering of large data streams. This algorithm generates such data about data termed as metadata along with some decision attributes which as generally estimated and predicted from the previous stored observations. The meta data that is in the tabular form is analyzed and recognized by which we can predict or detect the event [28].

2.9.2 Forest Fire Detection in Wireless Sensor Network Using Fuzzy Logic

Fuzzy Logic System is a technique by which we assign values ranging from 0 to 1 to all the variables involved.FLS structure consists of the following processes namely inference, defuzzification and fuzzification. Crisp inputs and converted into fuzzy inputs by the technique of fuzzification.

A function called membership function (MF) evaluates the truth degree for every input and output which are taken into consideration. The value of membership function is between 0 and 1 which ranges between the intervals of the chip. Usually triangle, trapezium and bell shaped curves are shapes which are used. This paper has taken five parameters into consideration in detection of fire [18]. These are smoke, temperature, carbon monoxide, ionisation and photoelectric effect. For the output, five attributes are considered: low, medium, high, very high, very low. For temperature, smoke and carbon monoxide there are three variables: high, medium, and low and MF for distance has three variables as close, average and far. The inputs of the fuzzy logic are considered which are based on fuzzy rules and give a fuzzy output. The fuzzy rule can be written as IF a1 is T1 and a2 is T2 . . . and an is Tn THEN b is bn . There are various steps for detection of fire using fuzzy rules. For each crisp input to be taken, there are a lot of variables which are defined. These are temperature having values from 0 to 120C and can have values V,M and H, smoke having values from 0 to 100ppm(categorized as L,M,H),humidity value ranges from 0 to 100ppm (categorized as L,M,H)light values ranging from 0 to 1000lux(categorized as L,M,H) and distance ranging from 0 to 80m categorized as Close, Medium and Far. By plotting the special input parameter along the X axis and projecting it on the vertical side on the side where the Mf lies. The outputs are achieved by studying the relations of the different input parameters. A very good example is that that the fire is low when we assume that the light, smoke and temperature is low, distance is far [29].

2.9.3 SWATS: Wireless Sensor Networks for Stream flood and Water flood Pipeline Monitoring

The system used has the main objective to permit the low cost and high accuracy monitoring of the water flood and the steam flood system. It identifies major problems (such as leakage, blockage, outside force that might cause the flow) and obstacles that would take place when developing pipeline networks so that systems with high reliability can be established .The system is made so that there is no false alarm that may take place due to various reasons such as pressure, humidity change, environmental effects or the phase change. To detect the anomalies in pipeline networks of steam and water is difficult and sometimes be erroneous because these sensors have inherent problems and the other environmental effects give an impetus to it thereby reducing the accuracy level of these sensors and sometimes generating false alarms thus making this field a challenging field in sensor detection. The problems are mainly due to the different size and the shapes of the pipes thus varied pressure that is applied on the pipes and complexity of the pipe topological properties such as merge, split etc. Water and steam is transient so a single sensor is not enough for the detection. All these problems are taken into account and they are solved by making a multisensory algorithm that employs multimodal sensing capacity. The sensor accuracy is increased by combining the inputs that are received from multiple sensors and also studying the data correlations that is present among the different attributes. This technique applies a novel method for application in water flood system and also include the localization and the identification of the steam and water [30].

2.9.4 Wireless Sensors and Neural Networks for Intruders Detection and Classification

The architecture used in this technique consists of both multi homing and multitier approaches .The lower tier comprises of the sensor nodes that are spread over a large region in the space where the event is to be monitored .The sensor nodes propagate the data to gateways which comprises the middle tier of the architecture. The lower tier consists of the architecture of the system utilizes both multitier and multihoming techniques. These gateways also deliver the packet to the central server. The middle and the upper tier communicate with each other by using the wireless fidelity (Wi-Fi). The cluster gateway, WSN cluster, the monitoring Client and the interface module are the basic networking modules. There are 20 modules in the WSN cluster. The sensor board is equipped with light and temperature sensor. An activation function as given in Eq.1 is applied to the ANN which is dot product so as to get the output of the system. The system involves a set of input values, a bias value plus synaptic weight. $U_j(x) = \sigma(x\text{ffl}w_j +_j) \quad (Eq.1)$

Where u_j is the output of the jth neuron, the activation function, x the input vector,w_j the synaptic weight vector of the j^{th} neuron, and $_j$ the bias associated with the j neuron. The activation function is usually a nonlinear function, e.g. hyperbolic-tangent or the logistic function[31].

2.9.5 Application of Wireless Sensor Networks in Forest Fire Detection under Uncertainty (Rough Set approach)

Attribute reduction mechanism and efficient feature selection process is established by rough set approach. In this approach the granularity and the aggregation is established by the equivalence relations sometimes also termed as indiscernibility relations on the set of objects. We require the formation of a system in the data representation of rough set which is a pair S = (U, A), where U is a nonempty, Finite set called the universe and A is a nonempty, finite set of attributes [32]. A is the decision attribute or class label. Using this technique the most critical aspects of forest fire is studied and rules are developed for its detection by building a robust model which also incorporates the missing values which might sometimes occur in real life situation .It has been also ensured that the performance will not go below a certain threshold in this system despite that the individual nodes fail. The sensors in a cluster are equipped with domain specific function procedures or lookup tables with limited computing capability.

For each cluster Sc= (U,A) is generated dynamically by the cluster and it basically consists of the observations that are acquired from the sensor nodes. These values are sent to the cluster head[33].

2.9.6 Distributed Event Detection in Wireless Sensor Networks for Disaster Management

A tree which uses discrete and continuous values as input to form a graph is known as decision tree. It is a greedy approach [22]. There are two phases in construction of decision tree: training phase and testing phase. A set of data is taken as input in training phase and minimum depth of the tree is found which will reduce the time complexity and memory space. Every sensor is involved in detection of events. The results of all the values sensed by all the nodes are collected and send to the voter which works on the basis of reputation based voting. A conlusion is made by finding the reputation of each and every node and the node with hoghest reputation value is selected. Firstly, it is assumed that all the events detected by sensors are correct. The detected values of all the nodes are sent to the neighbors detection value table and every node decisdes according to the value of its neighbor sensors. The difference between these values is calculated i.e. the sensor node value and the values of other nodes. The comparison of the difference is then compared with the threshold values which are predefined. The voter gives a positive vote if the value is smaller than the threshold value and viceversa. To make a decision the values of the NDVT tables are sent to the voter. The most difficult task is to choose the node with highest reputation value which will decide whether event has occured or not. This can be done by the use of voting techniques.

