International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-2 Number-4 Issue-6 December-2012

A Study of Data Mining Techniques for WSN Based Intellectual Climate System

Pravin C. Satpute[1], Deepti P. Theng[2]
M. E. Scholar, Department of Computer Science and Engineering, GHRCE, Nagpur, India[1]
Department of Computer Science and Engineering, GHRCE, Nagpur, India[2]

Abstract

In the Industrial environment there are various technical parameters which have to be maintained. If it is not maintained in the range then it will lead a large catastrophe. So, we need to maintain the climate. There is some parameter which is important like temperature, humidity, air flow etc. But, due to huge variation in the environment these parameters are vary continuously. If we analyze the variation we need to discover the meaning full data which will help us to generate the pattern and rule. There is large quantity of data received from sensor and stored in server. So, the data mining technique need to apply to get meaning full data and rule.

Keywords

Data mining, clustering: Hierarchical Clustering Algorithm, K-means Clustering Algorithm.

1. Introduction

Ambient Intellectualization is a vision where environment becomes intelligent, friendly, context-aware and responsive to any type of human requirements. In such a world, networking and computing technology coexist with people in a universal, friendly and pervasive way. Numerous miniature and interconnected intelligent devices create a new intelligence and interact with each other seamlessly.

In the industries, we know that there is much hardware equipment which is working in a proper environment. If the environment is not proper then there might be possible to lead causing important damages or in worth case human deaths. Temperature as well as humidity, air flow these are some important environment parameter which is needed monitoring and controlling time to time.

In the context of the industrial applications, the scope of this paper covers industrial processes such as pulp & paper, and petrochemical operations, with applications geared mainly towards process monitoring and control, process parameter value inference, detection of abnormal situation and their diagnostic and a general improvement of the process understanding through discovery of correlations between processes monitor [2-4].

Intellectual climate system can monitors all the environmental parameter but there is a variation among the data so, there are huge amount of data which is useless. For meaningful data extraction clustering is needed and for that different data mining techniques have to be applied which is used to discover the meaning full data to generate the patterns, correlations and changes in the data. There are some algorithms which are used for data mining. [5-7] this paper studied some algorithm for proposed applications for finding out the good one algorithm to get desired result and improve the performance of the applications.

In the WSN based intellectual climate system we are collecting the technical data from different nodes which are distributed in the different locations. The data collecting from different nodes are varied continuously whether it is temperature data or humidity data or anything else, some data which is useless but we need use full data from coming data from different nodes then we have to apply the data mining technique to get appropriate data or for generating the patterns meaningful data can be discovered or making some rules to maintain the climate for industries. In this paper use ZigBee protocol for transmission and receiving the data from different nodes. The most research work is to monitoring the application and the task is to implement the algorithms for different modules for collecting and monitoring the huge amount of data. [8-9]

2. Literature Review

A.

Apriori Algorithm:

In 2010, Yanxi Liu [10] proposed Apriori Algorithm which is an innovative way to find association rules on large scale, allowing presumption outcomes that consist of more than one item. Apriori is a crucial algorithm for finding frequent item sets using candidate generation. It is characterized as a level wise and complete search algorithm using anti-monotonicity of item sets. There are many Algorithms which is used to finding patterns such as decision tree, classification rules and clustering

techniques that are frequently used in data mining have been developed in industrial climate data research community.

In the Apriori algorithms have some drawbacks which are as follows,

- Needs several iterations of the data.
- Uses a unique minimum threshold support.
- Rarely occurring events are hard to find out.
- Optional methods rather than apriori can point out by using a non-unique minimum support threshold.
- There is comparatively approaches focus on partitioning and sample ling.

EM Algorithm:

In 1996, Cristophe Couvreur[11] proposed the Expectation- Maximization algorithm is one of the methods for data mining which is the choice for maximum-likelihood estimation. Due to asymptotic optional properties of Expectation-maximization, maximization-likelihood has become one of the preferred methods of estimation in different areas of application of statistics including pattern recognition and many others. The EM Algorithm is simple and versatile procedure for likelihood maximization in incomplete data problems. The EM Algorithm is easy to implement, numerically it is very stable and requirement of memory is less.

In EM Algorithm, There are also some drawbacks which are as follows

- The main drawback of EM Algorithm is its hopelessly slow convergence in some cases.
- Forward and backward probabilities have required.
- Significant implementation effort required compared to numerical optimization.
- Convergence may be slow if analytical expression for the M-step is not available since numerical optimization must be applied.
- Hessian must be calculated manually.

