International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

Classification of Data Using LAD

Aishwarya Jadhav[1], Vaishali Nandedkar[2]

[1, 2]Department of Computer Engineering, SavitribaiPhule Pune University, Pune, India

Abstract: ***One basic and important part in data analysis is data classification. Data mining is a technique used in various domain to give meaning to the available data classification is a data mining (Machine learning) technique used to predict group membership for data instances. The proposed approach is for producing fast and accurate data classification, learning from small set of record which is already classified. The concept is based on Logical analysis of data(LAD) framework, LAD is enriched with information obtained from statistical consideration on the data. The accuracy of the proposed approach is compared to the LAD algorithm, of support Vector machine on publicly available datasets of UCI repository.***

Keywords: Data mining, Machine Learning, Optimization, Classification, Neural Network.

1. Introduction

Information mining studies are worried with separating significant learning from substantial scale datasets. While there are different information mining systems, extensive number of standard ideas shows up in some structure in numerous information mining applications.

With the approach of new advancements research in different fields has been moved from theory headed to datadriven and characterization issue has ended up pervasive in some true applications that require segregation among predefined classes. Surely understood characterization calculations, for example, bolster vector machines (Burges 1998; Scholkopf and Smola2001)[1], neural systems (Fausett1994; Bishop 2007)[2], choice trees (Bishop 2007; Duda, Hart, and Stork 2001), k-Nearest Neighbor (Aiolli 2004; Mitchell and Schaefer 2001), and Naive Bayes (Duda, Hart, and Stork 2001; Bishop 2007), and so on., are intended to take care of parallel arrangement issues where a learning model is developed to partitioned perceptions in two predefined classes.

At the point when preparing information is extensive it involve more measure of Data so the classifier is more exact. Be that as it may, in numerous application information is not totallymarked it require to be order utilizing little arrangement of preparing information. On the other side unlabeled information is anything but difficult to gather. Thusly, procedures have been produced for enhancing an order by utilizing likewise a lot of unlabeled information, that is called approval set. Those systems can be brought into a few of the methodologies, acquiring semi-regulated classifiers. All above recorded methodologies for order are work best for various sort of dataset means no single calculation can give the best execution on the all datasets and this is the significant drawback. On the other side one Boolean approach for characterization of information is the Logical Analysis of Data(LAD). This is propelled by the mental procedure of human. People gain from short cases and little measure of data is adequate for them to group data. In this paper same methodology of order is utilized to execute LAD.

2. Related Works

In paper [1], The scientific establishment of LAD is indiscrete arithmetic, with an extraordinary accentuation on the hypothesis of Boolean capacities. To start with the assessment of the every cut point is assessed separate for numerical field and paired properties. Preparing sets are utilized to enhance the bolster set. In a related work, Boros et al. [5] consider the issue of discovering fundamental characteristics in double information, which again lessens to finding a little bolster set with a decent partition power. In this paper [2], Author have exhibited order of the test set by looking at weighted aggregate of the enacted examples to a suitable characterization limit. To minimize the mistakes creator propose to figure both the estimations of design weights and the estimation of arrangement limit. Design weights and arrangement edge are actually parameters for the characterization strategy. In this paper [3], Author has allude basically standard system as depicted in [6]. Model different such suitability criteria as halfway preorders characterized on the arrangement of examples. In this paper [4] Author demonstrates a few strategies for managing cut focuses on consistent fields who has typical (Gaussian) circulation, on discrete fields having binomial (Bernoulli) appropriation, or on general numerical fields having obscure dissemination. In this paper [5] After example are created, calculation of example weights and arrangement edge by utilizing the proposed blended whole number model is depicted. Utilizing publically accessible dataset of UCI storehouse information is ordered.

3. Proposed System

With the advent of new technologies research in various fields has been shifted from hypothesis-driven to data driven and classification problem has become ubiquitous in many real-world applications that require discrimination among predefined classes. Fig.1 shows the system architecture also known as general block diagram modules are as follows:

(A) Binarization
(B) Support Set
(C) Pattern Generation.

