Asian Journal of Information Technology 15 (20): 3943-3948, 2016
ISSN: 1682-3915

Knowledge Discovery in Big Data Using Symbolic Data Analysis

[1]S. Mythili, [1]R. PradeepKumar and [2]P. Nagabhushan
[1]Department of Computer Science and Engineering, United Institute of Technology, Anna university, Chennai, India
[2]Department of Computer Science and Engineering, University of Mysore, Mysore, India

Abstract: Recent developments in database technology have seen variety of data being stored in huge collections which are being referred to as generic database. The wide variety of data makes the analysis tasks of a generic database a strenuous task in Knowledge Discovery (KD). One approach to simplify the Strenuous Task of D is to summarize large data sets in such a waythat the resulting summary dataset is of manageable size and yet retains as much of the knowledge in the original dataset as possible. This process is termed as Symbolic Data Analysis (SDA). In SDA, Histogram has received significant attention as summarization object. This study demonstrates the approach of Symbolic Data Analysis (SDA) in very large database . This methodology analyzes large, very large datasets and extract glean useful information from within their massive confines. In this study, SDA uses histogram to summarize the data and linear regression to approximate the histogram. The application of SDA is illustrated in education environment using LMS Quiz data obtained from Ekluv-Ya . It discovers knowledge pattern in the dataset along with the help of data mining technique clustering. The application of this framework can serve as an assistance tool for managers of higher education institutions in improving the educational quality level.

Key words: Generic database, histogram, linear regression, knowledge pattern, clustering

INTRODUCTION

Today we live in the digital world. With increased digitization the amount of structured and unstructured data being created and stored is exploding. The data is being generated from various sources such as transactions, social media, sensors, digital images, videos, audios and clickstreams for domains including healthcare, retail, energy and utilities. In addition to business and organizations, individuals contribute to the data volume. For instance 3 billion content are being shared on Facebook every month; the photos viewed every 16 seconds in Picasa could cover a football field. The term "Big Data" was coined to address this massive volume of data storage and processing.

For most organizations, big data analysis is a challenge. Big data analytics refers to the process of collecting, organizing and analyzing large sets of data to discover patterns and other useful information. Not only will big data analytics help to understand the information contained within the data, but it will also help to identify the data that is most important to the business and future business decisions. Big data analysts basically want the knowledge that comes from analyzing the data. However Standard statistical methods don't have the power or flexibility to analyze these efficiently and extract the required knowledge. An alternative approach is to summarize a large dataset in such a way that the resulting summary dataset is of a manageable size and yet retains as much of the knowledge in the original dataset as possible. One consequence of this is that the data may no longer be formatted as single values, but be represented by lists, intervals, distributions and the like. These summarized data are examples of symbolic data (Billard and Diday, 2006). Hence, symbolic data arises through a process of aggregating a truly huge data set, perhaps one that is too large to conveniently store and analyze

The usual data mining model is based on two parts: the first concerns the observations, the second, contains their description by several standard variables including numerical or categorical. The Symbolic Data Analysis (SDA) model (Billard and Diday, 2006; Diday and Fraiture, 2008) needs two more parts: the first concerns units called concepts and the second concerns their description by symbolic data. The concepts are characterized by a set of properties called intent and by an extent defined by the set of observations which satisfy these properties. In order to take care of the variation of their extent, these

Corresponding Author: S. Mythili, Department of Computer Science and Engineering, United Institute of Technology, Anna university, Chennai, India

concepts are described by symbolic data which are standard categorical or numerical data but moreover intervals, histograms, sequences of weighted values and the like. These new kinds of data are called symbolic as they cannot be manipulated as numbers. Then, based on this model, new knowledge can be extracted by new tools of data mining extended to concepts considered as new kinds of observations.

Big data is a new phenomenon. Also a characteristic of modern large-scale data sets is that they often have a nontraditional form. New statistical methodologies with radically new ways of thinking about data are required. In that very important sense, Symbolic data analysis can be considered a method for Big Data (Lazar, 2013).

MATERIALS AND METHODS

Symbolic histogram data and linear regression: In symbolic data analysis, histograms have been projected as a feature type, possessing the generality to characterize all other feature types like single valued, multi-valued, interval valued, etc. It is noted that only histogram has the generic ability to effectively characterize most of the data types and the internal pattern of distribution of elements within the clusters can be easily depicted by histograms (Billard and Diday, 2006). Hence we use histograms as symbolic data to represent the concept descriptions.

