© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 07 Issue: 04 | July – September 2016
Paper is available at www.jrps.in | Email : info@jrps.in

Implementation Density Based clustering in Data Mining

[1]Shiwani, Department of Computer Engg , KUK Kurukshetra University Kurukshetra
[2]Abhishek Bhatnagar, : Assistant Professor, Department of Computer Engg

ABSTRACT: Data mining extraction of hidden predictive info from large database records, is a powerful new technology with great potential to help companies focus on most important info within their data value warehouses. Data mining utensils predict future trends &behaviors, permitting businesses to make taking the initiative, knowledge motivated decisions. Automated, prospective analyses offered by data mining transfer outside analyses of past events providing by retrospective utensils typical of decision support systems. data mining utensils may answer business questions that usually were too time consuming to resolve. They scour database record records for hidden patterns, finding predictive info that experts may miss since this lies outside their expectations.

[1] Introduction

Data mining is extraction of hidden predictive info from large database record records, is a powerful new technology within great potential to help by companies focus on most important info within their data value warehouses. Data mining utensils predict future trends & behaviors, permitting businesses to make taking initiative, knowledge motivated decisions. Automated, prospective analyses offered by data mining transfer outside analyses of past events providing by retrospective utensils typical of decision support systems. Data mining utensils may answer business questions that generally were too time consuming to resolve. They scour database record records for hidden patterns, finding predictive info that experts may miss since this lies outside their expectations.

Key methods of Data mining

Many dissimilar datamining, query model, processing model & data value collection methods are available. Which one do you use to mine your data & which one may you use within combination with your existing software & infrastructure? Examine dissimilar data mining & analytics methods & solutions & learn how to build them using existing software & installations. Explore dissimilar data mining utensils that are available, & learn how to determine whether size & complexity of your info might result within processing & storage complexities, & what to do.

Several core methods that are used within data mining describe type of mining & data value recovery operation. Unfortunately, dissimilar companies & solutions do not always share terms, which may add to confusion & apparent complexity.

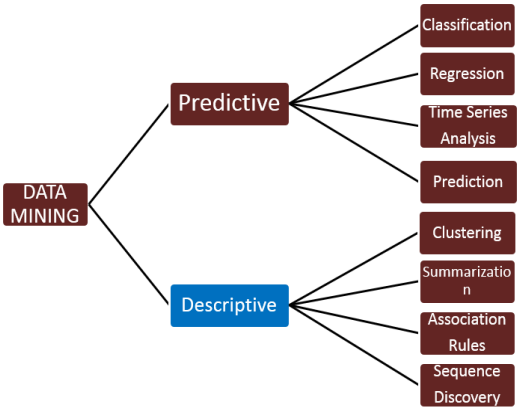

Fig 1 Types of data value Mining

Clustering

© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 07 Issue: 04 | July – September 2016
Paper is available at www.jrps.in | Email : info@jrps.in

By examining one or more attributes or classes, you may group individual pieces of data value together to form a structure opinion. At a simple stage, clustering is using one or more attributes as your basis for identifying a cluster of correlating results. Clustering is valuable to identify dissimilar info since this correlates with other examples so you may see where similarities & ranges agree. Clustering may work both ways. You may assume that there is a cluster at a certain point & then use our credentials criteria to see if you are correct .graph within Figure 3 shows a good example. within this example, a sample of sales data value compares age of customer to size of sale. this is not unreasonable to expect that people within their twenties (before marriage & kids), fifties, & sixties (when children have left home), have more disposable income.

[2] Literature Review

According to Berry and Linoff (2000), the resulting knowledge or information gained from data mining efforts can benefit an organization in increasing organizational profits by lowering costs and by increasing revenues. Data mining attempts to predict future business trends and customer behavior patterns from large data warehouses and, other forms of data resources. Data mining is clearly of value to any organization where there are large amounts of data, and something worth learning from that data.

According to Ma et al. (2000), data mining is critical to the enterprise that wants to exploit operational and other available data to improve the quality of decision making and gain critical competitive advantages. This is where data mining can play a crucial role, by disclosing important information in a cost effective and timely fashion.

Yada et al. (2000) proposed a new approach to discover the loyal customers in the future from newcomers as early as possible, using data mining tool, "C5.0" on real purchase data.

Dan Hopping (2000) emphasizes that the evolution of retailing reveals that technology has played a role as the primary enabler of change. The author discusses that there are many emerging technologies like knowledge management, data mining, customer relationship management, mathematical modeling, and data visualization that affect the future of retail. The decisions made by human affect the retailer's bottom line and the task of technology is to provide the information in a way that can help the decision makers to get the best way to get ready for the future.