Hierarchical Clustering:

In 2012, Manish Verma, Mauly Srivastava, et al [13] proposed Hierarchical Clustering method is one of the methods of cluster analysis which builds a cluster hierarchy. Child clusters available in every node. Hierarchical Clustering are of two types basically. No need to specify the number of clusters in advance. Generate smaller clusters which may be helpful for discovered the meaningful data.

Agglomerative: This is bottom up approach. Each and every observation starts from its own cluster and pairing of clusters are merged as one step moves up to the hierarchy. The complexity of agglomerative clustering is O (n3), which makes them too slow for large data sets.

Divisive: This is a top down approach, in this approach all the observations starts from any one cluster and split up which are performed recursively as one step moves down the hierarchy. Divisive clustering with an exhaustive search is O (n2) which is even worse.

Drawbacks of Hierarchical Clustering Algorithm are as follows,

- Objects may be incorrectly grouped at an early stage. The output should be examined to ensure it makes sense.
- Using different metrics for measuring distances between clusters may generate different results.
- Interpretation of output is an abstract.

K-Means Clustering Algorithm:

In 2012, Abhay Kumar, Ramnish Sinha,et al [14] proposed K-means algorithm in which predicting the likely behavior from observed behavior would be entirely legitimate if the relationship were found in the data. The most common data mining techniques for finding hidden patterns in data are clustering and classification analysis.

K-Means Clustering Algorithm is a data mining or machine learning algorithm used to cluster observations into related observations groups without any proper knowledge of those relationships. The K-Means Algorithm is one of the easiest and very important clustering techniques which are commonly used in changing field.

The K-means approach to clustering starts out with a fixed number of clusters and allocates all records into exactly the number of clusters. Another class of methods works by agglomeration. This method start out with each data point forming its own cluster and gradually merge them into larger and larger clusters until all points have been gathered together into big cluster [15].

The main Moto behind the data analysis is to discover the meaningful data. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters. The main idea is to define k centers, one for each point belonging to available data set associate to nearer center. When no point is pending, the first step is completed and an early group age is done. At this point we need to re-compute k new centroids as barycentrer of the clusters resulting from the previous step. After getting these new centroids, A new data is to be done between the same data set points and the nearest new center. We notice the generated loop that the k center change their location step by step until no more changes are done or in other words

centers do not move any more. Overall this algorithm aims at minimizing an objective function knows as squared error function given by:

J (v) =$\sum_{i=1}^{c}\sum_{j=1}^{ci}(||x_i - v_i||)^2$

'*Ci*' is the number of data points in i^{th} cluster.

'*C*' is the number of cluster centers.

STEPS FOR K-MEANS CLUSTERING ALGORITHM

Let X={x1,x2,x3,……,xn} be the set of data points and N={v1,v2,…….,vc} be the set of centers.

Randomly select 'c' cluster centers.

Compute the spaces between each data point and cluster centers.

Assign the data point to the center of the cluster whose distance from the center of the cluster is minimum of all the cluster centers.

Re-compute the new center of the cluster using:

Vi= (1/Ci)$\sum_{j=1}^{ci} x_i$

Where 'Ci' represents the number of data points in ith cluster.

Recomputed the distance between each data point and new obtained cluster centers.

If no data point was elevated then stop, otherwise repeat from step 3.

K-means is strongly related to fitting a mixture of k isotropic Gaussians to the data. Generally the measurement of all the distance to all Bergman divergences is related to fitting the data with a mixture of k components from the exponential family of the distributed area.

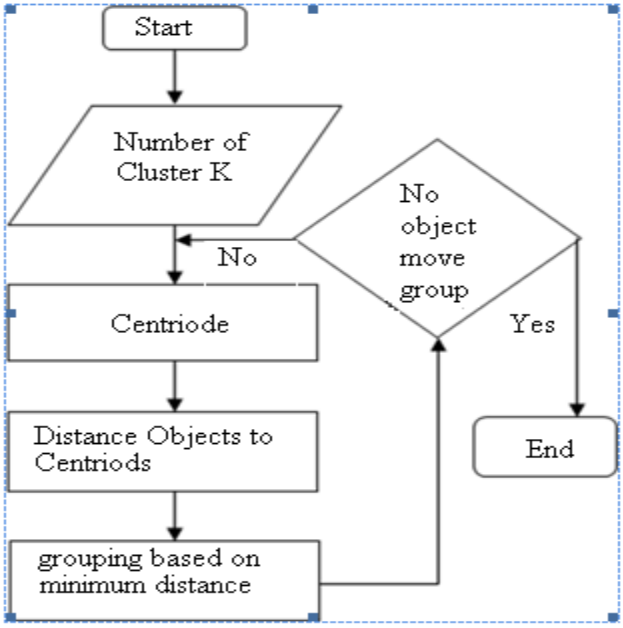


Fig.1: Flowchart for K-Means Algorithm Process

Advantages of K-Means Algorithm:

- It is relatively efficient.
- Fast, Robust and easy to understand.
- When data set are distinct or well separated from each other then it will produce best result.
- K-Means may produce strongest clusters than hierarchical clustering especially if the clusters are globular.