It is observed that algorithms for producing histograms have parameters such as number of bins and bin width. It is required to have same number of bins and same bin width for all datasets or all clusters within a dataset for the effective characterization of data into histograms. Hence the data has to be normalized to maintain the histogram spread between 0 and 1 and the number of bins could be fixed as 10 or 20 depending upon the required precision. Spread of the histogram is divided by the number of bins to get the bin width. It is noted that storage of histograms requires memory for number of bins, bin width and memory for the details of each bin (Kumar and Nagabhushan, 2006, 2007). It is further shown that histograms can be easily converted to regression lines. Hence the concept of transforming a histogram into a regression line and distance calculation between regression line was introduced.

Regression line is a powerful knowledge representative which holds more details about the data in a compressed form which requires only two parameters slope and intercept. Also regression line instead of a histogram reduces the memory requirement. Ultimately because of linear regression, the parameters for histograms such as bin width, number of bins etc will not be of concern. The computational complexities can be considerably reduced by adopting the regression based approach (Nagabhushan and Pradeep, 2007). So once the data is transformed into a symbolic description, most symbolic data analysis and mining tasks become easier.

Fitting linear regression to symbolic histogram: We recommend the following procedure for transforming histogram to linear regression.

Step 1: Generate frequency Histogram H:

$$H = \{b1b2b3b4b5b6...bn\} \quad (1)$$

Where b1,...., bn are the individual elements of a histogram denoted as H(j) where j is the location or index H(1) = b1, H(2)b = b2, H(3) = b3, H(4) = b4,....., H(n) = bn.

Step 2: Convert frequency histogram into cumulative frequency distribution Cj:

$$Cj=\sum_{i=1}^{j} H(i) \qquad j\in \{1,2,.....n\} \quad (2)$$

Then, normalized cumulative histogram given by:

$$\text{Cj}=\frac{Cj}{Cn}$$

Step 3: Fit linear regression:

$$y = ax+b \quad (3)$$

Distance measure between the regression lines: After having built the linear regression, now we have to define a distance measure for linear regression. The idea is to use the AB distance measure proposed by Kumar and Nagabhushan (2006, 2007). The distance introduced by Kumar and Nagabhushan (2006, 2007) is based on the "area" and "behavior difference" components between the regression lines.

Area between the regression lines: The area is computed by considering the region between the regression lines as a trapezoid. So area of trapezium characterizes the distance between two simple regression lines. For a pair of samples, two different cases are possible due to the line position. Therefore the respective area of the trapezium for two cases is summarized below:

Case 1: When pair of lines are not intersected shown in Fig. 1, area of trapezium A is given by the following formula where a and b are the lengths of the parallel sides and h is the perpendicular distance between the parallel sides and normalized to the range 0-1.

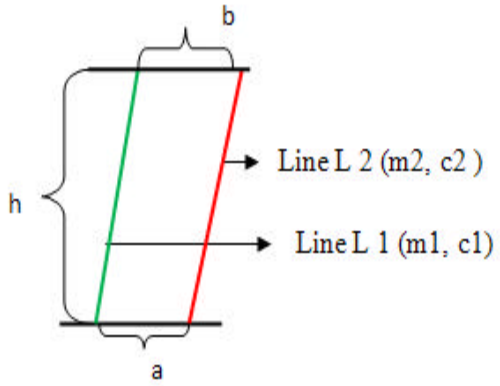

Fig. 1: Pair of lines are not intersected (Case 1)

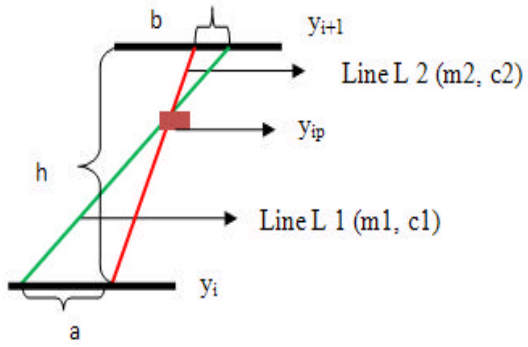

Fig. 2: Pair of lines are intersected (Case 2)

$$A = \frac{1}{2}(a+b)$$

$$a = \left\| \frac{(-c2)}{m2} - \frac{(-c1)}{m1} \right\|, \quad b = \left\| \frac{(1-c2)}{m2} - \frac{(1-c1)}{m1} \right\|$$

Where:

c1, m1 = The intercept and slope of simple linear regression line 1

c2, m2 = The intercept and slope of simple linear regression line 2

Case 2: When pair of lines areintersected shown in Fig. 2 area of trapezium A is considered as two triangles where their heights are computed based on the intersection point y_{ip} between two lines

$$A = \left(\frac{1}{2} * a * (y_{ip} - y_i)\right) + \left(\frac{1}{2} * b * (y_{i+1} - y_{ip})\right) \quad (5)$$

$$x_{ip} = \left(\frac{c2-c1}{m2-m1}\right), \; y_{ip} = m2 * x_{ip} + c2$$

Where X_{ip} and y_{ip} are intersection points between pair of samples.

