On the determination of important plants for ayurvedic formulas in Bangladesh using unsupervised machine learning approach

Accepted 24th January, 2019

ABSTRACT

Ayurveda is the most ancient traditional medicine system practiced in the sub-continents including Bangladesh. It is very popular in Bangladesh because its ingredients are cheap and the production process is much simpler than Allopathic medicines. Ayurvedic medicines are usually mixtures of various herbal plants and Bangladesh is very rich in bio-diversity. It has more than five hundred (500) medicinal plants species. Ayurvedic data was collected for analysis from literature. These data contain Ayurvedic Medicine Formulas with their plants or biological ingredients, non-plant ingredients like minerals and their application on different types of diseases. In our analysis, we considered diseases like Indigestion, stomach ache, hemorrhoid and dysentry etc. Formulas from different manufacturers were presented in the data. We generated two relationship matrices with one classifying disease classes and the other being a binary matrix. The binary matrix contains formulas in rows and plants in columns. Hierarchical clustering was applied into the binary matrix data set for accumulating similar formulas included in a cluster. Using different distance/similarity measures we chose ‘Tanimoto Coefficient’ for hierarchical clustering as it gives more clusters of similar efficacies. The thereafter, we organized a three column dataset which contain formulas, classes of diseases and cluster membership ID. Finally, we found out most effective plants against corresponding diseases from this three columns data set. Our research will help Ayurvedic medicine producers and practitioners to determine the best plant for an Ayurvedic formula for a specific disease class or classes of diseases.

Key words: Plants, ayurvedic, matrix, cluster, efficacy.

INTRODUCTION

Human biology from the onset has a fundamental need for food, water, clean air, shelter and relative climatic constancy. To cover their basic needs, they utilize plants. Myriads of plants have been used as medicinal plants. More than 70,000 medicinal plants are estimated around the world (Verpoorte et al., 2006). Plants are used as herbal medicines in many countries such as Bangladesh (Unani, Hamdard), India (Ayurveda, Siddha), Japan (Kampo medicine), China (traditional Chinese medicine) and Indonesia (Jamu). The usage of herbal medicines is growing dramatically and continues to grow worldwide whether the number of chemical medicines decreases every year as shown by the statistics (Ernst, 2003; Furnharm, 1996).

Ayurveda is one of the most renowned traditional systems of medicine that has survived and flourished from ages till date. It is the oldest healing science that originated in India more than 3,000 years ago (Wujastyk et al., 2008). Ayurveda contains two Sanskrit words where ‘Ayu’ means ‘life’ or ‘lifespan’ and ‘Veda’ means ‘knowledge’. In terms of worldwide popularity, Ayurvedic medicine is positioned as...
number three. People have recognized that from the common cold to serious ailments like arthritis and diabetes, through body detoxification and psychosomatic issues, the mind can create illness in the body and vice versa. A lot of research work is also going on in the field of Ayurveda (Aggarwal et al., 2006).

Many researchers are trying to systemize the formulations and develop the methods to find plant disease relations based on Jamu, Ayurveda and Unani formulas. For example, Afendi et al. (2013) developed a method to find relationships between Jamu plants and efficacy using Partial Least Squares Discriminant Analysis (PLS-DA) and bootstrapping technique. Previously, network clustering-based methods were developed to find plant disease relations based on Jamu and Unani formulas (Hossain et al., 2018; Wijaya et al., 2014). Many clinical trials on disease such as diabetes, schizophrenia, Alzheimer’s disease, Parkinson’s disease, hypertension, cardio vascular disease, osteoarthritis, psoriasis, stress and more have been carried out (Bhalerao et al., 2013).

Ayurvedic treatments provide comparatively slow but permanent relief where Allopathic treatments cannot provide permanent relief. Ayurvedic medicines are relatively cheaper than Allopathic medicines. For some serious ailments such as jaundice, piles and arthritis, etc, Allopathy cannot offer a permanent solution, but Ayurveda can successfully cure these diseases (Jawla et al., 2009).

The usage of medicinal plants is traditionally rooted in Bangladesh; the use of traditional medicine is an integral part of life and living of the common mass in Bangladesh due to economical and easy accessibility. Currently, Bangladesh has a market of about BDT3000 million worth herbal or traditional products per annum (Faroque et al., 2013).