Reputation Technique 1 It evaluates the values detected by each and every sensor. Then these values are taken into consideration by comparing with the neighbourhood. Then average of all the values is calculated and reputation value of each sensor node is multiplied which is taken as weight. It can be calculated as

$$W_s = R_s * Av_s$$

Where, R_s is the reputation value of the sensors,W_s is weight of sensors and Av_s is average of all the values.

Reputation Technique 2 In this technique, two threshold values are predefined as Q1 and Q2 and these values can be assigned manually and the reputation value is compared with these threshold values which helps to decide whether decision is perfect or not. If$R_s >= \theta_1$; decision is perfect. The decision is Ok if the value lies between both threshold values and poor if R is less than θ_2 [34].

2.9.7 Circle-based Approximation to Forest Fires with Distributed Wireless Sensor Networks using clustering

A disaster management system deals with situations in which various sensors are deployed in a distributed network in which data is collected from different sources. The main objective is to make decisions about the occurrence of events. In previous work the focus was on forest fires and EIDOS (Equipment Destined for Orientation and Safety)[35], system. Its main goal was to decrease the risk of occurrence of fires and increasing the detection accuracy. A large sensor network was used in affected area of forest fires which contributed a lot in gathering the information by the fire-fighters to increase the safety [36]. The aim of data gathered by the sensor nodes was to attain the location and position of fire fronts which were active at that time. The working of EIDOS system can be explained as follows. Firstly the multifunctional sensors are deployed in an environment. The fire- fighters carry devices such as smart phones which help to check the outcomes of the distributed algorithms implemented in sensor networks. The mobile sensors were able to display the fire map in form of Graphical user interface. In this the system was working correctly with the connectivity with the central node. Further, the topology of the nodes was irregular and unknown. One of the tasks was to detect the geographical position of every node. This was obtained by localization process. In this technique, the localization process was range free and connectivity information was used by sensor nodes to estimate their position. Assumption was that the parameters sensed must be within a range and threshold values were predefined. So, area burnt at a particular temperature ab, such that ab ¿ td, where td is the position after burning. To reduce the energy consumption, WSN nodes do not maintain any hierarchy and they do not have any predefined information. Within these limitations, EIDOS was implemented. Each node maintains and builds a local approximate value starting from the detection till occurrence of the fire [27].

2.9.8 Data Mining Approach to Predict Forest Fires using Meteorological Data.

This describes a novel data mining forest fire methodology in which real time and meteorological data was used. The real time data gathered from northeast region of Portugal was used which help to predict burnt area of forest fires [31]. There are various models with advantages and capabilities which have been used in regression task. Most easy and classical approach is the Multiple Regression (MR) model. Only linear mapping is learnt by it. Thus, to solve this disadvantage, nonlinear functions like neural networks and support vector machines should be used. Tree structures like decision trees and random forest can also be used but these are difficult to implement with large data. This approach considers four parameters like rain, wind, temperature and relative humidity[37]. It predicts the burnt area of fires in which majority of occurrences of fire were there. This help to make decisions regarding fire occurrences. A data set of regression consist of h 1, ...,K examples. Each maps an input vector (a1h......akh) to a predefined target b$_h.Theerrorisgivenby$:e$_h = b_h – \widehat{b^h}$, where b_h denotes the predicted value for the h input pattern. The performance of the overall task is calculated by a metric which is global and called as Mean Absolute Deviation (MAD) and Root Mean Squared Error (RMSE).

$$MAD = \frac{1}{k} * \sum_{i=1}^{k} |b_i – \widehat{b_i}|$$

$$RMSE = \sqrt{\sum_{i=1}^{k} \frac{b_i – \widehat{b_i}}{k}}$$

Lower the values of these better will be results. But RMSE is more prone to high errors. We can also compare the regression models by REC curve i.e. Regression Error Characteristic curve which gives the relation of error tolerance and predicted percentage of points of burnt area [31]. There are various other ways also by which we can solve make results for regression like on the basis of variance, MAE (Mean Absolute Error), Residual after regression fit, MAD (Mean Absolute Deviation), NMSE (Normal Mean Squared Error) and MAPE. The selection of best regression model can be evaluated by two procedures. First is the all possible regression method in which we evaluate R- squared, Adjusted R- squared, Mean Squared Residual and Mallow's Statistics. In other procedure called the sequential selection three steps are there: Forward selection, Backward selection and Step-wise selection [38].

S.No	Technique	ML Technique Used	Advantage	Event Detected	Basic Methodology
A	Distributed Event Detection in WSN for forest fires	Clustream And Support Vector Machine	1)Linear Complexity 2)Energy Efficient 3)Minimize delay in disseminating information.	Fire	1)One hop tree-minimize delay 2)SVM –prediction
B	Forest Fire Detection in WSN Using Fuzzy Logic	Fuzzy Logic	1)Deeper analysis can be done to observe effect of each input.	Fire	1)Uses five membership function (temperature,smoke,light,humidity and distance) as input.
C	SWATS: WSN for Steam flood and Water flood Pipeline Monitoring	Reinforcement Learning(Markov decision problem).	1) Reduction in false alarm reduction. 2)Quick localization,continious monitoring and Reliable	Blockage and leakage in stream flood and water flow pipelines in oil fields.	1) Use multimodal sensing and multisensor collaboration to exploit temporal and spatial patterns.
D	WSN and Neural Networks for Intruders Detection and Classification	Neural Network	1) Economical as it uses small cheap nodes that are self powering and self configuring. 2) Detect and classify intruders	Assets and national borders	1)Multi-homing technique allows reliable identification of intruders. 2) The implemented networking modules are the WSN cluster,the cluster gateway, the monitoring server, the monitoring client, and the interface module
E	Distributed Event Detection in WSN for Disaster Management	Decision Tree.	1)Voting mechanism to reach consensus.	Forest Fire.	1) Based on reaching a consensus of the detections made by various sensor nodes using decision tree classifiers.
F	Application of WSN in Forest Fire Detection	Rough set	1)Offers an attribute reduction algorithm and dependency metric for feature selection.	Forest Fire.	1)Unlike traditional compression algorithms, the rough set based dynamic feature selection algorithm allows the compression of data stream without altering underlying data semantics.

Table 2.1: Comparison of various techniques used for fire detection in wireless sensor network

2.10 Conclusion

Thus, SVM Regression technique has been applied in this to further implement the event detection using machine learning techniques and various simulation results have been formed thereafter the proper event detection techniques and various machine learning techniques and algorithms in this thesis.This will improve the performance of the various systems and the cost will be reduced accordingly.The functionality of detecting the event is helpful and important when we need to detect the natural calamity such as fire or earthquake. It has been observed that many real world activities exhibit certain pattern which can be detected by applying certain machine learning techniques. In this chapter, we have studied and discussed variety of methods applying machine learning techniques used for event detection and a comparative analysis of all of them are summarized.