The behavior of the regression lines: The behavior component plays an important role in charactering the difference between the regression lines. The behavior component is computed by taking the absolute difference between the regression lines.

A case study: knowledge pattern discovery: The database under investigation refers to the quiz data pooled by a Learning Management Systems (LMS) called EkLuv-Ya [Ekl0] is used. It is the product created by Amphisoft Technologies for revolutionizing engineering education through Automated Evaluation Systems in different branches of engineering.

A typical web based LMS such as moodle in its simplest form provides a platform for online learning where educators can post their learning materials, assignments, tests etc and monitor progress of their students who in turn have to log in to learn from the posted material. An LMS can keep record of all the student activities through their log files which pave way to gather large amount of data on a day to day basis. The amount of data which is gathered is proportional to the number of students registered for each course, number of times each student logs in for each of the registered courses and number of courses that are handled by an LMS. We can find streams of log records getting generated every moment, which makes the real time knowledge extraction almost impossible. Various data mining techniques can be applied to learn from the generated data of the LMS (Romero *et al*., 2008; Clari *et al*., 2009) and extract useful information instantaneously.

From the standpoint of the bigdata, the main objective of this case study is to show how the SDA summarizes the quiz data in a meaningful and intelligent fashion, to its important and relevant features. Also how it enhances the data mining technique to mine knowledge pattern in big data. Hence we need to use a clustering and classification technique in integration with SDA. Clustering is an unsupervised learning problem that group objects based upon distance or similarity. Each group is known as a cluster (Jain and Dubes, 1988; Flexer, 2001). While, Classification is a supervised learning problem of assigning an object to one of several pre-defined categories based upon the attributes of the object (Kaufinan and Rousseeuw,1990).

RESULTS AND DISCUSSION

The quiz data under consideration is the marks scored in different quizzes by the 52 students of sixth

