© INNOVATIVE RESEARCH THOUGHTS | Refereed | Peer Reviewed | Indexed

ISSN : 2454 – 308X | Volume : 06 , Issue : 05 | May 2020

Implementation Density Based clustering in Data Mining

[1]Poonam Rani,

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 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.

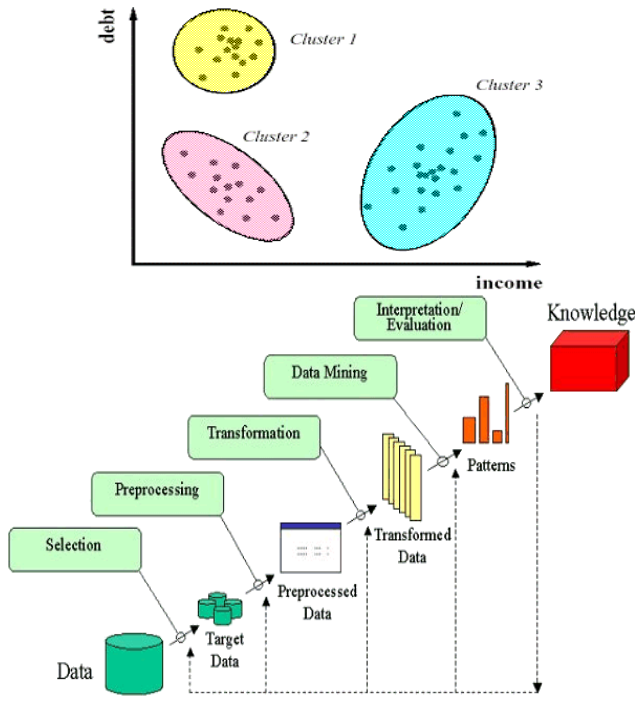

Fig 1 Data mining

Most companies already collect & refine massive quantities of data. data mining methods may be executed rapidly on existing software & hardware platforms to enhance value of existing info resources, & may be integrated with new products & systems as they are brought on-line. When executed on high performance client/server or parallel processing computers, data mining.

This white paper are delivers an over view to technologies of data value mining. Examples of gainful applications demonstrate its importance to today's business atmosphere as well as a basic description of how data value warehouse architectures may evolve to deliver value of data mining to end users.

Clustering

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

Fig 2 Clustering

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 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. Within this case, we've both hypothesized & proved our hypothesis with a simple graph that we may create using any suitable graphing software for a quick annual view. More complex determinations require a ful logical package, particularly if you want to automatically base decisions on nearest neighbor info. Plotting clustering within this way is a simplified example of so called nearest neighbor identity. You may identify individual customers by their literal proximity to each other on graph. It's highly likely that customers within same cluster also share other attributes & you may use that expectation to help drive, classify, & otherwise analyze other people from your data value set.

You may also apply clustering from opposite perspective; given certain input attributes, you may identify dissimilar artifacts. For example, a recent study of 4-digThis PIN numbers found clusters between digits within ranges 1-12 & 1-31 for first & second pairs. By plotting these pairs, you may identify & determine clusters to relate to dates (birthdays, anniversaries)

[2] Literature Review

Girish Punj & David Stewart (1983) have reviewed applications of cluster analysis in marketing problems & they have recommended a two stage cluster analysis methodology & preliminary identification of clusters via Ward's minimum variance method. authors have also discussed issues & problems related to use & validation of cluster analysis.

Agrawal et al. (1993), in their study, described that in recent past exploratory analysis in particular of large sets of market basket data has become topic of pertinent research due to various publications on data mining & knowledge in databases & generated association rules from market basket data, which describe relevant interrelations like "If a consumer purchases fruit juice, then, in 40% of cases they also purchase mineral water".

Piatetsky-Shapiro et al. (1996) surveyed a growing number of industrial applications of data mining. authors have examined existing data mining tools, described some representative applications like marketing, investment, manufacturing, fraud detection etc. & discussed issues for deploying successful application & their adoption by business

users. They also highlighted upon fact of widespread realization of potential value of data mining & a growing number of researchers & developers in this area.

