Data Mining Techniques Used for Email Mining

Pranjal S. Bogawar
Priyadarshini College of Engineering, Nagpur,
pbogawar@yahoo.com

Abstract:

Data mining or knowledge discovery refers to the process of finding interesting information in large repositories of data. The term data mining also refers to the step in the knowledge discovery process in which special algorithms are employed in hopes of identifying interesting patterns in the data. These interesting patterns are then analyzed yielding knowledge. Present work gives core ideas about data mining, its basic tasks, some most popular algorithms of data mining, applications of data mining and the short overview of email mining. Present study elaborates the idea of data mining algorithms, which may be useful for researchers to mine the email.

Keywords:
Data mining, Email mining, Neural Network, Support vector machine.

Introduction:

Data mining is more than just conventional data analysis. It uses traditional analysis tools (Like statistics and graphics) plus those associated with artificial intelligence (such as rule induction and neural nets). It is all of these, but different. It is a distinctive approach or attitude to data analysis. The emphasis is not so much on extracting facts, but on generating hypotheses. Data Mining is used everywhere as Data warehouse is huge. Data mining techniques are also used in email mining to extract useful information from the email. So, this paper gives the idea of email mining and the work of researchers who are using data mining techniques in email mining.

Data Mining:

Data mining is the exploration and analysis of large quantities of data in order to discover meaningful pattern and rules. Data mining comes into two flavors – directed and undirected. Directed data mining attempts to explain or categorize some particular target field such as income or response. Undirected data mining attempts to find patterns or similarities among between groups of records without the use of particular target field or collection of predefined classes [1].

Data mining models are either predictive or Descriptive. Predictive model makes the prediction about the values of data using known results found from different data. A descriptive model identifies patterns or relationships in data. Following figures gives the data mining model and its tasks [2].

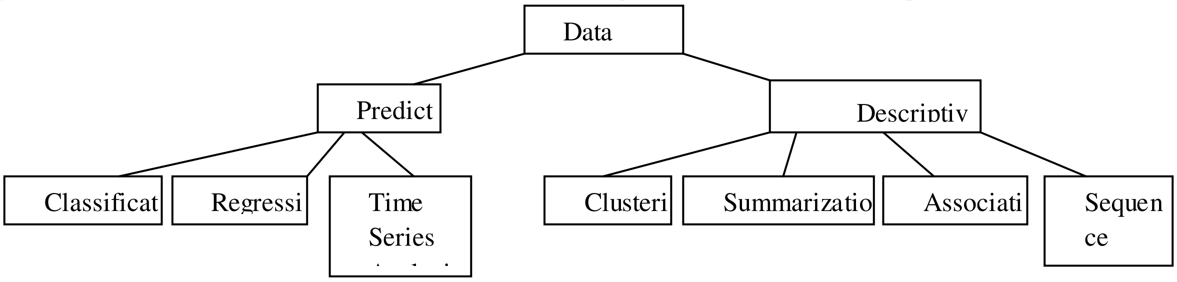

Figure. 1- Data Mining models and Tasks

Basic Data Mining Tasks:

Classification:

It maps data into predefined groups or classes. It is often referred to as supervised learning because the classes are determined before examining the data [2].

Types of classification models:

- Classification by decision tree induction
- Bayesian Classification
- Neural Networks
- Support Vector Machines (SVM):
- Classification Based on Associations [3]

Regression:

Regression (Technique for prediction) is used to map a data item to a real valued prediction variable. Regression assumes that the target data fit into some type of function (e.g. linear, logistic etc.) and determines the best function of this type that models the given data [2] Predication can be viewed as type of classification. The difference is that prediction is predicting a future state rather than a current state.

Types of regression methods

- Linear Regression
- Multivariate Linear Regression
- Nonlinear Regression
- Multivariate Nonlinear Regression[3]

Time Series Analysis:

With time series analysis, the value of an attribute is examined as it varies over time. The values usually are obtained as evenly spaced time points (daily, weekly, hourly etc.)[2].

Types of Time Series model

- Additive Model
- Multiplicative Model

Clustering:

It is similar to classification except that the groups are not predefined, but defined by the data alone. Clustering is also referred as unsupervised learning or segmentation [2].

