World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:9, No:3, 2015

Comprehensive Analysis of Data Mining Tools

S. Sarumathi, N. Shanthi

***Abstract*—**Due to the fast and flawless technological innovation there is a tremendous amount of data dumping all over the world in every domain such as Pattern Recognition, Machine Learning, Spatial Data Mining, Image Analysis, Fraudulent Analysis, World Wide Web etc., This issue turns to be more essential for developing several tools for data mining functionalities. The major aim of this paper is to analyze various tools which are used to build a resourceful analytical or descriptive model for handling large amount of information more efficiently and user friendly. In this survey the diverse tools are illustrated with their extensive technical paradigm, outstanding graphical interface and inbuilt multipath algorithms in which it is very useful for handling significant amount of data more indeed.

***Keywords*—**Classification, Clustering, Data Mining, Machine learning, Visualization

I. Introduction

THE domain of data mining and discovery of knowledge in various research fields such as Pattern Recognition, Information Retrieval, Medicine, Image Processing, Spatial Data Extraction, Business and Education has been tremendously increased over the certain span of time. Data Mining highly endeavors to originate, analyze, extract and implement fundamental induction process that facilitates the mining of meaningful information and useful patterns from the huge dumped unstructured data. This Data mining paradigm mainly uses complex algorithms and mathematical analysis to derive exact patterns and trends that subsists in data. The main aspire of data mining technique is to build an effective predictive and descriptive model of an enormous amount of data. Several real world data mining problems involve numerous conflicting measures of performance or intention in which it is needed to be optimized simultaneously. The most distinct features of data mining are that it deals with huge and complex datasets in which its volume varies from gigabytes to even terabytes. This requires the data mining operations and algorithms are robust, stable and scalable along with the ability to cooperate with different research domains. Hence the various data mining tasks plays a crucial role in each and every aspect of information extraction and this in turn leads to the emergence of several data mining tools. From a pragmatic perspective, the graphical interface used in the tools tends to be more efficient, user friendly and easier to operate, in which they are highly preferred by researchers [1].

Mrs.S.Sarumathi, Associate Professor, is with the Department of Information Technology, K. S. Rangasamy College of Technology, Tamil Nadu, India (phone: 9443321692; e-mail: rishi_saru20@rediffmail.com).

Dr.N.Shanthi, Professor and Dean, is with the Department of Computer Science Engineering, Nandha Engineering College, Tamil Nadu, India (e-mail: shanthimoorthi@yahoo.com).

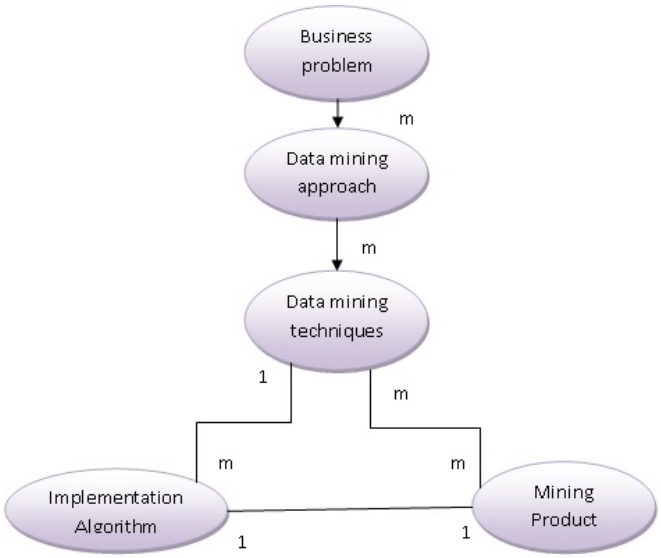

Fig. 1 A Data Mining Framework

Revolving into the relationships between the elements of the framework has several data modeling notations pointing towards the cardinality 1 or else m of every relationship. For these minimum familiar with data modeling notations.

- A business problem is studied via more than one classes of modeling approach is useful for multiple business problems.
- More than one method is helpful for any classes of model plus any known methods is used for more than one classes of models.
- There is normally more than one approach of implementing any known methods.
- Data mining tools may sustain more than one of the methods plus every method is supported by means of more than one vendor's products.
- For every known method a meticulous product supports a meticulous implementation algorithm [2].

II. Different Data Mining Tools

A. DATABIONIC

The Databionic Emergent Self-Organizing Map tool [3] is a collection of programs to do data mining tasks such as visualization, clustering and classification. Training data is a collection of points from a high dimensional space known as data space. A SOM contain a collection of prototype vectors in the data space plus a topology between these prototypes. Commonly used topology is a 2-dimensional grid where every prototype that is neuron has four direct neighbors and the locations on the grid from the map space. Additional two distance functions are necessary for each space. Euclidean

distance is normally used for the data space, then the City block distance in the map space. The function of SOM training is to adapt the grid of prototype vectors to the specified data generating a 2-dimensional projection that conserves the topology of the data space.

The update of a prototype vector using a vector of the training data is a dominant operation at SOM training. The prototype vector of the neuron is drawn nearer towards a given vector in the data space. The Prototypes in the district of the neuron are drawn in the similar direction with less emphasis. During training the emphasis and the size of the district are reduced. Online and Batch training are the two common training algorithms both searches the closest prototype vector for each data point. The best match is updated immediately in online training, but in Batch training first the best matches is collected for all data points then the updates if performed together.