Sadhana Aushadhalaya founded by Jogesh Chandra Ghosh, was the first to bring Ayurvedic medicine to Bangladesh. There are 68 sale centers in Bangladesh. There are about 450 different types of medicines, but presently 120 types of medicines are prepared (Upashana, 2014). Hamdard is also one of the major Ayurvedic pharmaceutical companies in Bangladesh. At present, both modern and traditional physicians recommend Hamdard medicine in their routine practice. Hamdard has introduced time honored presentations of herbal medicines like capsule, tablet, syrup, ointment and elixir in Bangladesh. Ayurveda Research and Health Center founded by Lady Sarwat Abed and Zahidalspahani in 2014, is the first ever authentic Ayurveda Panchkarma centre in Bangladesh. At ARHC, detoxification is done with the assistance of Kerala Panchkarma, and herbal supplements which are used for pacification (Adiba, 2018).

In Bangladesh, there is less research on Ayurveda and less information about traditional plants which are used to create Ayurvedic medicine. In our research, we tried to collect information regarding the traditional plants, and the type of plant used to cure diseases.

### MATERIALS AND METHODS

#### Data collection

The Ayurvedic data for this analysis is collected from a literature (Unani and Ayurvedic Board, 2011). This Ayurvedic data contain Ayurvedic medicine formulas with their plant or biological ingredients, non-plant ingredients like minerals and their application on different types of diseases. In these data, formulas regarding major and most common diseases such as paralysis, pain, fever, cold, male-female reproductive system diseases, gastric, intestinal problems, brain diseases, malaria and worm caused diseases are presented. This data contains information about 294 Ayurvedic formulas and 260 plants used in the formulas. This data is the basis of our analysis.

#### Data processing

For conducting our research and analysis, the Ayurvedic data previously mentioned and viewed needs to be organized. Because the data is a raw data that contains information about the formulas, their ingredients and applications but for our analysis we need some data processing. We have generated few relationship matrices by Microsoft Office Excel from the data and also created efficacy groups with similar diseases presented in the data for the formulas used. Finally, the efficacy group is generated.

#### Generation of efficacy group

In an efficacy group similar kind of diseases to be included in the same efficacy group or disease class was considered. We have considered 15 efficacy groups from NCBI and 2 others added (NCBI website, 2019). Here, comes the generation process of different relationship matrices for performing analysis in our research. We generated two types of data matrices and these are: Ayurvedic formula vs. plants matrix and 2) Ayurvedic formula vs. efficacy matrix.

#### Ayurvedic formula vs. plants matrix

Another matrix named “Ayurvedic Formula vs. Plants” for our data analysis and research was formed. In this matrix Ayurvedic Formulas are taken in rows in left of the matrix and plants in column. This is a 294 by 260 matrix of binary nature meaning it is a 1, 0 matrix. For better understanding it can be said that, the matrix is generated to show a binary relationship between formulas and plants. That means in a formula, the presence of a plant is indicated by 1 and the absence indicated by 0. If a plant is present in a formula then under the plants column for the formula it will be 1...
and if not it will be 0. The binary matrix is shown in Table 1. For example, in formula F₁, the value of plant P₁ is 0, because the plant is not used in the formula F₁ and the value of plant P₂ is 1 because the plant is used in the formula F₁.

### Ayurvedic formula vs. efficacy matrix

The third matrix generated for our research is named "Ayurvedic formula vs. efficacy matrix". This matrix is an indicator matrix for Ayurvedic formulas and their efficacies. That means an Ayurvedic formula is applicable for efficacy or efficacies that can be visualized from the matrix. It is a matrix that consists of 294 formulas and their efficacies. The Ayurvedic formula vs. efficacy matrix is shown in Table 2.

Referring to Table 2, it can be said that formula F₂ can be used for curing diseases under efficacy groups Y₂, Y₃, Y₅, and Y₇. From the matrix of Ayurvedic formula vs. efficacy, we can understand the formula that is applicable and for which efficacy or efficacies. We generated these matrices because their relationships will be necessarily used for our research and data analysis such as for Hierarchical clustering.

### RESULTS AND DISCUSSION

Hierarchical clustering was used as an unsupervised machine learning algorithm to analyze our data for accumulating similar formulas in a cluster. In addition, we can find out similar efficacies included in a cluster. More importantly, we can also find out the relationship between plants and diseases. The dataset called ‘Ayurvedic formula vs. plants’ is a binary matrix that contains 294 observations and 260 variables. Here, the observations are formulas, while the variables are plants. We used R programming language for analyzing and visualizing our dataset.

