Volume 7, Issue 3, March 2017 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Logical Analysis of Data-A Survey Paper

Himani Chauhan, Garima Saxena, Arpit Tripathi
Department of Computer Science, Galgotia College of Engineering & Technology, Greater Noida, Uttar Pradesh, India

Abstract- Logical analysis of data is a methodology used for analyzing observations and detecting structural information which provides a solution to problems such as classification, marketing, feature selection development of pattern-based decision support systems, detection of inconsistencies in databases etc. There are a number of areas such as statistics, clustering, machine learning etc that are parallel to problems evaluated by LAD. However, it plays a significant role in the field of data classification through systematic identification of 'patterns' in the datasets There are a large number of real world data analysis applications such as economics, oil exploration, medical diagnosis etc. that can be formulated using Logical analysis of data. Although, in the recent years medical and related disciplines have been the focus of LAD and there have been a considerable number of medical problems to which LAD is successfully applied. The goal of this paper is to provide an overview and a comparative study between different works related to Logical analysis of data (LAD) in the field of data classification. It gives an insight of how Logical analysis of data (LAD) has been successfully applied to various data analysis applications particularly medical science and related disciplines. It provides a brief description of some useful work done in the past in the field of data classification to provide an accurate data classification using a small training set. It particularly gives an overview about the work done by P.L Hammer and Renato Bruni and Gianpiero Bianchi in the field of data classification. It explores about the characteristics and challenges involved with the past research work and also contributes significant variations to them to provide a fast and accurate data classification.

Keywords- Data mining, classification algorithms, patterns, decision support, binarisation

I. INTRODUCTION

Logical analysis of data is a methodology used for analyzing observations and detecting structural information which provides a solution to problems such as classification, marketing, feature selection, development of pattern-based decision support systems, detection of inconsistencies in databases etc. There are a number of areas such as statistics, clustering, machine learning etc that are parallel to problems evaluated by LAD. However, it plays a significant role in the field of data classification through systematic identification of 'patterns' in the datasets [1] [2] [3]. There are a large number of real world data analysis applications such as economics, oil exploration, medical diagnosis etc. that can be formulated using Logical analysis of data. Although, in the recent years medical and related disciplines have been the focus of LAD and there have been a considerable number of medical problems to which LAD is successfully applied [4] [5] [6].

The fundamental concept of LAD is systematic identification of patterns from a supervised data set (training set).The available data set consists of feature vectors with a class label assigned to them i.e. the training set consists of an array of observations which are formerly categorized as positive and negative. LAD methodology comprises of detecting patterns from the training set to formulate a classifier which then is used to categorize new data points so far unseen. The classifier is learned using the information from the training set and is tested on a set of data known as test data set. It is ensured that the new records are classified in a way such that the result is consistent to the past classifications.

There are a number of other data analysis approaches that have emerged in the last decades and have been strongly applied to various data classification problems. The established ones include Support Vector Machine (SVM), Decision Trees, and Neural Networks, Nearest Neighbours etc which are based on different data models. A number of experiments have been performed onto different practical problems and it has been observed that LAD has a comparable accuracy to the most frequently used data analysis approaches. It has been found that results of data classification using LAD closely resemble to the results of four most frequently used approaches namely SVM, Nearest Neighbours, Decision Trees and Neural Networks. In the field of medical science and related disciplines LAD produces a better result which has established the use of LAD in many data analysis applications. During data classification, a new observation is either classified positive or negative depending on its relationship to one of the classes. The observations that exhibit a positive character are classified as positive and those which exhibit negative character are classified as negative. Most of the data analysis applications classify a new record without any explanation but on contrary to these methods, LAD provides an explicit justification of the reasons of this categorization and thus presents a good reason to stand among them [5].This justification of the classification into either negative or positive can be very useful to reach a decision in the field of medicine for everybody who is associated in the healthcare sector.

