International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 05 Issue: 06 | June 2018 www.irjet.net p-ISSN: 2395-0072

Scaling Distributed Associative Classifier Using Big Data

Bhanu Prakash H N[1], Dr. Niharika Kumar[2]

[1]*Research Scholar, Dept. of Computer Science and Engineering, B.N.M Institute of Technology, Karnataka, India*

[2] *Associate Professor, Dept. of Computer Science and Engineering, B.N.M Institute of Technology, Karnataka, India*

Abstract - *Scaling the dataset with the help of a special classifier called Distributed Associative Classifier. A decision tree is formed by defining the classifier and is compared with the state-of-the-art solution named as Random Forest. The Distributed Associative Classifier is proved to be more scalable compared to the Random Forest. Modules are implemented with the help of Apache Hadoop and Apache Pig.*

Key Words: Distributed Associative Classifier, Decision Tree, Random Forest, Apache Hadoop, Apache Pig.

1. INTRODUCTION

Dataset is the key parameter used to measure the classifying capability. Dataset can be defined with the help of two aspects. Size of the dataset and the number of features included in the dataset. Dataset that is being used to conduct experimentation is a Bluetooth Travel Sensor dataset. The dataset has a size of 1.8 Gb with one hundred and twenty-six features to define each column of the dataset called the disk storage. Volume of the dataset can be the number of records used in the Bluetooth sensor data. In order to obtain perfect evaluation of the module Big Data plays a very important role. It consists of large amount of data that can be used either to classify the data or cluster the necessary dataset. Domain of the dataset can vary from one module to another based on the characteristics of the data. Algorithms can be implemented on the dataset to achieve highest classification ratio. Features of the dataset rely on the parameters required for Bluetooth sensing technology. Data mining is the process of discovering patterns from the dataset. Patterns can either be frequent items or random item sets based on the type of algorithm used during implementation.

Distributed Associative Classifier [1] falls under the category of Supervised Learning Algorithms. In supervised learning algorithms data is always considered to be labelled, that is each feature of the dataset is defined by its own identity. Scalability problems can be tackled by using special algorithms in Machine Learning or by using dimensions that are mentioned above. Associative Classification can be defined as a process of utilizing associative rule mining to construct associative classes called associative classifiers, this way of segregating is called Associative Classification. Distributed Associative Classifier [2] distributes the things that are related to classification. Associative rule mining is made use for the purpose of classification that are widely used in the field of Machine Learning. High quality classification models and algorithms like Decision Trees are proved in the past by using this process.

Ordering amongst values can be defined by using discrete features in such algorithms. The algorithm used for implementation is a Frequent Pattern growth [3] algorithm. The complexity of the categorical variables increases with the absence of natural ordering. Massive datasets can also be used that consist of more than fifty giga bytes for implementing, but to reduce computations and to achieve the results quickly, a limited size of data that is more than one giga byte but less than two giga bytes is used. Experimental sessions are based on the type of modules and the module is readable as it makes use of association rules and can be tuned manually by the modification or the deletion process. Different categories are used in the dataset. Algorithm is effective compared to the existing algorithms, because it is designed in a systematic approach.

The main aim is to achieve high accuracy without reducing the quality of the algorithm. The model is trained, processed and the final tree is generated from the inputs taken. Features in the dataset has its own identification number and the metrics are shared to measure the significant growth [4] of the algorithm. State-of-the-art solutions are those which are available in the market, it can also be termed as existing technologies. The existing classifiers are decision trees, random forest, support vector machines and boosted trees. Both decision tree and random forest is used to determine the computational complexity of the classes. Data in the dataset can either be structured or unstructured based on the domain and dimensionality of the dataset. Gini Impurity is another measure used for the catalogs to test the models. The purity of the model can be determined with the help of Gini Impurity. If the output of the model is zero then the

purity is said to be normal, if the output of the model is one then the purity can be negligible. The quality is not sacrificed and the rules are pruned in such a way that the classifier can work under any circumstance and perform 'N' number of tasks compared to the state-of-the-art solution. The dataset consists of variables which are defined in each phase. DAC is used to train the model in a computing framework. Apache Spark is an in-memory framework to describe the cluster for the purpose of computations. Experimental session is performed in this framework. The feasibility of the system can be pruned, prevented from further damage and different schemes can be employed to determine the quality of the model.

2. RELATED WORK

Classification has been established as an effective mechanism wherein a fresh approach is used called as Association Rules that is relevant for various parameters thereby called as Classification that reckons on multiple aspects. The method focuses on recursive mining approaches which makes use of occurrences found in Frequent Patterns from a storage unit where two functions are modified. Storing and retrieving from a frequent itemset are some of the ideals used.

