CLUSTERING AND NOISE DETECTION FOR GEOGRAPHIC KNOWLEDGE DISCOVERY

Sneha N S[1] and Pushpa[2]
[1]PG Student, Department of Computer Science and Engineering,
Adichunchanagiri Institute of Technology, Chikmagalur, Karnataka, India
[2]Associate Professor & Head, Department of Computer Science and Engineering,
Adichunchanagiri Institute of Technology, Chikmagalur, Karnataka, India

ABSTRACT
Ample amount of geographic data has been collected with modern data acquisition techniques such as a Global Positioning System, high resolution remote sensing and internet based volunteered geographic information. Spatial datasets are large in size, multidimensional and have high complexity measures. To address these challenges Spatial Data Mining (SDM) for Geographic Knowledge Discovery (GKD) are the emerging fields for extraction of useful information and knowledge mining for many applications. This paper addresses the clustering and noise detection technique for spatial data. We considered multidimensional spatial data to provide feasible environment to place sensitive devices in a laboratory by using the data collected from the sensors. Various sensors were used to collect the spatial and temporal data. The GDBSCAN algorithm is used for clustering, which relies on density based notation of clustering and is designed to discover clusters of arbitrary shape and distinguish noise. The proposed work reduces the computation cost and increase the performance.

KEYWORDS: *Spatial Data, Temporal Data, Spatial Clustering*

I. INTRODUCTION

Due to the development of information technology, a vast volume of data is accumulated on many fields. Since automated methods for filtering/analyzing the data and also explaining the results are required, a variety of data mining techniques finding new knowledge by discovering hidden rules from vast amount of data are developed. In the field of geography, due to the development of technology for remote sensing, monitoring, geographical information systems, and global positioning systems, a vast volume of spatial data is accumulated. An automated discovery of spatial knowledge is required because of the fast expansion of spatial data and extensive use of spatial databases. Nowadays, the spatial data mining turn out to be more eminent and stimulating for the reason that abundant spatial data have been stored in spatial databases. The mining of meaningful patterns from spatial datasets is more knotty than mining the analogous patterns from conservative numeric and categorical data, due to the difficulty of spatial data types, spatial relationships and spatial autocorrelation. In various applications, spatial patterns have excessive demand. Since the spatial data has its own characteristics different from the non-spatial data, direct using of general data mining techniques incurs many difficulties .So there have been many studies of spatial data mining techniques considering the characteristics of the spatial data [1].

Spatial data are the data related to objects that occupy space. A spatial database stores spatial objects represented by spatial data types and spatial relationships among such objects. Spatial data carries topological and/or distance information and it is often organized by spatial indexing structures and accessed by spatial access methods. These distinct features of a spatial database pose challenges and bring opportunities for mining information from spatial data. Spatial Data mining or knowledge discovery in spatial database refers to the extraction of implicit knowledge, spatial relations, or other

patterns not explicitly stored in spatial databases. Spatial data mining refers to the extraction of knowledge, spatial relationships, or other interesting patterns not explicitly stored in spatial data sets. It is expected to have wide applications in geographic information systems, geo-marketing, remote sensing, image database exploration, medical imaging, navigation, traffic control, environmental studies, and many other application areas where spatial data are used. A crucial challenge to spatial data mining is the exploration of efficient spatial data mining techniques due to the huge amount of spatial data, and the complexity of spatial data types and spatial access methods. The extra features that distinguish spatial data from other forms of data are spatial co-ordinates, topological distance, and direction information. By inclusion of many features, query language has become too complicated. In contrast to the mining in relational databases, spatial data mining algorithms need to consider the objects that are near-by in order to extract useful knowledge as there is influence of one object on the neighboring object.

In Data analysis, Cluster analysis is very frequently used, which organizes a set of data items into groups (or clusters) so that items in the same group are similar to each other and different from those in other groups. Clustering methods can be broadly classified into Five groups they are Partitioning algorithms, Density based clustering, Hierarchical Algorithms, Grid-Based Methods and Model-Based Clustering Methods. Example algorithms of the above classification are K-Means, K-medoids, Density-based spatial clustering of applications with noise (DBSCAN) and Generalized Density-based spatial clustering of applications with noise (GDBSCAN), Chameleon. To consider spatial information in clustering, three types of clustering analysis are existing; they are spatial clustering, regionalization, and point pattern analysis. In this work only Density-Based clustering methods are considered.

In section 2, we discuss the related literature with respect to various clustering algorithms for geographic knowledge discovery using spatial data mining. The section 3 discusses data collection, GDBSCAN algorithm, spatial clustering and noise detection methods. Section 4 presents the results related to clustering and noise detection system.

