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

Decision Tree Clustering and Classification Based One-To-Many Data Linkage

K.Rajeshwar Rao[1], Dr.G.Charles Babu[2]

Asst. Professor, Dept. of CSE, MREC (A), Secunderabad, India

Professor, Dept. of CSE, MREC (A), Secunderabad ,India .

ABSTRACT: Data has its own place and its own value to a valued customer and makes to for strategic decision making. Technology makes us enabling to go next level decision management system in the perspective of the mining the information which is the great boom to the Industry of Automation In the real scenario , Implementation is not that much easy to compete with various vendors who come with various unique features to attract the client. In this paper , we generically giving emphasis on the techno world concept of the mining the data in terms of domain link either in the perspective map reduced one to meant techno mechanism in implementing the algorithmic approach which may in turn make us to give the robust, effective and performance oriented solution . In this Paper we have used the map reduce program to reduce time and effectiveness of the solution to the link various domain keyword on the umbrella of one domain. Hence, we have given the architectural solution to batch based process based on the user requirement to which extend to go forward for the shortest period with strategic decision making process.

KEYWORDS: clustering, classification, data matching, and decision tree induction

I. INTRODUCTION

Generalization is the most prevalent method for table anonymity, and it can also be valuable when inducing k-anonymous decision trees. In generalization, an attribute value is replaced with a less specific but semantically consistent value. In most works, attributes are generalized using a pre-determined hierarchy of values. For example, for a Zip Code attribute, the lowest level of the hierarchy could be a 5-digit zip code. This value could be generalized by omitting the rightmost digit to obtain a 4-digit zip code. Similarly, additional digits could be omitted up to a 1-digit zip code consisting only of the left-most digit. When this digit is omitted as well, the highest level of the generalization hierarchy is reached, with a single value (usually denoted *). Practically, replacing a value with * means suppression of the value. When the market grows, usually no one cares about the churn because the churn rate is low and the customer acquisition is very large.

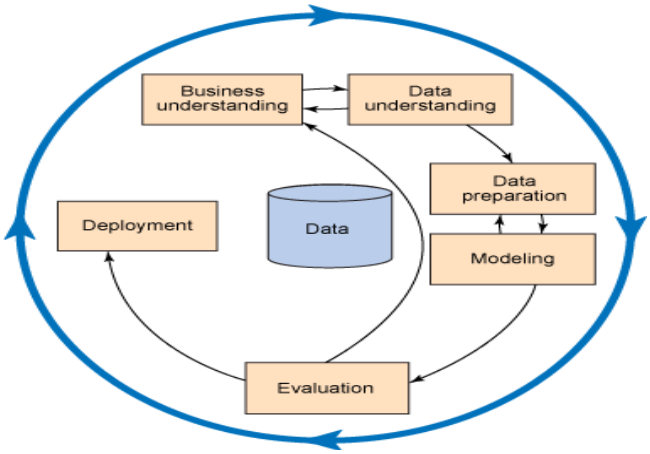

Fig.1.1. Illustration of the Cycle of the Cluster

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

The profile of each cluster is determined only by attributes which are different from population means; two clusters may have different at- tributes in their definitions. This method needs to have the data very well prepared.

II. RELATED WORK

It is often used in companies with very large data analysis activities to improve the common set of data analyses. The accepting or rejecting hypotheses task usually starts with some hypothesis (based on manager's intuition) and tries to find a suitable data which can say whether the hypothesis is correct or not. This task has various applications in most of major business activities in a organization. Classification and prediction tasks usually estimate some value for each record. The decision tree is a classifier with a very high capacity. Data mining methods require the data in a specified form, called a at table based hash mechanism. Some algorithms for the data mining have particular requirements on data. The data preparation includes all tasks which ensure that the data will be available in at table. Typical tasks solved by data analysis vary from organization to organization, from industry to industry and from country to country. The level of data analysis in a particular industry in a particular country usually depends on the level of data analysis of organization's competitors. But there are some common applications of the data analysis. The first typical application is customer retention. The churn is usually the big problem. The next clustering method is called the EM clustering, or the model based clustering. This method is based on the probability and it can handle more attributes. A minor improvement in the customer acquisition is usually much better than a churn analysis. But one day we can found that the market became exhausted and divided to conquer.