Bharat Rao (2000) specifies that information technology can be used as an effective tool in managing problems related to the retail supply chain. The author gives an overview of some of the common problems retailers face while evaluating alternative technology solutions and also presented a survey of available analytical tools. The author states that the Information technology solutions can reduce costs, increase flexibility and response and provide a more effective shopping experience for the customers. By utilizing and sharing technology, retailers can develop collaborative planning, forecasting, and replenishment programs.

Chris et al. (2002) give the overview of the concept of data mining and CRM. The authors offer a closer look at mainly two data mining techniques viz. Chi-Square Automatic Interaction Detection (CHAID) and Neural Networks. As a result of comparison of two techniques, the authors concluded that CHAID is much easier and quicker to construct and understand whereas neural networks provide more accurate models, especially for complex problems.

[3] Tools & Technology

© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 07 Issue: 04 | July – September 2016
Paper is available at www.jrps.in | Email : info@jrps.in

Design Methodology

Data mining is a collection of techniques for efficient automated discovery of previously unknown, valid, novel, useful & understandable patterns in large databases. patterns must be actionable so that they may be used in an enterprise's decision making process. It is usually used by business intelligence organizations, & financial analysts, but this is increasingly used in sciences to extract information from enormous data sets generated by modern experimental & observational methods.

HARDWARE & SOFTWARE REQUIREMENT

HARDWARE

- CPU 1Ghz or more
- HARDDISK (5GB Free space)
- DVD ROM
- MONITOR (HIGH RESOLUTION)
- KEYBOARD/MOUSE

SOFTWARE

- WINDOWS 7/8
- MATLAB
- DOT NET FRAMEWORK

MATLAB AS SIMULATION TOOL

MATLAB is a high-performance language for technical computing. It integrates computation, visualization, & programming with in an easy-to-use environment where problems & solutions are expressed with in familiar mathematical notation. Typical uses include: Math & computation.

MATLAB (**mat**rix **lab**oratory) is a multi-paradigm numerical computing environment & fourth-generation programming language. Developed by MathWorks, MATLAB allows matrix manipulations, plotting of functions & data, implementation of algorithms, creation of user interfaces, & interfacing with programs written with in other languages, including C, C++, Java, Fortran & Python.

Clustering

Clustering is a process of partitioning a set of data into a set of meaningful sub-classes, called clusters. Help users understand natural grouping or structure with in a data set. Used either as a stand-alone tool to get insight into data distribution or as a preprocessing step for other algorithms.

Cluster analysis or clustering is task of grouping a set of objects with in such a way that objects within same group (called a cluster) are more similar (in some sense or another) to each other than to those within other groups (clusters). It is a main task of exploratory data mining, & a common technique for statistical data analysis, used within many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, & computer graphics.

Types of Clustering Algorithms are:

1. K-means Clustering Algorithm

2. Hierarchical Clustering Algorithm

3. Density Based Clustering Algorithm

4. Self-organization maps (SOM)

5. EM clustering Algorithm

[4] PROPOSED WORK

Density Based clustering

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

© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 07 Issue: 04 | July – September 2016
Paper is available at www.jrps.in | Email : info@jrps.in

connects points that satisfy a density criterion, in original variant defined as a minimum number of other objects within this radius. A cluster consists of all density-connected objects (which could 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 database - & that it will discover essentially same results (it is deterministic for core & noise points, but not for border points) in each run, therefore there is no need to run it multiple times

Algorithm

DBSCAN requires two parameters: ε (eps) & minimum number of points required to form a dense region[a] (minPts). It starts with an arbitrary starting point that has not been visited. This point's ε-neighborhood is retrieved, & if it contains sufficiently many points, a cluster is started. Otherwise, point is labeled as noise. Note that this point might later be found in a sufficiently sized ε-environment of a different point & hence be made part of a cluster.

If a point is found to be a dense part of a cluster, its ε-neighborhood is also part of that cluster. Hence, all points that are found within ε-neighborhood are added, as is their own ε-neighborhood when they are also dense. This process continues until density-connected cluster is completely found. Then, a new unvisited point is retrieved & processed, leading to discovery of a further cluster or noise.

The algorithm could be expressed as follows, in pseudocode following original published nomenclature:[1]

```
DBSCAN(D, eps, MinPts) {
  C = 0
  for each point P in dataset D {
    if P is visited
      continue next point
    mark P as visited
    NeighborPts = regionQuery(P, eps)
    if sizeof(NeighborPts) < MinPts
      mark P as NOISE
    else {
      C = next cluster
      expandCluster(P, NeighborPts, C, eps, MinPts)
    }
  }
}

expandCluster(P, NeighborPts, C, eps, MinPts) {
  add P to cluster C
  for each point P' in NeighborPts {
    if P' is not visited {
      mark P' as visited
      NeighborPts' = regionQuery(P', eps)
      if sizeof(NeighborPts') >= MinPts
        NeighborPts = NeighborPts joined with NeighborPts'
    }
    if P' is not yet member of any cluster
      add P' to cluster C
  }
}

regionQuery(P, eps)
  return all points within P's eps-neighborhood (including P)
```

Note that algorithm could be simplified by merging per-point "has been visited" & "belongs to cluster C" logic, as well as by in lining contents of "expand Cluster" subroutine, which is only called from one place. These simplifications have been omitted from above pseudo code in order to reflect originally published version. Additionally, region Query function need not return P in list of points to

© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 07 Issue: 04 | July – September 2016
Paper is available at www.jrps.in | Email : info@jrps.in

be visited, as long as it is otherwise still counted in local density estimate.