II. LITERATURE SURVEY

N.Santhosh Kumar, V. Sitha Ramulu, K.Sudheer Reddy, Suresh Kotha, Mohan Kumar [2], presented how spatial data mining is achieved using clustering. Spatial data is a highly demanding field because huge amounts of spatial data have been collected in various applications, ranging from remote sensing, to geographical information systems (GIS), computer cartography, environmental assessment and planning, etc. Spatial data mining tasks include: spatial classification, spatial association rule mining, spatial clustering, characteristic rules, discriminant rules, trend detection. Cluster analysis groups objects (observations, events) based on the information found in the data describing the objects or their relationships.

All the members of the cluster have similar features. Members belong to different clusters has dissimilar features. Several clustering methods for spatial data mining include; Partitioning Around Medoid (PAM), Clustering LARge Applications(CLARA), Clustering LARge Applications based upon RANdomized Search (CLARANS), Spatial Dominant approach SD(CLARANS), Non Spatial Dominant approach NSD(CLARANS).

Ester M., Kriegel H.-P., Sander J. and Xu X.[3] in their paper provided a Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. They presented the clustering algorithm DBSCAN which relies on a density-based notion of clusters. It requires only one input parameter and supports the user in determining an appropriate value for it. They also performed a performance evaluation on synthetic data and on real data of the SEQUOIA 2000 benchmark. The results of these experiments demonstrated that DBSCAN is significantly more effective in discovering clusters of arbitrary shape than the well-known algorithm CLARANS. Furthermore, the experiments have shown that DBSCAN outperforms CLARANS by a factor of at least 100 in terms of efficiency.

Ng R.T., and Han J.[4] developed and Efficient and Effective Clustering Methods for Spatial Data Mining. They developed a new clustering method called CLAHANS which is based on randomized search. We also develop two spatial data mining algorithms that use CLAHANS. Their analysis and experiments show that with the assistance of CLAHANS, these two algorithms are very effective and can lead to discoveries that are difficult to find with current spatial data mining algorithms.

Furthermore, experiments conducted to compare the performance of CLAHANS with that of existing clustering methods show that CLAHANS is the most efficient.

III. METHODOLOGY

3.1. Geographic Knowledge Discovery System

Geographic knowledge discovery (GKD) is the process of extracting information and knowledge from massive geo-referenced databases. Spatial objects by definition are embedded in a continuous space that serves as a measurement framework for all other attributes. This framework generates a wide spectrum of implicit distance, directional, and topological relationships, particularly if the objects are greater than one dimension. Figure 1 gives the System Architecture of Spatial Data Mining System for Geographic Knowledge Discovery (GKD). The architecture is divided into three parts; they are, the Data Collection from various databases, the Processing stage which consists of spatial clustering method and noise detection and the analysis phase where the discovered patterns are analysed for Equipment feasibility.

