IJCSN International Journal of Computer Science and Network, Volume 4, Issue 6, December 2015
ISSN (Online) : 2277-5420 www.IJCSN.org
Impact Factor: 0.417

Internet-Sensor Information Mining Using Machine Learning Approach

[1] Shreya P. Amilkanthwar , [2] Poonam Railkar , [3] P. N. Mahalle

[1, 2, 3] Department of Computer Engineering, Savitribai Phule Pune University,
Pune, Maharashtra 411041, India

Abstract - Sensor networks are composed of multiple tiny, low power, low cost sensor nodes which are capable to collect data from environment i.e. pressure, temperature , weather , thermal etc and collaborate to forward it to their centralized backend such as sink or base station for further processing .There are lots of sensors & clusters of sensor which are connected to internet for the purpose of sharing , Communicating data over internet. In addition to this there are various web logs, sensor logs where big data posted by various users. It requires an efficient mining strategy to manage such a big data generated by sensors, also sensor networks collects data from dynamic environment that change over time. Such a dynamic behavior need machine learning techniques to provide appropriate, selective & useful information to users. This paper presents literature review of machine learning techniques used for mining purpose & proposed a high level algorithm for the unstructured, unsupervised dataset generated by sensor networks.

Keywords - *Sensor Networks, Data Mining, Internet Sensor Information, Machine Learning.*

1. Introduction

Data mining is an analytical process designed to explore the large amount of data typically market/business data in search of patterns[1].Ultimately data mining does prediction which is most common business application. There are 3 stages consist in process of mining: Exploration, Model building, Deployment

Machine Learning is a set of tools which allow us to teach our computer /system how to perform tasks by providing some intelligent methods. Machine learning programs detect the patterns from data & takes actions as per type of patterns. Machine Learning mainly concerned as the discovery of models, patterns and other regularities of data. The process of machine learning can be broken into two phases:

1. Training: A model is learned from large collection of training dataset.
2. Application: The model use to make decisions about some new test data.

Machine learning and data mining often employ the same methods and overlap significantly. They can be distinguished as follows:

- Machine learning focuses on prediction, based on known properties learned from the training data.
- Data mining focuses on the discovery of unknown properties from the data. This is the analysis step of Knowledge Discovery in Databases.

The two areas overlap in many ways: data mining uses many machine learning methods, but often with a slightly different goal in mind. On the other hand, machine learning also employs data mining methods as "unsupervised learning" or as a preprocessing step to improve e learner accuracy. Learning is a holistic activity, It takes place as objective for better decision making process. Learning usually results outputs from experienced person's inputs or one's own experience, and inference based on experience or past learning. Learning is not just knowledge gaining, it's a process of knowledge gain, augmentation & management. A machine learning studies computer algorithm to behave intelligently, to make accurate predictions & reactions on particular situations.

Types of machine learning:

- Supervised machine learning
- Un-supervised machine learning
- Reinforcement machine learning

In Supervised learning where algorithm generates a function that maps inputs to desired output. It falls under classification problem. It is learning based on labeled data. The computer/system is presented with example inputs & their desired outputs, given by a "teacher" i.e. algorithm and goal is to learn a general rule that maps input to output. Labeled examples are used for training in this learning such a set of labeled data are call training set. This type of learning is not only used for the purpose of classification, it is overall process that used for decision making. As depicted in fig 1.1 hyper plane is drawn after learning & separating two classes A & B, Each input is presented with input along their output instance.