Chapter 3

Regression and Support Vector Machines in Forest Fires.

3.1 Introduction

Forest fire detection has become a major issue now days. There are many reasons due to which forest fires occur. It may be due to human negligence or environmental changes. It was found that most of the reasons for occurrence of fires are humans as fires were more in week days as compared to weekends. Fires can be in forest, offices, home etc[39]. So, the detection of fires should be fast and accurate so as to save lives and avoid destruction. There are many machine learning techniques which help in fast detection but these should be accurate also[40].

There are various machine learning techniques by which these can be detected. The main advantage of using machine learning techniques is that these will act like differentiators in addressing event detection tasks. Theoretical and practical advancements in these have ability to handle a large number of applications like public health, disaster management, business, ecology, security etc. Since use of machine learning can cause class imbalance, so the data used by these should be correct[41].

There are many problems which can be solved by machine learning: Classification problem occurs when the data does not belong to a particular class it should belong to. Estimation problem occurs when there is doubt about the result of the proposed approach. The sensor nodes detect patterns and match them. So the discovery of patterns requires use of techniques called machine learning techniques [12].

There are various factors on which machine learning techniques depend upon. These can be accuracy, lift, ROC area, R-squared, cross entropy, squared error, average precision etc. Also, there are some factors which should be considered before opting any machine learning technique like memory space occupied, computational time, optimality, tolerance to topology changes, initial costs, add costs etc. The techniques should be light-weighted and accurate for decentralized event detection in a network[6].

Our proposed approach is regression, which not only increase the accuracy of detection of forest fires but also reduces the RMSE (Root Mean Squared Error). The work done by regression in this has taken less time for analysis as compared to other machine approaches. The details of the proposed approach are described in which regression technique is shown as the best technique for detection in terms of accuracy and RMSE. Also, a comparative analysis of few machine learning techniques has also been described.

3.2 Proposed approach

The proposed approach is based on increasing the accuracy of forest fires detection. In this approach the accuracy is increased and the error (RMSE) i.e. Root Mean Squared Error has been reduced. The dataset of forest fires has been taken from UCI repository. It contains 4 predictors (temperature, wind, rain, relative humidity) in conjunction with a SVM and it is capable of predicting fires, which constitute the majority of the fire occurrences. Also, FWI parameters were also used for predicting fires.

FFMC (Fine Fuel Moisture Code) denotes the moisture content surface litter and influences ignition and fire spread. DC (Drought Code) represents the moisture content of shallow and deep organic layers, which affect fire intensity. ISI (Initial Spread Index) is a score that correlates with fire velocity spread. BUI (Build up Index) represents the amount of available fuel. The Multiple Regression (MR) model is easy to interpret but it can only learn linear mappings. Use of Support Vector Machines (SVM) helps to solve this drawback since it uses non-linear mappings. In this approach basically we have taken a forest fire data set and divided that data set into 5 different levels of fires. If the level of the fire is more then only alarm will be generated i.e. if there are chances of very large fires. After dividing the data set into levels, the data set is further divided on the basis of months. The reason of dividing the dataset into months is that if there are chances of occurrence of fire in a particular month then the whole data set need not be traversed. Only the data set of that particular month will be traversed and the level of the fire will be quickly interpreted. On the basis of the level of the fire the alarm will be generated. For detection of forest fires and to save lives the detection of forest fires should be fast and accurate. So, detection of the forest fires becomes fast by division of data sets. Now, for increasing the accuracy of thee forest fire detection one important thing is that the parameters which are significant for the detection of fires should be considered. It was found that when only few parameters were considered the accuracy was less. But, when all the parameters which are important were considered, the accuracy was increased. This shows that the parameters necessary for the detection of forest fires should be considered.

3.2.1 Algorithm

A dataset based on collection of parameters acquired from sensor having n instances for which the accuracy has to be maximized.

MAX_SEN_ACC (*Lev*)
1)Initially all instances and attributes in the model .
2)Give appropriate level to class attribute according to $'Lev'$ given. where

$$maxlevValue_i = i * \frac{maxClassAttr - minClassAttr}{lev}$$

where maxlevValue is maximum value at level i.

$$minlevValue_i = maxlevValue_{i-1}$$

where minlevValue is minimum value at level i
3)Select the attributes that yield the highest significance/show highest variability.
4)Divide dataset based on months(eg fire could be likely/unlikely in specific month).
5)Make regression equation based on attributes selected in step 3 for each month.
6)Predict the level of fire and raise alarm based on the level of fire.

Output: A highly accurate and robust method of detecting fire

FIGURE 3.1: Proposed Graph Based Algorithm

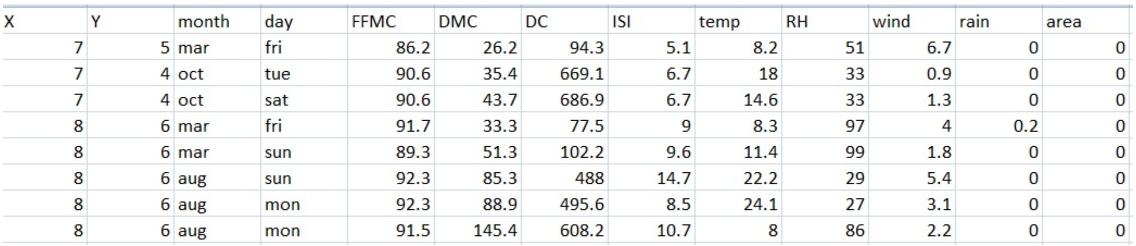
TABLE 3.1: Meaning of symbols used in the algorithm

Symbol	Meaning
MAX_SEN_ACC	Maximum Sensor Accuracy
$Maxlev_i$	Maximum value at level i
MaxClassAttr	Maximum value of the class attribute
MinClassAttr	Minimum value of the class attribute
Level	Level of the fire(L1,L2,L3,L4,L5)

The algorithm proposed was run on forest fire dataset to obtain a final model that would have the minimum mean square error and yield the maximum accuracy explaining the maximum variability in the model which would increase the predictive power.

3.2.2 Proposed approach and attribute selection

The following is the forest fire data set which is considered. In this dataset all the attributes were considered significant and were used as predictors for forest fire detection. Sometimes, measurement of a particular parameter becomes highly complex. The proposed model was done by taking input of the following dataset which is forest fire dataset according to months and days in Minitab. It was found that most of the fires were on week days as compared to week ends. In the proposed model, we can reduce the number of the attributes and yield an r-squared that explain the model approximately same. Thus, we can build a regression equation by the use of those attributes and predict the level of the fire. In this care all the attributes were considered significant and the area was taken as response attribute on the basis of which the levels were made as l1, l2, l3, l4, l5. These were the levels of fires.

