International Journal of Research in Engineering, Science and Management
Volume-1, Issue-12, December-2018
www.ijresm.com | ISSN (Online): 2581-5792

An Approach towards Hauling Out the Data Mining Techniques by the Side of Bayesian Classifier

T. Praveen[1], S. Vikash[2], S. Deepan Kumar[3]

[1,2]*Student, Dept. of Computer Science and Applications, Sri Krishna Arts & Science College, Coimbatore, India*

[3]*Professor, Dept. of Computer Science and Applications, Sri Krishna Arts & Science College, Coimbatore, India*

***Abstract*: Data mining helps to fetch the data from group of data to gain pure knowledge. It is the field which connects computer and mathematical (statistics) knowledge together. Any method which is used to extract the knowledge from the database is a form of data mining technique. Bayes classifier is more specialized in pattern recognition because it is based upon assumptions for prior and pattern distributions. Being a developing factor, data mining many futuristic scopes and it's going to rule the upcoming generation. This paper deals with the ideology of data mining and Bayesian classification technique.**

***Keywords*: Analyses of Bayesian classification, Bayesian classification, Data mining, Data mining techniques, data mining architecture with coupling,**

1. Introduction

Data mining concepts are used in many different sectors to maintain database records. Data are stored in data warehouse with the help of data mining techniques. Data mining comprises of numerous process. These processes involves in data storage, data modification and data retrieval. Data mining techniques are used in differentiating data according to their characteristics or using previous entries. And finally Bayesian classification is one of the influenced classifiers in data mining. Naïve Bayes has spent years and years to get this influence able classifier. Bayesian classifiers are seemed to have many advantages. This classifiers is used in many fields including mathematics especially statistics and our developing computer field. This module completely depicts data mining techniques, its architecture and the storage unit i.e. data warehouse or database. The main concepts depicted in this module are Bayesian classifiers with proper comparison over methods and discussions about Bayesians classifiers. This journal proposes data mining techniques and Bayesian classification in detail along with elaborate diagrams for user to understand the standards.

2. Data mining

Data mining is the progression of fetching data from group of data to gain knowledge. It is beneficial for the user when dealing with the large amount of data sets. For e.g. student details, their eligibility, eligibility criteria, their performance details in educational institutions, etc. Data mining can analyze the current trend of availability of resources data where modification and updating of data stand as a trending process. Data warehousing is the major technique hired to maintain past record more resourcefully. An institute with proper data mining and data warehousing can lead their mode towards improvement.

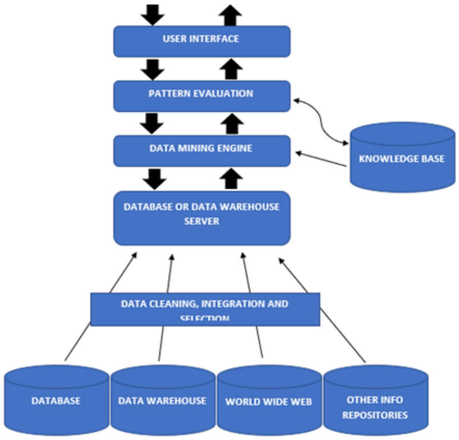

Fig. 1. Architecture of typical data mining

3. Data mining techniques

There were some data mining techniques used. They are:

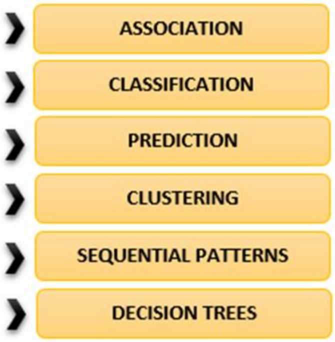

Fig. 2. Data mining techniques

A. Association

Association is one of the best-known data mining techniques. Here patterns of data are identified based on the relationship between the data. This technique is also called as relation technique. They are used in super markets to identify the products that are frequently bought by the customers to rise the productivity. Retailers also use this technique to find the customers buying habit.

B. Classification

Classification is based on machine learning. It is a classic data mining technique where it is used to classify data items into predefined groups or classes. It includes some of the mathematical techniques such as decision tress, linear programming, neutral networks and statistics.

C. Clustering

Clustering is a data mining technique used to classify data into clusters based on their characteristics. It defines the classes and assigns objects to those classes. Library management process uses clustering technique to assemble the books according to their characters.

