Journal of Engineering and Applied Sciences 12 (Special Issue 12): 9572-9575, 2017
ISSN: 1816-949X

An Algorithm for Spatial Data Mining Using Clustering

Karishma Vaswani and A.M. Karandikar
Department of Computer Science and Engineering,
Shri Ramdeobaba College of Engineering and Management, Nagpur, India

Abstract: Data mining is the process of analyzing large sets of data and then extracting useful and relevant data. Data mining has many tools for predicting the behaviour allowing systems to make proper decisions. It can answer questions that are strong to resolve. Therefore, they can be used to predict meteorological data which is called as weather prediction. Weather forecasting is an important application in meteorology and has been one of the most challenging problem around the world. Predicting the weather is important to help preparing for the best and the worst climate. Clustering is the common data mining technique for finding hidden patterns in data. Clustering tries to group a set of objects and find the similarity between those objects. In this study, we are going to apply clustering algorithms on spatial datasets to group together climatic data for weather analysis. We will measure the performance of various clustering algorithms and record their drawbacks. We will propose a model that will try to overcome these drawbacks thus giving effective results.

Key words: Spatial data mining, geographical conditions, weather and clustering, meteorological, algorithms, India

INTRODUCTION

Data mining is the process of collecting, searching and analyzing large amount of data in a database as to discover patterns or relationships. It helps businesses focus on the most important information in their data warehouses. Data mining tools predict trends and allowing business to make knowledge-driven decisions. It is also called Knowledge Discovery in Databases (KDD). The information obtained from data mining is hopefully both useful and new. Data mining is the process of finding correlations among dozens of fields in immense relational databases (Sheikh *et al*., 2016). Data mining techniques are used in weather analysis (Kalaiselvi and Geetha, 2016). Weather forecasting is used in many fields like agriculture, food, security, disasters and science. Weather is most effective environmental constraint in every phase of our life. In ancient years, we don't know weather conditions. So, we faced many problems in industry, agriculture and food management process. But, now we have many ways to find weather conditions. That is the main reason for applying data mining techniques to find the weather conditions.

Clustering tries to group a set of objects and find some relationship between those objects (Kanungo *et al*., 2002). There are two main branches of clustering: hierarchical and partitional. K-means is a popular partitional clustering method. It is also known as the generalized Lloyd algorithm in which euclidean distances are used to measure the dissimilarity between a data point and the cluster representatives. It is well known that the k-means algorithm suffers the demerit that it is dependent on the choice of the initial clusters and the instance order.

Literature review: Data mining is seen as an increasingly important tool to transform data into intelligent information for business benefit. Data preparation and filtering can take considerable amount of processing time. If there is irrelevant, redundant information or noisy and unreliable data then discovering new information during the training phase is difficult. Data preprocessing includes data cleaning, transformation, normalization, feature extraction and selection, etc. (Sheikh *et al*., 2016). The product of data preprocessing is the final training set.

K-means clustering is a data mining learning algorithm and also one of the simplest technique. It is used to cluster observances into groups of similar ones without any prior knowledge. It is commonly used in biometrics, medical and related fields.

WEKA tool: WEKA-Waikato Environment for Knowledge Analysis is a machine learning algorithm for data mining (Kalaiselvi and Geetha, 2016; Gunasekara *et al*., 2014).

Corresponding Author: Karishma Vaswani, Department of Computer Science and Engineering,
Shri Ramdeobaba College of Engineering and Management, Nagpur, India

It is well suited for developing new machine learning techniques. WEKA contains tools for data pre-processing, clustering, association rule, classification and visualization. The algorithms can either be applied directly to a dataset or called from your own java code. Each entry in a dataset is an instance of the Java class and each instance consists of a number of attributes.

MATERIALS AND METHODS

In this we apply the data mining technique k-means cluster algorithm on the data set which was modified in to suitable format from the raw format after preprocessing stage. Various disadvantages were observed and enhancement of algorithm is done using incremental k-means.

Implementation of proposed approach: Dataset is collected online and it contains several attributes and 735 instances as in Table 1. Here, we have just shown some rows and values of some instances. The dataset is that of the Chicago City in US. It has several attributes such as elevation, latitude, longitude, hourly visibility, sunrise, sunset, pressure, temperature and various other attributes regarding weather and climate conditions. The values for each of the attributes is noted on hourly basis and on these values the average is calculated by k-means.