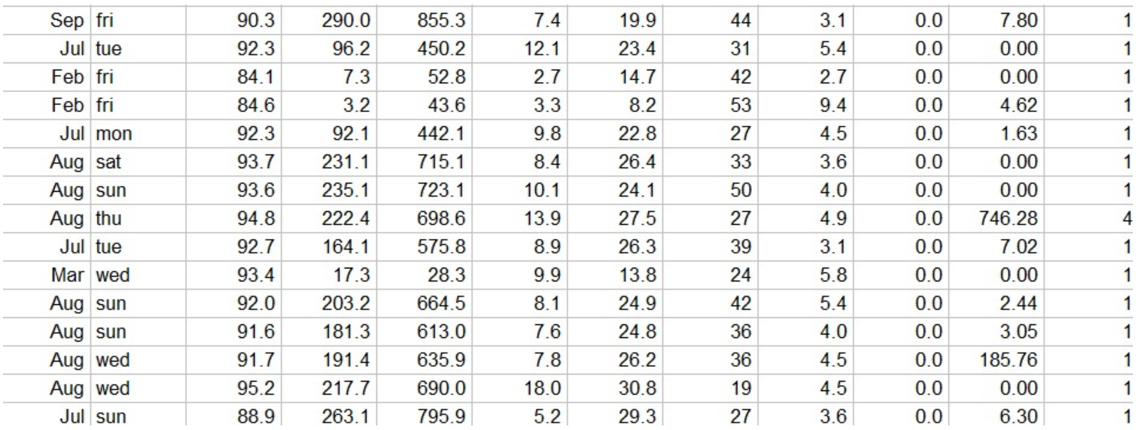


X	Y	month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area
7	5	mar	fri	86.2	26.2	94.3	5.1	8.2	51	6.7	0	0
7	4	oct	tue	90.6	35.4	669.1	6.7	18	33	0.9	0	0
7	4	oct	sat	90.6	43.7	686.9	6.7	14.6	33	1.3	0	0
8	6	mar	fri	91.7	33.3	77.5	9	8.3	97	4	0.2	0
8	6	mar	sun	89.3	51.3	102.2	9.6	11.4	99	1.8	0	0
8	6	aug	sun	92.3	85.3	488	14.7	22.2	29	5.4	0	0
8	6	aug	mon	92.3	88.9	495.6	8.5	24.1	27	3.1	0	0
8	6	aug	mon	91.5	145.4	608.2	10.7	8	86	2.2	0	0

Figure 3.2: Illustration of the proposed approach

The above is a small part of the forest fire dataset taken into consideration. In this we have taken all the attributes as significant attributes. After this the level of fire can be predicted by defining levels of fires. If more levels are defined then more accuracy is achieved. In this research work, we have considered five levels of fire.

In the below figure the division of dataset with respect to levels is shown in which five levels of forest fires were considered and according to the level of the fire the alarm was generated. If the level of the fire was 1, then there alarm would not be raised because the fires were less. But if the level is more like 4, then alarm would have been raised. The level of the data set is in increasing order as 1, 2, 3, 4, 5.

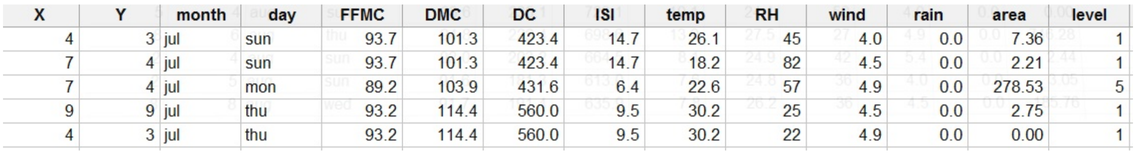


Sep	fri	90.3	290.0	855.3	7.4	19.9	44	3.1	0.0	7.80	1
Jul	tue	92.3	96.2	450.2	12.1	23.4	31	5.4	0.0	0.00	1
Feb	fri	84.1	7.3	52.8	2.7	14.7	42	2.7	0.0	0.00	1
Feb	fri	84.6	3.2	43.6	3.3	8.2	53	9.4	0.0	4.62	1
Jul	mon	92.3	92.1	442.1	9.8	22.8	27	4.5	0.0	1.63	1
Aug	sat	93.7	231.1	715.1	8.4	26.4	33	3.6	0.0	0.00	1
Aug	sun	93.6	235.1	723.1	10.1	24.1	50	4.0	0.0	0.00	1
Aug	thu	94.8	222.4	698.6	13.9	27.5	27	4.9	0.0	746.28	4
Jul	tue	92.7	164.1	575.8	8.9	26.3	39	3.1	0.0	7.02	1
Mar	wed	93.4	17.3	28.3	9.9	13.8	24	5.8	0.0	0.00	1
Aug	sun	92.0	203.2	664.5	8.1	24.9	42	5.4	0.0	2.44	1
Aug	sun	91.6	181.3	613.0	7.6	24.8	36	4.0	0.0	3.05	1
Aug	wed	91.7	191.4	635.9	7.8	26.2	36	4.5	0.0	185.76	1
Aug	wed	95.2	217.7	690.0	18.0	30.8	19	4.5	0.0	0.00	1
Jul	sun	88.9	263.1	795.9	5.2	29.3	27	3.6	0.0	6.30	1

Figure 3.3: Illustration of the proposed approach

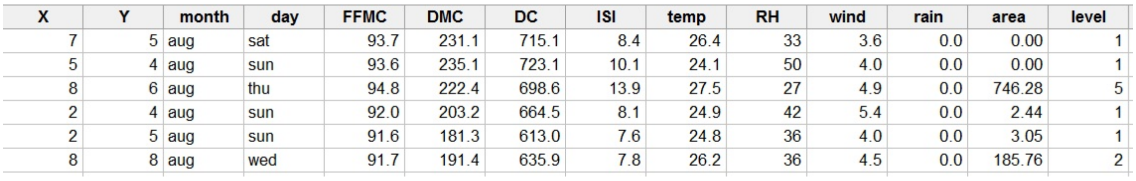
3.2.3 Division and better prediction

The algorithm divides the datasets based on the months (in some other application it could also be some other categorical predictor).This is done because the observations in a particular month would be similar in a particular way such as it would be more probable to have rain in august than in may. Thus there will be certain levels common to a particular month. This helps in easy prediction as one can estimate the value even without running algorithm for that particular value.