D. Prediction

Prediction data mining technique is used to predict the result before by analyzing the previous data. It determines the relationship between the independent variables and the dependent variables

E. Sequential patterns

Data mining uses sequential patterns to identify similar patterns or regular over a certain period. This pattern identification is used to determine the product reach, demand and success of the business.

F. Decision trees

Decision tree is the most frequently used data mining technique which helps the user to understand the decision tree based on the model. The root of the decision tree is a condition that has multiple answers and these answers then root us to different conditions to determine required data to conclude the result.

4. Data mining architecture

Data warehouse system helps to store the data, a data mining system should be designed in such a way used to couples and decouples with data warehouse system. There are four different techniques employed in data mining architecture:

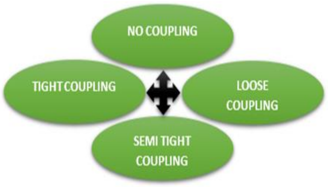

Fig. 3. Techniques in data mining architecture

A. No-coupling

In this system architecture data mining technique doesn't use any functionality of a data warehouse. Data mining algorithm is employed to fetch data from large data resources and the final data result is stored in the file system. This architecture is very efficient in organizing, storing, accessing and retrieving the data without any benefits of data warehouse.

B. Loose coupling

Loose coupling is used for effective data retrieval from data warehouse. The retrieved data stores the desired result in the system. This technique doesn't require high scalability and performance and memory based.

C. Semi-tight coupling

In this type of architecture, rather than using data mining features this uses database or data warehouse features. These features include sorting, indexing, aggregation etc. The results are stored as an intermediate in database or data warehouse.

D. Tight coupling

In this type of architecture both the database and data warehouse are treated as a normal function which are used for information retrieval. Both the features of database and data warehouse are used to perform data mining tasks. This improves the system scalability, high performance and integrated information. Tight coupling comprises of three layers. They are:

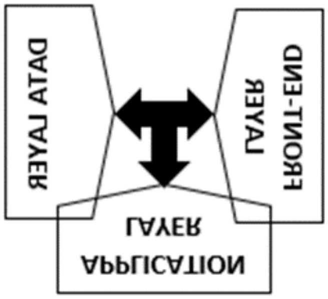

Fig. 4. Tight coupling layers

5. Introduction (Bayesian classification)

Bayesian classification, as the name influences that it is based on Bayes' theorem. Bayesian classifiers are the statistical classifiers. Membership probabilities are predicted using Bayesian classifiers. Naive Bayes has been studied extensively since the 1950's. It remains most popular method for categorizing text. They are scalable and require number of parameters. In both computer science and mathematical field Bayes classifiers are popularly used. Naive Bayes having an advantage that it requires a small number of trained data and that trained data is used to estimate necessary parameters. Naive Bayes classifications seem to be strong. Probability theory is used to classify data. Bayes theorem can be adjusted to new data introduced. It is not a single algorithm it is a family of algorithms joined together. Spam filters are also used with the help of Bayes classification. It is robust and easy to implement. It is used in many fields because it is that much accurate in classification and even faster.

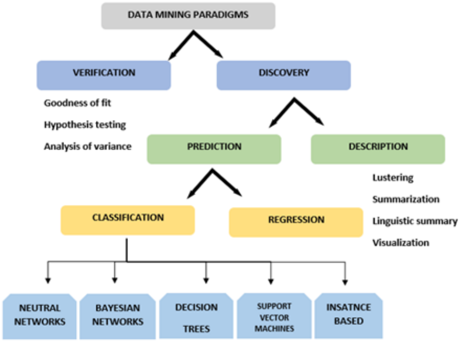

Fig. 5. Data mining paradigms based on Bayes theorem

Bayesian classification is a statistical technique which is used by Auto Class (an automatic classification program) for searching classes in the data. Instead of partitioning the data in the classes here data processes are descript in the form of probabilistic matter. With the help of the probabilistic analysis, one can determine that each object is a member of the class. Several advantages are provided due to the resultant classification system:

- Auto Class can able to determine the utmost likely number of classes and they are represented in the actual structure.
- Bayes' theorem helps to perform classification process without any need of ad hoc similarity measure, clustering, etc.… probability analysis used in Auto Class are carried out with the help of Decision theory.
- Description for the classes and the assigning objects to the classes are done with the help of probability distribution. Here" fuzzy" classes are used instead of categorical classification for sensing the common notation of class membership.
- Real and discrete valued attributes are mixed freely and there may be chances of loss of data or value. "Tree valued attributes can the effortlessly included in Auto Class.