Types of clustering methods

- Partitioning Methods:
- Density based methods
- Grid-based methods
- Model-based methods[3]

Summarization:

It maps data into subsets with associated simple description. Summarization is also called as characterization or generalization. It extracts or derives representative information about the database [2]. Models Based on Summarization are-

- Statistical Concepts(mean, variance, standard deviation, median,mode
- Graphical (Histogram, box plot, scatter diagram)[2]

Association Rules:

Link analysis, alternatively referred to as affinity analysis or association, refers to data mining task of uncovering relationships among data. An association rule is a model that identifies specific types of data association [2]. Types of association rule are-

- Multilevel association rule
- Multidimensional association rule
- Quantitative association rule

Sequence Discovery:

Sequential analysis or sequence discovery is used to determine sequential patterns in data. These patterns are based on a time sequence of actions [2].

Knowledge Discovery in Database:

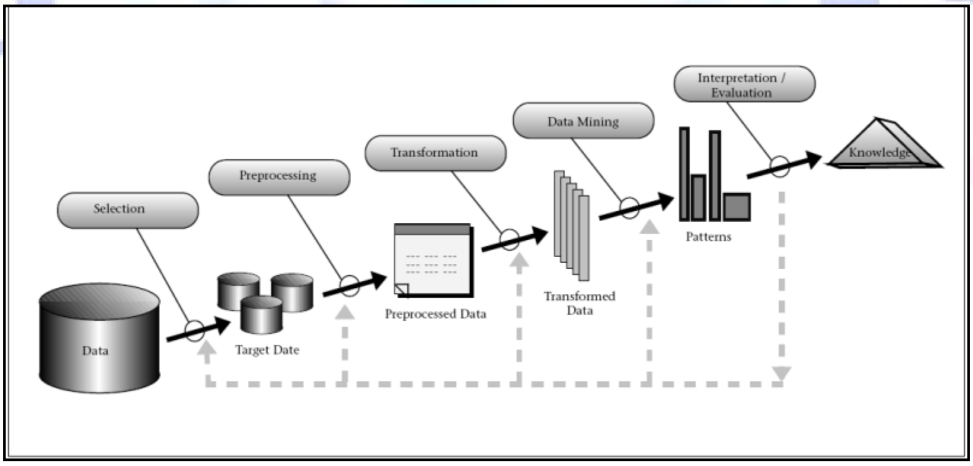

Figure. 2- Knowledge discovery process [3]

This is the process of discovering useful patterns in data: knowledge extraction, information discovery, exploratory data analysis, information harvesting and unsupervised pattern recognition. The process of knowledge discovery is shown in fig.1. So, the data mining is one step in knowledge discovery process.

Data Mining Techniques:

Here we are discussing some of the data mining algorithms.

Decision tree Induction Algorithm:

This technique is composed of three basic steps. They are

1) Generalization by attribute-oriented induction, to compress the training data. This includes storage of the generalized data in a multidimensional data cube to allow fast accessing,
2) Relevance analysis, to remove irrelevant data attributes, thereby further compacting the training data,
3) Multilevel mining, which combines the induction of decision trees with knowledge in concept hierarchies. The induction of decision trees is done at different levels of abstraction by employing the knowledge stored in the concept hierarchies [15].

Naïve Bayes Algorithm:

The Bayesian Classification represents a supervised learning method as well as a statistical method for classification. This Classification is named after Thomas Bayes (1702-1761), who proposed the Bayes Theorem. Bayesian reasoning is applied to decision making and inferential statistics that deals with probability inference. It is used the knowledge of prior events to predict future events. Baye's Theorem says that

$$P\left(\frac{h}{D}\right) = \frac{P(\frac{h}{D})}{P(D)} \text{.......................}(2)$$

where

P(h) : Prior probability of hypothesis h
P(D) : Prior probability of training data D
P(h/D) : Probability of h given D
P(D/h) : Probability of D given h[13,14]

Support Vector Machine Algorithm:

The original SVM algorithm was invented by Vladimir Vapnik[12]. SVM concept is based on the idea of structural risk minimisation which minimizes the generalization error. The advantage of SVM is that they do not require a reduction in number of features in order to avoid the problem of over fitting, which is useful when dealing with large dimensions as encountered in the area of text mining. It is a learning machine that classifies an input Vector X using decision function:

$$f\,(X)= <X,W> + b \quad \text{.........}(1)$$

SVMs are hyper plane classifiers and work by determining which side of hyper plane classifiers and work by determining which side of the hyper plane X lies. In the above formula given in eq. no. 1 the hyper plane is perpendicular to W and at a distance b/|| W|| from the origin. SVM maximize the margin around the separating hyper plane. The decision function is fully specified by a subset of training samples [11].