X	Y	month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area	level
4	3	jul	sun	93.7	101.3	423.4	14.7	26.1	45	4.0	0.0	7.36	1
7	4	jul	sun	93.7	101.3	423.4	14.7	18.2	82	4.5	0.0	2.21	1
7	4	jul	mon	89.2	103.9	431.6	6.4	22.6	57	4.9	0.0	278.53	5
9	9	jul	thu	93.2	114.4	560.0	9.5	30.2	25	4.5	0.0	2.75	1
4	3	jul	thu	93.2	114.4	560.0	9.5	30.2	22	4.9	0.0	0.00	1

Figure 3.4: Divided dataset for the month of July



X	Y	month	day	FFMC	DMC	DC	ISI	temp	RH	wind	rain	area	level
7	5	aug	sat	93.7	231.1	715.1	8.4	26.4	33	3.6	0.0	0.00	1
5	4	aug	sun	93.6	235.1	723.1	10.1	24.1	50	4.0	0.0	0.00	1
8	6	aug	thu	94.8	222.4	698.6	13.9	27.5	27	4.9	0.0	746.28	5
2	4	aug	sun	92.0	203.2	664.5	8.1	24.9	42	5.4	0.0	2.44	1
2	5	aug	sun	91.6	181.3	613.0	7.6	24.8	36	4.0	0.0	3.05	1
8	8	aug	wed	91.7	191.4	635.9	7.8	26.2	36	4.5	0.0	185.76	2

FIGURE 3.5: Divided dataset for the month of August

3.2.4 Regression technique for forest fire

Regression is a technique used to study the relationship between a dependent attribute such as level of fire which depends on attributes such as humidity, rain etc [42]. Regression along with SVM helps to decide the level of fire. We derive an equation that tells us the value of the level of fire which is :-

$$Y_i = \beta_1 X_{i1} + \cdots + \beta_p X_{ip} + \varepsilon_i$$

This equation can be written in matrix form as follows :- where Y is the vector of observation

X is the Regression variable

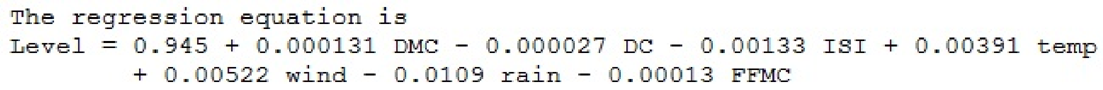
β is the vector of parameter

ε_i is the vector of error

β_1 is the Slope We can get a more precise and accurate value when we increase the number of levels. These levels help to decide whether there is danger or not. This approach can be extended to other applications also in the field of wireless sensor network.

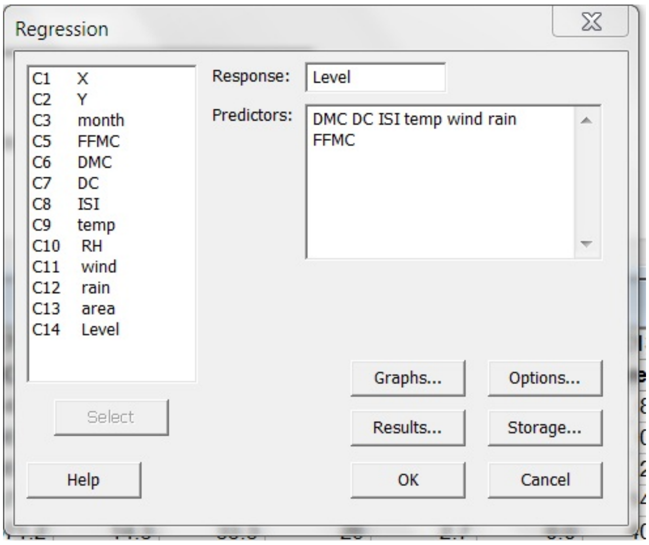
3.2.5 Predictor variation in the model

Sometimes measurement of a particular parameter involves highly complex experiment. In the model proposed we can reduce the number of the attributes and yield an r-squared that explain the model approximately same. Thus we can build a regression equation by the use of those attributes and predict the level of the fire.



```
The regression equation is
Level = 0.945 + 0.000131 DMC - 0.000027 DC - 0.00133 ISI + 0.00391 temp
        + 0.00522 wind - 0.0109 rain - 0.00013 FFMC
```

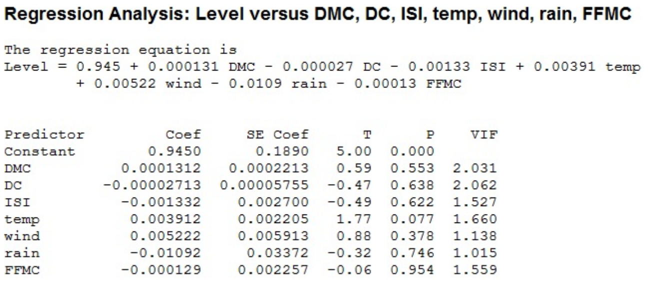
Figure 3.6: Results in minitab



Response: Level

Predictors: DMC DC ISI temp wind rain FFMC

Figure 3.7: Selecting the predictors in minitab



Regression Analysis: Level versus DMC, DC, ISI, temp, wind, rain, FFMC

```
The regression equation is
Level = 0.945 + 0.000131 DMC - 0.000027 DC - 0.00133 ISI + 0.00391 temp
        + 0.00522 wind - 0.0109 rain - 0.00013 FFMC


Predictor         Coef     SE Coef      T      P    VIF
Constant        0.9450      0.1890   5.00  0.000
DMC          0.0001312   0.0002213   0.59  0.553  2.031
DC         -0.00002713  0.00005755  -0.47  0.638  2.062
ISI          -0.001332    0.002700  -0.49  0.622  1.527
temp          0.003912    0.002205   1.77  0.077  1.660
wind          0.005222    0.005913   0.88  0.378  1.138
rain          -0.01092     0.03372  -0.32  0.746  1.015
FFMC         -0.000129    0.002257  -0.06  0.954  1.559
```

Figure 3.8: Regression Analysis

3.2.6 Simulation Setup

The simulation was done in Minitab in which a data set was taken as the input and taking area as class attribute, levels were created. After this, to find the level of the fire the dataset was divided according to months and regression above regression equation was formed by taking response as level as explained.

3.3 Conclusion

An algorithm for detection of fire has been proposed by using regression and dividing the datasets according to months. The algorithm achieves low root mean square error and high R-squared. The beauty of the algorithm lies in the way it can give the result without doing the computation on whole dataset. In future this approach can be extended by for other disasters as well. Application of certain transformation might also improve the model efficiency.