Emergency is the capacity of a system to improve higher level structures using the teamwork of various elementary processes. The structures evolve inside the system without external influences in self-organizing systems. Emergency is the form of high level phenomena which cannot be derived from the elementary processes. An emergent structure provides an abstract explanation of a complex system containing low level individuals. Transmitting the principles of self-organization to data analysis is achieved by allowing multivariate data points form themselves into homogeneous groups. Self-organizing Map is a well-known tool for this task that integrates the above mentioned principles. The SOM iteratively regulates to distance structures in a high dimensional space. That provides a low dimensional projection that reserves the topology of the input space as possible. The map is used in unsupervised clustering and supervised classification. The emergence of structure in data is frequently neglected by the power of self-organization. In the scientific literature this part is a misusage of SOM. The maps consist of some tens of neurons, which is used by some authors are commonly very small.

1) Features

a) Training of ESOM in dissimilar initialization methods, distance functions, ESOM grid topologies, training algorithms, neighborhood kernels and parameter cooling strategies
b) Visualization of high dimensional data space using p-Matrix, SDH and more
c) Animated visualization of the training process
d) Link ESOM
e) Scalable with more data to the training data, data descriptions and data classifications using clustering, interactive and explorative data analysis
f) Formation of ESOM classifier plus automated application of new data
g) Formation of non-redundant U-Maps from toroid ESOM
h) Databionic ESOM Analyzer
i) U-Max of hexa dataset

B. ELKI

ELKI is open source [4] (AGPLv3) data mining software written in Java. The aim of ELKI is research in algorithms, with an emphasis on unsupervised techniques in cluster analysis and outlier detection. ELKI provides huge data index structures like R* tree that offer main performance gains to attain high performance and scalability. ELKI is planned to be simple to extend for researchers and students in this domain plus welcomes contributions in specific of new methods. ELKI tries to provide that a huge collection of paraeterizable algorithms, that allows simple and fair evaluation and benchmarking of algorithms.

Data mining research directs to several algorithms for similar tasks. A fair and useful similarity of these algorithms is complex due to some reasons:

- Implementation of contrast, partners are not at hand.
- If implementations of dissimilar authors are available and a valuation in terms of effectiveness is biased to evaluate the hard works of dissimilar authors in well-organized program instead of estimating algorithmic merits.

An alternatively efficient data management tool similar to index-structures is able to show considerable impact on data mining tasks and is useful for a diversity of algorithms. Data mining algorithms and data management tasks in ELKI are divided and allow for a free evaluation. This division creates ELKI unique between data mining frameworks similar to Weka or Rapid Miner along with frameworks for index structures like GiST. Simultaneously ELKI is open to random data types, space or similarity measures, or file formats. The primary approach is the independence of fie parsers or database connections, data mining algorithms, distances, distance functions. They trust to serve the data mining and database research community usefully along with the development and publication of ELKI. The framework is open source for scientific usage. In application of ELKI in scientific publications which they would value credit in the form of a citation of the suitable publication that is, the publication related to the release of ELKI they were using.

The design goals are extensibility, Contribution, completeness, Fairness, Performance, progress. Extensibility in ELKI has a modular design and permits random combinations of data types, input formats, distance functions, index structures, algorithms and evaluation methods. Contributions in ELKI improve people contribute. By using a modular design that permits small charity like single distance functions plus single algorithms. They have students and external charity participate in the development of ELKI. Completeness for an exhaustive contrast of methods and they aspire at covering as available and qualified work as they can. Fairness is simple to do an unjust comparison by roughly implementing a competitor. They need to implement each method and publish the score code permit for external improvements. The aim to plus proposed improvements like index structures in earlier range and kNN queries. Performance is the modular architecture of ELKI that permits optimized versions of the algorithms and index structure for acceleration. Progress in ELKI is modified with each release.

To hold new features and enhance performance API breakage are necessary. They get a stable API with the 1.0 free.

1) Features

a) It offers various data index to attain high performance and scalability
b) User friendly to spread out for students and scholars
c) It has a high performance modular architecture

C. MALLET

MALLET is a Java-based package [5] for statistical natural language processing, clustering, information extraction and other machine learning applications to text. MALLET comprises sophisticated tools for document classification which offers an efficient routine for converting text to "features" with a wide range of algorithms plus code for calculating classifier performance by using several user metrics. It comprises tools for sequence tagging for applications as named-entity extraction from text. Algorithms consist of Maximum Entropy Markov Models, Hidden Markov Models and Conditional Random Fields. For finite state transducers, these methods are implemented in an extensible system. MALLET topic modeling is useful for analyzing large collections of unlabeled text and toolkit covers efficient, Pachinko Allocation, sampling-based implementations of Latent Dirichlet Allocation and Hierarchical LDA. Algorithms in MALLET based on numerical optimization. MALLET comprises an effective implementation of Limited Memory BFGS, between other optimization methods. In addition to sophisticated Machine Learning applications, MALLET contains procedures for converting text documents into numerical representations that can be processed effectively. This method is implemented through a flexible system of "pipes" which manage different task includes removing stop words, tokenizing strings and transforming sequences into count vectors. An additional package to MALLET is known as GRMM which includes support for implication in common graphical models plus training of CRFs with arbitrary graphical structure. The toolkit is Open Source Software then it is released in the Common Public License. The codes are used under the terms of the license for research and commercial purposes.

1) Features

a) It supports Java-based package
b) For document classification it compresses with sophisticated tools
c) It is convenient for evaluating huge collections of unlabeled text

D. ML-Flex

ML-Flex [6] makes use of machine-learning algorithms to derive models from independent variables with the determination of predicting the values of dependent variables. Sir Ronald Fisher introduced machine-learning algorithms have been applied to the Iris data set in 1936, which holds four independent variables such as sepal width, sepal length, petal length, petal width and one dependent variable are species of Iris flowers equal to virginica, setosa, or versicolor. Machine-learning algorithms can distinguish among the species with near-perfect accuracy.

