ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

A Parallel Study on Data Mining Techniques for Clustering to use Weather Dataset

Praveen Pappula[1], Rama B[2]

Research Scholar &Assistant Professor, SREC, Warangal, India[1]

Assistant Professor, Kakatiya University, Warangal, India[2]

ABSTRACT: In this paper, we give a survey on data mining techniques. More specially speaking, we talk about one important and basic data mining technique called Clustering, Clustering is one of the fundamental techniques in data mining. The primary objective of clustering is to partition a set of objects into homogeneous groups. An effective clustering needs a suitable measure of similarity or dissimilarity. So a partition structure would be identified in the form of natural groups. We are survey in data mining on Weather report, and it is a good reference for researcher on this topic.

KEYWORDS: *Clustering, K means, Hierarchical, Weather Dataset*

1. I.INTRODUCTION

The process of extracting useful patterns or information from large amount of data is known as data mining [1]. Most of the people think data mining as a synonym of knowledge discovery. But actually data mining can be considered as a step of knowledge discovery in databases (KDD). KDD process includes data cleaning (to remove noise and inconsistent data), data integration (where multiple data sources may be combined), data selection (where data relevant to the analysis task are retrieved from the database), data transformation (where data are transformed or consolidated into forms appropriate for mining by performing summary or aggregation operations), data mining (an essential process where intelligent methods are applied in order to extract data patterns, pattern evaluation (to identify the truly interesting patterns representing knowledge based on some interestingness measures) and knowledge presentation (where visualization and knowledge representation techniques are used to present the mined knowledge to the user) [1] [6].

Data mining has attracted great deal of attention in information industry as well as in business areas because of the need of turning large data into useful information [4] [10]. Data mining is useful in an explanatory scenario in which there are no predefined notions about what will constitute an interesting outcome [7]. The database system industry has witnessed an evolutionary path in the development of the following functionalities like data collection and database creation, data management and advanced data analysis that includes data warehousing and data mining. Prediction and description are considered as two primary goals of data mining. Predictive data mining which produces the model of system described by the given data set and descriptive data mining, which produces new, non trivial information based on the available data set. The goal of prediction and description are achieved through data mining tasks such as classification, discovering association rules and clustering [9].

II RELATED WORK

a. Association Rules

Association rule mining tries to find frequent item set among large data sets [3]. It aims to extract interesting correlations, frequent patterns, associations or casual structures among sets of items in the transaction databases or other data repositories and describes association relationship among different attributes. Such finding helps businesses

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

to make certain decisions like customer's behavior analysis. However the number of possible Association Rules for a given dataset is generally very large and most of them are usually of less value. Association mining is an important research area in data mining.

b. Classification

Classification is the most commonly applied data mining technique, which employs a set of pre-classified examples to develop a model that can classify the population of records at large [2]. It is the discovery of a predictive learning function that classifies a data item into one of the several predefined classes. Fraud detection and credit risk applications are particularly well suited to this type of analysis. The data classification process involves learning and classification. In Learning the training data are analysed by classification algorithm

c. Clustering

Clustering is one of the well known data mining technique and can be defined as the identification of similar classes of objects [5]. It is a common descriptive task in which one seeks to identify a finite set of categories or clusters. By using clustering techniques we can discover the overall distribution pattern and correlations among data attributes. Classification approach can also be used for effective means of distinguishing groups or classes.

Clustering is a type of unsupervised [8] classification which divide the data set into some number of subsets based on distance between each data. The clustering algorithms can be classified to hierarchical clustering and the non hierarchical clustering

Types of data in Cluster Analysis:

Data matrix

We take n in the number of objects and p variables (attributes) like n in person p is age, height, gender etc.

We create a relational table n-by-p matrix (nxp).

Dissimilarity matrix

This matrix stores a collection of proximities that are available for all pairs of n objects n by n (n x n)[2].

D(i, j) is the measured difference or dissimilarity between objects i & j,d (i, j) = d (j, i) i.e. d(2, 1) = d(1, 2) and d(i, i) = 0. It means d(1, 1) = d(2, 2) …. d(n, n) = 0.

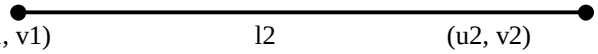
We find the distance between i to j d(i, j) we have most popular distance measure is Euclidean distance measure.