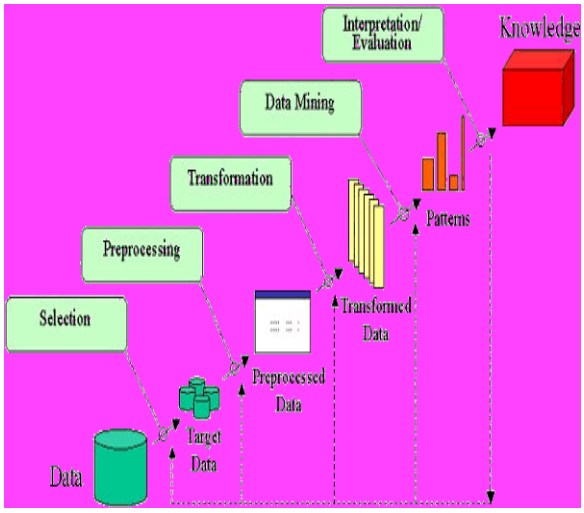

Fig.2.1. Index Based Data Mining in the Ranking Model

When we have an individual person which offers services to his customers, he usually knows his customers, he is able to estimate the profit from each individual customer, and he knows their preferences and so on. So he can estimate who is likely to churn, to whom offer a new product and other business tasks.

III. METHODOLOGY

These models try to estimate which customers are willing to buy some product. Application of propensity to buy models is in selecting the target group for direct marketing campaigns. The typical situation is that the organization has models for all key products. So it can say for each key product to who is good to offer this product. The more advanced application is that the organization estimates the best product for the customer and this is the one which the organization offers him in a direct marketing campaign. To apply the similar task in a organization which has several millions of customers is a very hard issue. The customer segmentation is dividing customers into groups, within these groups customers have similar properties like behavior, the value and preferences and between group customers have different properties. The strategy can be established for every individual segment. The effect of missing values on

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

model privacy depends on the way they have been treated during the model's induction. If, as is customary in many anonymity preserving algorithms, tuples with missing values are deleted from the database before the algorithm is executed, then the missing values will have no effect on privacy.

Algorithm 3.1 Node Mapping Cluster from One to Many Algorithms

Input: T – private dataset, A – public attributes, B – private attributes, C attribute, k – anonymity parameter
procedure MAIN()
  Create *root* node in *Tree*
  Create in *root* one bin b_c for each value $c \in C$, divide T among the bins
  if $C \in A$ then
    Create one span S_c for every value $c \in C$. $S_c.Bins \leftarrow \{b_c\}$, $S_c.Population \leftarrow b_c.Population$, $S_c.Nodes \leftarrow \{root\}$
    set *root.Spans* to the list of all spans
  else
    Create a single span s. Set $s.Bins$ to the list of all bins, $s.Population \leftarrow T$, $s.Nodes \leftarrow \{root\}$, $root.Spans \leftarrow \{s\}$
    if $0 < |s.Population| < k$ then
      return *nil*
    end if
  end if
  for $att \in A \bigcup B \setminus \{C\}$ do
    add $(root, att, \text{gain}(root, att))$ to *Queue*
  end for
  while *Queue* has elements with positive gain do
    Let $(n, a, gain) = \arg\max_{gain} \{Queue\}$
    if $n.sons \neq \emptyset$ then
      continue
    end if
    if $Breach(n, a, k)$ then
      if a has generalization a' then
        insert $(n, a', \text{gain}(n, a'))$ to *Queue*
      end if
    else
      Split(n, a)
    end if
  end while
  Set the Class variable in each leaf to the value with the largest bin.
  return *Tree*
end procedure

The suppression of tuples with missing values can be formalized by modeling it as splitting the root node on a binary public attribute that marks the existence of missing values. All tuples with missing values are routed to a distinct leaf in which no private information is provided. Since no private values are provided in this leaf, it comprises a span with a

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

single equivalence class, thereby maintaining anonymity regardless of the number of suppressed tuples. Suppression of tuples with missing values, however, might strongly bias the input and might also produce a poorer outcome as a result, especially when missing values are abundant. Another possible approach would be to treat a missing value as a value in its own right, and then analyze the privacy of the model using the methods described earlier for regular values.