In [1], the paradigms considered in designing a tree is comprehensive, and is termed as classifying the rules, which defines the set of rules that are made used by the algorithm and described in a tree. There are three functions that the tree performs.

A) Assurance of the product.

B) Collation of the product and

C) Driver Rage of the product.

Weight is one of the keyword that is taken into consideration. Classification of the tree is proportional to the product's weight. This shows that the product's weight is an important feature in decision tree. In the database, the whole process is tested and trained. The method focused is on the concepts of MapReduce. The software's can be risky, when it has to be released in the market. It can also cause problems to the user.

In [2], the approach is focused towards MapReduce concepts. The software's that release in the market (store) can be vulnerable when it comes to the working of the software.

The software's that are published can cause problems to the users. The software product goes through the MapReduce phase and is computed and manipulated simultaneously. The software is further processed which is the second phase of processing. The software is played in two steps namely: Map Phase and the Reduce Phase. Data in the software is stored in logs. The stored data is considered to be delicate and is kept confidential. The motivation of the software product is to work efficiently to the needs of the customer and process quickly.

Classifiers can also be parallelized in a phase where the training of the relational database or a dataset is mentioned and notified about the dangerous threats that a classifier can cause. Deviations can occur in such type of database and it becomes difficult to measure the scale of output that the datasets produce. In order to prove this action, suitable algorithms that match the needs of the classifiers are proposed, the proposed algorithm is based on MapReduce techniques and the approaches are gathered. In [3], two challenges are faced by the user. The selection of the software is the major challenge faced in the approach. The selection process can have its own differences and approaches, the user working in Hadoop should be aware of the software and the domain based on what characteristics the selection of the software is done. Further, the software is generated from the above outcomes and placed in the order that is comfortable for the user to manipulate and work in Hadoop. Software's are again divided and pruned to meet the requirements. The requirements are of two types, namely functional requirements of the software and nonfunctional requirements of the model. The setting of the software is another challenge faced by the user and it can be decided based on the order of placement that the software utilizes. The process can be proved to be scalable based on the division of the datasets.

In [4], the focus is towards the distribution algorithm that provides the data and makes use of a technique called Data Partitioning, which is the first phase of the process that occurs in the division. The data is then broken into segments and the splitting of the dataset is implemented, further the segments are partitioned. This defines the distribution algorithm which is an efficient technique to be followed and is very helpful in the field of programming. The distribution algorithm also makes use of clusters, wherein the clusters are divided and formed into groups. The approach is considered as one of the oldest and famous techniques that is always inculcated in many algorithms for the processing of the datasets. Distribution algorithm is a paradigm to many algorithms and it involves a phase called the analysis phase, where the information is processed and analyzed.

The analyzed results are tabulated and adequate graphs are plotted to define the scalability of the dataset. In this paradigm, the input is measured and read, that is the input data is available to read and once it is ready, there is an incremental processing of the input data. Distributed algorithm also makes use of processing and the type of processing that is operated here is processing of the data parallel and the next focus is on the fault tolerance of the dataset. The fault tolerance mechanism includes two types of processing, namely Real time processing of the data and Batch processing systems interlinked to the real time processing systems. The motivational factor of a distributed system is the number of executive logs that the system holds. Classification is inculcated in a distributed computing framework providing a friendly working environment for the users to work on. Load balancing is another concept that is dealt in this mechanism and the paradigm also plays an important role in MapReduce concepts, wherein mapping of dataset happens in the first phase and later comes the reduction of the model.

In [5], accurate and effective classifiers are formed from the paradigms that are employed in Associative Classification. The Rule mining mechanism makes use of a very large set of items, item sets that are also called as datasets. The lazy approach also makes use of pruning mechanism to divide the data into 'n' partitions. Data mining is another aspect that is used in this approach. The approach makes use of several representations and the representations are called as condensed representations, as they are divided, pruned to produce a light weight model.

The datasets are classified and the major challenge faced in this scenario is the misclassification of the dataset. Due to the misclassification of the dataset, the information that is stored in the dataset is lost. In some cases, the dataset can also be hidden. Therefore, the pruning technique that is employed is a lazy approach pruning mechanism.