D(i, j) = (xi1 – xj1)2 + (xi2 – xj2)2 + ….. + (xin – xjn)2

Where i = (xi1, xi2, … xin)

j = (xj1, xj2, …. xjn)

Distance between A1 to A2 = 5,d(A1, A2) =5.



l2 = ((u2 – u1)2 + (v2 – v1)2)1/2

Example A1 = (2, 10), A2 = (2, 5)

$$\sqrt{(2-2)^2 + (5+10)^2} = \sqrt{0+25} = 5$$

Clustering algorithms are generally categorized under two different categories – partitioned and hierarchical. Partitioned clustering algorithms divide the data set into non overlapping groups [11], Algorithms k-means, K-modes etc. fall under this category. This approach to group into a pre-determined k-number of clusters. min $\sum_{i=1}^{n} \|x_i - x_j\|^q$

Where xi the centre of jth cluster and is the center nearest to data object d, n is the number of elements in data set, q is an integer which defines the nature of the distance function (q = 2) for numeric valued data sets[1].

As opposed to partition clustering algorithms, hierarchical algorithms use the distance matrix as input and create a hierarchical set of clusters. These are categorized by two.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

1.Agglomerative: Start with a unique cluster which consisting all objects, here two closest objects are merged iteratively[8], till a final cluster which contains all data objects. It is followed to down approach.
2.Divisive:: Initial cluster containing all data objects which are split to contain cohesive sub clusters, till each object belongs to a unique cluster. Its followed bottom up approach

Various algorithms and data mining techniques like Classification, Clustering, and Association Rules etc are used for knowledge discovery [1] [6]. We have described the following data mining techniques.

IV. CLUSTERING ALGORITHMS

A . Non Hierarchical methods
K-means method

K-means is the simplest and most popular classical clustering method that is easy to implement the classical method[6] can only be used of the data about all the objects is located in the main memory..

Algorithm K-means

Decide K, the number of clusters that are needed finally
Given K, the K-means algorithm is implemented in four steps.
Step 1: Partition objects into K non empty subsets (randomly)
Step 2: Compute seed points as the centroids of the clusters of the current partitioning (the centroid is the center, i.e. mean point of the cluster.
Step 3: Assign each object to the cluster with the nearest seed point [8]
Step 4: Go back to step 2, stop when the assignment does not change.

B. Hierarchical Methods

When the data set S is set of n dimensional real vector x = {xi_1, xi_2, …., xi_n}, S = {x_i}, i = 1, 2, …. N and let a family of disjoint subsets of S at the clustering time step t be $a^{(t)} = \{a_k^{(t)}/a_k^t \in S, a_k^t \neq \phi, U\, a_k^t \cap a_k^t = \phi\}$[3],k, $k^1 = 1$ …. $|a^t|$, $k \neq k^1$, where akt call kth cluster the clustering time step t, the element of a_k^t be x_k^i, $k_i = 1$ …. Kn, $x_{ki} \in S$, the agglomerative hierarchical clustering algorithm[5] start with each element as separate cluster, such that $a_k^{(0)} = \{x_{ki}\}$, $k_i = 1$. $|a^t| = N$. Hierarchical clustering algorithm follows selecting clusters, merging clusters and updating distance[16], until the number of cluster $|a^t|$ equal to 1. Selecting clusters first select minimizes distance between $d(a_p^{(t)}, a_q^{(t)})$ is selected where p, q = 1 ….. $|a^t|$, $p \neq q$.
Next merging cluster, the selected p^{th} and q^{th} cluster are merged to new cluster $a_p^{(t)} \cup a^{q(t)}$. The new cluster $a_p^t \cup a_q^t$ is used instead of cluster a_p^t and a_q^t in the next time step t + 1. In the step of updating distance, the distances between each cluster are re-calculated based on set of cluster a_q^{t+1} at next time t+1[10].
The divisive hierarchical clustering algorithm is recursively take the procedure of the partitioned hierarchical clustering algorithm. Algorithm start with cluster included whole element of input data set such that $a^0 = S$ at the time step t.
The divisive hierarchical clustering algorithm [8] recursively divide the each or selected cluster a_k^t to k new cluster $A_k^t = \{A_{kl}^t/A_{kl}^t \in S, A_k^t \neq f, \underline{U}A_k^t = A_k, A_{kl}^t \cap A_{kl}^t = \phi,\}$, l, $l^1 = 1$, …..K, $l \neq l^1$, where K is the given number[4]. At the time step t +1 the new clusters A_{kl}^t is used instead of divided cluster a_k^t.