K-means clustering: K-means clustering is a data mining algorithm used to cluster observations into groups of similar observations without any prior knowledge of those relationships (Kalaiselvi and Geetha, 2016; Chakraborty and Nagwani, 2011; Gunasekara *et al*., 2014). K-means is the most popularly used clustering algorithm. User needs to specify the number of clusters (k) which serves as the input. Algorithm randomly selects k objects as cluster mean or center and k-means basic version researchers with numeric data only. It is a prototype based clustering technique defining the prototype in terms of a centroid which is considered to be the mean of a group of points and is applicable to objects in a continuous n-dimensional space. It is a method of vector quantization, really from signal processing that is famous for cluster analysis in data mining and cluster data using the k-means algorithm. It can use either the Euclidean distance. If the Euclidean distance is used then centroid are computed as the component-wise median rather than mean. The k-means clustering algorithm is a partition-based cluster analysis method (Kalaiselvi and Geetha, 2016). Following steps in this process:

Step 1: Initialization step: initialize K centroids
Do
Assignment step: assign each data point to its centroid
Re-estimation step: Re compute centroid (cluster centers) while (there are still changes in the centroid) select k objects as initial cluster centers. The dataset is partitioned into K clusters and the data points are randomly set to the clusters resulting in clusters that have roughly the same number of data points

Step 2: For each data point calculate the euclidean distance from the data point to each cluster, The euclidean distance is the straight-line distance between two pixels

$$... = v(..1 - ..2)2 + (1 - ..2)2$$

where (..1, .., 1) and (..2, .., 2)) are two data points

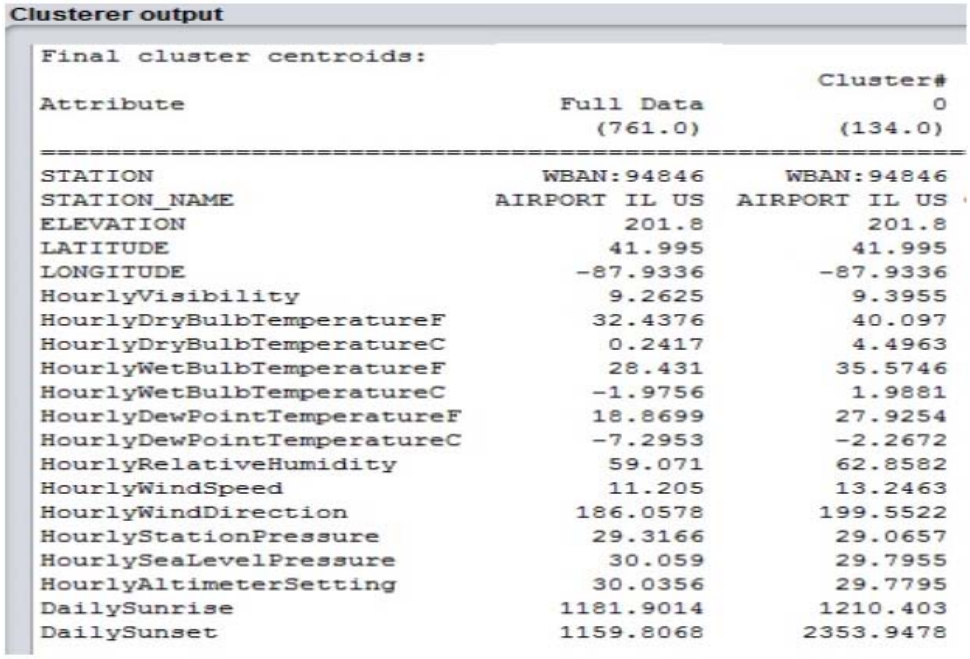
Step 3: If the data point is closest to its own cluster then no change. If the data point is not closest to its own cluster, move it into the nearest cluster. Repeat the above step until an entire iteration through all the data points results in no point transferring from one cluster to another. At this point the clusters are stable and the clustering process ends

Step 4: Repeat this process until the criterion function converged and square error criterion for clustering. The choice of initial partition can greatly affect the end clusters that result in terms of inter-cluster and intra-cluster distances and cohesion.