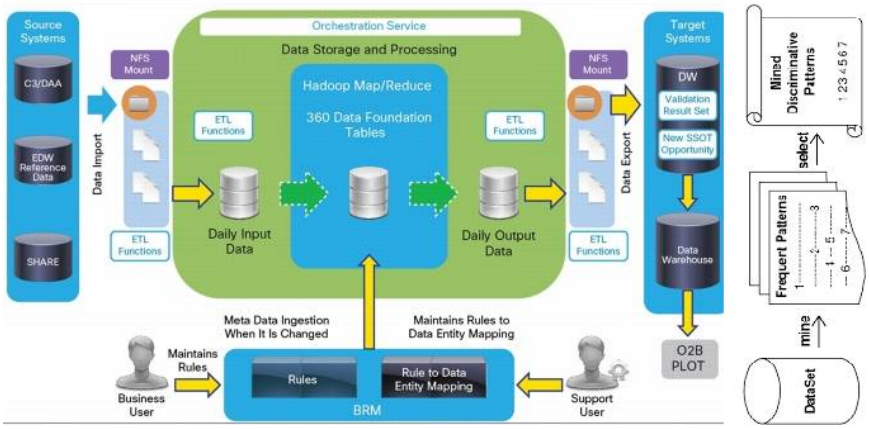

Fig.3.1. Architectural Model View of the Map Reduce Design

We should better focus on high value segments; ask how to convert customers from low value segments into high value segments, to develop products for segments and so on. Many of these results become to be regular reports. The data description task has various applications, including a costs analysis and a customer behaviour analysis. The Data Exploration task is a task where the only goal is to find some- thing that will help to solve some (usually given) business problem. The data exploration is a very interesting task, but it is not so typical, because there is no guarantee that investments into this data analysis will bring any results/benefits. The analysis of the best product for the customer is a very advanced analysis. It usually requires all propensities to buy models for individual products, profit analyses for products and the other information which also comes from the strategy of the organization. Ensuring that results of modeling are usable for the business application is a very crucial thing. We should focus on data sources available. In this phase, we should explore, which data sources are available, to obtain the data, to profile the data, attributes and their values, to identify problems with the data (quality, availability, checking whether the data are up-to-date, periodicity of data refreshment) and to exactly formulate the problem in the language of the data (to extend the problem definition with specific attribute names and values). In the Data Description task, the goal is usually to organize and visualize the data, maybe in a form in which the data were not displayed and this can bring a new knowledge to the organization. The data description task is a typical task of the data analysis. The K-means clustering is a widely used method in a situation, where we have only a few attributes (less than 10) and all attributes are important in all cluster profiles. Applications of the K- means clustering include a demographical segmentation and supplementary segmentations for prediction models.

3.1 Evaluation and Analysis

It has been chosen as the first method intentionally decision trees are what this work is about. Here we provide only a brief description. Decision trees will be de- scribed in a more detail later. The decision tree induction is a method based on two principles. The first principle is called divide and conquer. This means, that in every step, the dataset is split into two or more parts and the algorithm continues recursively on individual parts. The second principle is a greedy principle. That means that the splitting is based only on little information. Decision trees are used mainly for predictions, classifications and descriptions.

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

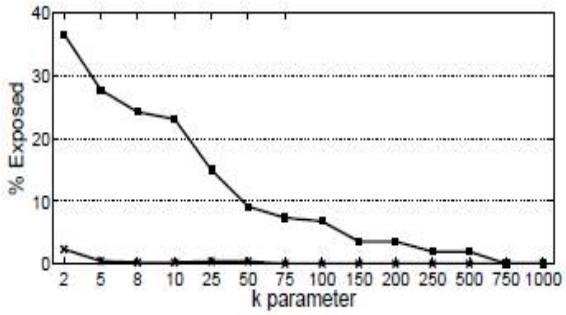

Fig.3.1.1. Comparison of the Nodes

The evaluation phase also deals with consequences of data quality problems. This phase should find any mistakes we could have made during all previous phases.

IV .CONCLUSION AND FUTURE WORK

Technology has its own limitation to extend which needs future research. A data owner who wishes to publish a decision tree while maintaining customer anonymity can use this technique to induce decision trees which are more accurate than those acquired by anonym zing the data first and inducing the decision tree later. This way, anomy is done in a manner that interferes as little as possible with the tree induction process. To the best of our knowledge, this problem was not studied before in the context of anatomization and privacy. According to our experiments, the utility of handling missing values may change in accordance with the specific data set at hand and the anonymity parameter. Nevertheless, the inability to handle missing values might render data sets with abundant missing values unusable.

