Available online at www.sciencedirect.com

ScienceDirect

Materials Today: Proceedings 4 (2017) 9039–9048

materialstoday: PROCEEDINGS

www.materialstoday.com/proceedings

ICAAMM-2016

Featured Based Pattern Analysis using Machine Learning and Artificial Intelligence Techniques for Multiple Featured Dataset

Annaram Soujanya[a*], O. Subhash Chander Goud[b], Sai Prasad. K[c], G Prabhakar Reddy[d]

[a] *Asst Prof, IARE, JNTUH, Dundigal, Hyderabad, Telangana, India.*
[b] *Asst Prof, NIZAM College, OU, Hyderabad, Telangana, India*
[c] *Asst Prof, MLRIT, JNTUH, Dundigal, Hyderabad, Telangana, India*
[d] *Asst Prof, MLRIT, JNTUH, Dundigal, Hyderabad, Telangana, India*

Abstract

Data mining is a process of extracting patterns from a large datasets. We are trying to uncover the data bonded features which are hard to visualize and if many feature exists for the data then it becomes difficult to analyze the data. The main aim of the scheme of work is to classify the features and over which try to classify the data sets. Machine learning is one of the techniques of Artificial Intelligence which is used for extracting valuable knowledge from large data base. Machine Learning is also used for extracting patterns, models in data. In this paper we are trying to group the data based on multi-dimensional feature classification. Clustering process makes the similar features to form into one group and or else multiple groups, here in we try to group the features which are similar and form multiple groups. The US schooling data is in the form of flat files. Classification process is performed on the raw data. Classification is performed according to hierarchical clustering. Filtration process is used in order to obtain non-zero values. The saturation points are generated by performing clustering. Based on the clusters obtained the patterns can be extracted. Attribute based classification and hierarchical clustering is performed on the data. The attributes obtained are named as income and expenses. By combining the income and expenses attributes patterns can be identified. By using some combination of the two attributes some patterns have been obtain. By performing clustering on all the combinations of each attribute we can identify the patterns. Then according to this process we find the patterns generated by the feature classification or clustering process against the data classification or clustering. This new methodology tries to give the co-relation between the data and properties of a data set and how they behave.

Selection and Peer-review under responsibility of the Committee Members of International Conference on Advancements in Aeromechanical Materials for Manufacturing (ICAAMM-2016).

Keywords: Classification, DBSCAN Clustering, Multi-dimensional clustering, US Schooling data, Pattern Extraction.

Selection and Peer-review under responsibility of the Committee Members of International Conference on Advancements in Aeromechanical Materials for Manufacturing (ICAAMM-2016).

1. Introduction

Text clustering is one of the techniques used for obtaining meaning data from large amount database. There are some traditional clustering algorithms which uses some of the approach like BOW (Bag of Words) approach. WordNet and based lexical chains semantics approaches are used for solving documents clustering. It removes ambiguity, overcomes the high dimensionality of the clusters [1]. The process of mining is extracting meaning data from huge amount of spatial data is known as Spatial Data mining. One of the famous density based clustering algorithm is DBSCAN which performs clustering by increasing high density area and finds any kind of shape of clustering [13]. The two parameters used in DBSCAN clustering are epsilon (eps) and MinPts. The clustering starts from unvested arbitrary starting point. A cluster is formed only when number of neighbor are greater equal to MinPts. After cluster is formed the starting point is marked as visited and neighbor added into that particular cluster. This process is repeated until all the neighbors are visited. If the number of MinPts is less than neighbor then that point is marked as Noise. Clustering are mostly used for identify classes in spatial database [2]. K-means clustering is used in various fields like image and audio compression, system modeling and neutral network structures [3]. Sequential pattern mining, maximum pattern mining are some of the pattern mining [4]. Among them most of the methods are proposed for developing efficient mining algorithms in order to find particular patterns within a reasonable and acceptable time [5].

1. Methodology

In our paper, we have taken the Public School System Finances database for performing clustering and classification process. The Public School System Finances database consists of fifteen thousand records of various schools. The data contains the financial information like revenue, expenditure, debt and assets for public elementary-secondary school collected as a part of the Census of Governments. Classification process is performed by removing the zero values from the 15869 records. The below table consists of various school databases like STATE, IS, SUPID, NAME, 19H, 61V….W61. Table 1 shows the US schooling database.

Table 1: US Schooling Database

STATE	ID	SUID	NAME	V33	TOTAL REV	.....	.....	.....	....	61V	66W	W01	W31
01	015001001	503	AUTAUGA CO SCH DIST	7137	21039	....	....	....	....	0	0	211	1603
01	015002001	403	BALDWIN CO SCH DIST	17983	67403	....	....	....	....	0	0	1146	10687
01	015003001	303	BARBOUR CO SCH DIST	2255	8554	....	....	....	....	0	0	122	575
01	015003002	103	EUFAULA CTY SCH DIST	2997	10906	....	....	....	....	0	0	0	1825
01	015004001	203	BIBB CO SCH DIST	3564	11548	....	....	....	....	0	0	138	0
01	015005001	103	BLOUNT CO SCH DIST	6045	21990	....	....	....	....	0	0	47	1842
01	015005002	903	ONEONTA CITY SCH DIST	1105	3674	....	....	....	....	0	0	0	0
01	015006001	003	BULLOCK CO SCH DIST	2023	7361	....	....	....	....	0	0	0	5
01	015007001	903	BUTLER CO SCH DIST	4476	14780	....	....	....	....	0	0	19	104