**Journal of AL-Qadisiyah for computer science and mathematics**
**Vol.6 No.1 Year 2014**

**Lamia.A**

**Page 52-56**

# Predictive Model based on Logical Analysis of Data

**Lamia Abed Noor , College of Computer & Mathematics Sciences**

**Recived :18\11\2013 Recived : 29\12\2013 Accepted :31\12\2013**

**Abstract –**Logical Analysis of Data is a technique that used in finding a specific pattern. In Data mining, using this technique through learning process with available data set in order to extract a useful pattern. Then use it in different tasks. So in this paper, we apply this framework in construction a predictive model based on the useful extracted patterns concerned with the pattern sets generation. The practical work testing this predictive model, would be performed based on data set of disease that convenient with the nature of this technique. The data would be obtained from UCI web site for experimented data set.

**Keywords:** data analysis, data mining, patterns, classification, predictive model, LAD

## I. INTRODUCTION

Predictive modeling, which is perhaps the most-used subfield of data mining, draws from statistics, machine learning, database techniques, pattern recognition, and optimization techniques. The data analysis (or mining) algorithms can be divided into three major categories based upon the nature of their information extraction. These include predictive modeling, clustering (segmentation), and frequent pattern extraction[1].

Data mining involves the use of sophisticated data analysis tools to discover previously unknown, valid patterns and relationships in large data set. These tools can include statistical models, mathematical algorithm and machine learning methods. Consequently, data mining consists of more than collection and managing data, it also includes analysis and prediction[2]. Predictive analytics is the branch of data mining concerned with the prediction of future probabilities and trends. The central element of predictive analytics is the predictor, a variable that can be measured for an individual or other entity to predict future behavior. There are different techniques used in predictive, classification is one of advanced predictive analytics techniques is used attributes in data to assign an object to a predefined class or predict the value of a numeric variable of interest [3]. Although the LAD was developed specifically for classification model building, it is shown that the LAD patterns can effectively be used as input variables to other classification methods[4].

Logical analysis of data (LAD) is a special data analysis methodology
which combines ideas and concepts from optimization, combinatory, and Boolean functions. The central concept in LAD is that of *patterns*, or rules, which were found to play a critical role in classification, ranked regression, clustering, detection of subclasses, feature selection and other problems. The research area of LAD was defined and initiated by Peter L. Hammer, who was the catalyst of the LAD oriented research for decades, and whose consistent vision and efforts helped the methodology to move from theory to data analysis applications, to achieve maturity and to be successful in many medical, industrial and economics case studies[5].

The key features of the Logical Analysis of Data (LAD) are the discovery of minimal sets of features necessary for explaining all observations and the detection of hidden patterns in the data capable of distinguishing observations describing "positive" outcome events from "negative" outcome events. Combinations of such patterns are used for developing general classification procedures[6].

Mining small, useful, and high-quality sets of patterns has recently become an important topic in data mining. The standard approach is to first mine many candidates, and then to select a good subset. However, the pattern explosion generates such enormous amounts of candidates that by post-processing it is virtually impossible to analyze dense or large databases in any detail[7].

## II.LOGICAL ANALYSIS FOR DATA

The Logical Analysis of Data (LAD) is a methodology developed since the late eighteens, aimed at discovering structural information in data set. LAD was originally developed for analyzing binary data by using the theory of partially defined Boolean function[8]. LAD deals with data set $R^n$, consists of positive and negative observations containing (n) binary features. The basic idea of LAD is to apply a "decomposition/aggregation" approach to a subspace of $R^n$ containing the positive and negative observations. In the "decomposition" step, a family of small subsets(patterns) of $R^n$, each having the same simple structural properties and having strong positive or negative characteristics, is identified. In the "aggregation" step, unions of some of such positive (respectively, negative) subsets are proposed as approximations of the areas of $R^n$ containing the positive (respectively, negative) "new" and "old" observations[5].

In the case of a binary dataset Ω consists of two groups of observations, i.e. subset $\Omega^+$ contains observations labeled positive and subset $\Omega^-$ contains observations labeled negative. So $\Omega = \Omega^+ \cup \Omega^- \subseteq \{0, 1\}^n$, with $\Omega^+ \cap \Omega^- = \theta$ [4]. By assuming each observation is represent as a vertex in a cube, a *pattern* is a simply a homogeneous subcube, i.e., a subcube having (i) a nonempty intersection with one of the sets Ω+ or Ω−, and (ii) an empty intersection with the other set (Ω− or Ω+, respectively). We recall that a subcube consists of those points of the *n*-cube for which a subset of the variables is fixed to 0 or 1, while the remaining variables take all the possible 0,1 values[9]. $\Omega^+_S$ covers with a family of (possibly overlapping) homogeneous subsets of the reduced real space, each of these subsets having a significant intersection with $\Omega^+_S$, but being disjoint from $\Omega^-_S$. The only subsets considered by LAD are intervals of $IR|S|$ with faces parallel to the axes; these intervals are called "positive patterns.". A similar construction is applied to $\Omega^-_S$ for finding "negative patterns."[5].

## III.PATTERNS GENERATION

In LAD, pattern generation process is guided by two natural objectives: On the one hand, we give preference to the generation of short patterns (simplicity principle) and, on the other hand, the attempt to cover every positive observation by at least one pattern (comprehensiveness principle)[10].