[5] Result and Discussion

5.1 Density Based Spatial Clustering

Density Based Spatial Clustering of Applications with Noise is of Partitional type clustering where more dense regions are considered as cluster and low dense regions are called noise. Obviously clusters are define on some criteria which is as follows

Core: Core points lie in the interior of density based clusters and should lie within Eps (radius or threshold value), MinPts (minimum no of points) which are user specified parameters.

Border: Border point lies within the neighbour hood of core point and many core points may share same border point.

Noise: The point which is neither a core point nor a border point

Directly Density Reachable: A point r is directly density reachable from s w.r.t Eps and MinPts if a belongs to NEps(s) and |NEps (s)| >= MinPts

Density Reachable: A point r is density reachable from r point s wrt.Eps and MinPts if there is a sequence of points r1….rn, r1 = s, rn = s such that ri+1 is directly reachable from ri.

DBCLASD

DBCLASD (Application Based Clustering Algorithms for Mining in Large Spatial Databases) Basically DBCLASD is an incremental approach. A point is assigned to a cluster that processed incrementally without considering the cluster.

Algorithm

1. Make set of candidates using region query
2. If distance set of C has expected distribution then point will remain in cluster
3. Otherwise insert point in list of unsuccessful candidates
4. In the same way expand cluster and check condition
5. Now list of unsuccessful candidates is again checked via condition.
6. If passes then put in cluster otherwise remain in that list

There are two main concepts in DBCLASD. First one is generating candidates and candidate generation is done on the basis of region query that specifies some radius for circle query to accept candidates. Second one is testing the candidates which is done through chi square testing. Points that lie under the threshold value are considered right candidates while those lies above threshold are remain in unsuccessful candidates‟ list. In last unsuccessful candidate list is again checked and every point go through test and points passes test are considered in cluster while left remains in unsuccessful candidates‟ list

DENLUE

DENCLUE [4] (Density based clustering) Main concepts are used here i.e. influence and density functions. Influence of each data point can be modelled as mathematical function and resulting function is called Influence Function. Influence function describes the impact of data point within its neighbourhood. Second factor is Density function which is sum of influence of all data points. According to DENCLUE two types of clusters are defined i.e. centre defined and multi centre defined clusters .In centre defined cluster a density attractor. The influence function of a data objects y Є F is a function. Which is defined in

© INTERNATIONAL JOURNAL FOR RESEARCH PUBLICATION & SEMINAR
ISSN: 2278-6848 | Volume: 07 Issue: 04 | July – September 2016
Paper is available at www.jrps.in | Email : info@jrps.in

terms of a basic influence function F, $F(x) = -F(x, y)$.

Algorithm

1. Take Data set in Grid whose each side is of 2σ
2. Find highly densed cells i.e
3. Find out the mean of highly populated cells
4. If d(mean(cl), mean(c2)) < 4a then two cubes are connected.
5. Now highly populated or cubes that are connected to highly populated cells will be considered in determining clusters.
6. Find Density Attractors using a Hill Climbing procedure.
7. Randomly pick point r
8. Compute Local 4 σ density
9. Pick another point (r+1) close to previous computed density.
10. If den(r) < den(r+1) climb
11. Put points within (σ /2) of path into cluster
12. Connect the density attractor based cluster

Comparative analysis of three density based algorithms

Name Of the Algo-rithm	Comp Lexity	Shape Of Clust ers	Input Para-meters	Hand -ling Of noise	Clus-Ter Quali-Ty (F Disag)	Ru n Tim e(m s)
DBSC-AN	$O(n^2)$	Arbit-rary	Two Input Para-meters	Not Very well	8.7%	50
DBCL-ASD	$O(3n^2)$	Arbit-rary	No Input Para-meters	Good	5.8%	130
DENC -LUE	O(log|D|)	Arbit-rary	Two Input Para-meters	Very well	15.94 %	31

Fig 2 Comparative analysis of three density based algorithms Above table shows that run time of DENCLUE algorithm is lowest while DENCLUE having highest run time. In terms of cluster quality DBCLASD leads while DENCLUE is lacking behind.