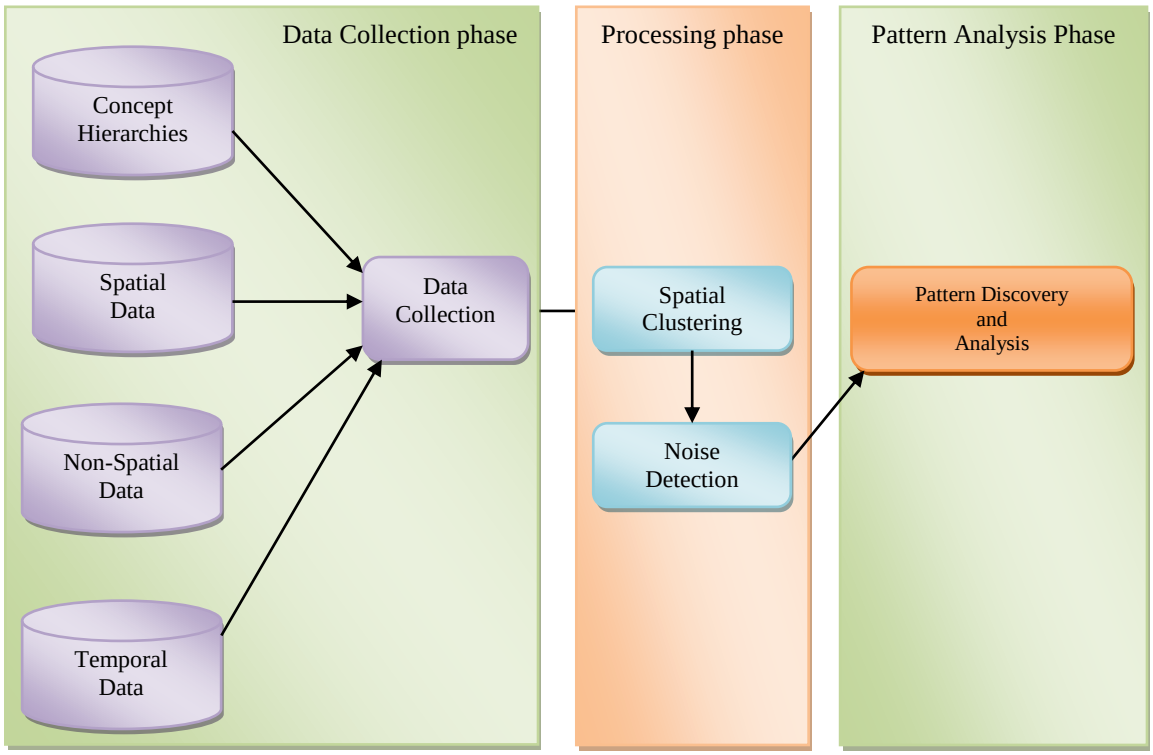

Figure 1: System Architecture of GKD

3.2 Data Collection

Spatial data has positional and topological data that do not exist in general data, and its structure is different according to the kinds of spatial data. Temporal data are the data which explicitly refer to time. The spatial data consisting of Temperature, Light, Humidity, Voltage, Location information,the Temporal data consists of date, the non-spatial data consists of Sensor ID are collected.The spatial dataset used in this work consists of multidimensional data of size 145.133 MB. The dataset includes 23,03,450 Records. The Table 1 gives ten sample records of the spatial data used in the proposed work. These data are stored in the database in a well-defined format in form of table. The database consists of set of similar tables to store data pertaining to location of the sensors and the spatial data specifications of the equipments to be placed.

The attributes used for storing spatial data as shown in Table 1 are Date, Sens ID, Temp, Humid, Light,Volt. Date gives the date on which these data are entered, Sens ID, gives the ID of the sensors for which the data is entered, Temp, consists the temperature of the particular sensor defined by the

sensor ID. Similary Humid, Light, Volt gives the humidity, light and voltage information of the sensors selected. All these data are collected and stored.

Table 1: Spatial data Table of SDM for GKD

Date	Sens ID	Temp	Humid	Light	Volt
18/04/14	S1	10.25	20.12	50.15	1.256
18/04/14	S18	24.26	14.26	78.24	6.456
18/04/14	S19	41.25	31.29	136.1	5.698
18/04/14	S2	8.25	16.25	36.24	2.020
18/04/14	S20	30.24	10.24	99.12	5.240
18/04/14	S21	45.78	35.65	187.2	4.250
18/04/14	S22	49.25	39.45	146.2	3.450
18/04/14	S25	14.36	26.24	151.4	2.456
18/04/14	S28	47.54	36.24	100.5	4.500
18/04/14	S30	34.25	61.24	200.1	1.458

3.3. Spatial Clustering and Noise Detection

Spatial Clustering is interpreted as the task of collecting the objects of a spatial database into meaningful detectable subclasses (i.e. clusters) so that the members of a cluster are as similar as possible whereas the members of different clusters differ as much as possible from each other.

3.3.1 Density Based Spatial Clustering

Density based algorithms typically regard clusters as dense regions of objects in the data space that are separated by regions of low density. The main idea of density-based approach is to find regions of high density and low density, with high-density regions being separated from low-density regions. These approaches can make it easy to discover arbitrary clusters. A common way is to divide the high-dimensional space into density-based grid units. Units containing relatively high densities are the cluster centers and the boundaries between clusters fall in the regions of low-density units. In density-based clustering, clusters are defined as areas of higher density than the remainder of the data set. Objects in these sparse areas - that are required to separate clusters - are usually considered to be noise and border points.

The most popular density based clustering method is DBSCAN. In contrast to many newer methods, it features a well-defined cluster model called "density-reachability". Similar to linkage based clustering; it is based on connecting points within certain distance thresholds. However, it only connects points that satisfy a density criterion, in the original variant defined as a minimum number of other objects within this radius. A cluster consists of all density-connected objects (which can form a cluster of an arbitrary shape, in contrast to many other methods) plus all objects that are within these objects' range. Another interesting property of DBSCAN is that its complexity is fairly low - it requires a linear number of range queries on the database - and that it will discover essentially the same results (it is deterministic for core and noise points, but not for border points) in each run, therefore there is no need to run it multiple times. The key idea of density-based clustering is that for each object of a cluster the neighborhood of a given radius Ɛ has to contain atleast a minimum number of μ objects, i.e. the cardinality of the neighborhood have to exceed a given threshold.

3.3.2 Generalized Density-Based Spatial Clustering of Applications with Noise (GDBSCAN)

The clustering algorithm DBSCAN relies on a density-based notion of clusters and is designed to discover clusters of arbitrary shape as well as to distinguish noise. In this work generalized version of this algorithm is used. The generalized algorithm - called GDBSCAN - can cluster point objects as well as spatially extended objects according to both, their spatial and their non-spatial attributes.