In performing a machine-learning experiment main important aspect to consider is the validation strategy. Using the wrong kind of validation approach biases is introduced and they may appear as an algorithm which has the extra predictive ability than it has. Most commonly used validation strategy is the cross validation that helps us to avoid such biases. The data instances are partitioned into a number of groups "k" and each group is held separate as "test" instances. The Algorithm develops a model by the remaining "training" instances, and then the model is applied to the test instances. The performance of the algorithm is evaluated by using how well the test instances predictions concur with the predicted actual values. A value for "k" is 10 that are ten-fold cross validation. Variations are leaving-one-out cross validation where "k" is equal to the number of data instances and a simple training or testing split where the data are partitioned and part of the data are used for testing. An additional variation is to use nested cross-validation inside each training set. Cross validation is used to optimize the model previously it is applied to the "outer" test set. Cross validation step is repeated multiple number of times on the same data set to evaluate the robustness of their results as data instances are assigned randomly to folds. Such validation strategies are useful and computationally intensive mainly for large data sets. ML-Flex addresses their enabling analyses to be split into multiple threads on a single or multiple computers which leads to shorter execution times.

Cross validation is applied carefully or biases are introduced. Training sets should never overlap with test sets which are complicated in the nested cross validation and model optimization can be applied to the training sets, for example applying feature-selection algorithms to find the relevant features. Some algorithms are more than enough to collect on subtle variations in a data set and accuracy is attained if the approach is skilled on the full data set. This phenomenon is a harmful case of over fitting the data. ML-Flex architecture prevents such biases.

Machine-learning algorithms are developed using programming languages. They offer incompatible ways of interfacing with them. ML-Flex interface with any algorithm which provides a command-line interface. This flexibility allows users to process machine-learning experiments along with ML-Flex as a harness, whereas applying algorithms that are developed in dissimilar programming languages or else they offers dissimilar interfaces.

1) Features

a) Integrate with additional machine-learning packages
b) Generate HTML based reports results
c) Execute in parallel over multiple threads

E. Scikit-Learn

Scikit-learn [7] offer a wide range of supervised plus unsupervised learning algorithms through a consistent

interface in python. It is licensed under a permissive simplified BSD license and is scattered under multiple Linux distributions and commercial use. The library is developed on the SciPy that is Scientific Python that should be installed earlier they can use Scikit-learn. These stacks consist of NumPy is Base n-dimensional array package, Scipy is a fundamental library for scientific computing, Matplot lib is the comprehensive 2D or 3D plotting, I Python is enhanced interactive console, Sympy is symbolic mathematics, Pandas is data structures and analysis. Extension or else the modules for SciPy care conservatively named as SciKits. Such the module offers learning algorithms and is named as Scikit-learn. The version of the library is a stage of robustness and that support necessary use of prediction systems. It heads a deep center on concerns like code quality, performance, easy to use, documentation and collaboration. Though the interface is python, c-libraries are influences for a performance like LAPACK, numpy for arrays, plus matrix operations, LibSVM and the use of python.

1) Features

a) It offers both supervised and unsupervised learning algorithms
b) Easy to use
c) It has high performance

F. Shogun

Shogun is an open source and free toolbox [8] written in C++. It provides many algorithms and data structures for machine learning problems. It is licensed below the condition of the GNU General Public License version 3. The center of shogun is on kernel machines like classification problem, support vector machines for regression. Shogun provides full implementation of Hidden Markov models. The main aim of shogun is written in C++ and other interfaces like R, Java, C#, etc., since 1999 shogun is in under active construction. There is a vibrant user community using shogun as a base for research and education and contributing to the core package.

Shogun supports Support Vector machines, Linear discriminant analysis, Kernel Perceptron's, K-Nearest Neighbors, Hidden Markov Models, Clustering algorithms like k-mean and GMM, Dimensionality reduction Embedding like Iso map, PCA, Linear Local Tangent Space Alignment etc., kernel Ridge Regression, Support Vector Regression, etc., Most dissimilar kernels are implemented and ranging from kernels from numerical data to kernel on special data. At present implemented kernels for numeric data contain polynomial, linear, sigmoid kernels, Gaussian. The kernel supports for data consist of weighed Degree, Spectrum, and Weighted Degree with Shifts. In later group of kernels permit processing of arbitrary sequences among fixed alphabets like DNA sequences with full e-mail texts.

1) Features

a) It supports the pre-calculated kernels
b) It is likely to use a mutual kernel that is a kernel, consisting of a linear mixture of arbitrary kernels
c) It provides a multiple kernel learning functionality
d) The coefficient or else weights of the linear mixture are learned

G. FITYK

FITYK [9] is an agenda for nonlinear fitting of analytical functions to data. The main brief description is peak fitting software. There are group using it to take away the baseline from data or else to show data only. It is used in Raman spectroscopy, crystallography and so on. Authors has a common understanding of experimental techniques other than powered diffraction and like to create it more useful to as many groups as possible. FITYK provides a variety of nonlinear fitting methods and simple background subtraction and other manipulations to the datasets, support for analysis of series of datasets, easy placement of peaks and changing of peak parameters, and more. Flexibility is the major advantage of the program which parameters of peaks can be arbitrarily bound to each other. For example the width of a peak is an independent variable and the same as the width of other peak or can be given by difficult formula.

1) GUI vs. CLI

The program is divided into two versions such as Graphical User Interface which is comfortable for users and Command Line Interface version named as cfityk.