Anand et al. (1996) focusedon organizations' need to investment in data mining solutions because of phenomenal expansion of data space & resulting sharp increase in size of typical data base. There is no alternative to heavy reliance on computer programs set to discover patterns for themselves with relatively little human intervention. authors also proposed a general framework of data mining based on Evidence Theory that consisted of methods for representing data & knowledge, & methods for data manipulation & knowledge discovery.

According to Richard A. Spinello (1997), Wal-Mart captures point-of-sale data from over 2,900 stores in 6 countries & transmits this data to its massive 7.5-terabyte data warehouse & uses it to identify customer-buying patterns, to manage local store inventory & identify new merchandizing opportunities using data mining techniques.

Peter Spiller & Gerald Lohse (1997), in their paper, present a classification of on-line retail stores based upon convenience sample of 137 Internet retail stores. Cluster & factor analysis identified five distinct Web catalog interface categories which provide a better understanding of strategies pursued in Internet-based marketing.

[3] 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. A typical example for a data mining scenario may be "In context of a super market, if a mining analysis observes that people who buy pen tend to buy pencil too, then for better business results seller could place pens & pencils together."

Data mining strategies could be grouped as follows:

• **Classification**- Here given data instance has to be classified into one of target classes which are already known or defined [19, 20]. One of examples could be whether a customer has to be classified as a trustworthy customer or a defaulter with in a credit card transaction data base,given his various demographic & previous purchase Characteristics.

• **Estimation**- Like classification, purpose of an estimation model is to determine a value for an unknown output attribute. However, unlike classification, output attribute for an estimation problem are numeric rather than categorical. An example could be "Estimate salary of an individual who owns a sports car.

• **Prediction**- It is not easy to differentiate prediction from classification or estimation. only difference is that rather than determining current behaviour, predictive model predicts a future outcome. output attribute could be categorical or numeric. An example could be "Predict next week's closing price for Dow Jones Industrial Average". [53, 65] explains construction of a decision tree & its predictive applications.

• **Association rule mining** -Here interesting hidden rules called association rules with in a large

transactional data base is mined out. For e.g. rule {milk, butter->biscuit} provides information that whenever milk & butter are purchased together biscuit is also purchased, such that these items could be placed together for sales to increase overall sales of each of items.

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. Cluster analysis itself is not one specific algorithm, but general task to be solved. It could be achieved by various algorithms that differ significantly within their notion of what constitutes a cluster & how to efficiently find them. Popular notions of clusters include groups with small distances among cluster members, dense areas of data space, intervals or particular statistical distributions. Clustering could therefore be formulated as a multi-objective optimization problem. appropriate clustering algorithm & parameter settings (including values such as distance function to use, a density threshold or number of expected clusters) depend on individual data set & intended use of results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial & failure. It is often necessary to modify data preprocessing & model parameters until result achieves desired properties.

[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 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.