GDBSCAN algorithm is based on center-based approach. In the center-based approach, density is estimated for a particular point in the dataset by counting the number of points within a specified radius, eps of that point. This includes the point itself. The center-based approach to density allows us to classify a point as a core point, a noise or border point. A point is core point if the number of points within eps, a user-specified parameter, exceeds a certain threshold, $MinPts$, which is also a user-specified parameter taken as 3 in this work. Any two core points that are close enough within a

distance *eps* of one another are put in the same cluster. Likewise, any border point that is close enough to a core point is put in the same cluster as the core point. Noise points are discarded.
To find a density-connected set, GDBSCAN starts with an arbitrary object p and retrieves all objects density-reachable from p with respect to Eps and MinPts. If p is a core object, this procedure yields a density-connected set with respect to Eps and MinPts. If p is not a core object, no objects are density-reachable from p and p is assigned to NOISE. This procedure is iteratively applied to each object p which has not yet been classified. Thus, clusters are formed and the noises are detected.
The Algorithm of the function is[5],

```
GDBSCAN (SetofPoints, Eps, MinPts,)
// SetofPoints is UNCLASSIFIED
   ClusterId:= 1;
FOR i FROM 1 TO SetofPoints.size DO
   Sens: = SetofPoints.get (i);
  IF Sens.ClId = UNCLASSIFIED THEN
    IF ExpandCluster (SetofPoints, Sens, ClusterId,Eps, MinPts) THEN
          ClusterId: =nextId (ClusterId)
     END IF
  END IF
END FOR
END; // GDBSCAN
```

SetofPoints is either the whole database or a discovered cluster from a previous run. Eps and MinPts are the global density parameters whose parameters are considered as 3 and 115.0 respectively. ClusterIds are from an ordered and countable datatype where UNCLASSIFIED < NOISE < "other Ids", and each object is marked with a clusterId Sens. The function nextId (clusterId) returns the successor of clusterId in the ordering of the datatype. The function SetofPoints.get (i) returns the i-th element of SetofPoints.
A call of SetofPoints.Region(Sens, Eps) returns the Eps-neighborhood of Point in SetOfPoints as a list of objects. Obviously the efficiency of the above algorithm depends on the efficiency of the neighborhood query because such a query is performed exactly once for each object in SetofPoints which satisfies the selection condition.
The clusterId of some objects p which are marked to be NOISE because Eps (p)) < MinPts may be changed later if they are density-reachable from some other object of the database. This may happen only for border objects of a cluster. Those objects are then not added to the seeds-list because we already know that an object with a ClusterId of NOISE is not a core object, i.e., no other objects are density-reachable from them.
In algorithm, function Expand-Cluster constructing a density-connected set for a core object Object is presented in more detail next[5] :

```
ExpandCluster (SetofPoints, Sens, ClId, Eps,MinPts): Boolean;
        seeds:=SetofPoints.Region(Sens,Eps);
        IF Count (seeds) < MinPts THEN // no core point
           SetofPoints.changeClId(Sens, NOISE);
           RETURN False;
        END IF
// still here? sens is a core object
SetofPoints.changeClIds (seeds,ClId);
seeds.Remove(Sens);
        WHILE seeds  > 0 DO
           currentSens := seeds.first();
           result := SetofPoints.Region(currentSens, Eps);
```

```
        IF Count(result) ≥ MinPts THEN
          FOR i FROM 1 TO result.size DO
          P: = result.get (i);
          IF P.ClId IN {UNCLASSIFIED,NOISE} THEN
              IF P.ClId = UNCLASSIFIED THEN
              seeds.Add(P);
              END IF;
          SetofPts.changeClId(P,ClId);
          END IF; // UNCLASSIFIED or NOISE
          END FOR;
        END IF; // MinPts
      seeds.Remove(currentSens);
      END WHILE; // seeds Empty
 RETURN True;
END; // ExpandCluster
```

IV. RESULTS AND DISCUSSION

4.1 Experimental setup

For the proposed work, the data is collected from various data sources such as sensors which are deployed in a research laboratory. These sensors are required to collect time stamped topology information, along with humidity, temperature, light and voltage values. The x and y coordinates of the sensors are considered in meters. The data collected includes more than 2.3 million records. Sensor ids range from 1-54; Temperature is in degrees Celsius. Humidity is temperature corrected relative humidity ranging from 0-100%. Light is Lux, voltage is expressed in terms of volts.

4.2. Data Collection Form

The data collection is the main part of data mining. This section deals with the various data collection forms used in our project.

4.2.1 Data Entry Forms

Figure 2 shows the form used for obtaining the geographic details pertaining to sensors such as co-ordinate location. The data entered are displayed in the grid format. The density based clustering is done based on these location details. The sensor Id takes the Ids of each sensors and the location x, location y takes the x and y coordinate values of the location of sensor. On clicking the Add button these data are added to the database and these sensors entered are displayed in the grid-format. The Sensor details can be edited or deleted. The Reset button is used to clear all the textbox values.