2) Features

a) Instinctive interfaces like graphical and command line
b) Support for data file formats and, thanks to the xylib library
c) A group of build-in functions and maintain for user-defined functions
d) Equality constraints
e) Appropriate systematic errors of the x coordinate of points
f) Manual, graphical placement of peaks and auto-placement by peak detection algorithm
g) Several optimization methods
h) Handling serious of data sets
i) Computerization with macros and embedded Lua for extra complex scripting
j) From NIST the precision of nonlinear regression confirmed with reference datasets
k) An append for powder diffraction data
l) Modular architecture
m) Open source license (GPL)

H. PyBrain

A modular Machine Learning Library for Python is PyBrain [10]. Its aim is to provide flexible, easy-to-use yet still algorithms for Machine Learning Tasks as well as a range of predefined environments to test and compare your algorithms. PyBrain is Python-Based Reinforcement Learning, Neural Network Library and Artificial Intelligence. In general, it came up with the name first and later reverse-engineered this rather descriptive "Backronym".

1) How Is PyBrain Different?

There are some machine learning libraries out there where PyBrain goals are to easy-to-use modular library that can be used by entry-level students, but they provide the flexibility and algorithms for state-of-the-art research. They are continually working on quicker algorithm and developing fresh environments plus improving usability.

2) What PyBrain Can Do?

PyBrain is a tool for real-life tasks. It has algorithms for neural networks for unsupervised learning, for reinforcement learning and evolution. Most of the problems contract with continuous state and action spaces and function approximates have to use to manage with the large dimensionality. The library is constructing around neural networks in the kernel plus most of the training techniques allow a neural network like the to-be-trained instance.

3) Features

a) Supervised Learning
b) Black-box optimization/ Evolutionary Methods
c) Reinforcement Learning
d) Tasks and Benchmarks
e) Compositionality

I. UIMA

UIMA [11] is an Unstructured Information Management Application. UIMA is a software system that analyzes large volumes of unstructured information in turn to find out knowledge that is related to an end user. Example of UMI application may ingest plain text and recognize entities like organizations, persons, etc., UIMA allow applications to be decayed into components. Each component implements interface defines the framework plus offers self-describing metadata through XML descriptor files. The framework maintains these components plus the data flow between them. Components are written in Java or C++ and the data that flows between components is planned for efficient mapping among these languages.

UIMA offers capabilities to cover components as network services plus they can scan to big volumes through replicating processing pipelines over a cluster of networked nodes. Apache UIMA is an Apache-licensed open source execution of the UIMA requirements. Frameworks, Components and Infrastructure are licensed under the apache license.

The frameworks are obtainable for both C++ and Java and they run the components. The Java Framework supports both Java and non-Java components. The C++ framework supports annotator in C or C++ and also supports Perl, Python, and TCL annotators. The UIMA-AS plus UIMA-DUCC is Scale out Frameworks and they are add-ons to the Java framework. The UIMA-AS is flexible scale out capability depends on Java Messaging Services plus Active MQ. The UIMA-DUCC expands UIMA-AS By giving cluster management services to scale-out of UIMA pipelines through computing clusters. The frameworks support configure plus running pipelines of Annotator components. These components do the analyzing the formless information. They can write or else configure and use pre-existing annotators. Few annotators are available and others are in several repositories on the internet. Extra infrastructure support components contain a simple server which receives REST requests plus revisit annotation outcomes.

1) Features

a) Platform independent data representations and interfaces for text and multi-modal analytics
b) Analyze unstructured information

J. Natural Language Toolkit (NLTK)

NLTK is a most important platform for structuring Python programs to work with human language data. It offers easy-to-use interfaces to above 50 corpora, plus lexical resources like Word Net with a matching set of text processing libraries for tagging, classification, semantic reasoning, tokenization, parsing, stemming and an active discussion forum. NLTK is fitting for researchers, linguists, etc., NLTK are obtainable for Linux, Windows, OS X and Mac. NLTK is an open source, free and community-driven project. NLTK is called as a tool for working in computational linguistics by Python plus a library to play with natural language. It offers a programming for language processing. NLTK direct the basics of writing Python programs and other [12].

NLTK plans with few main goals.

a) Simplicity

To offer an instinctive framework with considerable building blocks which gives sensible information of NLP with no receiving bogged down in the boring housekeeping typically associated along with processing annotated language data.

b) Consistency

To offer a uniform framework along with reliable interfaces plus data structures and effortlessly guessable method names.

c) Extensibility

To offer a formation in which novel software modules can be effortlessly accommodated which containing option executions plus rival approaches to the same work.

d) Modularity

To offer components that is used separately with no needing to realize the rest of the toolkit .

NLTK is developed at the University of Pennsylvania in combination with a computational linguistics course in 2001. It is planned with three applications like projects, assignment and demonstrations.

1) Assignments

NLTK supports assignments of complexity and scope. With the easiest assignments, students experiment with components to do a variety of NLP tasks. As students become more well-known with the toolkit and change existing components or else they make the entire system out of existing components.

2) Demonstration

The NLTK's interactive graphical display is confirmed to be helpful for students learning NLP concepts. The display gives a step-by-step implementation of algorithms which demonstrating the current state of data structures.

3) Projects

NLTK offers students with a supple framework for higher projects. Usual projects offer applies an algorithm, applying a task, or else developing a component. They select Python because it has its syntax; semantics are transparent plus shallow learning curve and also has string-handling functionality. As an interpreted language the Python makes easy interactive exploration. As an object-oriented language the Python allows data plus methods to be encapsulated and reused. Python has a standard library which includes tools for graphical programming plus numerical processing. The syntax builds it easy to make interactive implementations of algorithms [13].