3. PROPOSED MODEL

The characteristics defined from all the mechanisms focuses on a keyword called Distributed Associative Classifier (DAC), which is the proposed technique in the project. The classifier observes a set of features derived from the dataset and it makes use of associative rule mining to achieve this objective. Features can be of different classes from the minute parts of the set. The classification takes a set of items and there is a simplification that occurs between the classes, further combining of the items is proposed which reduces the convolutions faced in the process. When there is a division in the model, concatenation is another keyword that the user faces, wherein the contents of one class from the dataset is copied to the other phase of the dataset.

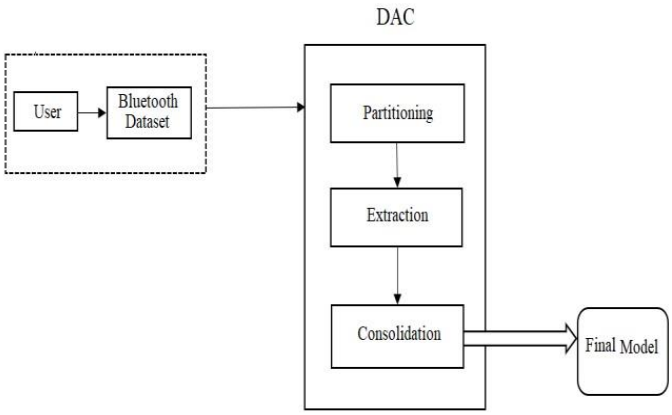

Fig -1: Architecture

The dataset is termed as a Bluetooth dataset of zones. Each sample in the database is followed to be as an instance. Computations takes place at each phase of the dataset and its scalability is measured. There is an equidistant gap between the positions of each classes and the data is represented pictorially. The dataset plays a very important role because that is the first phase of the project, if there is any flaw in choosing the correct dataset then there will be no guarantee of getting the output. In the extraction phase, an algorithm is employed which is a frequent occurrence of pattern growth algorithm. The last stage is the reduction process called as consolidated process, which is a full-fledged programming stage.

DAC process and a consolidation mechanism is conducted to form the last model. Similarly, in level one there is a detailed description regarding the DAC process. DAC is split up into 'N' partitions and a partitioned model is obtained, further there is an extraction that occurs and the extracted data is consolidated to produce the end process. In the third level, the extraction process is split and an algorithm is introduced called the pattern algorithm. Organizations have a set of guidelines and respective deadlines that the client has to fulfil in order to meet the requirements of the organization. Personal success is

connected with the success of an individual in an organization. Personal success makes use of stakeholders, project managers, product managers, domain experts, executive sponsors, business analysts, onsite customers.

The volume of the project can vary from one dataset to another dataset based on the domain of the project. Associative Classifiers makes use of another motivational technique called voting. Voting is related to the dataset of the classification. Voting is a unique technique followed in Associative

Classification which enhances the scalability of the classifiers. The complex portion of voting relies on the selection of the dataset.

The dataset has to contain huge amount of data in order to partition the patterns. Eventually, the partition occurs in the second stage of the product whereas voting happens in the final stage of implementation. The distinguishing feature of classification is the scalability of both Distributed Associative Classifier and the Random Forest. Quora is an official link which provides the features and appropriate details regarding the classification of algorithms in Machine Learning. The potential of Distributed Associative Classifier can be computed and measured by defining the scalability of the dataset as to how scalable the dataset is. Scalability of the dataset is the measure of the dataset. The next measure to be considered is the utilization of a Random Forest. The key objective is to compare the Distributed Associative Classifier with the Random Forest and to prove that the Distributed Associative Classifier is 1 percent more scalable compared to the Random Forest. The single element that has to be represented and obtained at the end of each phase is termed as 'Frequent Patterns'.

In the recent generation the technology involved in distinguishing the 'Frequent Patterns' has become much more efficient and fast. The origin of 'Frequent Patterns' occurs from a large amount of data and the absolute amount is indefinite. The extracted patterns can be accessed and implemented in CentOS and the accurate results can be obtained, but the metrics have to be considered in proper intervals of time. To acquire this CentOS plays a major role in its operation and the actions are implemented. As it is known that Big Data makes applies and utilizes a very large dataset. Chunks, huge volume of data is stored in a unit called Data Warehouse. Big Data is very useful in the field of business and its industry application. Information is processed and the plethora of data is utilized in business appliances.

4. IMPLEMENTATION

The model is far more comprehensive compared to the other support vector machines specified in the existing system. Partitioning is the key parameter in the proposed methodology, here the classes from the set is partitioned into 'N' ways for the future implementation of the model. Particulars of the model can be emphasized based on the formulated graph that is obtained from a tree. The greater half of the project is done by considering an algorithm called as the Frequent Pattern growth algorithm.