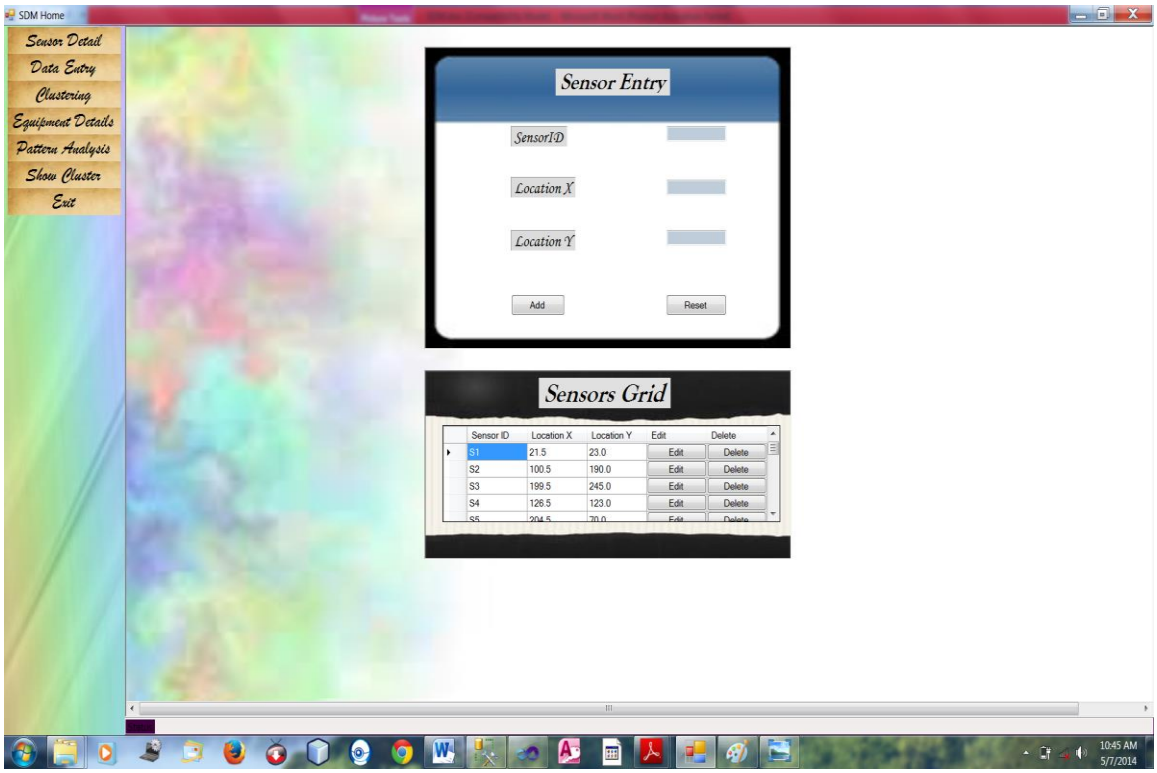

Figure 2. Sensor Location Entry

Figure 3 shows the form used to obtain the spatial and temporal details of the sensors such as the date of entry, temperature, humidity, light and voltage. These data entered are represented in the grid form, which can be updated if necessary. The data must be entered within a specified threshold. The sensor Id lists the Id's of the sensors entered in the previous page and the required sensor to which the data must be entered should be selected. The date for which the data needs to be entered must also be selected. By default the current date will be considered. The spatial data will be entered by the administrator. On clicking the Add button, these data entered will be stored into data base if they are within the specified threshold. The entered data will be displayed in the grid format in the page. These data can be edited or deleted.