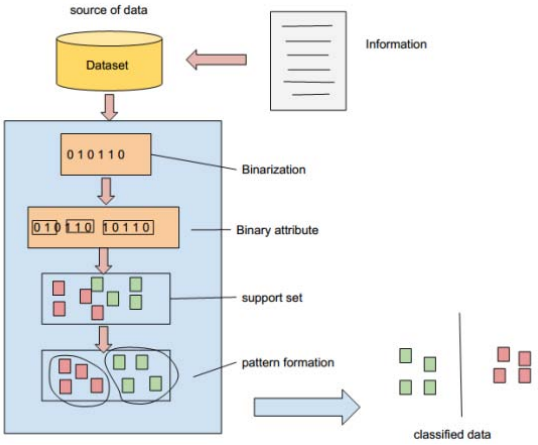


Figure 1: Proposed System Architecture

Each block of the above system is described in detail as follows

(A)Binarization: The logical analysis of data originally developed for the analysis of data sets whose attributes take only (0-1) values. The purpose of binarization is the transformation of a database of any type into a "Boolean Database". Hence data is encoded into the Binary form by means of discretization process. The key process in discretization is the selection of intervals which can be determined by an expertise in the field. Also this done by using training set for computing specific values for each field called cut points, that cut points in case of numerical fields split into Binary attributes.

(B)Support Set: The selected Binary attribute constitute a support set and are combined for generating logical rules called Patterns. Basically support set used to reduce the size of Binary archive by eliminating as many redundant attributes as possible, Those set of binary attribute is called as support set if the archive obtained by the elimination of all the other attributes will remain "contradiction free" means it will not contain simultaneously true and false observations.

The main reason to eliminate the variables and use smaller support set is to reduce the computational complexity of pattern and theory generation.

It is clear that the smaller the chosen support set, the less information we keep, and therefore the less classification power we have. It is therefore important to keep a healthy balance between the discriminating power and the computational cost of later steps. Not only that the discriminating power of generated support set is not expected to be smaller, but also the computational cost of an approximation algorithm is much small than that of an exact method.

(C)Pattern generation: Binary attributes are combined for generating logical rules called Patterns. Patterns are used to classify each unclassified record on the basis of the sign of weighted sum of the pattern activated by the record.

The structure of records, called record scheme R, consists of a set of fields fi, with i = 1, . . ., m. A record instance r, also imply called record, consists of a set of values vi, one for each field. A record r is classified if it is assigned to an element of a set of possible classes C. Positive record instance is shown by r+ and negative is by r−. Training set „S" of classified record is given. Can be denoted as |s| + and |s| −. Set of records used for evaluating the performance of the learned classifier is called test set T. Real classification of each record t €T.

Lad methodology starts by converting all fields to binary form. This process called binarization. Here converts each non binary field fi into binary. Thus, Set of binary attributes a_j^i where j = 1 . . . ni.A binarized record scheme Rb.Binary record instance rb.
Set of binary values b_j^i €0, 1
Rb={a_1^1,a,a_m^1,…,$a_m^n m$}
rb={b_1^1,b,b_m^1,…,$b_m^n m$}

Cut-points a_j^i should be set at values representing some kind of watershed for the analyzed phenomenon. Generally, a_j^i are placed in the middle of specific couples of data values $v_i^{'}$ and $v_i^{''}$

$$a_j^i = v_i^{'} + v_i^{''}/2$$

Positive patters are generated from top down removing one after one literals from the conjunction of literals covering single positive record until no negative record cover. Set of indices H+ positive pattern Set of indices H− negative pattern
Finally each record is get classified into positive and negative value of the weighted sum.

4. Algorithm Used

A. Binarization Algorithm:

Algorithm 1: Binarization

Input: Observation data or training data.
Output: Binarized data table/matrix of mxq
Method:
Step 1: check for the attribute(x) data type for nominal or numerical.
Step 2: if the attribute is nominal then set the count for values Vs. Check the condition x = Vs if yes return 1, if no return 0.
Step 3: If attributes are numerical select the cut point. Check for the type of Boolean variable.
Level variables : b(x ,t) = { 1 if x ≥ t , 0 if x < t.}
Interval variables :b(x, t`, t``) = { 1 if t` ≤ x < t``, 0 Otherwise.}
Step 4: If variable are Interval type then select the interval cut point and check the condition for return binary value.
Step 5: Formation binary data matrix

Support Set Algorithm:

Algorithm 2: Support Set

Input: m x q matrix.
Output: Inpainted Video
Method:
Step 1: Start.