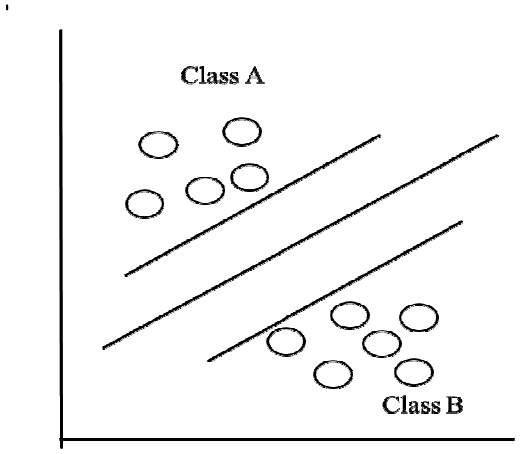

Fig. 1.1 Supervised Learning

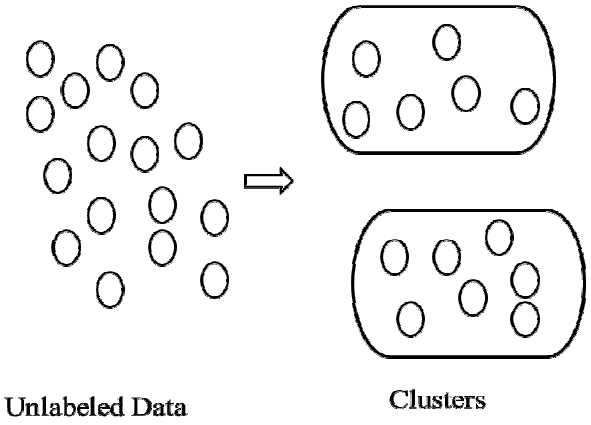

Fig 1.2 Unsupervised learning

In un-supervised learning the goal is to have a system to learn how to do something that "teacher" i.e. algorithm don't tell how should do. AS shown in fig 1.2 this type of learning uses unlabeled data , here unlabeled data clustered on the basis of similarities & differences among data nothing else.

In reinforcement learning where the algorithm learns a policy of how to act, given an observation of the world. Every action has some impact on environment & environment provides feedback that guides learning algorithm. In this system there is a decision maker & environment. A decision maker or learner performs action by taking appropriate decision in association with environment. In such scenario intelligent learning agent is doing a key role of decision making. There is also a set of trial & error runs, from that IS agent chooses best possible action to achieve goal & solves problem. the rest of paper is organized in the following way: Section 2 explains the idea of paper with current trends. Section 3 contains brief about internet sensor networks & information collected over there. In Section 4 we review 9 paper regarding machine learning approaches used in data mining. Section 5 summarizes the gap analysis of 9 papers by comparing them with parameters like approaches used , instance types ,type of algorithm & nature of algorithm. Section 6 gives proposed architecture with algorithm work flow. Finally we conclude with future work.

2. Motivation

As we know until few years ago there are only laptops & computers are capable devices of using internet. Now a day's mobile phones, tablets and televisions some more kind of devices uses internet, so the use of internet grows on day by day. Information or data generated by WSN networks goes on increasing. Sensor networks are used in various applications, in various domains for measuring physical entities & natural event, which allows machines to capture the characteristics of physical objects.

Wireless sensor network monitors dynamic environmental changes which occurs rapidly. Such dynamic behavior may caused by external physical factors or initiated by system designers .To get know such type of dynamic conditions & changes, sensor network needs intelligent machine learning techniques.

There are lots of sensors & clusters of sensors which are connected to internet, which generates big data over internet. As we studied various types of machine learning algorithms, it seems that currently systems using lightweight supervised machine learning approach in internet information mining .The unsupervised approach could be more interesting to be used in internet sensor information mining. Since we know how the IoT's are growing and data formed by such networks are also growing day by day. Such data should get mined in intelligent way so we are going for designing an unsupervised machine learning algorithm.

3. Internet Sensor Information

Sensor network could be wireless or wired. But wireless solutions allow safety guidelines wirelessly connected sensors & unique identification objects via IPV6 addressing will enable new possibilities to make world smarter. E.g. "where is my car key?" such questions can be answered by search engines in future. Sensor information is generated by sensors & actuators, sensors as part of network are usually devices measuring physical parameters such as temperature, pressure, humidity, energy consumption, acceleration, water level which are transforming into electrical signals. Such type of information is called internet sensor information.

Other complex sensors are also possible i.e. (Global Positioning systems) GPS,(Radio Frequency Identification) RFID through to intelligent camera sensors. Once the connectivity of sensor reaches to network (internet) by crossing border of local network, worldwide unique identification & addressing comes into play. Few years ago, the border elements of internet were computers & laptops only. Now it founds various internet capable devices such as mobiles, tablets, navigation systems, televisions etc. WSNs are unusual in real-life systems, which a network of individual smart sensing devices that are discrete sources of data and information, and that transmit over a shared ad-hoc wireless network.

4. Literature Survey

The paper [2] proposed data mining approach to build an RVM classifier for classification of Learning Style(LS) based on Learning object(LO) according to student preference used of LO.RVM is used for classification of learners, It is better in terms of complexity & accuracy as compared with Support vector machine & Neural Network. Relevance vector machine is typically used for solving classification & regression problem. Learning object can be digital, non digital objects ,which can be used ,re-used, referenced during support learning. Learning Style is an inventory in which is 1st level cycle consist 4 way classification i.e. Concrete Experience (CE), Reflective Observation(RO), Abstract Conceptualization (AC), Active Experience (AE). 2nd level cycle consist of 2 way classification: Diverging (CE&RO),Assimilating(RO&AC),Converging(AC&AE), Accommodating(CE&AE).