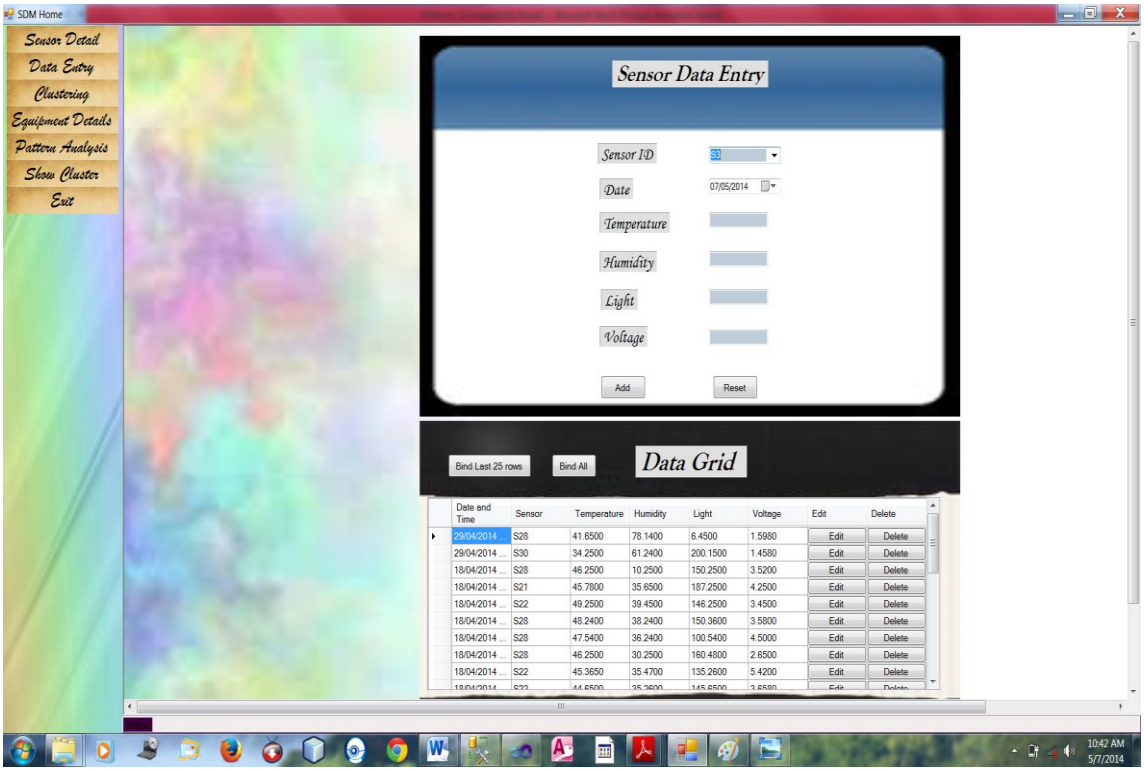

Figure 3. Sensor Spatial Data Entry

Figure 4, shows the equipment data form. This form is used to obtain the spatial data of the equipment such as temperature, humidity, light and voltage specifications of the equipment to be placed. These data are added to the database only if they do not exceed a specified threshold. These data are represented in a grid format. These data can be updated if required.

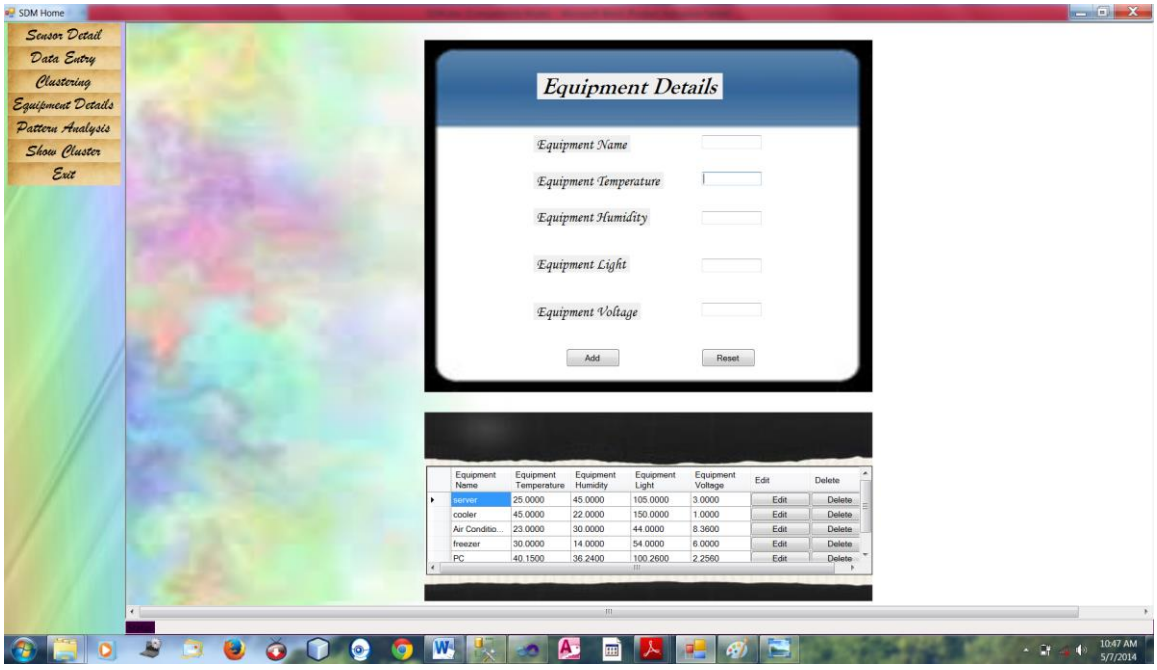

Figure 4. Equipment Data Entry

4.3. Clustering and Noise Detection Forms

The clustering and noise detection is done through the density based spatial clustering algorithm with noise called GDBSCAN. Figure 5. This page gives a grid representation of various clusters formed

along with the sensor ID's of those sensors that fall under each of the clusters. The sensors which constitute noise are also detected and are displayed in the noise grid. The clustering is done based only on the location information of the sensors entered through the sensor location entry page.