II. LITERATURE REVIEW

The concept of logical analysis of data initially involved the case of binary data i.e. the attributes that take only values '0' and '1', however it was later realized that most of the real world data analysis applications involved more of categorical as well as real data. So, this approach was further extended to non binary data in the research work done by P.L Hammer [1] [2]. A number of research works has been done in this field afterwards, and indeed P.L. Hammer was the motivation whose true endeavor encouraged the methodology to find practical utility in data analysis applications. As described in the work of P.L Hammer, the non binary data first has to be encoded into binary form through the process of binarisation. Each field (attribute) in the training set is partitioned into binary attributes by means of some specific values called 'cut-points'. These cut-points assign several binary attributes to each numerical value of feature vector. A subset of binary attributes is selected to form a support set such that the positive and negative observations are disjoint. The size of the chosen support set both decides the computational cost of the algorithm as well as the classification or discrimination power. The small size of set implies less computational cost as well as less discriminating power. So, the main focus here was to keep a healthy balance between the classification power and computational cost of the algorithm.

The fundamental aspect of LAD involves the exploration of hidden patterns from the training set. A combination of such patterns is used to formulate a classifier which classifies new records on the basis of past observations. The main focus during pattern generation involves a rational selection of patterns such that each and every positive (negative) observation is covered by at least one positive (negative) pattern .However, it has to be ensured that the number of patterns selected must not be too large. A combination of large number of patterns leads to an ineffective use of patterns, and also there are some patterns that do not occur frequently and they become redundant when large number of patterns is considered. To classify new data points a weighted sum of positive and negative patterns known as 'discriminant' is considered; the weight of the pattern simply implies the relative importance of the patterns. The positive (negative) value of the discriminant indicates that the new observation exhibits a positive (negative) nature. There may be certain situations when there is no strong evidence about the character of patterns, in such cases the classification does not solely depends on the value of discriminant. For the classification of such records a threshold value is taken into account such that no observation is wrongly categorized. This ensures that the number of classification errors is minimal i.e. the accuracy of the classifier is maximized.

A number of researches have been done emphasizing on the work of P.L Hammer [3] [4] [8] and determining the accuracy of techniques for logical analysis of data [7]. However LAD can be applied to a number of data analysis applications; a significant amount of work has been done in the field of medical science and related disciplines. A number of combined studies have been done with researchers from National Institutes of Health (NIH), New Jersey Center for Biomaterials, Robert Wood Johnson Medical School, and University of Grenoble in an effort to apply LAD to a series of medical data. Presently, a large number of additional works is ongoing in this area such as research work in NIH Clinical Center for Radiology and Imaging Sciences, the NIH Clinical Proteomic Applications Center, Semelweiss Medical University (Budapest) etc. so that LAD can be effectively used for the analysis of medical data

One of the recent works in relation to Logical Analysis of data has been initiated by Renato Bruni and Gianpiero Bianchi [8]. The proposed work deals with the improvement of standard LAD algorithm through statistical considerations on data. The framework of proposed methodology is based on the usual LAD algorithm proposed by P.L Hammer with slight variations in the concept of support set minimization. A set of binary attributes (support set) is created in a similar fashion with the use of cut-points through the process of binarisation. However, the support set is minimized by eliminating redundant attributes such that the quality information is upholded. To ensure this, quality of each attribute is computed by calculating the quality of subsequent cut-points. The quality of each cut-point denotes its separating power i.e. how well a given cut-point fairly classifies observations as positive or negative. The weighted sum of selected attributes is minimized where weights being the reciprocal of quality, such that the 'uselessness' of the cut-points is minimized. A number of tests are performed which has shown that the concept of evaluating quality values improves the performance of standard LAD algorithm. However the calculation of quality values requires some additional time but it becomes irrelevant when performance of algorithm is taken into account. A bottom-up-top-down approach is employed to detect the hidden patterns from the given supervised data set. Firstly a bottom-up approach is used to find a combination of patterns such that each observation is covered by at least one pattern but it could leave some observations that remain uncovered by any pattern. So, to enclose these uncovered observations, a top-down approach is used. The unclassified records are then classified using these patterns. This methodology uses a similar criterion of 'discriminant' to decide the class label of ambiguous observations. The proposed algorithm produces a fast and accurate data classification using a small training set and gives sound results when tested.

