AK Miner: An Online Data Mining Tool

Kshitij Garg[*], Akshay Shah, Kajal Gupta, Rama Devi
Department of Information Science and Engineering
PES Institute of Technology
Bangalore, India
[*]Email: kshitij.2054 [AT] gmail.com

Abstract—**The world today is moving faster in the software industry than anywhere else. To keep data mining tools accessible to everyone is a Herculean task. We plan to solve this problem in this paper and tool. AK Miner is an online data mining tool that is easily accessible by anyone with a computer. The one of the advantages of AK Miner over the already pre-existing Data mining Tools is that it is not a single step producer for the output, allowing students and industry-grade scientists to understand how the final output is achieved. A visual representation of the output makes it easier to understand the system. There are five types of algorithms present, namely, DBScan, K Means, FP Growth, Apriori, and K Nearest Neighbour. Basically the user can choose which of these algorithms are required to be used to provide the requested output.**

Keywords- **Apriori algorithm, Frequent Patterns, K Means Algorithm, K nearest neighbor algorithm, data mining, mining component**

I. Introduction

The existing Data Mining Tools like R, Orange, weka; all have a single step procedure of taking in the input and providing the required output. The output provided is in the form of data. To have a more illustrated experience, AK Miner is being developed so the output is well explained and is provided output in graphical representation.

And for the people finding something more interesting, our whole web application is a single webpage load design.

Our application provides user a cutting edge benefit as it provides:

- Greater speed as website is single webpage design
- Vast choices to choose user chosen datasets to examine
- Wide variety of mining algorithms providing cutting edge support and mining functionalities
- Better layout of the mining results providing a helping hand for users to analyse the mining results in more depth.
- Animations in regards to mining algorithms to get better insight into mining techniques.

This Data Mining Tool is present online and can be easily accessed by student, companies, etc. without any requirement of downloading. The input is in the form of an arff, csv and xsl, which is taken into the system and the output that is provided in the form of a graph for visual understanding of the output.

AK Miner is a tool that can be used in captivating a large audience, with its visual graphic display the user can easily understand what the final output looks like. Instead of reading long output and trying to understand it, this tool automatically makes a graph and displays the output. AK Miner focuses on the basic requirement of the user and deals with them keeping in mind the vast variety of algorithms implemented. (Such as Apriori, FP Growth, K-means, K Nearest Neighbour and DBSCAN).

II. Data Mining Tools and algorithms

A. *Data Mining*

Data mining process used by companies to turn raw data into useful information. By using software to look for patterns in large batches of data, businesses can learn more about their customers and develop more effective marketing strategies as well as increase sales and decrease costs. Data mining depends on effective data collection and warehousing as well as computer processing.

B. *Existing Data Mining Tools*

There are various tools that are existing are R Tool, Orange, weka, etc. The major disadvantage of these tools is that the output provided is data, which is difficult to understand. It lacks graphical representation, as the visual understanding becomes easier to work with.

Weka is a workbench that contains a collection of visualization tools and algorithms for data analysis and predictive modelling, together with graphical user interfaces for easy access to this functionality.

The major disadvantage of Weka is that when used by students it is a single step procedure where the various inner steps in the algorithm are not seen. These are limited in number; however, there is a library that provides access to the MOA data stream software containing state-of-the-art algorithms for large datasets or data streams. Note also that non-incremental learning algorithms can be applied to large datasets by subsampling the data. Reservoir sampling is an incremental sampling method implemented in Weka that can be used for this purpose.

R is a programming language and software environment for statistical computing and graphics. The R language is widely used among statisticians and data miners for developing statistical software and data analysis. The usage of R is widely essential when it comes to ggplot visual representation of resulting data. The major disadvantage with R is that it is a one-step procedure; the output is not well defined as a step by step analysis is not done.

Orange is a component-based data mining and machine learning software suite, featuring a visual programming front-end for explorative data analysis and visualization, and Python bindings and libraries for scripting. Orange library is a hierarchically-organized toolbox of data mining components. The major disadvantage is accessing the Tool is extremely difficult and makes it complicating for students to use.

C. Algorithms Used

There are five algorithms used in the Data Mining Tool namely, DB Scan, FP Growth, Apriori, K Means and decision tree.