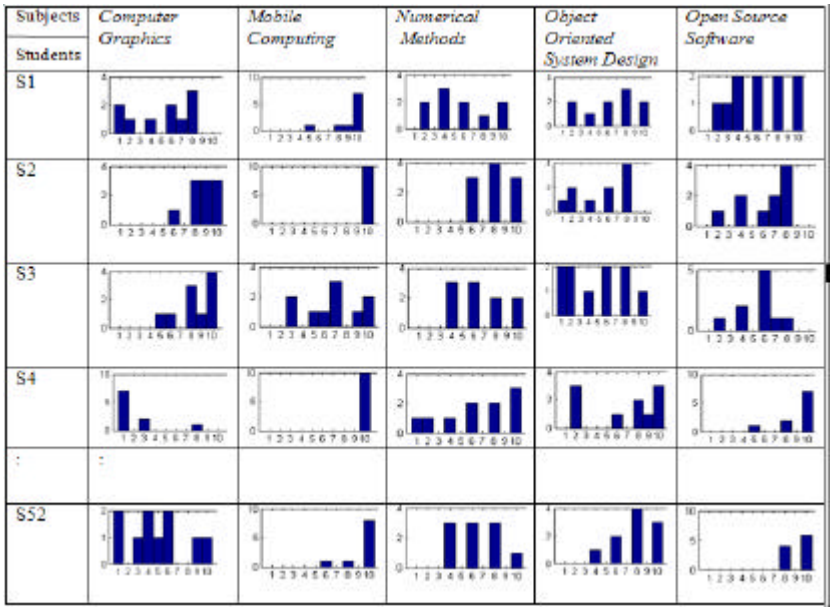

Fig. 3: Summarization of very large data sets into symbolic histogram feature

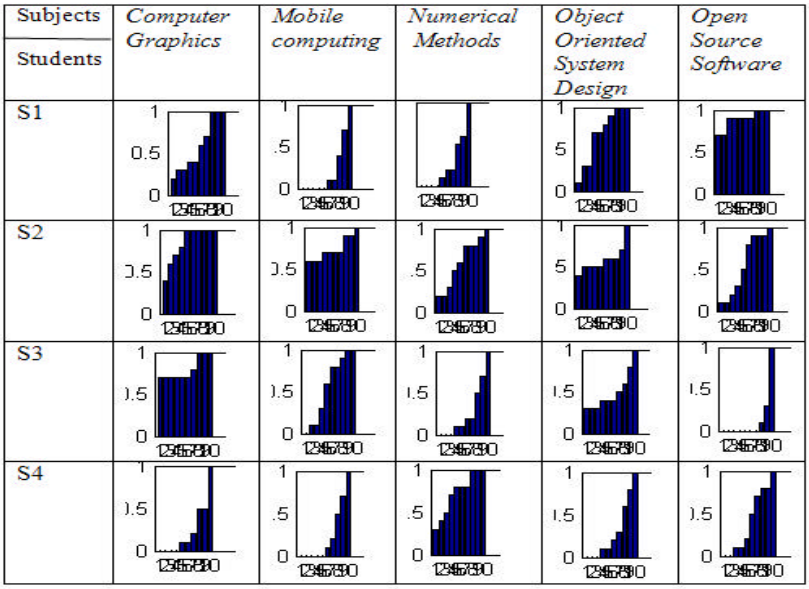

Fig. 4: Bar graphs of cumulative histogram of quizzes 1-10 of student 1-4 for all subjects

semester of Bachelor of Engineering from United Institute of Technology, Coimbatore (Tamil Nadu), India. The subjects of study in the order of appearance are: Computer Graphics, Mobile Computing, Numerical Methods, Object Oriented System Design and Open Source Software. For the sake of experimentation, minimum marks for each quiz is 2. maximum marks for each quiz is 10. Marks of students who could not take up the quiz in a particular subject have been marked as '1'.

The first step is to compute the histogram matrix by summarizing the distribution of marks of ten quizzes under each of five subject for every student as illustrated in Fig. 3. Then, corresponding cumulative histogram Figure 4 is approximated by a regression line (Fig. 5)

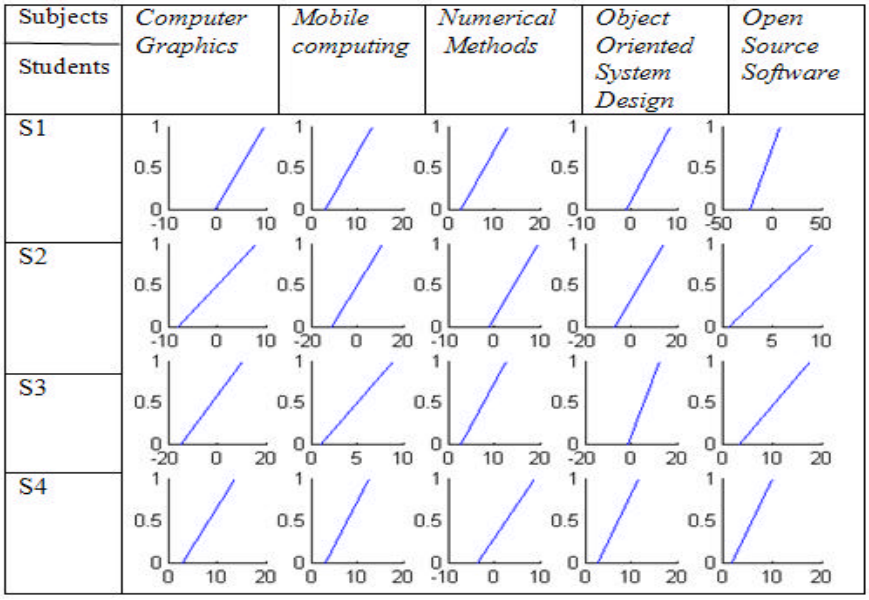

Fig. 5 : Regression Features representing students performance

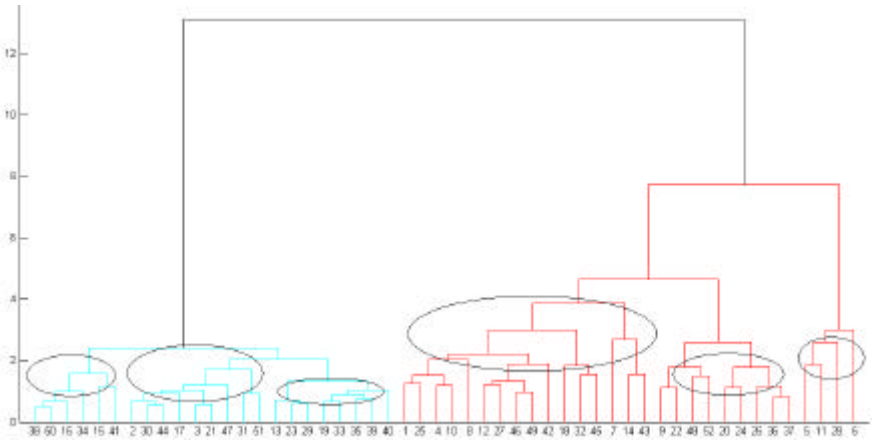

Fig. 6 : Dendrogram obtained through complete linkage clustering using AB distance measure