The patterns can be generated conventionally by a greedy algorithm, i.e. all the possible sets of attributes are generated. The most straightforward approach to pattern generation is based on the use of combinatorial enumeration techniques. In view of the existence of various possible measures of quality of any given pattern, it is important for the pattern generation procedure not to miss any of the "best" patterns. Pattern generation techniques can follow a "topdown " or a "bottom-up" approach. The top-down approach starts by associating to every positive observation its characteristic term. Such a term is obviously a pattern and it is possible that, even after the removal of some literals, the resulting term will remain a pattern. The top-down procedure systematically removes literals one-by-one until arriving to a prime pattern. The bottom-up approach starts with terms of degree one that cover some positive observations. If such a term does not cover any negative observations, it is a pattern. Otherwise, literals are added to the term one by one as long as necessary, i.e., until generating a pattern[10].

However, It has been observed in empirical studies and practical applications that some patterns are more “suitable” than others for use in LAD. A heuristic algorithm[8] would be characterized by selection the patterns with determined supported observation and testing error. Another approach to find the LAD patterns adopt linear programming As optimization-based, the proposed MILP-based approach can generate patterns of all different degrees with equal ease and, hence, do not require redundant artificial interval variables for analyzing data[11].

## IV. PRACTICAL WORK

The logical Analysis of data would be applied on experiment data in order to generate the patterns that candidate to build the predictive model; using SPECT heart data from UCI repository data. This data consists of two group training (40 instances) and (187 instances) with 22 attributes and one attribute that refers to the binary classification of each instance. It is binary data so it is convenient with LAD technique. The work was implemented to generate the small set of patterns from training data then using these patterns in order to predicate the class of each instance in test data so compute the accuracy.

## V.RESULTS

The results are shown through generation the sets , starting with set(one variable) and then with two, three and at last four as shown in table(1).

**Table(1) Time of patterns generation**

| Size of patterns | Time of generation |
|---|---|
| One variable | 6.74E-9 |
| two variable | 1.85E-07 |
| three variable | 3.59E-07 |
| four variable | 2.55E-04 |

These patterns would be used in prediction the class of test instances. The algorithm of prediction start from small pattern sets. Table(2) shows the no of patterns that used and the no. of predicated instances with these groups of pattern sets.

**Table(2) no. of patterns used in prediction**

| Size of patterns | No. of patterns | No. of predicated instances |
|---|---|---|
| One variable | 2 | 46 |
| two variable | 19 | 92 |
| three variable | 18 | 27 |
| four variable | 8 | 18 |

The generation patterns with 4 groups achieve answer correct class for 183 instance so the accuracy that is 0.978 as shown in details with table(3).

**Table(3) details of predicted instances in test data**

| Type of instances class | No. of instances | No. of predicated instances |
|---|---|---|
| Positive instances | 15 | 15 |
| Negative instances | 172 | 168 |

## VI. Conclusion

1-The generation of small sets for patterns is more convenient in time as shown in table(1).
2- The results show that small sets give answer for a lot of prediction instances class. There is no invalid answer of prediction because the patterns used are prime patterns.
3- High score of accuracy for this technique encourage to improve the work in order to use in more application of classification problems.
4- The average prediction instances to the patterns sets as shown in Table(2) is high for the (one variable sets). However this is promise for optimization of the patterns generation.

**REFERENCES**

[1] C. V. Apte, S. J. Hong, R. Natarajan, E. P. D. Pednault, F. A. Tipu, S. M. Weiss," Data-intensive analytics for predictive modeling", 17 (2003) Mathematical Sciences at 40.

[2] N. P. Thair, "Survey of Classification Techniques in Data Mining", Proceedings of the International Multi Conference of Engineers and Computer Scientists 2009 Vol I IMECS 2009, March 18 - 20, 2009, Hong Kong.

[3] M. Debahuti, K. D. Asit , Mausumi, and M. Sashikala, 2010, Int. J. of Computer and Communication Technology, Vol. 2, No. 1.

[4] H. Jeong, K. Norman, K. J. Myong, Y. Bong-Jin, 2011," Comparisons of classification methods in the original and pattern spaces", Expert Systems with Applications 38 (2011), p.p. 12432–12438.

[5] A. Gabriela, A. Sorin, O. B. Tibérius, K. Alexander, "Logical analysis of data – the vision of Peter L. Hammer", Ann Math Artif Intell (2007) 49, p.p. 265–312.

[6] B. Endre, L. H. Peter, I. Toshihide, K. Alexander, M. Eddy, and M. Ilya, "An Implementation of Logical Analysis of Data", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 12, NO. 2, MARCH/APRIL 2000, p.p. 292-306.

[7] L. H. Peter, K. Alexander, S. Bruno, and S. Sandor , " Pareto-Optimal Patterns in Logical Analysis of Data ", RUTCOR RESEARCH REPORT, RRR 7-2001, January, 2001.

[8] B. Endre Boros and others :"Logical Analysis of Numerical data", RUTCOR RESEARCH REPORT, RRR 04-97, February 1997.

[9] L. H. Peter, K. Alexander, S. Bruno, S. Sundor, "Pareto-optimal patterns in logical analysis of data, "Discrete Applied Mathematics 144 (2004) 79 – 102.

[10] B . Endre, L. H. Peter, I. Toshihide, K. Alexander, M. Eddy, and M. Ilya, "An Implementation of Logical Analysis of Data", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 12, NO. 2, MARCH/APRIL 2000.

[11] S. R. Hong, J. In-Yong, "MILP approach to pattern generation in logical analysis of data I", Discrete Applied Mathematics, 157 (2009) 749_761.