International Journal of Engineering & Technology, 7 (1.1) (2018) 406-410

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Data mining techniques for herbs

J. Satish Babu [1] *, M. Niveditha [1], V. Bhavya [1], K. Gowthami [1]

[1] *Computer science and engineering, Koneru Lakshmaiah Educational Foundation, Vaddeswaram, Guntur, Andhra Pradesh- 522502*
**Corresponding author E-mail: jampanisatishbabu@gmail.com*

Abstract

The most important source of ingredients in the discovery of new drugs are Natural products. Moreover Nagoya protocol is most commonly used in selection of herbs based on similar efficiency, Later scientists have voiced their concern on protocol also proved it as less effective therefore, this project uses data mining classification approaches, novel targeted Selection which makes use of MED - LINE(Medical Literature Analysis and Retrieval system online) database that consists of biomedical information to identify herbs of same efficacy .Neural network technique among all classification techniques is inspired by biological nervous system. AS neural network is successful on wide array of noisy object selection of herbs is done effectively. SOM (self-organizing map) is most popular Neural Network provides a topology preserving mapping from the high dimensional space to map units. The main objective of this project is to survey on various data mining methods and their techniques and to conclude the suitable algorithm.

Keywords: *Data Mining; Classification Techniques; Mesh Techniques; Artificial Neural Networks; SOM Algorithm*.

1. Introduction

DATA MINING: Definition, history, goals, phases and methods:
In this section, we first present definitions of data mining followed by its history and then goals of data mining

1.1. Data mining definition

The term Data Mining is used to discover patterns and relations among data in large dataset. Data mining also called as discovery of knowledge from data, used to extract interesting information from large datasets. Firstly, the data to be mined is selected and undergoes data cleaning (it takes 60% effort). Find required and useful features in- order to reduce the data. Search for suitable patterns and evaluate them. Thus, the knowledge is discovered from data.

1.2. History of data mining

Following its definition, we move to understand the history of data mining, considering the saying "we are present in age of information", we are in the era where use of vast amount of data is done daily such data must be analyzed and the need of data mining provides tools to discover knowledge from data [1]. Data mining is everywhere, data mining is everywhere, at its story begins before a considerable length of time before Money ball and Edward Snowden. Those milestones would major turning points.

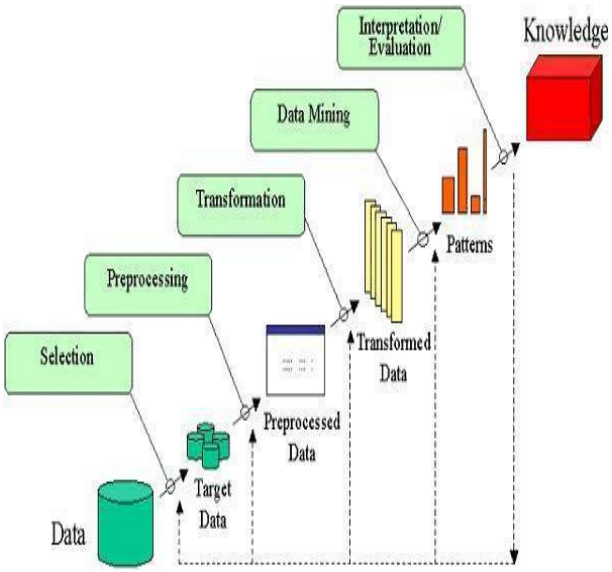

Fig. 1: Step Wise Procedure of Discovering Knowledge from Data.

Identifying dependencies complexities and patterns. 1763 Thomas Bayes' paper may be distributed posthumously viewing related to current likelihood with former likelihood. It understands complex probability values.

Copyright © 2018 J. Satish Babu et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.



Fig. 2: History of Data Mining.

1.3. Goals of data mining

The two main goals of DM are, Prediction includes utilizing the data among datasets and how those attributes are further represented or used in future. Description concentrates with respect to discovering human-interpretable examples describing the information. Those relative fact that prediction and description to specific in formation mining vary respectably.
By using some commonly used Data Mining tasks we can achieve the required result of prediction and description.

1) Classification is used to classify data based on attribute type also based on the category it categorizes the data.
2) Identification Its difficult to identify existence of item among vast data, by using identification we can easily identify an event also an activity.
3) Regression Unlike classification regression models continuous valued functions.