[6] Future Scope

The problem of physical design of data warehouses should rekindle interest in the well-known problems of index selection, data partitioning and the selection of materialized views. However, while revisiting these problems, it is important to recognize the special role played by aggregation. Decision support systems already provide the field of query optimization with increasing challenges in the traditional questions of selectivity estimation and cost-based algorithms that can exploit transformations without exploding the search space (there are plenty of transformations, but few reliable cost estimation techniques and few smart cost-based algorithms/search strategies to exploit them). Partitioning the functionality of the query engine between the middleware and the back end server is also an interesting problem. The management of data warehouses also presents new challenges. Detecting runaway queries, and managing and scheduling resources are problems that are important but have not been well solved. Some work has been done on the logical correctness of incrementally updating materialized views, but the performance, scalability, and recoverability properties of these techniques have not been investigated. In the short-term, the results of data mining will be in profitable, if mundane, business related areas. Micro-marketing campaigns will explore new niches. Advertising will target potential customers with new precision.

REFERENCE

[1] Mr. Dishek Mankad "The Study on Data Warehouse Design and Usage" International Journal of Scientific and Research Publications , Volume 3, Issue 3, March 2013 ISSN 2250- 3153

[2] Surajit Chaudhuri wrote on An Overview of Data Warehousing and OLAP Technology (Appears in ACM Sigmod Record, March 1997).

[3] Manjunath T. N. wrote on Realistic Analysis of Data Warehousing and Data Mining Application in Education Domain

[4] Weiss, Sholom M.; and Indurkhya, Nitin (1998); Predictive Data Mining, Morgan Kaufmann

[5] Kimball, R.The Data Warehouse Toolkit. John Wiley, 1996.

[6] Barclay, T., R. Barnes, J. Gray, P. Sundaresan, "Loading Databases using Dataflow Parallelism." SIGMOD Record, Vol.23, No. 4, Dec.1994.

[7] Blakeley, J.A., N. Coburn, P. Larson. "Updating Derived Relations: Detecting Irrelevant and Autonomously ComputableUpdates." ACM TODS, Vol.4, No. 3, 1989.

[8]Gupta, A., I.S. Mumick, "Maintenance of Materialized Views: Problems, Techniques, and Applications." Data Eng. Bulletin, Vol. 18, No. 2, June 1995. 9 Zhuge, Y., H. Garcia-Molina, J. Hammer, J. Widom, "View Maintenance in a Warehousing Environment, Proc. Of SIGMOD Conf., 1995.

[9] Roussopoulos, N., et al., "The Maryland ADMS Project: Views R Us." Data Eng. Bulletin, Vol. 18, No.2, June 1995.[11] O'Neil P., Quass D. "Improved Query Performance withVariant Indices", To appear in Proc. of SIGMOD Conf., 1997.

[10] O'Neil P., Graefe G. "Multi-Table Joins through BitmappedJoin Indices" SIGMOD Record, Sep 1995.

[11] Harinarayan V., Rajaraman A., Ullman J.D. " Implementing Data Cubes Efficiently" Proc. of SIGMOD Conf., 1996.

[12] Chaudhuri S., Krishnamurthy R., Potamianos S., Shim K."Optimizing Queries with Materialized Views" Intl.Conference on Data Engineering, 1995.

[13] Levy A., Mendelzon A., Sagiv Y. "Answering Queries Using Views" Proc. of PODS, 1995. 16 Yang H.Z., Larson P.A. "Query Transformations for PSJ Queries", Proc. of VLDB, 1987

[14] Witten, Ian H.; Frank, Eibe; Hall, Mark A. (30 January 2011). Data Mining: Practical Machine Learning Tools and Techniques (3 ed.). Elsevier. ISBN 978-0-12-374856-0.

[15] Ye, Nong (2003); The Handbook of Data Mining, Mahwah, NJ: Lawrence Erlbaum

[16] Cabena, Peter; Hadjnian, Pablo; Stadler, Rolf; Verhees, Jaap; Zanasi, Alessandro (1997); Discovering Data Mining: From Concept to Implementation, Prentice Hall, ISBN 0-13-743980-6

[17] M.S. Chen, J. Han, P.S. Yu (1996) "Data mining: an overview from a database perspective". Knowledge and data Engineering, IEEE Transactions on 8 (6), 866–883

[18] Feldman, Ronen; Sanger, James (2007); The Text Mining Handbook, Cambridge University Press, ISBN 978-0-521-83657-9

[19] Guo, Yike; and Grossman, Robert (editors) (1999); High Performance Data Mining: Scaling Algorithms, Applications and Systems, Kluwer Academic Publishers

[20] Han, Jiawei, Micheline Kamber, and Jian Pei. Data mining: concepts and techniques. Morgan kaufmann, 2006.