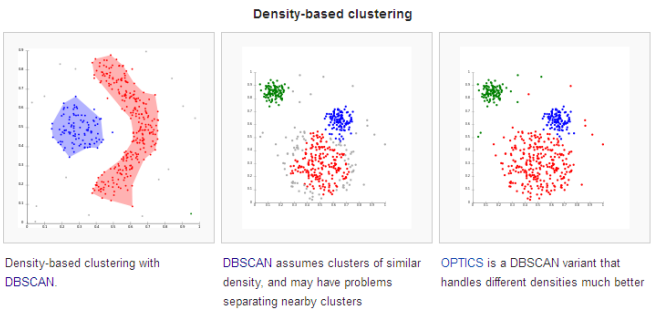

Fig 3 Density based clustering

Another interesting property of DBSCAN is that its complexity is fairly low - it requires a linear number of range queries on database - & that it would 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. OPTICS is a generalization of DBSCAN that removes need to choose an appropriate value for range parameter & produces a hierarchical result related to that of linkage clustering. DeLi-Clu, Density-Link-Clustering combines ideas from single-linkage clustering & OPTICS, eliminating parameter entirely & offering performance improvements over OPTICS by using an R-tree index. key drawback of DBSCAN & OPTICS is that they expect some kind of density drop to detect cluster borders. Moreover, they cannot detect intrinsic cluster structures which are prevalent in majority of real life data. A variation of DBSCAN, EnDBSCAN,[14] efficiently detects such kinds of structures. On data sets with, for example, overlapping Gaussian distributions - a common use case in artificial data - cluster borders produced by these algorithms would often look arbitrary, because cluster density decreases continuously. On a data set consisting of mixtures of Gaussians, these algorithms are nearly always outperformed by methods such as EM clustering that are able to precisely model this kind of data.

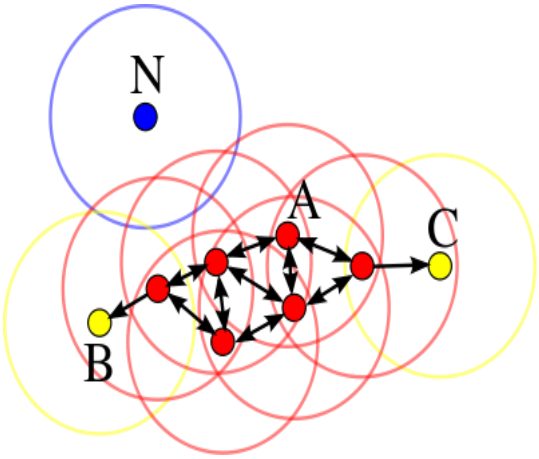

Fig 4 In this diagram, minPts = 4. Point A & other red points are core points, because area surrounding these points in an ε radius contain at least 4 points.

[5] Result & Discussion

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 & low dense regions are called noise. Obviously clusters are define on some criteria which is as follows

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

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

Noise: 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 & MinPts if a belongs to NEps(s) & |NEps (s)| >= MinPts

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

Algorithm

Steps of algorithm of DBSCAN are as follows

1. Arbitrary select a point r
2. Retrieve all points density-reachable from r w.r.t Eps & MinPts.
3. If r is a core point, cluster is formed
4. If r is a border point, no points are density-reachable from r & DBSCAN visits next point of database

5. Continue process until all of points have been processed

Noise points: A noise point is a point that is neither a core point nor a border point.

formal definition of DBSCAN algorithm is illustrated below:

1. Eliminate noise points
2. Perform clustering on remaining points
3. *current_cluster_label := 0*

for all core points **do**
If core point has no *cluster_label* **then**
current_cluster_label := current_cluster_label +1
Assign current core point *current_cluster_label*
end if
For all points within radius **do**
If point does not have a cluster_label **then**
Label point with *current_cluster_label*
end if
end for
end for

We could implement this algorithm using following codes in MATLAB:

dbscan.m

```
function [class,type]=dbscan(x,k,Eps) % x = dataset, k = no. of %points within radius & Eps as radius

[m,n]=size(x);

if nargin<3 || isempty(Eps)

  [Eps]=epsilon(x,k);

end

x = [[1:m]' x];

[m,n] = size(x);

type = zeros(1,m);

no = 1;

touched = zeros(m,1);

for i = 1:m

  if touched(i) == 0;

    ob = x(i,:);

    D = dist(ob(2:n),x(:,2:n));

    ind = find(D<=Eps);

    if length(ind)>1 && length(ind)<k+1

      type(i) = 0;

      class(i) = 0;

    end

    if length(ind)==1

      type(i) = -1;

      class(i) = -1;

      touched(i) = 1;

    end


    if length(ind)>= k+1;

      type(i) = 1;

      class(ind) = ones(length(ind),1)*max(no);


      while ~isempty(ind)
```

```
        ob = x(ind(1),:);
        touched(ind(1)) = 1;
        ind(1) = [];
        D = dist(ob(2:n),x(:,2:n));
        i1 = find(D<=Eps);

        if length(i1)>1
          class(i1) = no;
          if length(i1) >= k+1;
            type(ob(1)) = 1;
          else
            type(ob(1)) = 0;
          end

          for i = 1:length(i1)
            if touched(i1(i)) == 0
              touched(i1(i)) = 1;
              ind = [ind i1(i)];
              class(i1(i)) = no;
            end
          end
        end
      end
      no = no+1;
    end
  end
```

```
  end

  i1 = find(class == 0);
  class(i1) = -1;
  type(i1) = -1;
```