This paper [3] used machine learning algorithm based on combinatorial optimization, here referred as GRNCOP (Gene Regularity Network inference by Combinatorial Optimization).It is designed for the inference of putative GRNs, also it obtains an optimal classifier that represents potential interaction relationship between genes. Currently it has ability to infer potential regularity rules with one to one cardinality, in which precedent contains only one gene. But phenomenon say comprise relationships described by many to one cardinality.

This paper [4] proposed recommender system to reduce human efforts for performing domain analysis. Here they are used Novel incremental diffusive algorithm to extract features online product description & employ association rules & KNN machine learning methods. In future they focus on mining additional software repositories & directories to broaden the knowledge of current system, improve interaction between user & system as well as enhancement in navigation & visualization of recommended features.

This [5] system proposed the framework with dynamic machine learning model which dynamically generate according to query input. OSML (On-site model for legend binding sites prediction) model is designed. Current dynamic learning framework can perform protein legend binding site prediction for 10 different types on 3 different levels. But can be modified to regularly update OSML by incorporating new legend types by improving prediction capability.

In this [6] study of machine learning approach used for detection of harmful Algal blooms. The study based on spatio -temporal data mining approach. Support vector machine (SVM) used for dynamic machine learning approach which determines harmful Algal blooms. It describes dynamic behavioral change of HABs across space & time. This machine learning based spatio-temporal data mining approach gave a very satisfactory performance improvement over empirical algorithm in HAB detection over large spatio-temporal datasets. This approach proved success in reducing the false alarm rate.

This paper [7] describes a semi-supervised pattern discovery approach that uses the by-products of complete validation studies on experimental setups for gene profiling. Class prediction and feature selection are two learning methodologies that are mostly combined in the search of molecular profiles from microarray data. They originate sample-tracking profiles as aggregated off-training evaluation of SVM models of increasing gene panel sizes. Genes are ranked by E-RFE, an entropy-based variant of the recursive feature elimination for support vector machines (RFE-SVM).A Dynamic Time Warping (DTW) algorithm is then applied to define a metric between sample-tracking profiles. In future it can be automate the existing system by applying unsupervised learning.

In this paper [8] they have discussed various machine learning approaches used in mining of data, which are distinguish between symbolic and sub-symbolic data mining methods. Here they proposed hybrid method with the combination of Artificial Neural Network (ANN) and Cased Based Reasoning (CBR) in mining of data.CBR uses knowledge of past experience when dealing with new cases.

In this paper [9] the authors propose an architecture that makes the real-time big data processing and analysis possible. The proposed architecture is based on two main components: a stream processing engine called Apache Storm and a framework called Yahoo SAMOA (Scalable Advanced Massive Online Analysis) allowing to perform data analysis through distributed streaming machine learning algorithms. The machine learning algorithm used is "vertical Hoeffding Tree" i.e. the distributed version of Hoeffding tree algorithms. The designed architecture is used for Skype traffic recognition within network traffic generated by several Personal Computers in a streamed way. Their future plan is to analyze dataset with unsupervised machine learning approach.

In this paper[10], author propose a novel scheme of learning nonlinear distance functions with side information, which aims to learn a Bregman distance function using a nonparametric approach that is similar to Support Vector Machines. Also they incorporated the Bregman distance function into the k-means clustering algorithm and hierarchical clustering algorithms. But the statistical t-test on both k-means clustering and hierarchical clustering results showed that the proposed distance function outperforms the other regular distance metrics significantly in most of the cases.

The growing prevalence of network attacks is a well-known problem which can impact the availability, Confidentiality, and integrity of critical information for both individuals , businesses and enterprises. Here in this paper [11] they propose supervised machine learning technique for real time intrusion detection.

5. Gap Analysis

Here I compared various types of approaches used in several referred papers, to get know which approach is more efficient to perform efficient mining of internet data. Similarly with approaches we also analyze in which type they fall under. Next on which dataset the mining algorithms are applied and the purpose of using particular algorithm for particular data is also studied by comparing all theses papers. As well as nature of used algorithms is also matters, to verify role of operation.

Following table no. 5.1 shows summarized view and comparison among several referred papers. In such a way we compared all these papers to know which method is more preferable to apply on our dataset in terms of nature, complexity, approach etc.