In order to minimize the support set an approach based on correlation analysis can also be used such that all the relevant attributes can be maintained while the size of support set is kept minimal. Correlation analysis between the attributes and the result (class) or between the attributes gives an interpretation of how the attributes are related to each other and what impression they have on the result or outcome. Using this information, the irrelevant or redundant attributes can be removed and the support set is minimized such that there is no loss of relevant information.

III. CONCLUSION

The standard LAD algorithm as described when tested on a number of standard datasets gives a comparable performance to that of the best techniques reported. Its wide acceptability, high classification accuracy, robustness with respect to noise and strong explanatory powers makes it stand among the well established approaches in this area. Improvements have been done in the standard LAD algorithm for the betterment of its accuracy and significant variations

have been proposed to produce a fast and accurate classification. The discussion here revolves around providing a fast and accurate data classification from possibly a small training set with high accuracy and low computational cost. The work of P.L Hammer focused on maintaining a balance between the two but there was an opportunity for further advancement in support set minimization which was well seized by Renato Bruni and Gianpiero Bianchi. It is computationally necessary to select a small support set but that excluding relevant attributes means losing information which is not feasible. The computation of quality of each attribute ensures that no relevant information is lost but it also has its own concerns. It can be observed that the attributes that have poor isolated effect are discarded irrespective of their combined effect with other attributes during generation of patterns.

A number of researches have been done since P.L Hammer and there is still a scope of further improvements in the LAD algorithm to get an accurate and feasible data classification. Different variations have been proposed to minimize the size of the support set to get a computationally fast and accurate formulation of LAD algorithm, however there is still an opportunity to seize LAD has found its way in the medical and related disciplines and there is still a lot to be achieved in other areas.

ACKNOWLEDGEMENT

The authors are thankful to teachers of Galgotia's College of engineering and technology particularly Mr. Lucknesh Kumar and Mr. Manish Singh, Asst. Professor, Department of computer Science and Engineering for their constant support and guidance throughout the implementation of work. We are also thankful to the God almighty for all the blessing received during this endeavor. Last, but not the least we are very thankful to all our friends for the support and encouragement they have given us during the course of our work.

REFERENCES

[1] E.Boros, P.L. Hammer, T.Ibaraki, and A.Kogan (1997), *Logical analysis of numerical data*. Mathematical Programming,79(1-3):163-190.

[2] E.Boros, P.L.Hammer, T.Ibaraki and A.Kogan, An *implementation of logical analysis of data*,IEEE transactions on knowledge and data engineering, Vol.12, No. 2, March/April 2000

[3] G. Alexe, S.Alexe, T.O.Bonates, *A.Kogan, Logical Analysis of Data – a vision of P.L.Hammer*, Annals of Mathematics Artificial Intelligence, April 2007, Volume 49

[4] A.Reddy, H.Wang, Hua Yu3, T.O. Bonates, V. Gulabani, J. Azok, G. Hoehn, P. L Hammer, A. E Baird and King C Li, *Logical Analysis of Data (LAD) model for the early diagnosis of acute ischemic stroke*, BMC Medical Informatics and Decision Making 2008, 8:30

[5] Peter L. Hammer · Tib´erius O. Bonates, *Logical analysis of data—An overview: From combinatorial optimization to medical applications*, Annals of operational research,November 2008,volume 128

[6] Sorin Alexe, Eugene Blackstone, Peter L. Hammer, Hemant Ishwaran, Michael S. Lauer,Claire E. Pothier Snader, *Coronary Risk Prediction by Logical Analysis of Data*, Annals of operational research March 2003, Volume 119, Issue 1, pp 15–42

[7] Martin Anthony, *Accuracy of techniques for the logical analysis of data*, Department of Mathematics London School of Economics Houghton Street London WC2A2AE, UK

[8] Renato Bruni and Gianpiero Bianchi. *Effective Classification using a small Training Set based on Discretization and Statistical Analysis*, IEEE transactions on knowledge and data engineering