This algorithm observes the occurrences of patterns that occur quite often in a dataset. The next stage of the mechanism is to reduce the size of the model which is termed as the 'Reduction Phase'. Eventually the size is decremented and can be used for the next stage. Therefore, the two events that happen is the recording of frequent patterns from the symbiosis and the reduction of the model. This process of reducing the model is called the consolidation of the model, where each class is consolidated with the feature vector of another vector. The methodological model proposed is done in a software called Apache Pig and the programming language that is adopted is JAVA programming language.

A graph is plotted from all the samples against another set of samples, the process is called as a distance weighted mechanism. The complexity of the feature is reduced when each class is plotted. Each class is in a relationship and is congruent to another class. If the classification falls into one of the classes that is either class one or class two then it is said to be true else it is termed false. Therefore, the output is based on two types of probabilities that is prior probability and posterior probability. Distributed Associative Classifier follows all the above criterion and provides a true result and the model produced is rephrased and a consolidated model with respect to the sample corresponding to the features present in the data is obtained.

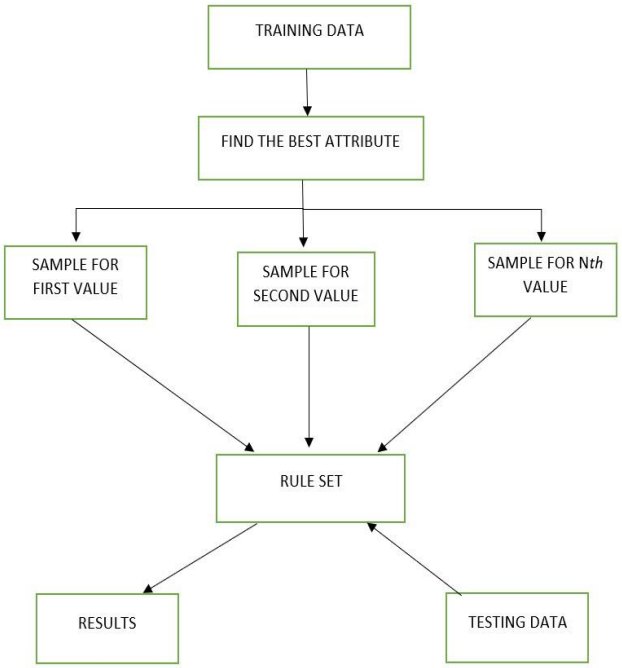

Fig -2: Flow Chart

During the implementation phase, as shown in figure 2 there exists a planning strategy and the first planning paradox used is an algorithm and the algorithm used is a decision tree algorithm. The above flowchart demonstrates each repository machine where the project implementation occurs. Each repository is locked and a paradox is created and the box is updated to meet the requirements of the dataset. The flowchart prevents the cause and the change that any other individual wants to access. Therefore, it provides necessary protection which can also be called as the locking system. The details present in the flowchart can be checked and be further reverted to the previous stage. It is a one step to another step flow of information from one state to next generating step.

Both the architectures follow the same design cycle but the difference lies in its presentation. The Associative method has two types of Architectural design called as Type I Architecture and Type II Architecture. Type II Architecture which is in the form of a flowchart is defined as a pictorial representation of information and the gap between a flowchart and a data flow diagram. There is a retrospective flow and it is released in each module.

The flow also contains latest changes allowing the user to access the data conveniently. The content can also be split. There is also a comparative study which is explained briefly in the coming unit. Multiple changes from the flowchart can be resolved in order to prevent further conflicts. It provides a feature of rolling back and processing each and every phase of the flowchart.