Table No. 5.1

	Approach	Supervised/ Unsupervised	Instance type	Nature of algorithm
[1]	RVM(Relevance Vector Machine)	Unsupervised	Learning style preferences by students	Sequential
[2]	GRNCOP(Gene Regularity Network inference by Combinatorial Optimization)	Supervised	a Saccharomyces cerevisiae gene expression data set.	
[3]	Association rules & KNN(K-nearest neighbor)	Unsupervised	Enterprise dataset	Linear
[4]	OSML(On-site model for legend binding sites prediction)	Supervised	protein-ligand interaction database	
[5]	Support vector machine (SVM)	Semi-supervised	spatio-temporal dataset	Linear
[6]	recursive feature elimination for support vector machines (RFE-SVM)	Semi-supervised	bioinformatics databases.	Linear
[7]	combination of Artificial Neural Network (ANN) and Cased Based Reasoning (CBR)	Hybrid	Any dataset(e.g. hospital information database)	Distributed
[8]	Vertical Haeoffding Tree	Supervised	Streamed data such as (skype traffic)	Distributed
[9]	Bregman distance function similar to SVM	Semi-supervised	Distance function	Non–linear

6. Proposed Work

In our system we are going to propose our contribution in doing data mining approach at dataset which is formed by various local networks such as access network, ad-hoc network and convergence network. There are lots of sensors & clusters of sensors which are connected to internet. Our main Objective is to enable individuals or cluster of sensors to share, cooperate & communicate data over internet. In addition to this there are various web logs, server logs where big data posted by various users, such logs are referred to as sensor nets, s-logs. New approaches & schemes are required to integrate sensor nets & internet. Objectives are summarizes as follows:

1) To design scheme or protocol which enables simple integration of sensor net & internet .They should be lightweight algorithm which will connect sensor net to internet through wireless network.
2) To design efficient machine learning algorithm to provide appropriate, selective & useful information to user.
3) To design new algorithm to manage sensor data discloser in order to ensure confidentiality of sensor data.

I am going to apply machine learning approach of data mining over there. As we saw there is currently using supervised machine learning approach, which is having predefined set of inputs & relevant assumed outputs. We are going to design an algorithm on unsupervised approach, which would be dynamic in nature. As it does not set for particular set of input , it would be able to respond on time. Which approach would be real time in nature?

As shown in fig. no.6.1. At the top in architecture diagram there are users of internet such as laptop users, desktop users, and mobile users and so on. They requested some queries to service providers which are staying at application layer. Application layer gets data from internet, where data comes from various networks. There are 3 types of networks such as ad-hoc network which consist of WSN, Converge networks such as IOT, CPS, RFID etc. Access network such as 802.11, 802.14 generates data, which I am going to analyze and mine for sending it to internet. Here before sending to internet we are preprocessing on it, we are going to design an algorithm using Machine learning approach. This algorithm works as mediator between internet & dataset. To design an algorithm we have studied various data mining algorithms, various machine learning approaches, and light weight supervised machine learning approach. Such as Support vector machines, Relevance vector machine, decision trees & so on.

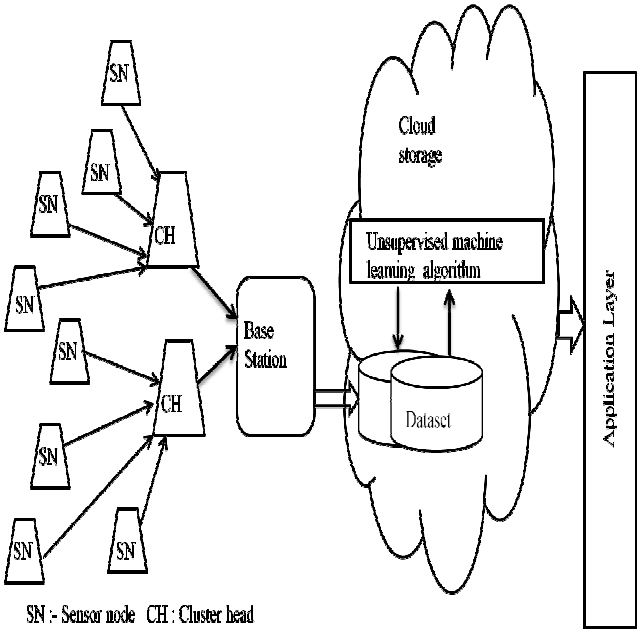

Fig 6.1 Architecture of proposed work

In our system we require to perform machine learning on dataset generated by multiple sensor nodes, a dataset may contains raw data, unsupervised data , or may be supervised & labeled data. So we are going to generate an algorithm which could help to manage datasets & provide required output to the application users. There are various steps as follows required for data processing of sensor data as shown in fig 6.2.