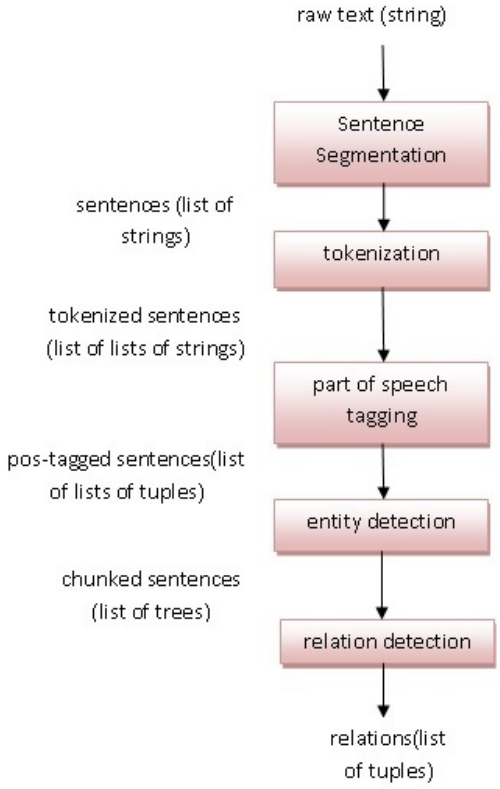

Fig. 2 NLTK Framework

4) Features

a) Platform for structuring Python programs
b) Offers programming for language processing

K. Dlib

Dlib is a cross-platform C++ library planned using agreement programming plus C++ techniques. It is free software plus licensed under the Boost Software License [14].

1) Containers

Containers are enthused by the work of the Reusable Software Research Group at Ohio State. Many of the objects do not maintain copies in any other form only exchanging is permitted. When objects are added or else removed from these containers they will exchanged in and out but not copied. This permit them to do things have containers of containers of containers with no come across the overhead of the huge replication that would outcome if they did the similar thing with the STL. They can store objects which are not duplicated inside these containers that is not they can perform with the SLT prior to C++11. These container swap () plus operator< do not throw. It will not function if this statement is broken and the built in type like int, char, etc., and std::string will not report operator<or else swap () to throw. The majority of the containers comes into from the enumerable interface. The member function comes from enumerable are defined in the enumerable class plus documentation is not frequent in every container documentation. These contain the size () member function in every container [15].

2) Image Processing

The page documents the functions present in the library which contract with the management plus manipulation of images and there is no explicit image object. The lot contract with array2d objects that include several types of pixels, or else, user defined generic image objects [16].

a) Pixel Types

Most of the image conduct routines in dlib will allow images containing pixel type. It is completely possible by defining a traits class, pixel_traits for every pixel type. It traits class enables image processing customs to decide to handle every type of pixel and only pixels that have a pixel_traits is used. The following list defines all the pixel kinds that come with pixel_traits definitions.

- *RGB*

In dlib there are RGB pixel type's rgb_pixel plus bgr_pixel. It defines a 24bit RGB pixel type. The bgr_pixel is the same to rgb_pixel apart from that lays the color channels down in memory in BGR order before RGB order plus helpful for interfacing with other image processing tools which expect this format.

- *RGB ALPHA*

The rgb_alpha_pixel has an 8bit per channel and RGB pixel with an 8bit alpha channel.

- *HSI*

The hsi_pixel is a 24bit pixel that symbolizes a point in the Hue Saturation Intensity color space.

- *Gray Scale*

Built in scalar type is used as a gray scale pixel type. For example, int, unsigned, etc.

3) Containers

An API wrapper offers a transportable object oriented interface for networking, file browsing, multithreading and GUI development. Program written is compiled in POSIX or else MS Windows platforms with no changes in the code [17].

4) Graph Tools

There are two different types of graph representations in dlib. One is, some graphs depend on an object which summarizes the entire graph like graph plus directed_graph objects. Alternatively, there are graphs that are standing for simple vectors of edges. Here, they use vectors of ordered_sample_pair or else sample_pair objects for direct plus undirected graphs [18].

5) Machine Learning

The most important design goal of this part of the library is to offer an extremely high modular and easy architecture for dealing with kernel is parameterized to permit a user to offer either one of the predefined dlib kernels or as a new user defined kernel. In addition to the functioning of the algorithms are totally divorced from the data on which they work. This creates the dlib functioning generic enough to work on any type of data, images, be it column vector, or else a number of other forms of structured data. Each and every one that is essential is an appropriate kernel [19].

6) Features

a) Documentation
b) High Quality Portable Code
c) Threading
d) Networking
e) Graphical User Interfaces
f) Numerical Algorithms
g) Machine Learning Algorithms
h) Graphical Model Interference Algorithms
i) Image Processing
j) Data Compression and Integrity Algorithms
k) Testing
l) General Utilities

L. Jubatus

Jubatus is the first open source platform for online machine learning, distributed computing framework on the data streams of Big Data. Jubatus has multiple features like regression, classification, data mining, etc. [20].

In large databases where computer science will face fresh challenges in Big Data applications like nation-wide M2M sensor network analysis, real-time security observation on the raw Internet traffic, and online advertising optimization for millions of customers. It is useless to be relevant normal approaches for data analysis on little datasets by amassing all data into databases and examining the data in the databases as a batch-processing plus visualizing the summarized results.

The future of data analytics platform should enlarge to three directions at the same time they are maintaining bigger data, performing in real-time and applying deep analytics. There is no such analytics platform for huge data streams of constantly generated Big Data with distributed scale-out architecture. They use a loose model sharing architecture for well-organized training plus sharing of machine learning methods using three fundamental operations like Mix, Update and Analyze and same as with the Map plus decrease operations in Hadoop. The point is to minimize the size of the models plus the number of the Mix operations as maintaining high accuracy while Mix large models for several times causes high networking cost plus high latency in the distributed environment.