5. RESULTS

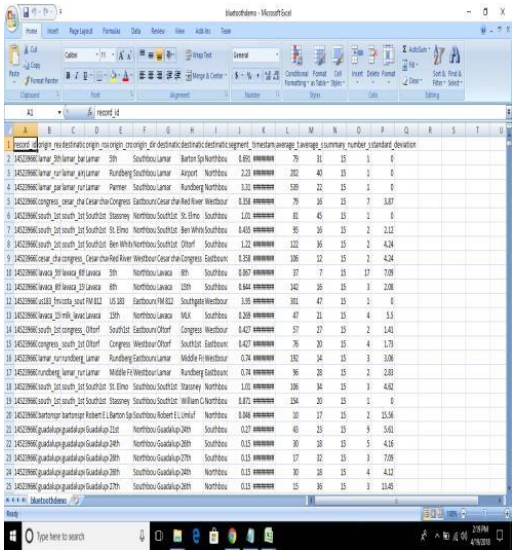

Fig -3: Dataset

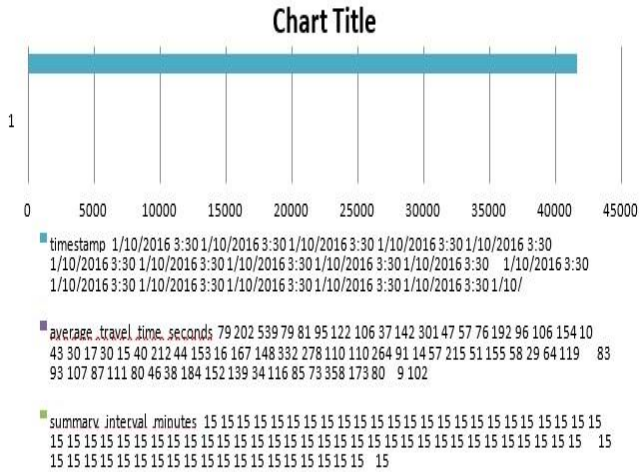

Fig -4: Chart

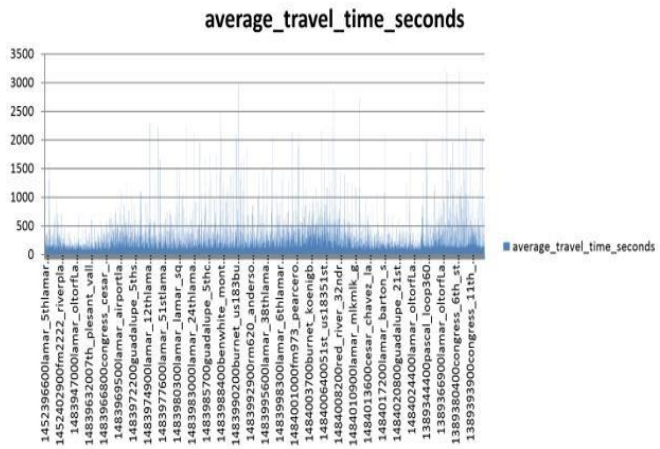

Fig -5: Classified Graph

CONCLUSION AND FUTURE WORK

Distributed Associative Classifier is proved to be more scalable, Bluetooth dataset played a very important role during the classification phase, the final graph is formed from the dataset taken. Distributed Associative Classifiers make use of Associative Rule Mining, associative rule makes use of two variables 'A' and 'B' and is defined as 'A' yields to 'B' represented as A=>B. It can also be used for variables 'M' and 'N' and is represented as M=>N. Figure 3 describes the input data, figure 4 and figure 5 shows the output of DAC. DAC can be further used by taking a dataset of more than five to ten giga bytes. Datasets can also be downloaded from the Data.gov website. Thus, the Distributed Associative Classifier can scale and be more efficient, accurate compared to the state-of-the-art solutions.

REFERENCES

[1] Fabien Dubosson Stefano Bromuri, and Michael Schumacher "Machine Learning Algorithms and features evaluation", 2016.

[2] Kiran Chavan, Priyanka Kulkarni, Pooja Ghodekar "Frequent Itemset Mining for Big Data", Inf Sci. 2016.

[3] Baralis E, Chiusano S, Garza "A lazy approach to associative classification". P. IEEE Transactions on knowledge and Data Engineering. Vol. 20, No. 2, . 2015.

[4] Li W, Han J, Pei J. "Accurate and efficient classification based on multiple class-association rules", CMAR: Proceedings IEEE international conference on, data mining, 2015. New York: IEEE; 2015. p. 369–76. conference on, data mining, New York: IEEE; 2017 p. 369–76.

[5] Bechini A, Marcelloni F, Segatori "A Boosted Random Forest for associative classification", Inf Sci. 2016.

[6] Luca Venturini, Elena Baralis and Paolo Garza "Scaling Associative classification for very large datasets". In: ICDM 2001, Proceedings IEEE international conference on, data mining, New York: IEEE; 2017 p. 369–76.

BIOGRAPHIES

Mr. Bhanu Prakash H N pursuing Master of Technology in Computer Science and Engineering at B.N.M Institute of Technology, Bengaluru, Karnataka 560070.

Dr. Niharika Kumar received the Ph.D. degree in Computer Science from the Visvesvaraya Technological University (VTU), Belgaum, India. Area of Specialization is Networking and Wireless Communication. She has over fifteen years of teaching experience. She has published many articles in National, International, Journals and Conferences. She has also published a book chapter in International Publication.