Partitioning Algorithm:

The algorithm executes in two phases. In the first phase, the Partition algorithm logically divides the database into a number of non-overlapping partitions. The partitions are considered one at a time and all large item sets for that partition are generated. At the end of phase I, these large item sets are merged to generate a set of all potential large item sets. In phase II, the actual support for these item sets is generated and the large item sets are identified. The partition sizes are chosen such that each partition can be accommodated in the main memory so that the partitions are read only once in each phase [16].

Hierarchical Agglomerative (divisive) methods:

If it is given a set of N items to be clustered and a N*N distance (or similarity) matrix then the basic process of agglomerative hierarchical clustering can be done iteratively following these four steps:

1. Start by assigning each item to a cluster. Let the distances (similarities) between the clusters are the same as the distances (similarities) between the items they contain;
2. Find the closest (most similar) pair of clusters and merge them into a single cluster;
3. Compute distances (similarities) between the new cluster and each of the old clusters;
4. Repeat step 2 and 3 until all items are clustered into a single cluster of size N.

Step 3 can be different because of the varieties in the definition of the distance (or similarity) between clusters:

- Single linkage clustering (nearest neighbor technique) – here the distance between two clusters is given by the value of the shortest link between clusters. At each stage the two clusters for which the distance is minimum are merged;
- Complete linkage clustering (farthest neighbor) – is the opposite of the single linkage i.e. distance between groups is defined as the distance between the most distant pair of objects, one from each group.
- Average linkage clustering – the distance between two clusters is defined as the average of distances between all pairs of objects, where each pair is made up of one object from each group. At each stage the two clusters for which the distance is minimum are merged;
- Average group linkage clustering – with this method, groups once formed are represented by their mean values for each variable, that is their mean vector and inter-group distance is defined in terms of distance

between two such mean vectors. At each stage the two clusters with minimum distance are merged.

- Ward's hierarchical clustering – Ward (1963) proposed a clustering procedure seeking to form the partitions *P* ,..., *P*1 *n* in a manner that minimizes the loss associated with each grouping and to quantify that loss in a form that is readily interpretable. At each step the union of every possible cluster pair is considered and the two clusters whose fusion results in minimum increase in "information loss" are combined. Information loss is defined by Ward in terms of an error sum-of-squares criterion.

Expectation Maximization (EM):

The EM algorithm was explained and given its name in a classic 1977 paper by Arthur Dempster, Nan Laird, and Donald Rubin.

EM is an iterative optimization method to estimate some unknown parameters Θ, given measurement data U. However, we are not given some "hidden" nuisance variables J, which need to be integrated out. In particular, we want to maximize the posterior probability of the parameters Θ given the data U, marginalizing over J:

$$\Theta = \operatorname{argmax}_{\Theta} \sum_{J \in \mathcal{J}^n} P(\Theta, J|U)$$

The intuition behind EM is an old one: alternate between estimating the unknowns Θ and the hidden variables J. This idea has been around for a long time. However, instead of finding the best J ∈ $\mathcal{J}$ given an estimate Θ at each iteration, EM computes a distribution over the space $\mathcal{J}$[17].

K-Means:

K-means clustering (MacQueen, 1967) is a method commonly used to automatically partition a data set into k groups. It proceeds by selecting k initial cluster centers and then iteratively refining them as follows:

1. Each instance di is assigned to its closest cluster center.
2. Each cluster center Cj is updated to be the mean of its constituent instances [18].

Bisecting K-means:

It produces the clusters of the similar sizes and with smaller entropy than K-means [19].

Ada Boost algorithm:

It works as follows. First, it assigns equal weights to all the training examples *(**x**i , yi)(i* ∈ {1, . . . ,*m*}*)*. Denote the distribution of the weights at the *t*-th learning round as *Dt* . From the training set and *Dt* the algorithm generates a *weak* or *base learner ht* : *X* → *Y* by calling the base learning algorithm. Then, it uses the training examples to test *ht*, and the weights of the incorrectly classified examples will be increased. Thus, an updated weight distribution *Dt*+1 is obtained. From the training set and *Dt*+1 AdaBoost generates another weak learner by calling the base learning algorithm again. Such a process is repeated for *T* rounds, and the final model is derived by weighted majority voting of the *T* weak learners, where the weights of the learners are determined during the training process [5].