6. Comparison with other methods

Many other techniques also use automatic classification of data. Auto Class has its own features comparatively with them:

A. *Clustering analysis*

- Clustering analysis and Auto Class differ fundamentally in their goals and techniques. Clustering analysis works on the concept of grouping of data points by seeking the classes and assign points to those classes whereas Auto Class works based on description of data in the classes and never assigns points to the classes.
- Secondly, clustering analysis state a class indirectly by providing criteria for the analyses of clustering hypothesis i.e. by maximizing the intra-class similarity whereas Auto Class technique state the class explicitly with the help of model function and Bayes theorem is used for class separation process.

B. *Conceptual clustering*

The similarity lies between Auto Class and conceptual clustering is both the methods deals with description of classes instead of partitioning the objects. Auto Class uses logical description language and later it is overcoming by probabilistic description of classes. Logical way approach is used effectively even when the data are too noisy or overlapped. Conceptual clustering technique specifies the class definition with the criteria "clustering quality" which works based on the desired clustering rather than actual clustering.

C. *Maximum likelihood mixture separation*

Auto Class seems very likely when compared with maximum likelihood mixture separation as they are used to separate finite mixtures with recognition of statistical analysis. The main difference between them is Auto Class analysis removes the singularities from the search space and provides effectiveness in determine the number of classes than the existing method testing.

D. *Minimum message length method*

Minimum message length method gathers the classification which then can be encoded into fewer bits; here encoded data is of two parts: Information required defining the class parameters and the information required to encode the data in the class parameters. This method helps to reduce the length of message to fit to the data. Similar method formulation made by Bayesian classification and minimum message length criteria is a similar clash to that of Bayesian criteria.

7. Discussions

Rejection of Bayesian classification is often taken place due to the prior distribution, it happens because the analysis is based on personal basis. Usage of prior are uninformative and completely insecure able. Informative prior is simpler than the uninformative prior in the mathematical format, so that Auto Class uses an informative prior that introduces a small bias. It can be extended to gain strong prior knowledge if is available so. Auto Class are useful to gain more knowledge from examples, in other words supervised learning combined with unsupervised learning in same system. 'Auto Class III' progressions include many exponential distributions for attributes in real time. Automatic selection of class distributions took advantage of increasing model complexity with the estimation additional parameters. This type selection of models is likely to selection of classes.

8. Conclusion

We have completed one of the module in the computer field cloud computing techniques along with the most influence able classifier i.e. Bayesian classifier. We had come across data storage units this journal helps to learn about the techniques used in data storage and data retrieval. This concept provides us brand new view of data classification by means of data mining and Bayesian classification. The information depicted is apparent.

References

[1] Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). Classification and Regression Trees. Wadsworth.

[2] Han, J. and Kamber, M. (2006). Data Mining: Concepts and Techniques. The Morgan Kaufmann series in data management systems. Elsevier San Francisco (Calif.), Amsterdam, Boston, Heidelberg.

[3] Zaki, M. J. and Jr, W. M. (2014). Data Mining and Analysis: Fundamental Concepts and Algorithms. Cambridge University Press, New York, NY, USA.

[4] Maragaret H Dunham, data mining introductory and advanced topics, Pearson education (2009).

[5] P. Cheeseman, J. Kelly, M. Self, J. Stutz, and W. Taylor, “Machine Learning, 1988, Elsevier.

[6] Qiong Wang, George M. Garrity, James M. Tiedje, and James R. Cole, “Naïve Bayesian Classifier for Rapid Assignment of rRNA Sequences into the New Bacterial Taxonomy, Appl. Environ. Microbiol. Aug 2007, 73 (16), 5261-5267.

[7] H Sun - ChemMedChem: Chemistry Enabling Drug Discovery, 2006 - Wiley Online Library.

[8] P LeVan, E Urrestarazu, J Gotman - Clinical Neurophysiology, 2006 – Elsevier.

[9] Arun K Pujari; Data Mining Techniques, University Press (2009).

[10] Jiawei Han, and Micheline Kamber, “Data Mining Concepts and Techniques,” Morgan Kaufmann Publishers, Second Edition (2007).

[11] David Hand, Hekki Mannila, and Padhraic Smyth, “Principles of Data Mining,” Prentice-Hall of India (2009).

[12] K. P. Soman, Shyam Diwakar, V. Ajay, “Insight into Data Mining Theory and Practice,” Prentice-Hall of India (2006).

[13] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, “Introduction to data Mining,” Pearson Education (2009).

[14] Alex Berson, Stephen J. Smith, “Data Warehousing, Data Mining & OLAP,” Tata McGrawhill (2009).