3. Conclusion

In the context of data mining the main objective is to make discoveries from the available data. We analyzing different algorithms and we studied different factors and situation, we can get following conclusion are as follows:

- The number of cluster improves the Performance of K-Means algorithm.
- The response of K-Means algorithm is better than other algorithm.
- When clustered the data, all the algorithms have some ambiguity in data.
- When we are working on huge dataset then K-Means algorithm is faster than other clustering and it will produces quality clusters.

With the help of K-means algorithm we will discover the meaningful data and according to that we can adjusted the climate parameters value to avoid any threats.

It will also help to find out the conductivity of the product which is manufactured in the industries.

So, we can choose the K-means clustering algorithm on industrial applications which is used to discover the meaningful data which help manufacturer to maintain the climate.

It also Implement an alert system which consists of SMS in order to send the information to the concern person of the industries or company about the different situations in the industries.

References

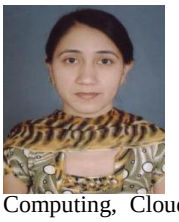
[1] M. Popa and H. Ciocarlie, "Distributed Intelligent Climate System for Indoor Locations", 2011.

[2] M. Popa, A. S. Popa, et al," Remote Temperature Monitoring and Regulating System for Indoor Locations", 2009.

[3] Eliza Mazmee Mazlan, Shazana Rahman et al," Development of a Manufacturing Industry Success Rate Analyzer Using Data Mining Technique", 2010.

[4] Rana A baalkhail, Mauricio Orozco,et al,"Home Energy saving for heating/cooling system by distributed intelligent energy controller",2012.
[5] Parna Khot, Ashok K. Krishnamurthy et al," A Parallel Data Mining Toolbox Using MatlabMPI", 2006.
[6] Kollukuduru Sravanthi, Dr.K.S.Rajan," Spatio-Temporal Mining of Core Regions: Study of rainfall patterns in Monsoonal India", 2011.
[7] Li Wang," Application of Data Warehouse Technology in Digital Mine Information System", 2011.
[8] J.A. Ferre, A. Pawlowski, et al," A Wireless Sensor Network for Greenhouse Climate Monitoring", 2010.
[9] Hero Modares, Rosli Salleh, et al,"Overview of security issies in wireless sensor networks", 2011.
[10] Yanxi Liu," Study on Application of Apriori Algorithm in Data Mining", 2010.
[11] Cristophe Couvreur,"The EM Algorithm: A Guided Tour", 1996.
[12] Narendra Sharma, Aman BajpaiRatnesh Litoria,"Comparison the various algorithm of weka tools" ISSN 2250-2459, Volume 2, Issue 5, May 2012.
[13] Manish Verma, Mauly Srivastava, et al. "A comparative study of various Clustering Algorithms in data mining", IJERA, Vol. 2, Issue 3, May-Jun 2012.
[14] Abhay Kumar, Ramnish Sinha,et al," Modeling using K-Means Clustering Algorithm",2012.
[15] Ruhaizan Ismail, Zalinda Othman and Azuraliza Abu Bakar," Associative Prediction Model and Clustering for Product Forecast Data", 2010.

Mr. Pravin C. Satpute received the Batchelor of Engineering degree in Computer Technology from Yashwantrao chawhan college of engineering, Nagpur, India in 2008.Presently he is pursuing his PG(M.E.) in Embedded System and Computing from G.H. Raisoni college of engineering, Nagpur, India. His research interest includes data mining and embedded system. He is having 3 years of teaching and 7 years of industrial experience.



Ms. Deepti Theng received Master of Technology from G.H.Raisoni College of Engineering, Nagpur, India and currently working as an Assistant Professor in the Department of Computer Science and Engineering at GHRCE, Nagpur. Her area of interest includes Distributed Computing, Cloud Computing, Computer Architecture and Parallel Processing.