1.4. Phases/steps

1) Define the problem: To process the data, we must identify business goals and data mining goals.
2) Identify required data: As in data mining we find large amount of unrelated data, we must identify required data.
3) Prepare and preprocess: Data we found is not clear consists of errors, missing values etc. process the data before modeling.
4) Model the data: Based on the input that we give selection of model is to be done in this phase.
5) Train and Test: Consider some of the sample datasets test and iterate results that obtained for the dataset used.
6) Verify and deploy: This phase involves visualization, transformation and removing patterns.

Method	General Characteristics
Partitioning methods	– Find mutually exclusive clusters of spherical shape – Distance-based – May use mean or medoid (etc.) to represent cluster center – Effective for small- to medium-size data sets
Hierarchical methods	– Clustering is a hierarchical decomposition (i.e., multiple levels) – Cannot correct erroneous merges or splits – May incorporate other techniques like microclustering or consider object "linkages"
Density-based methods	– Can find arbitrarily shaped clusters – Clusters are dense regions of objects in space that are separated by low-density regions – Cluster density: Each point must have a minimum number of points within its "neighborhood" – May filter out outliers
Grid-based methods	– Use a multiresolution grid data structure – Fast processing time (typically independent of the number of data objects, yet dependent on grid size)

Fig. 3: Phases/Steps in Data Mining.

1.5. Methods

1) Partitioning methods: Not often used method only suitable for small to medium size of datasets. Makes use of mean and medoid to represent cluster center.
2) Density based method: Density based method consists of exclusive clusters but not fuzzy clusters.
3) Hierarchical methods: It creates hierarchical decomposition of given set of data objects. It is classified in to two types either, agglomerative or divisive depends on how hierarchical decomposition is formed.
4) Grid based methods: Makes use of multi resolution data structure. Procession time is fast and suitable to any dataset.

2. Literature work

Table 1: Characteristics of Methods

S.No	AUTHOR	TITLE OF PROJECT	JOURNAL	METHOD	ALGORITHM	INPUT	OUTPUT	FUTURE SCOPE
[1]	(Sang-Jun Yea, 2016)	Selection of herbs using any of the existing datamining approaches.	Journal of Ethnopharmacology	Targeted selection method, Random selection method	SOM algorithm	Single Herb	Different herbs of same type	The SOM algorithmic approach is used to explore natural biproducts.
[2]	(A.Linda Sherin, 2017) (Brenda n Coolsaet , 2013)	SIMILAR HERB SLECTION USING DATA MINING	International Journal of Scientific & Engineering Research	An Algorithm based on similarity based index and MEDLINE databases are used.	SOM Algorithm,	The basic input for this proposed work is the scientific name or botanical name and the common name of the various herbs and plants.	Similar input of the given input herb	- -
[3]	(Brenda n Coolsaet , 2013)	The Challenges for Implementing the Nagoya Protocol in a Multi-Level Governance Context		implementation process is based on a normative institutionalist approach	Access and benefit sharing using Nagoya protocol.	To Prove Belgium as a Multi-Faceted Case of Multilevel Governance	As there is no proper input for Access and benefit sharing this cannot be proven.	Overcome the disadvantages of triple implementation deadlock.
[4]	(kasa., 2016) (Juha Vesanto, 1999)	Uses of medicinal plants in Ethiopia	International Journal of Advanced Research (2016)					
[5]	(Juha Vesanto, 1999) (Rajdev Tiwari, 2010)	SOM in MATLAB		Self-organizing map known as SOM and unsupervised learning techniques.	SOM training Algorithm	Input must be in the following format SD = som_read_data('data.txt');	SD=som_data_struct (D)	Quantitative analysis of self-organizing mapping is more concentrated.
[6]	(Rajdev Tiwari, 2010)	Correlation-based Attribute Selection using Genetic Algorithm	International Journal of Computer Applications (0975 – 8887)	method for selecting optimal attribute subset based on correlation	Genetic algorithm (GA)	two attributes, X and Y. Where X and Y are the two features/attributes	Correlation(y)	–
[7]	(SKASKI, 1996)	Exploratory Data Analysis by The Self-Organizing Map Structures of Welfare and Poverty in The World		Exploratory Data Analysis	Self-Organizing Map(SOM)	Based on 39 Statistical indicators chosen to describe the Standard of living.	Representation as decision support system.	Not only make use of SOM Algorithm but also represents the datasets and their states.
[8]	(Arezou Rezaei)	An Encyclopedia of Herb-Disease, a Quick Shortcut for Herbal Research: A Comprehension Based on Iranian Herbal Studies.	Journal of Biomedical Sciences ISSN 2254-609X	Inclusion and exclusion criteria, Data Extraction and classification	Botanical scientific nomenclature	Investigated subjects and diseases have been categorized in 18 groups	Only 69 of 560 studied (12.5%) Herbs are identified	There is much more to be done about herbal plants identification and selection of targeted herbs.