K-means clustering algorithm is used for finding maximum and minimum weather condition detail in overall process. It is the most important flat clustering algorithm. Its objective is to reduce the average squared Euclidean distance of documents from their cluster centers is defined as mean or centroid (Kumar *et al*., 2012). There are two clusters like cluster instance 0 and cluster instance 1. These two clusters show final centroid cluster display the maximum and minimum values of overall weather attributes.

Table 1: Weather dataset

Stations	Station name	Elevation	Latitude	Longitude	Hourly visibility	Hourly dry bulb temperature (F)	Hourly dry temperature (eC)	Hourly wetbulb temperature (eF)	Hourly wetbulb temperature (C)
WBAN: 94846	Chicago airport	201.8	41.995	-87.9336	10.00	30	-1.1	26	-3.4
WBAN: 94846	Chicago airport	201.8	41.995	-87.9336	10.00	30	-1.1	26	-3.4
WBAN: 94846	Chicago airport	201.8	41.995	-87.9336	6.00	29	-1.7	26	-3.2
WBAN: 94846	Chicago airport	201.8	41.995	-87.9336	0.75	28	-2.2	26	-3.3
WBAN: 94846	Chicago airport	201.8	41.995	-87.9336	10.00	26	-3.3	24	-4.5
WBAN: 94846	Chicago airport	201.8	41.995	-87.9336	10.00	25	-3.9	22	-5.3
WBAN: 94846	Chicago airport	201.8	41.995	-87.9336	5.00	23	-5.0	21	-6.4

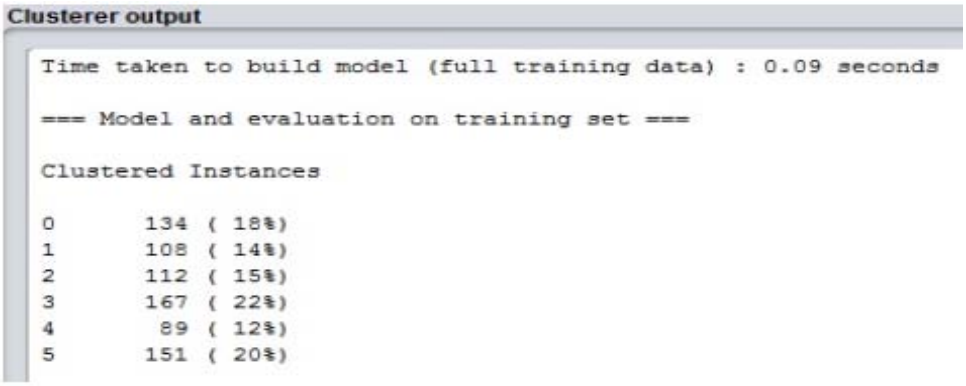


Clusterer output

Final cluster centroids:

Attribute	Full Data (761.0)	Cluster# 0 (134.0)
STATION	WBAN:94846	WBAN:94846
STATION_NAME	AIRPORT IL US	AIRPORT IL US
ELEVATION	201.8	201.8
LATITUDE	41.995	41.995
LONGITUDE	-87.9336	-87.9336
HourlyVisibility	9.2625	9.3955
HourlyDryBulbTemperatureF	32.4376	40.097
HourlyDryBulbTemperatureC	0.2417	4.4963
HourlyWetBulbTemperatureF	28.431	35.5746
HourlyWetBulbTemperatureC	-1.9756	1.9881
HourlyDewPointTemperatureF	18.8699	27.9254
HourlyDewPointTemperatureC	-7.2953	-2.2672
HourlyRelativeHumidity	59.071	62.8582
HourlyWindSpeed	11.205	13.2463
HourlyWindDirection	186.0578	199.5522
HourlyStationPressure	29.3166	29.0657
HourlySeaLevelPressure	30.059	29.7955
HourlyAltimeterSetting	30.0356	29.7795
DailySunrise	1181.9014	1210.403
DailySunset	1159.8068	2353.9478

Fig. 1: Clustering done on WEKA



Clusterer output

```
Time taken to build model (full training data) : 0.09 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       134 ( 18%)
1       108 ( 14%)
2       112 ( 15%)
3       167 ( 22%)
4        89 ( 12%)
5       151 ( 20%)
```

Fig. 2: Clustering instance on WEKA

RESULTS AND DISCUSSION

Results were obtained on WEKA tool. Initially when the dataset was large and unfiltered, WEKA could not give the output. Hence, preprocessing was done as mentioned in the above steps and clustering was done where we compared the results for any number of clusters which was provided as the input.