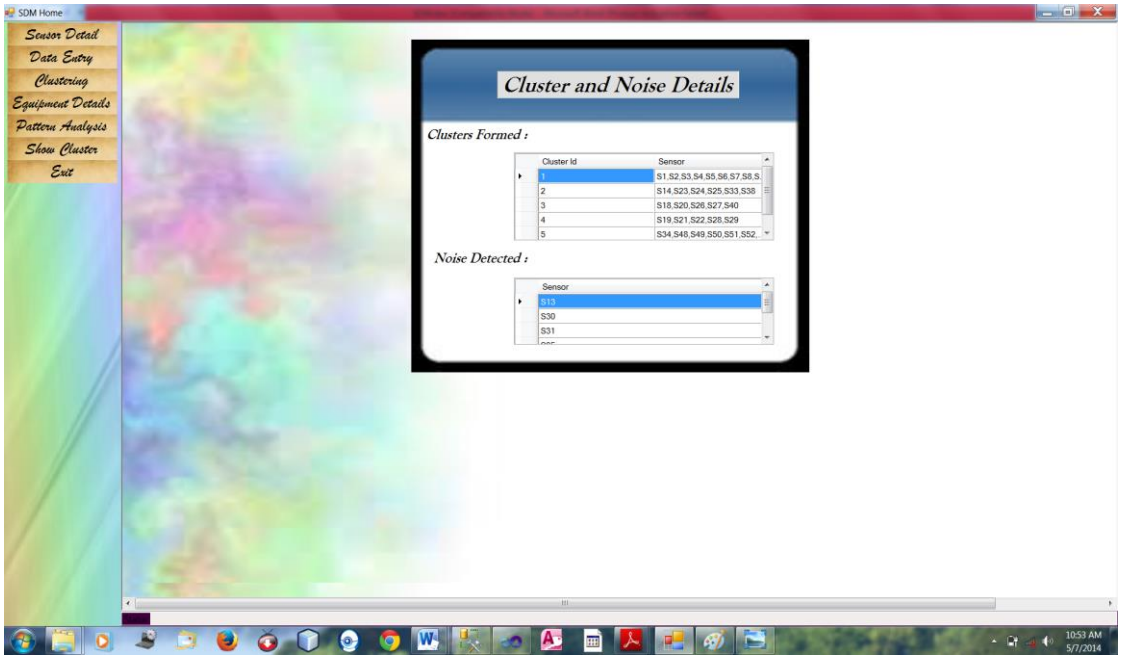

Figure 5. Cluster and Noise Grid

Figure 6 shows the graphical representation of the sensors and their respective clusters. The noises are those sensors which are not included in any cluster formations. The purple dots represent the sensors; the red lines are used to group the sensors of a particular cluster together. On clicking the show button present in the top of the screen we get the sensors which form the clusters and those sensors which form noise.