Then our improvement team contains component, researchers will merge the latest advances in online machine learning, randomized algorithms and distributed computing to offer well-organized machine learning features for Jubatus. At present it supports essential tasks containing regression, outlier detection, etc., a demo system for chirp classification of quick Twitter data streams is obtainable.

1) Scalable

It supports scalable machine learning processing. They can maintain multiple data per second through commodity hardware clusters. It is planned for clusters of commodity and shared-nothing hardware.

2) Real-Time

It updates a model instantaneously for receiving data plus simultaneously analyzes the data.

3) Deep Analysis

It supports several tasks for deep analysis containing graph analysis, clustering, regression, etc.

4) Difference from Hadoop and Mahout

- There are several points among Hadoop or Mahout and Jubatus. They are scalable plus working on commodity hardware. Hadoop is not prepared with sophisticated machine learning algorithms as the majority of the algorithms do not fit its Map Reduce paradigm. While Apache Mahout is a Hadoop-based machine learning platform and online processing of data streams is out of the scope.
- Jubatus procedures, data in an online manner plus obtain high throughput and low latency. To attain these features it makes use of unique loosely model synchronization for scale out plus quick model sharing in distributed environments.
- They procedures all data in memory plus focus on operations for data analysis [21].

Jubatus is a scalable dispersed computing structure designed for online machine learning. The starting point of the name is the Latin word for that agile animal the cheetah. The foremost aim of Jubatus is to make possible speedy and profound analysis of stream-type big data. The processing consists of desires such as profound, fast, and large volume. Jubatus satisfies both profound analyses along with scalability. Profound analysis is the mechanical categorization of formless information planned for human beings like natural language processing plus mechanical multi-category classification by

maximum speed without lag of a data stream. These three requirements have a trade-off relationship plus inherently difficult to gratify each of them concurrently.

Jubatus gratify both profound analyses also scalability. Everywhere profound analysis is the mechanical categorization of formless information intended for human beings like natural language. Furthermore, it replaces human labor for indistinctly formulated processing work like prediction, etc. Technical perspective comprises challenges in the area of machine learning, etc. Scalability encompasses the matters of (1) enlarges in processing requests (2) enlarges in data size. Issue (1) it is additional classify into throughput plus response. In widespread batch processing focuses on throughput, whereas real-time processing focuses on response. Approaches to issue (2) either process the data without waiting or else dividing plus storing it.

Separate the profound analysis functionality from the scalability, non-functionality. A profound analysis, design the logic of online machine learning to an engine or else CPU that can be constantly upgraded like removable analysis modules. A scalable design is a general infrastructure motherboard that can be scaled by means of installing analysis modules into the general framework. The final aim of Jubatus is to offer everyone with scalable machine learning. The Major strategy is to proffer an extensively easy-to-understand online machine learning framework for large data which is easy to use with hardware which scales out cheap good servers to enable massively parallel distributed processing plus software that is not restricted to a few specialists, data scientists, along with programmers.

5) Applicable Areas

An important classification of Twitter information is the appliance of Jubatus. The information depictions in natural language plus presented with a minimum number of characters, etc. were incorporated into the analysis subject theme. Jubatus hastily classifies the 2000 tweets per second into their matching business categories pus provisions information to an analysis application. This functioning utilizes an online mechanism leaning technique called Jubatus classifiers or else multi-valued classification.

6) Architecture and Functions

Jubatus is collected of a cluster of machine learning engines plus a high-speed framework which ropes them. In dissimilarity to other machine learning engine units that typically handled small-to medium-scale information plus required batch processing plus individual development. Jubatus has a large variety of engines installed in a high-speed framework plus an improved mechanism with general stipulation for high-speeding large data processing with faults within a permitted array tolerated. The categories of machine learning that support Regression, classification, and statistics, etc. Jubatus is a useful application that needs speedy judgment. It is usual to discover plus analysis, creative relationships between the data volumes from dissimilar domains.

7) Distributed Processing Architecture

The flow of large data streams starting left to right. Clients are organized of a number of user processes along with proxy processes. The proxy processes convey the clients' request to the server processes that enables the servers to be transparent to the user processes. User processes are applied by means of using the Jubatus client application programming interface (API) and they are written in a common programming language or else in a scripting language. The communication amongst proxy processes plus server processes are depending on MessagePack remote procedure calls. Non-block input or output enables more well-organized communications plus synchronization control. The server processes perform the training and prediction processing and also learning model synchronization that has linear increasing performance by means of the quantity of servers. Zookeeper procedures, administer the cooperation among proxy plus server processes and the balancing between distributed servers, the selection of new leaders, plus the monitoring of server state that is alive or dead. In parallel processing amid distributed servers major techniques for satisfying jointly profound analysis plus scalability is mix processing. In the mix processing consider are resembling a collection study meeting for self-teaching plus synchronization with other [22].

8) Features

a) Multi-classification algorithms
b) Regression algorithms
c) Feature extraction method for natural language

M. SCaVis

A SCaVis is a background in scientific computation, data visualization plus data analysis planned for students, scientists plus engineers. The program integrates multiple open-source software packages interested in a coherent interface by the idea of dynamic scripting. SCaVis software is used all over the place where an analysis of huge numerical data volumes, statistical analysis, data mining, plus math computations are necessary like modeling, natural science, analysis of financial markets plus engineering. Scarves program is completely multiplatform and works on all platforms where Java is installed. Like a Lava application, SCaVis obtains the complete advantage of multicore processors.

SCaVis is used with various scripting languages for the Java platform likes BeanShell, Groovy, Jython, JRuby. This carries more and more power plus simplicity for scientific computation. The programming is also being completed in native Java. At last symbolic calculations is completed by using Matlab or Octave high-level interpreted language.