Here, number of clusters taken are 6. WEKA then performs k-means clustering over the attributes of the dataset mentioned in Fig. 1. It then calculates the mean of all the values of a particular attribute. This procedure is done on all the attributes. After the centroid is fixed it puts the instance in that particular cluster and finally it shows that how many instances are placed in each cluster by giving the number of cluster and the percentage as seen in Fig. 2 and 3. Similarly results were seen for other number of clusters. Visualization of instances between any 2 attributes could be seen in WEKA in Fig. 3. Here, one attribute is hourly dry bulb temperature and the other is hourly altimeter setting. The orange and blue dots are the instances of the above attributes, respectively. It shows that how many instances have close values, so that, they could be accommodated in the nearby cluster. Similarly one can see the visualization of any the 2 attributes in WEKA.

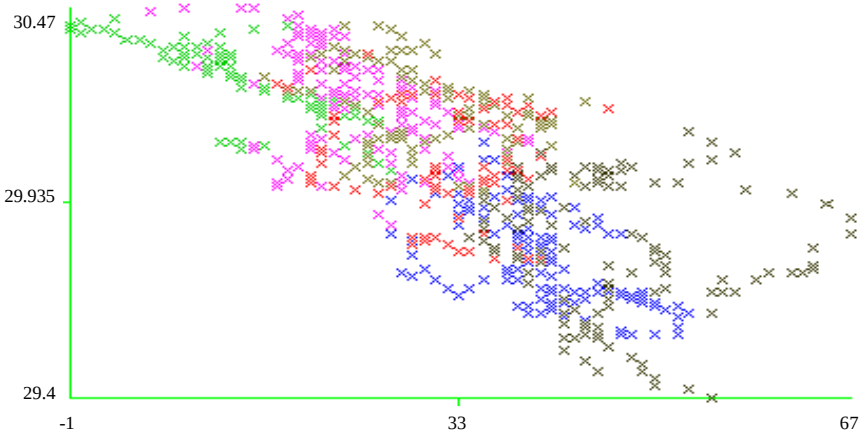


Fig. 3: Visualization of 2 attributes

CONCLUSION

It was observed that the weather data keeps on updating on daily basis. Hence, it is necessary to maintain the whole data and perform clustering on regular basis for weather analysis.

RECOMMENDATIONS

Our future research will focus on performing clustering on updated datasets considering the number of clusters by taking into consideration incremental k-means concept. It will also focus on the accuracy and misclassification. The results will be compared with the results of WEKA and enhancement will be done, so that, incremented data do not take more time to perform clusters.

REFERENCES

Chakraborty, S. and N.K. Nagwani, 2011. Analysis and study of incremental K-means clustering algorithm. Proceeding of the 2011 Conference on High Performance Architecture and Grid Computing (HPAGC'11), July 19-20, 2011, Springer, Chandigarh, India, pp: 338-341.

Gunasekara, R.P.T.H., M.C. Wijegunasekara and N.G.J. Dias, 2014. A study on how to improve the perfromance of K-mean data mining algorithm in a parallel environment. J. Eng. Appl. Sci., 9: 441-446.

Kalaiselvi, P. and D. Geetha, 2016. Weather prediction using J48, EM and K-means clustering algorithms. Intl. J. Innovative Res. Comput. Commun. Eng., 4: 20889-20895.

Kanungo, T., D.M. Mount, N.S. Netanyahu, C.D. Piatko, R.S. Angela and Y. Wu, 2002. An efficient k-means clustering algorithm: Analysis and implementation. IEEE Trans. Pattern Anal. Mach. Intell., 24: 881-892.

Kumar, A., R. Sinha, V. Bhattacherjee, D.S. Verma and S. Singh, 2012. Modeling using K-means clustering algorithm. Proceedings of the 1st International Conference on Recent Advances in Information Technology (RAIT'12), March 15-17, 2012, IEEE, Dhanbad, India, I SBN: 978-1-4577-0694-3, pp: 554-558.

Sheikh, F., S. Karthick, D. Malathi, J.S. Sudarsan and C. Arun, 2016. Analysis of data mining techniques for weather prediction. Indian J. Sci. Technol., 9: 1-9.