3. Related work

3.1. Classification

Classification is one of the techniques in data mining. Which deals with the different types to collect the data? Such that to analyse which is important and not. This is also known as decision tree as classification has several methods which is used to get the analysis of large and heavy data sets.

3.2. Why classification

In this used technique is classification because when com-pared to clustering technique classification technique has best properties rather than the properties in clustering technique in data mining. Such as priori knowledge is supportable to classification but not to clustering technique. While coming to the data needs classification is labelled samples from a set of classes but where as in clustering there are un-labelled samples. So, we preferred classification instead of clustering technique [2].

4. Existing system

This Nagoya protocol is the primary used technique or the method. This method origin came back from early days. The Nagoya Protocol on accessing resources in Genetic and Sharing of Benefits from Utilization to the Convention on Bio-logical Diversity which aims at sharing the benefits utilization of genetic resources.
At the initial state this Nagoya protocol is important. This protocol created greater certain legality and transparency for the user and for providers of genetic resources are as follows ac-cording to the survey.
For accessing to genetic resources needs to Establish certain more predictable conditions.
Ensuring benefit-sharing when providing the genetic resources by leaving the country.



Fig. 4: Nagoya Protocol Representation.

The Nagoya Protocol creates encourages to protect and in a way that can be maintained at a certain rate or level. This method is time taking process and it is also less efficient. There are various methods for selecting natural products can be categorized into two groups:

1) Random selection

It is a process which selects the products randomly without depending on any other constraints. For Example, for random selection is as follows: Consider, when chosen a target herb and with the similar herbs which are named as candidate herbs. To acquire similar composition herb this random selection method randomly selects the herbs which are like the targeted herb. In this way random selection process selects product [4].

2) Targeted selection

It is one of the important data mining technique which is used to target the customers on which they have shown the interest on a product (or) item in past.

3) Segmentation method

The key of the data mining technique is segmentation. It deals with the group of consumers react in the same way to a marketing approach [6].

5. SOM Algorithm

This Self-organizing is a neural networks algorithm which are used as input patterns into groups of similar patterns. This Algorithm is the most powerful algorithm. With this we can acquire efficiency when compared to other algorithms [5]. As this algorithm imposes on topological patterns so they are known as “maps”. As this algorithm has proven to be most powerful in visualization of data. Usually, patterns which are given as input are usually multi-dimensional. For Example: When considered herb the target herb is selected and there are more than one candidate herbs. Such that out of those they selected which are like targeted herb.

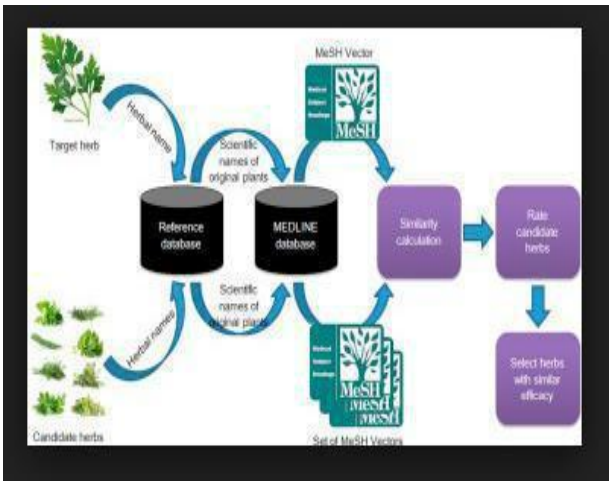

Fig. 5: SOM Flow Diagram.

6. Applications of SOM algorithm

SOM algorithm has various good properties. These representation on the string or on the grid is easy to interpret.
The topology conservations give different order in its class and it is possible to use the data sets with some missed values and we use different classification algorithms with respective their context and these are quick and efficient.
Similar herbs selection using SOM algorithm is the main task which we are reviewing in this paper. The concept is we select a target herb and we also find the composition of the targeted herb. We find the candidate herb from candidate data base with the help of their biological names of both targeted and candidate herb. By using novel selection methods and application of SOM algorithm we find the similar herb with same type of composition and these are used for production of drugs.
SOM can be used in business fields for example cotton yielding and spinning [7].
Classification techniques which are applied on cotton lint with the help of data analysis methods such as SOM algorithm, k-means techniques and clustering techniques and Probabilistic Neural Network (PNN). This technique can also be used for classifying cotton lint based on their specific characteristics. This set of data is used for classifying cotton bales using a Self-Organizing Maps (SOM) which helps in visualizing the high dimensional cotton lint.
SOM algorithm is used in two -level clustering approach and Fuzzy partitioning.
The SOM algorithm which is a feature map acts as ideal map for data type identification because the similarities between data types are to be known in priori. This Self organizing map is used to implement data type identifications in various ways and the differences produced in the results are different in most cases [9,10].
Natural language processing can be done using SOM algorithm in different ways.