Neural Network:

Artificial Neuron appeared in 1943, but the computer usage of neural networks begins in 1980.Neural network also referred as Artificial Neural Network. Neural network is a directed graph as shown in fig.2 with many nodes (processing elements) and arcs (interconnections) between them. The NN approach, like decision tree requires graphical structure to be built to represent the model and then structure is applied to data. The NN is a directed graph with source (input), sink (output), and internal (hidden) nodes. To perform the data mining task, a tuple is input through input nodes and output nodes exist in output layer. The NN has one input node for each attribute value to be examined to solve the data mining function. The NN may be changed to improve future performance.

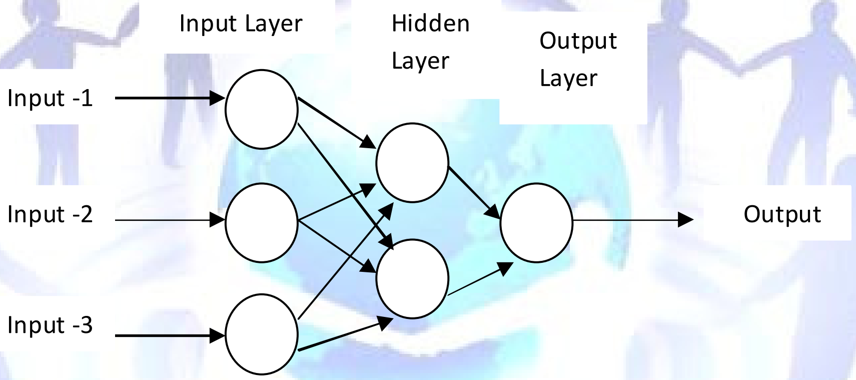

Figure. 2- Neural Network

Genetic Algorithm:

It is a computational model consisting of five parts:

1. Starting set of individuals, P
2. Crossover Technique
3. Mutation algorithm.
4. Fitness function.
5. Algorithm that applies the crossover and mutation techniques to P iteratively using the fitness function to determine the best individuals in P to keep. The algorithm replaces the predefined number of individuals from the population with each iteration when some threshold is met.

Data Mining Applications:

Data mining is used for a variety of purposes in both private and public sectors. Industries such as banking, insurance, medicine and retailing commonly use data mining to reduce costs, enhance research, and increase sales. It is used in Web Mining. It is also used in email mining for finding various characteristics of email.

Email Mining:

Email mining can be considered as mining of data embedded in header and or body of the email message. Various text mining techniques which extract unknown and useful information from huge set of emails can be employed to achieve email mining. Email Mining can be considered as an application of the upcoming research area of Text Mining (TM or also known as Knowledge Discovery from Textual Data) on email data.

However, there are some specific characteristics of email data that set a distinctive separating line between Email and Text Mining:

2. Email includes additional information in the headers of email that can be exploited for various email mining tasks.
3. Text in email is significantly shorter and, therefore, some Text Mining techniques might be inefficient in email data.
4. Email is often cursorily written and, thus, linguistic well-formedness is not guaranteed [6]. Spelling and grammar mistakes as well as nonstandard user acronyms also appear frequently.
5. Email is personal and therefore generic techniques are difficult to be effective to individuals.
6. Email is a data stream targeted to a particular user and concepts or distributions of target classes of the messages may change over time, with respect to the messages received by that user.
7. Email will probably have noise. HTML tags and attachments must be removed in order to apply a text mining technique. In some other cases, noise is intensively inserted. In spam filtering for example, noisy words and phrases are inserted, in order to mislead machine learning algorithms.
8. It is rather difficult to have public email data for experiments, due to privacy issues. This is a drawback especially for research since comparative studies cannot be conducted without public available datasets. An exception to the above statement is the Enron Corpus (Klimt & Yang, 2004), which was made public after a legal investigation concerning the Enron Corporation [6, 7].

Data mining Algorithms Used for Email Mining:

Various researchers worked on email to find language [8, 9], gender [8], writeprints[10], keyword based matching for finding the data. To find the gender and language (EFL and ESL) Oliver de Vel et.al. used support vector machine algorithm. Nouf Al Fe'ar et. Al. classified Arabic and English emails by using naïve bayes algorithm. Farkhund Iqbal et.al. used expectation maximization, K-means, bisecting K-means clustering algorithms to classify the author of email using their writeprints (stylistic characters, grammatical mistakes etc.) . Appavu alias Balamurugan et.al., introduced the new Ad Infinitum algorithm to classify the threatening messages. Ad infinitum algorithm for decision tree induction was used which is a greedy algorithm that constructs decision trees in a top-down recursive divide-and-conquer manner. The algorithm is a version of ID3, a well-known decision tree induction algorithm. John Yearwood et. al., used the structural characteristics of the emails received by persons and the information derived on hyperlinks from 'Whois Database' for profiling of phishing emails. They used AdaBoost. MH and AdaBoost.MR to solve multi-label classification problems. The results of classification algorithm were used for the profile generation.