SCaVis works on Linux, Windows, and Mac plus Android operating systems. The Android application is known as A Work. Thus the software signifies the final analysis framework that is used on hardware like netbooks, desktops, laptops, android tablets plus production servers.

SCaVis are a transportable application. SCaVis needs no installation. Easily download and open the package and run it. Once run it from a hard drive through a USB flash drive or

else from some media. An individual can carry in the region of with a transportable device plus use on any Linux, Mac, plus Windows computer [23].

1) Features

a) Used for huge numerical data volumes, statistical analysis etc.
b) It has no installation process

N. CMSR

CMSR stands for Cramer Modeling, Segmentation and Rules. It is a data miner and rule-engine suite containing rule-engines as a unique feature. Rule-engines offer rule-based predictive model assessment. By Cho Ok-Hyeong, Rosella Software CMSR is particularly planned for business applications with database focus. It supports several algorithms, helpful for business databases like SOM-based neural clustering, cross table with deviation analysis, neural network, visualization charts, hotspot drill-down analysis and so on.

Cramer approaches from the CMSR decision tree make use of Cramer coefficients as splitting criteria. A unique characteristic of CMSR is rule-engine. Rule-engine offers rule-based predictive model assessment. There are two different types of rule-engines are supported. First is the sequential rule engine. Another is Rete-like rule-engine. These tasks depend on forward chaining plus rule activation. Together supports CMSR models like neural clustering, regression models, decision tree, and neural network. Rules make use of a subset of SQL-99 database query language. This creates it simpler for SQL-users to study [24].

CMSR Data Miner Suite [25] offers an incorporated environment for predictive modeling, rule-based model assessment, etc. It also offers incorporated analytics plus rule-engine environment for higher power users.

1) Features

a) Self-organizing Maps (SOM) -Neural Clustering
b) Neural network predictive modeling
c) (Cramer) Decision tree classification plus segmentation
d) Hotspot drill-down plus profiling analysis
e) Regression
f) Radial Basis Function (RBF) by means of the rule engine
g) Business rules – Predictive Expert systems shell engines
h) Powerful charts: 3D, etc.,
i) Segmentation, plus gains analysis
j) Response plus profit analysis
k) Correlation analysis
l) Cross-sell Basket Analysis
m) Drill-down statistics
n) Cross tables with deviation or hotspot analysis
o) Group by tables with deviation or hotspot analysis
p) SQL query or batch tools

O. Vowpal Wabbit

Vowpal Wabbit is a free source [26] fast out-of-core learning system library plus program urbanized initially at Yahoo! Research, plus at present at Microsoft Research. John Langford guide and initiate VW. VW is famous as a well-organized, scalable functioning of online machine learning plus support for a numerous machine learning reductions like importance weighting plus a selection of dissimilar loss functions and optimization algorithms.

1) The VW Program Supports

- Multiple supervised plus semi-supervised learning problems like regression, classification.
- Multiple learning algorithms like OLS regression, Single layer Neural net, etc.,
- Multiple loss functions like a hinge, squared error, etc.,
- Multiple optimization algorithms like BFGS, SGD, etc.,
- Regularization
- Flexible input like Numerical, binary, categorical.
- Other features like bootstrapping.

2) Scalability

VW is used to study a tera-feature dataset on 1000 nodes in an hour. Its scalability is supported by various factors:

- Out-of-core online learning is no need to fill every data into memory.
- The hashing trick is the feature identities are transformed into a weight index through a hash.
- Making use of multi-core CPUs that parsing of the input plus learning are completed in separate threads.
- Compile C++ code.

III. Summarization of Data Mining Tools

Table I illustrates the comparative view of different data mining tools based on their compatibility characteristics and its application domains. The main aim of this comparison is not to scrutinize which is the best Data Mining tool, but to exemplify the usage paradigm and the awareness of tools in several fields.

IV. Conclusion

Numerous Data Mining tools were expanded along with their usage of various tasks in this paper. Each and every sub task of data mining tends to be a highly essential reinforcement process for skillful information extraction. This requisite paves the way for the development of many data mining tools. These tools have the widespread technical paradigm, outstanding graphical interface and inbuilt multipath algorithms in which it is very useful for handling substantial amount of data more indeed and legibly. Thus the main role of this survey is to improve knowledge about the data mining tools and its appliance in several industries which will be very useful for the readers and also it meets the needs of data mining researchers to innovate more advanced tools in future.

TABLE I
COMPARISON OF DATA MINING TOOLS

Name of the Tools	Mode of Software	Applications	Categories	Languages
DATABIONIC	Commercial	Visualization, Clustering and Classification	Information Analysis, Visualization	Java
ELKI	Free & Open Source	Outlier detection, Visualization and Clustering	Data Mining and Machine learning Software	Java
MALLET	Open Source	Statistical natural language processing, Document classification, Cluster analysis, Information extraction	Free Artificial Intelligence application, Software stubs, Natural language processing toolkits	Java
ML-FLEX	Free & Open Source	Machine learning analyses	Data Mining and Machine learning Software	Java / Other Programming
SHOGUN	Free & Open Source	Bioinformatics	Free Software Programmed in C++, Data Mining and Machine learning Software, Free Statistical Software	C++
FITYK	Open Source	Chromatography, Spectroscopy, Power diffraction	Regression and Curve fitting software, Data analysis software, Software that uses wxwidgets	C++
PYBRAIN	Open Source	Supervised, unsupervised and reinforcement learning	Support Vector Machines, Neural Networks	Python
UIMA	Open Source	Text Mining, Information Extraction	Software architecture, Data Mining and Machine learning software	Java with C++
NLTK	Free & Open Source	Artificial Intelligence, Information Retrieval, Machine learning	Natural language parsing, Python libraries, Data analysis	Python
DLIB	Open Source	Data Mining, Image processing, Numerical optimization	Computer vision software, Data Mining and Machine learning software	C++
JUBATUS	Open Source	Classification, Regression, Anomaly Detection	Data Mining and Machine learning software, Computing stubs, Free software stubs	C++
SCAVIS	Open Source & Commercial	Data analysis and Data visualization	Statistical software, Numerical programming languages, Infographics	Java, Jython
CMSR DATA MINER	Open Source	Predictive modeling, segmentation, data visualization, statistical data analysis, and rule-based model evaluation	Data Mining and Machine Learning Software	Java
VOWPAL WABBIT	Open Source	Classification, Regression	Data Mining and Machine Learning Software	C++