1) Document clustering
2) Document retrieval

3) Automatic query

SOM algorithm plays an important role in Image segmentation in which the similar property variables are clustered together and gives more efficiency in understanding the similar properties.

The algorithm also includes clear explanation about K-means Algorithm basic target is to find churner which means same type or similar groups in the given data set [6].Breast cancer data can be visualized using SOM algorithm.The presentation of component plane is of integrated self-organized map is a powerful AI tool for analysis of big, complex, biological databases. This allows for displaying of multi-dimensional mapping output disease data bases in different sample specific presentation and providing with various benefits in visual inspection of biological significance of features which are clustered in each unit. This helps to analyse its behaviour of the different attributes having breast cancer and their correlation between them through their component planes.

SOM algorithm is also used in applications for ground penetrating Radar Imaging Systems. Self-organizing maps helps in working of automated mapping in hydrographic systems which are exported from satellite Imagery.

By using self-organizing maps in land -cover classifications can be clustered with both spectral and spatial information.

Meteorology and Oceanography also uses self-organizing maps for understanding its properties of variables and tells about its effects [11].

Image searching in a visual feature space with SOM-based clustering and modified inverted indexing [8].

SOM algorithm is used in facial expression recognition model and use adaptive learning capability.

Effective Web sites mining by combining SOM and ontologies.

7. Conclusion

In this project, the identification of different types of data mining techniques which are implemented on selection of herbs. The various herbs which has its own scientific name are collected from vivid resources. Using these scientific names, the articles which hold the phytochemicals were in the MesH database. These articles are stored in the cloud which helps in easy retrieval. Using these phytochemicals the candidate herbs of a target herb is found out. However, upon all the comparative study made on different techniques we can conclude that using SOM algorithm phytochemical structures present in the herbs are also screened.

References

[1] A. Linda Sherin, D. (2017). SIMILAR HERB SELECTION USING DATA MINING. Ijser.

[2] Aiping Lu, M. J. (n.d.). An integrative approach of linking traditional Chinese medicine pattern classification.

[3] An Encyclopedia of Herb-Disease, a Quick shortcut for herbal research. (n.d.). Jbs.

[4] Arezou Rezaei, A. F. (n.d.). An Encyclopedia of Herb-Disease, a Quick Shortcut for Herbal Research: A Comprehension Based on Iranian Herbal Studies.

[5] Ashish K Sharma, R. K. (n.d.). problems associated with clinical trials of ayurved medicines.

[6] Brendan Coolsaet, T. D. (2013). The Challenges for Implementing the Nagoya Protocol in a Multi-Level Governance Context: Lessons from the Belgian Case. Mdpi.Gori, F. F. (n.d.). Herbal Medicine Today: Clinical and Research Issues.

[7] Jon C Tilburt a, T. J. (n.d.). Herbal medicine research and global health: an ethical analysis.

[8] Juha Vesanto, J. H. (1999). Self-organizing map in MATLAB: the SOM Toolbox.

[9] kasa, T. (2016). Uses of medicinal plants in Ethiopia. Ijar. https://doi.org/10.21474/IJAR01/1012.

[10] Lobo, V. J. (n.d.). Application of Self-Organizing Maps to the Maritime Environment.

[11] Mwasiagi, J. I. (n.d.). Self-Organizing Maps - Applications and Novel Algorithm Design.

[12] P. Venkatesan, M. (n.d.). Visualization of Breast Cancer Data by SOM Component Planes.

[13] Quanquan Gao, T. R. (n.d.). A Classified Herbs Method and Searching Algorithm of Classification Tree Used for Mathematical Measurement of Effect.

[14] Rajdev Tiwari, M. P. (2010). Correlation-based Attribute Selection using Genetic Algorithm. Ijca.

[15] Sang-Jun Yea, B. Y. (2016). A data mining approach to selecting herbs with similar efficacy: Targeted selection methods based on medical subject headings (MesH). Jep.