Chapter 4

Results and Discussions

4.1 Data Collection

The dataset used in the thesis has been obtained from the UCI repository which has datasets and data generators of various fields which can be used for performing machine learning techniques.

Forest Fire Data Set: This multivariate dataset has real type of attributes has 517 samples and 13 features, aims to predict burnt area [9]. There are 11 predictors in this dataset used for prediction of forest fires.This dataset has collected data according to months. Each month has different readings but the predictors are same. These 11 predictors are temperature, relative humidity, rain, FFMC (Fine Fuel Moisture Code), DC (Drought Code), ISI (Initial Spread Index), BUI (Build up Index). All the predictors in this dataset are significant. For increasing the accuracy of forest fires and for fast detection R-squared and RMSE has been calculated. In these results we have calculated R-squared as

$$R - Sqrd = \frac{SS_{res}}{n-k-1}$$

Table 4.1: Comparison of machine learning techniques

Machine Learning Technique	R -square	RMSE	Analysis Time
Decision Tree	16.40	0.204558	00:00:19
Linear Regression	2.30	0.221132	00:00:17
GRNN	0.92	0.222693	00:06:45
SVM	0.003	.2237171	29:44:49
Proposed Approach	69.21	0.079785	00:00:12

Table 4.1 shows the comparison between the simple linear regression, decision tree, general regression, neural network and support vector machine on the basis of root mean square error. In the above table, decision tree has accuracy of 16.40 percent, General Neural Networks are 0.92 percent accurate, Support Vector Machines are 0.003 percent accurate. Thus SVM is least accurate as compared to other machine learning techniques which have been taken into consideration. SVM is least accurate and it cannot be used for large fires. The RMSE of decision tree is 0.204558, GRNN is 0.222693 and SVM is .2237171. In this there are more chances of occurrence of errors in case of decision trees. The decision trees cannot be used for large datasets.

Thus it can be observed from the figure that the proposed algorithm builds a model that always has least Root Mean Square Error and has ease in prediction of fire as it defines some particular level that can be a final concept class for a particular month.

Thus, if only the month is given we can tell the fire levels that are common without even running the algorithm for that particular observation by using the value obtained by algorithm for previous values. Since, the proposed approach has least Root Mean Squared Error as shown in Fig 4.1 , so it is more accurate as compared to other techniques.

In the above figure, it can be seen that the RMSE(Root Mean Squared Error) of the proposed approach is very less as compared to other techniques. The accuracy and the analysis time are more and less respectively of the proposed approach. The datasets are divided after defining 5 levels of fires. This division strategy helps in

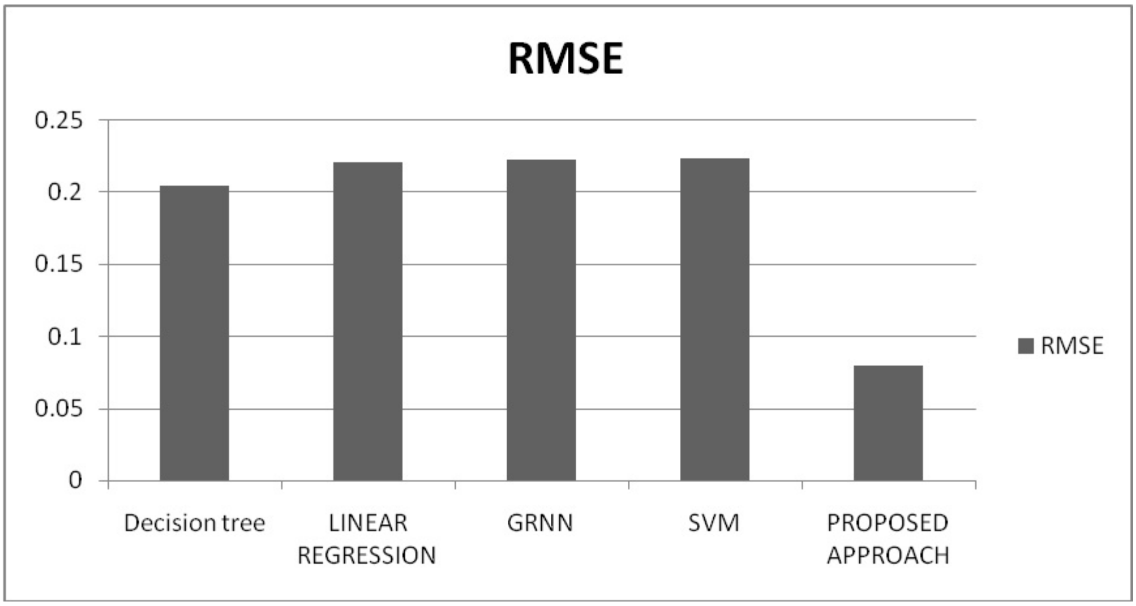


FIGURE 4.1: Graph of performance analysis

fast and accurate detection which helps to detect the forest fires very quickly and accurately because only the dataset of that particular month in which detection is done will be traversed. . Thus, there is no need to traverse the whole dataset for detection of the fires and it can be detected very fast as compared to other machine learning techniques. This adds costs and optimality of the proposed approach is better as compared to other techniques and thus, it is best in terms of prediction of fires or any other natural disasters.

In the Fig 4.2 it can be seen that its R-squared is also high. This means that the detection accuracy of the proposed approach is very large as compared to other machine learning techniques which show that proposed approach is best for accurate forest fire detection. It can also be observed that the proposed approach shows the maximum variability in the model as it has the highest R squared value. The selection of the predictor is also responsible for increasing the accuracy of the detection technique. The below is the detection accuracy when all the parameters were considered as significant parameters for detecting forest fires in a particular