1. let D be Dataset.
2. we require 1st clustering algorithms to be applied on unsupervised data to get similar components in one group.
3. After performing clustering it will be in the form of labeled data, these clusters need classification further.
4. The output of process will give classes of data.

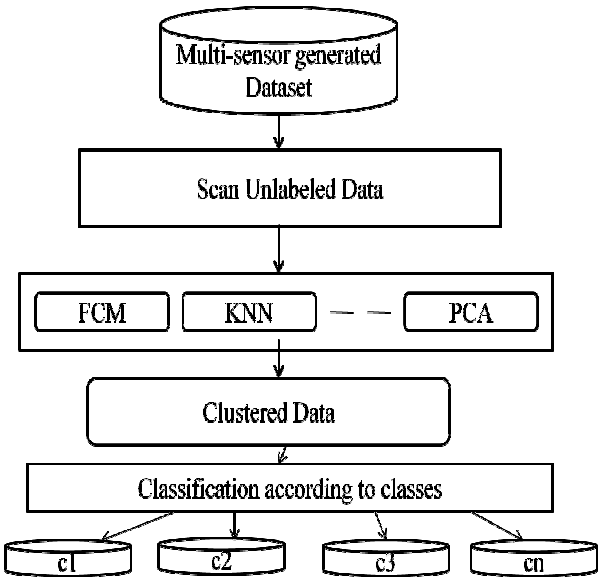

Fig 6.2 Algorithm Workflow

7. Conclusions & Future Work

As our objective is to enable clusters of sensor to share data over internet, and to apply machine learning algorithm for data mining of such data formed by sensors . Here paper presents study of various data mining algorithms as well as machine learning approaches used in data mining. Along with this here studied lightweight supervised machine learning algorithm. As we go through various types' machine learning data mining algorithms, it seems that the unsupervised machine learning approach is more efficient. Since Supervised machine learning algorithms have some limitations, So we are going to design an Un-supervised machine learning algorithm for mining of Internet Sensor information. Our future work is to build a proposed work &to implement it.

References

[1] www.statsoft.com/textbook/data-mining-techniques.

[2] Nor LiyanaMohdShuib, HarunaChiroma, Rukaini Abdullah, Mohammad Hafiz Ismail ,Ahmad SofiyuddinMohdShuib&NurFaizahMohdPahme*"Data Mining Approach: Relevance Vector Machine for the Classification of Learning Style based on Learning Objects"* 2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation.

[3] Ignacio Ponzoni, Francisco J. Azuaje, Juan Carlos Augusto, and David H. Glass "*Inferring Adaptive Regulation Thresholds and Association Rules from Gene Expression Data through Combinatorial Optimization Learning"*

[4] Negar Hariri, Carlos Castro-Herrera, Mehdi Mirakhorli, *Student Member, IEEE*, Jane Cleland-Huang, *Member, IEEE*, BamshadMobasher, *Member, IEEE"Supporting Domain Analysis through Mining and Recommending Features from Online Product Listings"*

[5] Dong-Jun Yu , *Member, IEEE*, Jun Hu, Qian-Mu Li, Zhen-Min Tang, Jing-Yu Yang, and Hong-Bin Shen**"Constructing Query-Driven Dynamic Machine Learning Model With Application to Protein-Legend Binding Sites Prediction"*

[6] BalakrishnaGokaraju, Surya S. Durbha, *Member, IEEE*, Roger L. King, *Senior Member, IEEE*, and Nicolas H. Younan, *Senior Member, IEEE "A Machine Learning Based Spatio-Temporal Data Mining Approach for Detection of Harmful Algal Blooms"*

[7] CesareFurlanello, Maria Serafini, Stefano Merler, and Giuseppe Jurman*"Semisupervised Learning for Molecular Profiling"*

[8] Jyothibellary,BhargaviPeyakunta,SekharKonetigari*"Hybrid Machine Learning Approach In Data Mining"*

[9] Mario Di Mauro, Cesario Di Sarno*"A framework for Internet data real-time processing:a machine-learning approach"*

[10] Lei Wu, Steven C.H. Hoi, Member, IEEE, Rong Jin, Jianke Zhu, and Nenghai Yu*"Learning Bregman Distance Functions for Semi-Supervised Clustering"*

[11] Phurivit Sangkatsaneea , Naruemon Wattanapongsakorn a , ChalermpolCharnsripinyob *"Practical real-time intrusion detection using machine learning approaches"*