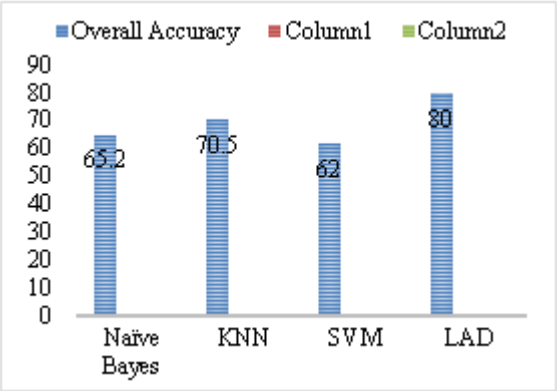
Step 2: Examine the true / false point of the matrix table. If Boolean variable y_i to attribute bi has y_i = 1 attribute is in support set, if y_i = 0 then attribute is omitted.
Step 3: Define true and false vectors p` and p`` the subset I(p`, p``). select the smallest set using set covering problem.
Step 4:assure the distinguish of true and false point by right side constraints
Step 5: Estimate the discrimination capability of individual attribute, using weights in the objective function.
Step 6: store the support set in q - vector.
Step 7: Stop

B. Pattern Generation Algorithm:

Algorithm 2: Pattern Generation

Input: Set of positive and negative Boolean points. Maximum degree for patterns to be generate
Output: prime pattern
Method:
Step 1: Start.
Step 2: examine the Cd term in the lexicographic order, generate each term only once.
Step 3: Generate term of degree d+1 from Cd by addition of all possible ways with term T ∈ Cd.
Step 4: Assume indices of of literal in T be i_d,check T` can be obtained by adding to T.
Step 5: check T``∉ Cd, if prime term or Candidateterm.
Step 6: If T`` ∈ Cd then it is already accepted.
Step 7: Output prime pattern
Step 10: Stop

5. Results



6. Conclusion

To arrange in brief times with a decent levelof precision on the premise of little preparingsets is required in an assortment of down to earth applications. Sadly, getting these three alluring elements together can be exceptionally troublesome. We consider here the structure of the Logical Analysis of Data (LAD), and propose a few improvements to this system in view of measurable contemplations on the information.

7. Possible Future Work

The research on LAD is at its beginning, and that much further reaserch is needed for a better understanding of the mathematical and computational aspects of LAD, as well as its domains of applicability

8. Acknowledgment

I am grateful to the principal for encouragement to carry out this work and alsoI take this golden opportunity to owe deep sense of gratitude to Prof. V.S. Nandedkar, for her instinct help and valuable guidance with a lot of encouragement throughout the research. I would like to thank you for encouraging my research.

References

[1] Renato Bruni, Gianpiero Bianchi, "Effective Classification using a small traning set based on Discretization and Statistical Analysis," IEEE Transactions on Knowledge and Data Engineering, 2015
[2] L. Wang, T. I slam, T. Long, A. Singhal, and S. Ja jo dia, "An Efficient Method for Internet Traffic Classification and Identification using Statistical Features", IJERT, 2015
[3] Alexander J. Stimpson, Mary L. Cummings, "Assessing Intervention Timing in Computer-Based Education Using Machine Learning Algorithms", IEEE Access, 2014.
[4] S.Neelamegam, Dr.E.Ramara j, "Classification algorithm in Data mining: An Overview", IJPTT, 2013
[5] Nanda Gopal Reddy, Roheet Bhatnagar, "Data Mining Techniques for logical Analysis of Data in Content Based Image Retrieval System ",IJCSEE,2013.

Author Profile

Miss. Aishwarya B. Jadhav Completed her B.E. from Pune University, Interested in Data Mining Area, M.E. pursuing from Dept. of Computer Engineering, PadmabhooshanVasantdadaPatil, Institute of Technology, Bavdhan, Pune 028, Maharashtra, India.

Prof. V. S. Nandedkar is Assistant Professor, Deptartment of Computer Engineering, PadmabhooshanVasantdadaPatil, Institute of Technology, Bavdhan, Pune-21, Maharashtra, India.