epsilon function

```
function [Eps] = epsilon(x,k) % Analytical calculation of rad if %not given
[m,n] = size(x);
Eps = ((prod(max(x)-min(x))*k*gamma(.5*n+1))/(m*sqrt(pi.^n))).^(1/n);
```

Dist function

```
function [D] = dist(i,x) %Distance Calculation
[m,n] = size(x);
D = sqrt(sum((((ones(m,1)*i)-x).^2)'));
if n == 1
  D = abs((ones(m,1)*i-x))';
end
```

plot_cluster.m

```
data = [2,1;2,2;2,3;3,2;4,4;10,3;10,7;12,4;11,5;11,3.4];
[class,type] = dbscan(data,3,3);
for k = 1:length(data)
  if(type(k)==1)
```

```
    plot(data(k,1),data(k,2),'ks');hold on;
elseif(type(k) ==0)
        plot(data(k,1),data(k,2),'bo');hold on;
else
        plot(data(k,1),data(k,2),'rx');hold on;
end
text(data(k,1),data(k,2),num2str(class(k)))
end
```

Result:

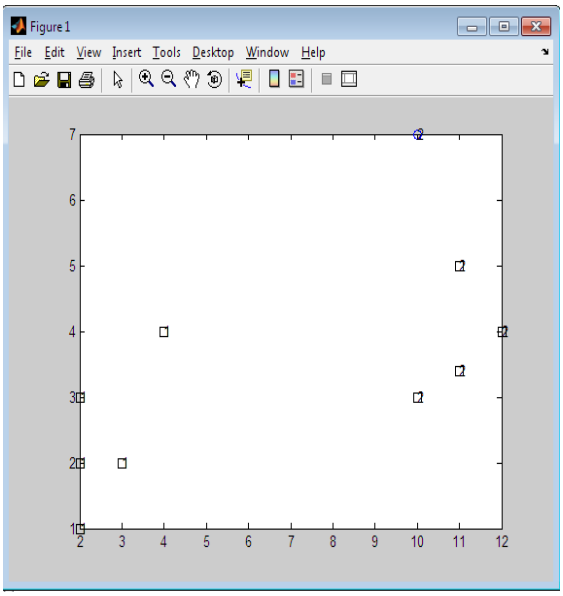

Fig 5 Results

[6] Conclusion

Before data mining algorithms could be used, a target data set must be assembled. As data mining could only uncover patterns actually present within data, target data set must be large sufficient to contain these patterns while remaining concise enough to be mined within an acceptable time limit. A common source of data is considered data mart or data warehouse. Pre-processing is essential to analyze multivariate data sets before data mining. target set is then cleaned. We have described substantial technical challenges in developing & deploying decision support systems. While many commercial products & services exist, there are still several interesting avenues for research. We would only touch on a few of these here. Data cleaning is a problem that is reminiscent of heterogeneous data integration, a problem that has been studied for many years. But here emphasis is on data inconsistencies instead of schema inconsistencies. Data cleaning, as we indicated, is also closely related to data mining, with objective of suggesting possible inconsistencies. In particular, failure & check pointing issues in load & refresh in presence of many indices & materialized views needs further research. adaptation & use of workflow technology might help, but this needs further investigation.

REFERENCE

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

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

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

[4] Weiss, Sholom M.; & 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 & Autonomously ComputableUpdates." ACM TODS, Vol.4, No. 3, 1989.

[8]Gupta, A., I.S. Mumick, "Maintenance of Materialized Views: Problems, Techniques, & 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 & Techniques (3 ed.). Elsevier. ISBN 978-0-12-374856-0.

[15] Ye, Nong (2003); 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 & data Engineering, IEEE Transactions on 8 (6), 866–883

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