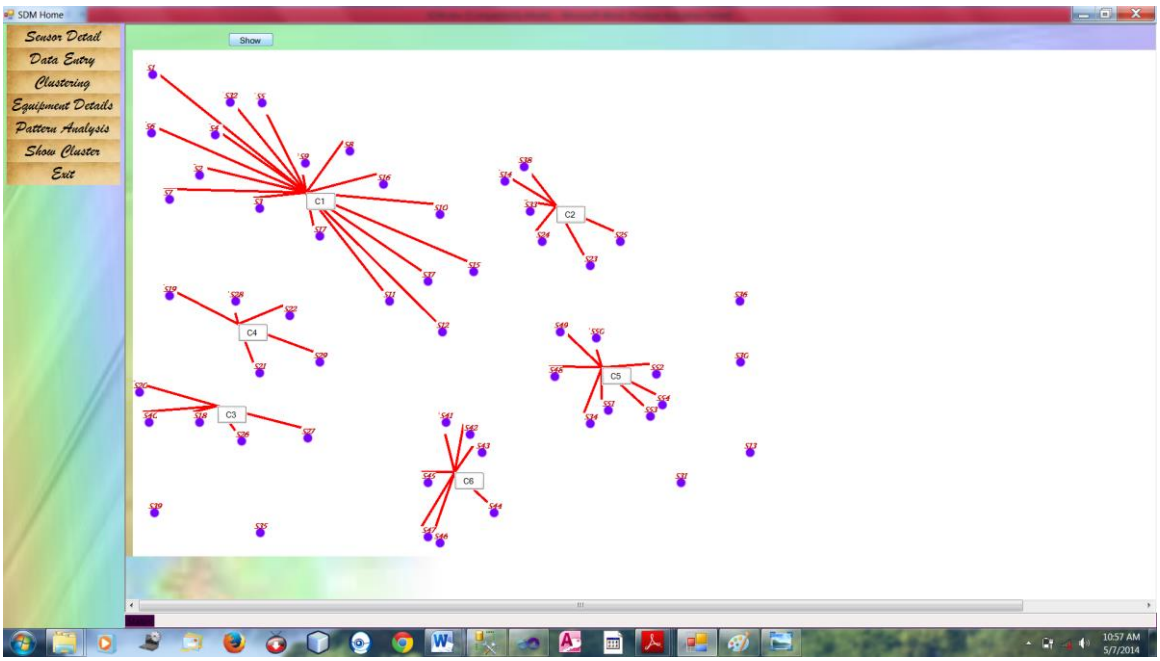

Figure 6. Cluster Formation and Noise Detection Representation

V. CONCLUSION

We considered 54 sensors to capture the spatial data, out of which 6 sensors do not belong to any of the clusters, so that are treated as outliers. Hence 11% of the noise is detected. This method is found to be feasible for Spatial Data Mining (SDM) in Geographic Knowledge Discovery (GKD). The density based GDBSCAN algorithm is used for clustering. It estimates the density for a particular point in the dataset by counting the number of points within a specified radius. There are 6 clusters formed for 54 sensors, named as C1, C2, C3, C4, C5, C6. The number of sensors in each cluster, after using GDBSCAN algorithm are such that, C1 consists of 17 sensors, C2 consists of 6 sensors, C3 consists of 5 sensors, C4 consists of 5 sensors, C5 consists of 8 sensors and C6 consists of 7 sensors. By this analysis cluster C1 has more number of sensors and it senses more spatial data and coverage of the location is also high so that we can place more sensitive devices in C1 compared to other clusters.

VI. FUTURE WORK

Clustering and Noise Detection method for the GKD will help in placing Laboratory devices in a feasible location, so that the damage to the devices due to high temperature, high humidity and high voltage can be avoided. By using this technique maintenance cost can be reduced and increase the lifetime of the devices. Further this work can be enhanced for pattern analysis, so that computed scoring for the temperature, light, humidity and voltage values can increase the speed of analysis process.

REFERENCES

[1]. Duck-Ho Bae, Ji-Haeng Baek, Hyun-Kyo Oh, Ju-Won Song, Sang-Wook Kim "SD- Miner: A Spatial Data Mining System", Proceedings of IC-NIDC 2009.
[2]. N.Santhosh Kumar, V. Sitha Ramulu, K.Sudheer Reddy, Suresh Kotha, Mohan Kumar, "Spatial Data Mining using Cluster Analysis", International Journal of Computer Science & Information Technology (IJCSIT) Vol 4, No 4, August 2012.
[3]. Ester M., Kriegel H.-P., Sander J. and Xu X. 1996. "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise". Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining. Portland, OR, 226-231.
[4]. Ng R.T., and Han J. 1994. "Efficient and Effective Clustering Methods for Spatial Data Mining". Proc. 20th Int. Conf. on Very Large Data Bases. Santiago, Chile, 144-155.
[5]. Jörg Sander, Martin Ester, Hans-Peter Kriegel, Xiaowei Xu, "Density- Based Clustering in Spatial Databases: The Algorithm GDBSCAN and its Applications", Data mining and knowledge discovery 2,169-194(1998).
[6]. Ertöz L., Steinbach M., Kumar V.: "Finding Clusters of Different Sizes, Shapes, and Densities in Noisy, High Dimensional Data", SIAM International Conference on Data Mining (2003)
[7]. Sayal M., Scheuermann P.: "A Distributed Clustering Algorithm for Web-Based Access Patterns", in Proceedings of the 2nd ACM-SIGMOD Workshop on Distributed and Parallel Knowledge Discovery, Boston, August 2000
[8]. Fayyad U., Piatetsky-Shapiro G., and Smyth P. 1996. "Knowledge Discovery and Data Mining: Towards a Unifying Framework". Proc. 2nd Int. Conf. on Knowledge Discovery and Data Mining, Portland, OR, 82-88.
[9]. Ester M., Kriegel H.-P., and Xu X. 1995. A Database Interface for Clustering in Large Spatial Databases, Proc. 1st Int. Conf. on Knowledge Discovery and Data Mining, Montreal, Canada, 1995, AAAI Press, 1995.
[10]. Jain A. K., Dubes R.C.: "Algorithms for Clustering Data", Prentice-Hall
[11]. Dr. Mohammed Otair," Approximate K-Nearest Neighbour Based Spatial Clustering Using K-D Tree", IJDMS Vol.5, No.1, February 2013.
[12]. Lovely Sharma1, K. Ramya, " A Review on Density based Clustering Algorithms for Very Large Datasets", IJETEA, Volume 3, Issue 12, December 2013
[13]. Richa Sharma Bhawna Malik Anant Ram," Local Density Differ Spatial Clustering in Data Mining", IJARCSSE, Volume 3, Issue 3, March 2013

AUTHORS

Sneha N S, is currently Pursuing 4th Semester, Master of Technology in Computer Science and Engineering at AIT, Chickmagalur. She has completed her Bachelor of Engineering from Srinivas Institute of Technology, Mangalore. She had published a paper. Her areas of interests include data mining and Information Security.

Pushpa, Associate Professor in the Department of Computer Science and Engineering, AIT, Chickmagalur and Research Scholar at R.V.College of Engineering, Bangalore. She had 12 years of teaching experience. She has completed her Master of Technology from Visvesvaraya Technological University. She had publishes research papers in international journals. Her research interests are in the area of data mining and Computer Networks.