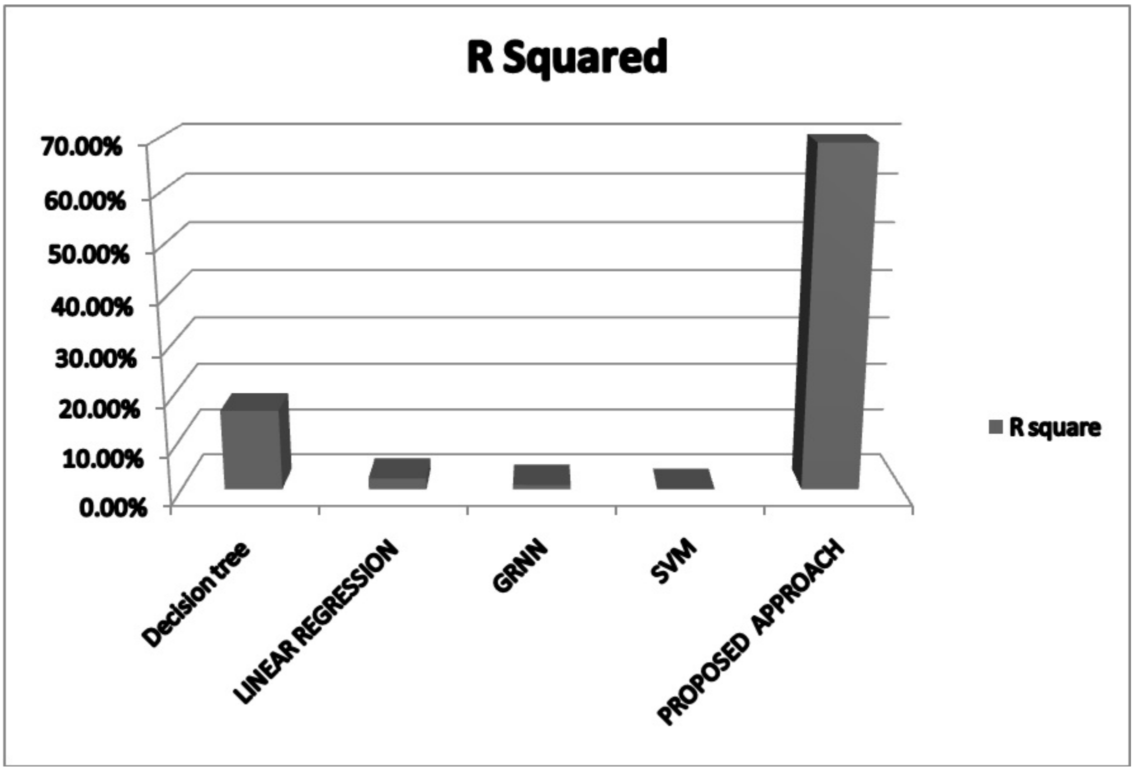


Figure 4.2: R-squared values comparison

month.It was observed that when only few parameters were considered the detection accuracy was less. But, when all the parameters were considered significance as shown by the tool, the detection accuracy was increased. SVM and generalised Neural Network have least accuracy as compared to other techniques according to the proposed methodology. Thus, the proposed approach is more accurate because the R-squared value of the proposed approach is very large as compared to other machine learning techniques. R- Squared value is in increasing order as $SVM < GRNN < LinearRegression < Decisiontree < proposedapproach$.

4.2 Conclusion

In this thesis, the detection of forest fires has been addressed. The system has fulfilled the drawback of accurate detection. The detailed analysis of various machine learning techniques have been done in this research work and these techniques have been applied on various event detection applications. Machine learning helps in fast and accurate detection.

We have proposed a novel detection algorithm which helps to increase the detection accuracy of forest fires and the errors have been reduced. The detection is fast and accurate. The division of the dataset makes the detection of forest fires fast because there is no need to traverse the whole dataset. Only the dataset of that particular month can be traversed.

The comparative analysis of the proposed approach with other machine learning techniques has also been done in this research which proves that the proposed approach is better than other techniques. The main issue of accurate and fast detection has been fulfilled by RMSE and R-squared calculation respectively. The beauty of the algorithm lies in the way that it can give the results without doing the computation on whole dataset. The complexity of the proposed approach is linear which comes out to be O (n).

In future, this approach can be extended by for other disasters as well. Application of certain transformation might also improve the model efficiency. Further, the accuracy can also be improved by applying some transformations to the proposed approach.

References

[1] S. B. Kotsiantis, “Supervised machine learning: a review of classification techniques.” *Informatica (03505596)*, vol. 31, no. 3, 2007.

[2] S. DATA, V. R. B. O. W. LINEAR, and P. CODING, “i vol. 1, issue 4, pp. i-iii,” *SPACE*, vol. 160, 1963.

[3] Z.-J. Zhang, J.-S. Fu, H.-P. Chiang, and Y.-M. Huang, “A novel mechanism for fire detection in subway transportation systems based on wireless sensor networks,” *International Journal of Distributed Sensor Networks*, vol. 2013, 2013.

[4] A. Phani Kumar, A. M. Reddy V, and D. Janakiram, “Distributed collaboration for event detection in wireless sensor networks,” in *Proceedings of the 3rd international workshop on Middleware for pervasive and ad-hoc computing.* ACM, 2005, pp. 1–8.

[5] O. Salem, Y. Liu, and A. Mehaoua, “Anomaly detection in medical wireless sensor networks,” *Journal of Computing Science and Engineering*, vol. 7, no. 4, pp. 272–284, 2013.

[6] S. Jarupadung, “Distributed event detection and semantic event processing,” in *The 6th ACM International Conference on Distributed Event-Based Systems (DEBS 2012)(Doctoral Symposium)*, 2012.

[7] J. R. Quinlan, “Induction of decision trees,” *Machine learning*, vol. 1, no. 1, pp. 81–106, 1986.

[8] G. J. Pottie and W. J. Kaiser, "Wireless integrated network sensors," *Communications of the ACM*, vol. 43, no. 5, pp. 51–58, 2000.

[9] Y. Zhang, N. Meratnia, and P. Havinga, "Outlier detection techniques for wireless sensor networks: A survey," *Communications Surveys & Tutorials, IEEE*, vol. 12, no. 2, pp. 159–170, 2010.

[10] D. Margineantu, W.-K. Wong, and D. Dash, "Machine learning algorithms for event detection," *Machine Learning*, vol. 79, no. 3, pp. 257–259, 2010.

[11] A. W. Chickering and Z. F. Gamson, "Development and adaptations of the seven principles for good practice in undergraduate education," *New directions for teaching and learning*, vol. 1999, no. 80, pp. 75–81, 1999.

[12] C. T. Vu, R. A. Beyah, and Y. Li, "Composite event detection in wireless sensor networks," in *Performance, Computing, and Communications Conference, 2007. IPCCC 2007. IEEE Internationa.* IEEE, 2007, pp. 264–271.

[13] M. Angeles Serna, A. Bermudez, and R. Casado, "Circle-based approximation to forest fires with distributed wireless sensor networks," in *Wireless Communications and Networking Conference (WCNC), 2013 IEEE.* IEEE, 2013, pp. 4329–4334.

[14] P. Domingos, "A few useful things to know about machine learning," *Communications of the ACM*, vol. 55, no. 10, pp. 78–87, 2012.

[15] M. Aurangabad, "Classification of lung tumor using svm," *Editorial Board*, p. 1254.

[16] A. Föerster and A. L. Murphy, "Machine learning across the wsn layers," 2010.