Conclusion:

This paper gives the idea of data mining, basic data mining task, top most algorithms which are used for data mining. The paper discusses the various applications of data mining. One main application area of data mining is email mining. In this paper we gave the brief overview of email mining and the work of researchers who worked in the field of email mining.

References:

[1] **Michael J. A. Berry, Gordon S. Linoff,** "Data Mining techniques"

[2] **Margrate H. Dunham, S. Sridhar,** "Data mining- Introductory and Advanced topics"

[3] **Mrs. M. Bharati Ramgiri,** "Data mining techniques and its applications", Indian Journal of Computer Science and Engineering,Vol.1 No4301-305, pp-301-305

[4] **Jeffrey W. Seifert,** " Data Mining: An Overview", CRS Report for Congress, order code RL31798, pp-1-16

[5] **XindongWu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang, Hiroshi Motoda, Geoffrey J. McLachlan , Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Steinbach , David J. Hand , Dan Steinberg,** "Top 10 algorithms in Data mining", Knowl Inf Syst (2008) 14:1–37 DOI 10.1007/s10115-007-0114-2, pp-1- 37

[6] **Ioannis Katakis, Grigorios Tsoumakas, Ioannis Vlahavas,** "Email Mining: Emerging Techniques for Email Management", 2006, pp- 1-32

[7] **Enron Email Dataset**, http://www.cs.cmu.edu/~enron/

[8] **Oliver de Vel, Malcolm Corney, Alison Anderson, and George Mohaly,** " Language and Gender Author Cohort Analysis of Email for Computer Forensics", In Proc. Of digital forensic workshop(2002),pp-1-16

[9] **Nouf Al Fe'ar, Einas Al Turki, Asma Al Zaid, Mashael Ai Duwais,** "E-Classifier: A Bi-Lingual Email Classification System", Information technology, ITSim2008, IEEE,Vol-2, pp-1-4

[10] **Farkhund Iqbal, Rachid Hadjidj, Benjamin C.M. Fung, Mourad Debbabi,** "A novel Approach of mining write prints for authorship attribution in email forensic", Digital Investigation(Elsevier Journal)2008,pp42-51

[11] **Steve Gunn,** "Support Vector Machines for Classification and Regression", ISIS Technical Report http://www.svms.org/tutorials/Gunn1998.pdf, 14 May 1998, pp:1-52

[12] **V. Vapnik.** The Nature of Statistical Learning Theory. Springer-Verlag, New York. 1995.

[13] **Naive Baye's** Classification Algorithm, h ttp://software.ucv.ro/~cmihaescu/ ro/ teaching/ AIR/ docs /Lab4-NaiveBayes.pdf ,pp1-17

[14] **Elaine Rich, Kevin Knight,** "Artificial Intelligence"

[15] **Devi Prasad Bhukya and S. Ramachandram,** "Decision Tree Induction: An Approach for Data Classification Using AVL-Tree", International Journal of Computer and Electrical Engineering, Vol. 2, No. 4, August, 2010 1793-8163, pp-660-665

[16] **Ashok Savasere, Edward Omiecinski, Shamkant Navathe,** "An Efficient Algorithm for Mining Association Rules in Large Databases", Proceedings of the 21st VLDB Conference Zurich, Swizerland, 1995, pp-432-444

[17] **A. P. Dempster, N. M. Laird and D. B. Rubin,** "Maximum Likelihood from Incomplete Data via the EM Algorithm", Journal of the Royal Statistical Society. Series B (Methodological),Vol. 39, No. 1 (1977), pp. 1-38

[18] **Kiri Wagstaff,Claire Cardie, Seth Rogers Stefan Schroedl,** "Constrained K-means Clustering with Background Knowledge", Proceedings of the Eighteenth International Conference on Machine Learning, 2001, p. 577-584.

[19] **Sergio M. Salveresi and Daniel Boley,** "A comparative analysis on the bisecting K-means and PDPP clustering algorithms", http://www-users.cs.umn.edu/~boley/publications/papers/ savaresi04.pdf, pp1-18

[20] **Appavu alias Balamurugan, Rajaram,Muthupandian and Athiappan,** " Automatic mining of threatening e-mail using Ad Infinitum algorithm, In International Journal of Information Technology,Vol14. No.2,2008