according to the algorithm. Using AB distance measure dissimilarities between individual objects are calculated. The obtained distance matrix is given as input to cluster analysis to classify them into groups. Classical method as hierarchical clustering can be suitably used for this purpose. Hierarchical cluster methods produce a hierarchy of clusters, ranging from small clusters of very similar items to larger clusters of increasingly dissimilar items.

Symbolic data analysis represents the quiz data in the form of symbolic Fig. 3, thereby providing users with the ability to visually analyze and explore large, complex datasets. Linear regression Fig. 5 characterize the relationships and dependencies that exist within the histogram. The strength of the distance measure is portrayed by the dendrogram of the obtained data set.

The dendrogram shows the merging of samples into clusters at various stages of the analysis and the similarities at which the clusters merge with the clustering displayed hierarchically. There are 2 major classes obtained in the structure (Fig. 6). The students grouped in cluster 1 have secured mark above 70% which indicates good knowledge in subjects. Cluster 2 students below 70% who have poor knowledge in subjects. In order to validate, btained results when compared with teachers handling the subjects match exactly with the expected

results.This experimentation will help the teacher in finding out the students of a particular group and counsel them as the information about serial number of students is retained as knowledge.

CONCLUSION

To conclude, in this research we have shown how symbolic data analysis is used in mining big data. The construction of symbolic histogram proved to be a necessary precursor to statistical analyses when the size of the original dataset is too large for classical analyses. Also a methodology based on regression line features helps for further reducing the computational complexity of this symbolic histogram. To support our proposed theory we have done a case study in educational environment for better understanding of student failed courses andthe assessment of the student learning process in every subject .This in turn helps to improve teaching quality of the subject teacher and to improve students' academicperformance and trim down failure rate. Hence it plays an important role in strengthening the management of higher education institutions. It is expected that this work can be successfully used in different areas including financial, banking.

REFERENCES

Billard, L. and E. Diday, 2006. Symbolic Data Analysis: Conceptual Statistics and Data Mining. John Wiley, Hoboken, New Jersey, USA.,.

Clari, P.M., M.A. Herraez and V.C. Lleo, 2009. Data analysis as a tool for optimizing learning management systems. Proceedings of the 2009 9th IEEE International Conference on Advanced Learning Technologies, July 15-17, 2009, IEEE, New York, USA., ISBN: 978-0-7695-3711-5, pp: 242-246.

Diday, E. and M.N. Fraiture, 2008. Symbolic Data Analysis and the SODAS Software. John Wiley and Sons, Hoboken, New Jersey, USA., Pages: 188.

Flexer, A., 2001. On the use of self-organizing maps for clustering and visualization. Intell. Data Anal., 5: 373-384.

Jain, A.K. and R.C. Dubes, 1998. Algorithms for Clustering Data. Prentice-Hall, Englewood Cliffs, New Jersey, ISBN: 013022278X.

Kaufman, L. and P.J. Rousseeuw, 1990. Finding Groups in Data: An Introduction to Cluster Analysis. John Wiley and Sons, New York, ISBN: 0471878766, Pages: 342.

Kumar, R.P. and P. Nagabhushan, 2006. AB distance based histogram clustering for mining multi-channel EEG data using wavesim transform. Proceedings of the 2006 5th IEEE International Conference on Cognitive Informatics, July 17-19, 2006, IEEE, New York, USA., ISBN: 1-4244-0475-4, pp: 467-477.

Kumar, R.P. and P. Nagabhushan, 2007. An approach based on regression line features for low complexity content based image retrieval. Proceedings of the International Conference on Computing: Theory and Applications, March 5-7, 2007, IEEE, New York, USA., ISBN: 0-7695-2770-1, pp: 600-604.

Lazar, N., 2013. The big picture: Symbolic data analysis. Chance, 26: 39-42.

Nagabhushan, R.P. and K. Pradeep, 2007. Histogram PCA advances in neural networks-ISNN 2007. Lect. Notes Comput. Sci., 4492: 1012-1021.

Romero, C., S. Ventura and E. Garcia, 2008. Data mining in course management systems: Moodle case study and tutorial. Comput. Educ., 51: 368-384.