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

IV. EXPERIMENTAL REPORT

K-means

Run information ===

Scheme: weka.clusterers.SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10
Relation: weather
Instances: 14
Attributes: 5
outlook
temperature
humidity
windy
Ignored:
play
Test mode: Classes to clusters evaluation on training data

=== Clustering model (full training set) ===

K-means
======

Number of iterations: 3
Within cluster sum of squared errors: 11.119394136333145
Missing values globally replaced with mean/mode

Cluster centroids:

	Cluster#		
Attribute	Full Data	0	1
	(14)	(9)	(5)
outlook	sunny	sunny	overcast
temperature	74	76.5556	69.4
humidity	81.7692	84.3077	77.2
windy	FALSE	FALSE	TRUE

Time taken to build model (full training data) : 0 seconds

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

Clustered Instances

0 9 (64%)
1 5 (36%)

Class attribute: play

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

Classes to Clusters:

```
 0 1  <-- assigned to cluster
 6 3 | yes
 3 2 | no
```

Cluster 0 <-- yes
Cluster 1 <-- no

Incorrectly clustered instances : 6.0 42.8571 %

Run information ===

Scheme: weka.clusterers.HierarchicalClusterer -N 2 -L SINGLE -P -A "weka.core.EuclideanDistance -R first-last"
Relation: weather
Instances: 14
Attributes: 5
outlook
temperature
humidity
windy
play
Test mode: split 66% train, remainder test

=== Clustering model (full training set) ===

Cluster 0
(((1.0:1.04053,1.0:1.04053):0.01021,1.0:1.05074):0.0377,(1.0:0.73521,1.0:0.73521):0.35324)

Cluster 1
(((0.0:0.3674,0.0:0.3674):0.76884,(0.0:0.8919,0.0:0.8919):0.24433):0.00411,((0.0:0.96006,(0.0:0.61441,0.0:0.61441):0.34565):0.09176,(0.0:1.04002,0.0:1.04002):0.0118):0.08853)

Time taken to build model (full training data) : 0.02 seconds

=== Model and evaluation on test split ===
Cluster 0
(((1.0:0.74741,1.0:0.74741):0.35958,1.0:1.10699):0.11182,(0.0:1.1667,0.0:1.1667):0.0521)

Cluster 1
((0.0:0.37012,0.0:0.37012):0.87832,(0.0:1.1555,0.0:1.1555):0.09295)

Time taken to build model (percentage split) : 0 seconds
Clustered Instances

0 4 (80%)
1 1 (20%)

ISSN(Online) : 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 4, Issue 4, April 2016

V. CONCLUSION

In this paper, we surveyed the list of existing Clustering Algoritms.The topic of discovering Clusters has been studied over couple of decades. Most of the foundation researches have been done. A lot of attention was focus on the performance and scalability of the algorithms, but not enough attention was given to the quality(interestingness) of the rules generated. In the coming decades, the trend will be to turn the attention to the application of these researches in various areas of our lives, e.g. genetic research, medicine, homeland security, and Weather forecasting etc.

REFERENCES

[1] Jiawei Han MichelineKamber, *Data Mining concepts and techniques*, 2nd Edition.
[2] Jing Ding, Shanlin Yang, ―*Classification Rules Mining Model with Genetic Algorithm in Cloud Computing*‖, International Journal of Computer Applications (0975 – 888), Volume 48– No.18, June 2012.
[3] Ling Juan Li, min Zhang, ―*The strategy of mining association rule based on cloud computing*‖, International conference on business computing and global information, 2011.
[4] VenKatadri .M, Dr. Lokaanaathaa C.Reddy, ―*A review on data mining from past to future*‖, International Journal of Computer Applications (0975 – 8887) Volume 15– No.7, February 2011.
[5] A. Mahendiran, N. Saravanan, N. Venkata Subramanian and N. Sairam, ―*Implementation of K-Means Clustering in Cloud Computing Environment*,‖ Research Journal of Applied Sciences, Engineering and Technology 4(10): 1391-1394, 2012.
[6] *http://en.wikipedia.org/wiki/Data_mining/*
[7] A.K.Jain, M.N. Murty, P.J. Flynn, Data clustering: a review, ACM Computing Surveys 31 (3) (1999) 264-323.
[8] Arun K Pujari, *Data Mining Techniques*, second edition.
[9] D.Charalampidis, A modified K-means algorithm for circular invariant clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence 27 (12) (2005) 1856-1865.
[10] M. Li, M.K. Ng, Y.M. Cheung, Z. Huang, Agglomerative fuzzy K-means clustering algorithm with selection of number of clusters, IEEE Transaction on Knowledge and Engineering 20 (11) (2008) 1519-1534.
[11] Iaurent Galluccio,ilivier Michel, Pierre Comon,MarkKligee,AlfredO,Hero, InformationScience, Volume 2.51,1 December 2013,pages96-113.