REFERENCES

[1] S. Sarumathi, N. Shanthi, S.Vidhya M. Sharmila. "A Review: Comparative Study of Diverse Collection of Data Mining Tools". International Journal of Computer, Information, Systems and Control Engineering Vol:8 No:6, 2014
[2] M Ferguson. "Evaluating and Selecting Data Mining Tools"InfoDB, Vol:11 No:2
[3] Data Bionics Research Group, University of Marburg: Databionic esom tools, Website (2006), http://databionic-esom.sourceforge.html/
[4] ELKI: Environment for Developing KDD-Applications Supported by Index-Structures, (online). Available at: http://elki.dbs.ifi.lmu.de/
[5] McCallum, Andrew Kachites. "MALLET: A Machine Learning for Language Toolkit", 2002.
[6] ML-Flex, "Introduction to ML-Flex (online). Available at: http:// mlflex.sourceforge.net/tutorial/index.html
[7] Jason Brownlee, "A Gentle Introduction to Scikit-Learn: A Python Machine Learning Library" 2014.
[8] Soren Sonnenburg et al., "The SHOGUN Machine Learning Toolbox", Journal of Machine Learning Research 11, 1799-1802, 2010.
[9] M. Wojdyr, J. Appl. Cryst, " Fityk" 2010.
[10] Tom Schaul, Justin Bayer, Daan Wierstra, Sun Yi, Martin Felder, Frank Sehnke, Thomas Rückstieß, Jürgen Schmidhuber, "PyBrain" 2010.
[11] UIMA, "The Apache Software Foundation",2006.
[12] S. Bird, E. Steven, Edward Loper and Ewan Klein, "Natural Language Processing with Python" O'REILLY.
[13] Steven Bird, Edward Loper, "NLTK: The Natural Language Toolkit" 2002.
[14] Davis E. King, "Dlib C++ Library", 2015, Dlib (online). Available at: http://dlib.net/
[15] Davis E. King, "Containers", 2013, Dlib (online). Available at: http://dlib.net/containers.html
[16] Davis E.King, "Image Processing", 2015, Dlib (online). Available at: http://dlib.net/imaging.html
[17] Davis E.King, "API Wrappers", 2015, Dlib (online). Available at: http://dlib.net/api.html
[18] Davis E.King, "Graph Tools", 2013, Dlib (online). Available at: http://dlib.net/graph tools.html
[19] Davis E.King, "Machine Learning", 2015, Dlib (online). Available at: http://dlib.net/ml.html
[20] Jubatus (online). Available at: http://www.predictiveanalyticstoday.com/ top-40-free-data-mining-software/
[21] Jubatus, PFN & NTT, 2011. (online). Available at: http://jubat.us/en/overview.html
[22] Satoshi Oda, Kota Uenishi, and Shingo Kinoshita," Jubatus: Scalable Distributed Processing Framework for Realtime Analysis of Big Data", NTT Technical Review.
[23] SCaVis, Scavis community, 2014.
[24] Cho Ok-Hyeong, "CMSR Data Miner",2014.
[25] CMSR Data Miner Data Mining & Predictive Modeling Software, Rosella Predictive Knowledge and Data Mining,2005
[26] Vowpal Wabbit: Fast Learning on Big Data, n13,2014

Mrs. S.Sarumathi received B.E. degree in Electronics and Communication Engineering from Madras University, Madras, Tamil Nadu India in 1994 and the M.E. degree in Computer Science and Engineering from K.S.Rangasamy College of Technology, Namakkal Tamil Nadu, India in 2007. She is doing her Ph.D. programme under the area Data Mining in Anna University, Chennai. She has a teaching experience of about 17 years. At present she is working as Associate professor in Information Technology department at K.S.Rangasamy College of technology. She has published 5 reputed International Journals and two National journals. And also she has presented papers in three International conferences and four national Conferences. She has received many cash awards for producing cent percent results in university examination. She is a life member of ISTE.

Dr.N.Shanthi received the B.E. degree in Computer Science and Engineering from Bharathiyar University, Coimbatore, Tamil Nadu, India in 1994 and the M.E. degree in Computer Science and Engineering from Government College of Technology, Coimbatore, Tamil Nadu, and India in

2001. She has completed the Ph.D. degree in Periyar University, Salem in offline handwritten Tamil Character recognition. She worked as a HOD in department of Information Technology, at K.S.Rangasamy College of Technology, Tamil Nadu, India since 1994 to 2013, and currently working as a Professor & Dean in the department of Computer Science and Engineering at Nandha Engineering College Erode. She has published 29 papers in the reputed International journals and 9 papers in the National and International conferences. She has published 2 books. She is supervising 14 research scholars under Anna University, Chennai. She acts as the reviewer for 4 international journals. Her current research interest includes Document Analysis, Optical Character Recognition, and Pattern Recognition and Network security. She is a life member of ISTE.