K Means is a partition clustering approach [1], which is the basic algorithm is very simple. With each cluster there is an associated centroid. Each point is allocated to the cluster with the closest centroids. The number of clusters, k must be specified. The first centroid is randomly selected. The centroid is the mean is the cluster.

DBSCAN clustering is basically clustering of data items from vector array or distance matrix. As name suggests, it is Density-Based Spatial Clustering [4] of Applications with Noise .It helps in finding the core samples of high density and expands clusters from them, perfect for data which contains clusters of resembling density.

Apriori is one of the most popular data mining approaches is to find frequent item sets from a transaction dataset and derive association rules. Finding frequent item sets (item sets with frequency larger than or equal to a user specified minimum support) [2] are not trivial because of its combinatorial properties. Once frequent patterns are obtained, it is effortless to generate association rules with confidence larger than or equal to a user specified minimum confidence.

FP-Growth: allows frequent item set discovery without candidate item set generation [3] [4]. It comprises of two steps: Firstly, building a compact data structure called the FP-tree and then building using 2 passes over the data-set.1. Finally, extracting frequent item sets directly from the FP-tree.

K-Nearest Neighbours This algorithm [5] [6] provides a pattern study of how actually are the items distributed around the item we are focusing. It suggests and appropriates the neighbours which are considered to be the nearest neighbours of the item in focus. It also tells how exactly are the items of interest are distributed and the distribution pattern.

III. LITERATURE SURVEY

As we move towards the age in which every bit of thing seems to be done automatically and the way we like, more important is the thing is it is being done in which each individual like, it’s like every bit seems to be done in so called my-way. How this could be even possible? To answer this, we move towards something called as big data and to understand something out of big-data, we need data mining tools. Hence, in today’s era, where mining of important data is being done at every step to help predict the needs and favorable points of every individual, the want for a better and more efficient, accurate mining algorithm is on hunt and we take you to the making of an industry changing software, which provides a better insight to data mining algorithms and better depiction of results.

What’s different and great is:

- Our software is available to use online as a website so no more hassles of downloading and installing and non-supportive OS environments problem
- Our software website is a Static-Loading, i.e. website is loaded just once and all other things will follow without any other different page required to load.
- Our software has self-explanatory diagrams which help in getting better insight to the selected mining technique.

Now, we move towards to explain the mining techniques we have chosen to incorporate in this version of our software ‘AK Miner’:

A. Apriori Algorithm

Apriori algorithm is a frequent pattern mining algorithm that is best suited for some applications like market data analysis, which products are generally purchased together? Apriori is one of the most popular data mining approaches is to find frequent item sets from a transaction dataset and derive association rules. Finding frequent item sets (item sets with frequency larger than or equal to a user specified minimum support) are not trivial because of its combinatorial explosion. Once frequent item sets are obtained, it is straightforward to generate association rules with confidence larger than or equal to a user specified minimum confidence.

B. FP Growth Algorithm

FP Growth allows frequent item set discovery without candidate item set generation. Two step approach, build a compact data structure called the FP-tree and built using 2 passes over the data-set and extracts frequent item sets directly from the FP-tree.

C. K Nearest Neighbours Algorithm

Each describes an instance and gives the class to which it belongs. As before, we’ll assume instances are described by a set of attribute-value pairs, and there is a finite set of class labels L. So the dataset comprises examples of the form h{A1 = a1, A2 = a2, . . . , An = an}, class = cli.

D. K Means Clustering

K-means clustering is a method of vector quantization, originally from signal processing, that is popular for cluster analysis in data mining. *K-means* clustering aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. The problem is computationally difficult (NP-hard); however, there are efficient heuristic algorithms that are commonly employed and converge quickly to a local optimum. These are usually similar to the expectation-maximization algorithm for mixtures of Gaussian distributions via an iterative refinement approach employed by both algorithms. Additionally, they both use cluster centres to model the data; however, *k*-means clustering tends to find clusters of comparable spatial extent, while the expectation-maximization mechanism allows clusters to have different shapes.