[17] M. Di and E. M. Joo, "A survey of machine learning in wireless sensor netoworks from networking and application perspectives," in *Information, Communications & Signal Processing, 2007 6th International Conference on.* IEEE, 2007, pp. 1–5.

[18] L. Kotthoff, I. P. Gent, and I. Miguel, "An evaluation of machine learning in algorithm selection for search problems," *AI Communications*, vol. 25, no. 3, pp. 257–270, 2012.

[19] R. Caruana and A. Niculescu-Mizil, "An empirical comparison of supervised learning algorithms," in *Proceedings of the 23rd international conference on Machine learning.* ACM, 2006, pp. 161–168.

[20] C. C. Bonwell and J. A. Eison, *Active Learning: Creating Excitement in the Classroom. 1991 ASHE-ERIC Higher Education Reports.* ERIC, 1991.

[21] G. Wittenburg, N. Dziengel, C. Wartenburger, and J. Schiller, "A system for distributed event detection in wireless sensor networks," in *Proceedings of the 9th ACM/IEEE International Conference on Information Processing in Sensor Networks.* ACM, 2010, pp. 94–104.

[22] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," *IEEE transactions on systems, man, and cybernetics*, vol. 21, no. 3, pp. 660–674, 1991.

[23] L. Schoen and L. D. Fusarelli, "Innovation, nclb, and the fear factor the challenge of leading 21st-century schools in an era of accountability," *Educational Policy*, vol. 22, no. 1, pp. 181–203, 2008.

[24] L. Rucco, A. Bonarini, C. Brandolese, and W. Fornaciari, "A bird's eye view on reinforcement learning approaches for power management in wsns," in *Wireless and Mobile Networking Conference (WMNC), 2013 6th Joint IFIP.* IEEE, 2013, pp. 1–8.

[25] M. Sa and A. K. Rath, "A simple agent based model for detecting abnormal event patterns in distributed wireless sensor networks," in *Proceedings of the 2011 International Conference on Communication, Computing & Security.* ACM, 2011, pp. 67–70.

[26] S. Ortmann, M. Maaser, and P. Langendoerfer, "Adaptive pruning of event decision trees for energy efficient collaboration in event-driven wsn," in *Mobile and Ubiquitous Systems: Networking & Services, MobiQuitous, 2009. MobiQuitous' 09. 6th Annual International.* IEEE, 2009, pp. 1–11.

[27] Y. Li, Y. Wang, and G. He, "Clustering-based distributed support vector machine in wireless sensor networks," *Journal of Information & Computational Science*, vol. 9, no. 4, pp. 1083–1096, 2012.

[28] F. Silva, T. Olivares, F. Royo, M. Vergara, and C. Analide, "Experimental study of the stress level at the workplace using an smart testbed of wireless sensor networks and ambient intelligence techniques," in *Natural and Artificial Computation in Engineering and Medical Applications.* Springer, 2013, pp. 200–209.

[29] Y. Singh, S. Saha, U. Chugh, and C. Gupta, "Distributed event detection in wireless sensor networks for forest fires," in *Computer Modelling and Simulation (UKSim), 2013 UKSim 15th International Conference on.* IEEE, 2013, pp. 634–639.

[30] P. Bolourchi and S. Uysal, "Forest fire detection in wireless sensor network using fuzzy logic," in *Computational Intelligence, Communication Systems and Networks (CICSyN), 2013 Fifth International Conference on.* IEEE, 2013, pp. 83–87.

[31] S. Yoon, W. Ye, J. Heidemann, B. Littlefield, and C. Shahabi, "Swats: Wireless sensor networks for steamflood and waterflood pipeline monitoring," *Network, IEEE*, vol. 25, no. 1, pp. 50–56, 2011.

[32] X. Zhu, "Semi-supervised learning literature survey," *Computer Science, University of Wisconsin-Madison*, vol. 2, p. 3, 2006.

[33] M. A. Rassam, M. Maarof, and A. Zainal, "A survey of intrusion detection schemes in wireless sensor networks." *American Journal of Applied Sciences*, vol. 9, no. 10, 2012.

[34] S. Mal-Sarkar, I. U. Sikder, and V. K. Konangi, "Application of wireless sensor networks in forest fire detection under uncertainty," in *Computer and Information Technology (ICCIT), 2010 13th International Conference on.* IEEE, 2010, pp. 193–197.

[35] M. Bahrepour, N. Meratnia, M. Poel, Z. Taghikhaki, and P. J. Havinga, "Distributed event detection in wireless sensor networks for disaster management," in *Intelligent Networking and Collaborative Systems (INCOS), 2010 2nd International Conference on.* IEEE, 2010, pp. 507–512.

[36] Y. Li and L. E. Parker, "Detecting and monitoring time-related abnormal events using a wireless sensor network and mobile robot," in *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on.* IEEE, 2008, pp. 3292–3298.

[37] M. A. Rassam, A. Zainal, and M. A. Maarof, "Advancements of data anomaly detection research in wireless sensor networks: A survey and open issues," *Sensors*, vol. 13, no. 8, pp. 10 087–10 122, 2013.

[38] P. Cortez and A. d. J. R. Morais, "A data mining approach to predict forest fires using meteorological data," 2007.

[39] L. Yu, N. Wang, and X. Meng, "Real-time forest fire detection with wireless sensor networks," in *Wireless Communications, Networking and Mobile Computing, 2005. Proceedings. 2005 International Conference on*, vol. 2. IEEE, 2005, pp. 1214–1217.

[40] N. Dziengel, M. Ziegert, M. Seiffert, J. Schiller, and G. Wittenburg, "Integration of distributed event detection in wireless motion-based training devices,"

in *Consumer Electronics-Berlin (ICCE-Berlin), 2011 IEEE International Conference on.* IEEE, 2011, pp. 259–263.

[41] N. Dziengel, G. Wittenburg, and J. Schiller, "Towards distributed event detection in wireless sensor networks," in *Adjunct Proc. of 4th IEEE/ACM Intl. Conf. on Distributed Computing in Sensor Systems (DCOSS'08), Santorini Island, Greece*, 2008.

[42] N. R. Draper and H. Smith, *Applied Regression Analysis (Wiley Series in Probability and Statistics).* Wiley-Interscience, 1998.

LIST OF PUBLICATION

1) Aditi Kansal, Yashwant Singh, “Survey on Machine Learning Techniques for Event Detection in Wireless Sensor Networks” in International Journal of Modern Computer Science (IJMCS) Volume 2, Issue No1, 2014, pp. 2320-7868.

2)Aditi Kansal, Yashwant Singh, “Accurate Detection of Forest Fires using Machine Learning Technique” (Communicated in Third International Conference on Advancing in Computing, Communications and Informatics, IEEE).