REFERENCES

[1] I.P. Fellegi and A.B. Sunter, "A Theory for Record Linkage," J. Am.Statistical Soc., vol. 64, no. 328, pp. 1183-1210, Dec. 1969.
[2] M. Yakout, A.K. Elmagarmid, H. Elmeleegy, M. Quzzani, and A.Qi, "Behavior Based Record Linkage," Proc. VLDB Endowment, vol. 3, nos. 1/2, pp. 439-448, 2010.
[3] J. Domingo-Ferrer and V. Torra, "Disclosure Risk Assessment in Statistical Microdata Protection via Advanced Record Linkage," Statistics and Computing, vol. 13, no. 4, pp. 343-354, 2003.
[4] F. De Comite´, F. Denis, R. Gilleron, and F. Letouzey, "Positive and Unlabeled Examples Help Learning," Proc. 10th Int'l Conf.Algorithmic Learning Theory, pp. 219-230, 1999.
[5] M.D. Larsen and D.B. Rubin, "Iterative Automated Record Linkage Using Mixture Models," J. Am. Statistical Assoc., vol. 96, no. 453, pp. 32-41, Mar. 2001.
[6] S. Ivie, G. Henry, H. Gatrell, and C. Giraud-Carrier, "A Metric-Based Machine Learning Approach to Genealogical Record Linkage," Proc. Seventh Ann. Workshop Technology for Family History and Genealogical Research, 2007.
[7] A.J. Storkey, C.K.I. Williams, E. Taylor, and R.G. Mann, "An Expectation Maximisation Algorithm for One-to-Many Record Linkage," Univ. of Edinburgh Informatics Research Report, 2005.
[8] P. Christen and K. Goiser, "Quality and Complexity Measures for Data Linkage and Deduplication," Quality Measures in Data Mining, vol. 43, pp. 127-151, 2007.
[9] P. Langley, Elements of Machine Learning. Morgan Kaufmann, 1996.
[10] H. Blockeel, L.D. Raedt, and J. Ramon, "Top-Down Induction of Clustering Trees," ArXiv Computer Science e-prints, pp. 55-63, 1998.
[11] D.J. Rohde, M.R. Gallagher, M.J. Drinkwater, and K.A. Pimbblet, "Matching of Catalogues by Probabilistic Pattern Classification," Monthly Notices of the Royal Astronomical Soc., vol. 369, no. 1, pp. 2-14, May 2006.
[12] L. Gu and R. Baxter, "Decision Models for Record Linkage," Data Mining, vol. 3755, pp. 146-160, 2006.
[13] P. Christen and K. Goiser, "Towards Automated Data Linkage and Deduplication," technical report, Australian Nat'l Univ., 2005.
[14] E. Frank, M.A. Hall, G. Holmes, R. Kirkby, and B. Pfahringer, "WEKA - A Machine Learning Workbench for Data Mining," The Data Mining and Knowledge Discovery Handbook, pp. 1305-1314, Springer, 2005.
[15] J.R. Quinlan, C4.5: Programs for Machine Learning. Morgan Kaufmann, 1993.
[16] O. Benjelloun, H. Garcia, D. Menestrina, Q. Su, S. Whang, and J. Widom, "Swoosh: A Generic Approach to Entity Resolution," The VLDB J., vol. 18, no. 1, pp. 255-276, 2009.
[17] S.E. Whang and H. Gercia-Molina, "Joint Entity Resolution," technical report, Stanford Univ., 2009.

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

BIOGRAPHY

[1] **Mr. K. RAJESHWAR RAO** (ISTE,CSI Life Member), working as Asst. Professor in Department of Computer Science and Engineering, Malla Reddy Engineering College, MREC (A) ,Secunderabad , Telangana State. He has rich teaching experience and his research domains are Data mining, cloud computing, Computer Networks ,software engineering.

[2]**Dr. G.Charles Babu** completed B.Tech from KLCE in 1997 and M.Tech from JNTU in 1999.He Completed his Ph.D in ANU, Guntur in 2015.He has 17 Years of Teaching Experience and Presently Working as a Professor in MREC(Autonomous) , Hyderabad. His Research areas are Data Mining, Cloud Computing, Networks and Image Processing.