E. DB SCAN Algorithm

DBSCAN requires two parameters: ε (eps) and the minimum number of points required to form a dense region. It starts with an arbitrary starting point that has not been visited. This point's ε-neighborhood is retrieved, and if it contains sufficiently many points, a cluster is started. Otherwise, the point is labeled as noise. If a point is found to be a dense part of a cluster, its ε-neighborhood is also part of that cluster. Hence, all points that are found within the ε-neighborhood are added, as is their own ε-neighborhood when they are also dense. This process continues until the density-connected cluster is completely found. Then, a new unvisited point is retrieved and processed, leading to the discovery of a further cluster or noise.

IV. IMPLEMENTATION

This web application is largely implemented using python. The main programs of algorithms that carry out the mining operations effortlessly are crafted using python.

This web application projects the following mining algorithms i.e. Apriori, FP Growth, DB Scan, K-means, K-nearest-Neighbors and Decision Trees.

This design provides the user with much needed speed by loading the web page only once and the whole web application is mounted on that single page means faster rendering of results. Our web application is also made to teach and we encourage learning, hence, comes loaded with lots of exciting self-learning and interactive diagrams that help the user to get a better insight into the mining techniques used. This application is made into a web application with the use of web programming languages such as JavaScript, CSS and Flask.

This application is designed in 3 modules:

Firstly, uploading of the dataset that is chosen by the user.

Secondly, choosing of the appropriate mining algorithm with respect to type of uploaded dataset.

And lastly, analyzing the results generated by the mining technique and better visualization using graphical projection techniques.

By designing such an application, we try provide the world of mining, something to consider and something from which everyone can learn.

V. CONCLUSION

Software, to find important information from data using mining tools, has been developed successfully. User requirements and basic functionalities are implemented and tested.

Application provides web interface for end user. Web interface has one user type:

1. Unregistered

It has functionality for the user to choose his/her own dataset file which is valid with the particular algorithm and take advantage of mining algorithms on own data.

This software is a static-loading webpage that is the page needs load only once and after that no new webpage is opened, everything on single webpage. It is a single webpage software designed.

At last, this software provides adequate space for the user to choose own dataset files and then choose the relevant option of the algorithm to apply upon the data file generating a whole new and more explanatory view of mining the data and finding relevant and important information.

Software also provides help regarding mining algorithms used so user can go and check it out for a greater insight into mining techniques for a better understanding using interactive diagrams.

This tool is helpful to a broad variety of people, students, start-up companies, etc. all these individuals require a step by step or a visual analysis of data which is provided by AK Miner.

VI. FUTURE ENHANCEMENT

- Software can be launched as separate applications for android, iOS and windows devices.
- More algorithms can be added to the software. More graphical outputs with interactive schemes for users.
- Algorithms can be adaptive to user’s data request.
- Increasing the number of algorithms within the tool improves the variety and the demand of the product. Also putting it on cloud, makes it more accessible to people all around the world.

REFERENCES

[1] Shu-Chuan Chu;Roddick, J.F.;Tsong-Yi Chen;Jeng-Shyang Pan “Efficient search approaches for k-medoids-based algorithms”,2002 IEEE Region 10 Conference on Computers, Communications, Control and Power Engineering,Feb 2003.

[2] Rakesh Agrawal and Ramakrishnan Srikant “Fast algorithms for mining association rules in large databases”. Proceedings of the 20th

International Conference on Very Large Data Bases,Santiago, Chile, September 1994 pp. 221-230.

[3] K. Alsabti, S. Ranka, and V. Singh, “An Efficient k-means Clustering Algorithm,” Proc. First Workshop High Performance Data Mining, Mar. 1998. Pp. 201-220

[4] Bayardo Jr, Roberto J. (1998). "Efficiently mining long patterns from databases". ACM Sigmod Record pp. 120-123.

[5]] Tepwankul, A. ; King Mongkut"s Univ. of Technol. Thonburi, Bangkok, Thailand ; Maneewongwattana, S. “U-DBSCAN : A density-based clustering algorithm for uncertain objects” Data Engineering Workshops (ICDEW), 2010 IEEE 26th International Conference on 1-6 March 2010 pp. 1042-1050.

[6] Viswanath, P. ; Dept. of Comput. Sci. & Eng., Rajeev Gandhi Memorial Coll. of Eng. & Technol., Nandyal, India ; Sarma, T.H. “An improvement to k-nearest neighbor classifier” Recent Advances in Intelligent Computational Systems (RAICS), 2011 IEEE, 22-24 Sept. 2011 pp. 10-20