### Hierarchical clustering

We used the most common type of hierarchical clustering known as agglomerative or bottom up clustering based on Ward’s method. Initially, each of the observation is treated as its own cluster. Thereafter, the two similar clusters were fused together based on their distance. This process continues iteratively until all of the observations belong to one single cluster (James et al., 2013). Applying hierarchical clustering into the binary matrix produced a cluster dendrogram where each cluster has its own cluster membership ID. A total of 100 classes for cluster were observed as we divide our cluster dendrogram by 100. Attaching these cluster membership IDs into our dataset called ‘Ayurvedic formula vs. efficacy matrix’ (two columns file) enabled us to observe the formula belonging to each cluster including their efficacies and after the attachment, we named this as ‘Three column file’. Each cluster contains one or more than one formula based on their similarity. Table 3 and Figure 1 show the ‘Three columns file’ and the cluster dendrogram. These three clusters marked ‘circle’ contain formulas of similar efficacy (Nutritional and Metabolic Diseases) greater or equal to 50%.

### Hierarchical clustering of Ayurvedic formulas using different distance/similarity measures

The results of hierarchical clustering depend on distance/similarity measure used in the clustering process. There are many different types of distance/similarity measures regarding binary vectors (Wijaya et al., 2016). In the current work, we applied three different types of distance/similarity measures namely “Euclidean distance”, “Pearson correlation coefficient” and Tanimoto similarity coefficient”.

The Euclidean distance between two points “a and b” with k dimensions is defined as:

\[
    d = \sqrt{\sum_{i=1}^{k} (a_i - b_i)^2}
\]

(1)

Here, d is called distance or Euclidean length which is the distance between a and b in the k-dimensional Euclidean space. The lesser the distance between points, the more similar they become.

The correlation between two points a and b with k dimensions is calculated as:

\[
    \frac{\text{cov}(a,b)}{\text{std}(a) \times \text{std}(b)}
\]

(2)
Table 3: Three columns file.

<table>
<thead>
<tr>
<th>Ayurvedic formula</th>
<th>Efficacy</th>
<th>Cluster membership ID</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anqaruya Kabir</td>
<td>Muscle and bone, the digestive system</td>
<td>1</td>
</tr>
<tr>
<td>Guruchyadi Kwath</td>
<td>The digestive system</td>
<td>10</td>
</tr>
<tr>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Darusataka Lep</td>
<td>Muscle and bone, the digestive system</td>
<td>99</td>
</tr>
</tbody>
</table>

Figure 1: Hierarchical dendrogram plot of formulas.

Where covariance and standard deviation are:

\[
\text{cov}(a, b) = \frac{1}{k} \sum_{j=1}^{k} (a_j - \bar{a}) \times (b_j - \bar{b}) \tag{3}
\]

\[
\text{std}(a) = \sqrt{\frac{1}{k} \sum_{j=1}^{k} (a_j - \bar{a})^2} \tag{4}
\]

\[
\bar{a} = \frac{1}{k} \sum_{j=1}^{k} a_j \tag{5}
\]

This correlation is called “Pearson Product Momentum Correlation”, simply referred to as Pearson's correlation. It ranges from +1 to -1, where +1 is the highest correlation. Complete opposite points have correlation -1.
### Table 4: Most important plants effective against correspondent disease classes.

<table>
<thead>
<tr>
<th>Diseases class</th>
<th>Plants</th>
</tr>
</thead>
<tbody>
<tr>
<td>Nutritional and metabolic diseases</td>
<td>Aconitum heterophyllum, Cyperus rotundus, Hollerrhena ant dysentricra, Cinnamomum camphora, Sida cordifolia, Aegle marmelos, Woodfordia fruticosa, Apium graveolens, Butea monosperma,</td>
</tr>
<tr>
<td>Skin and connective tissue</td>
<td>Terminalia chebula, Trichosanthes dioica, Azadirachta indica</td>
</tr>
<tr>
<td>Male-specific diseases</td>
<td>Cinnamomum zeylanicum, Anacyclus pyrethrum, Piper longum, Zingiber officinale, Pterocarpus santalinus, Myristica frangrans, Papaver somniferum, Crocus sativus</td>
</tr>
<tr>
<td>The digestive system</td>
<td>Piper longum, Zingiber officinale, Emblica officinalis, Trachyspermum ammi, Carum copticum, Myristica frangrans, Glycyrrhiza glabra, Nigella sativa, Piper nigrum, Piper longum, Acorus calamus, Hollerrhena ant dysentricra, Operculina turpethum</td>
</tr>
<tr>
<td>Muscle and bone</td>
<td>Piper nigrum, Aconitum ferox</td>
</tr>
<tr>
<td>Blood and lymph diseases</td>
<td>Aconitum ferox</td>
</tr>
<tr>
<td>Mental and behavioral disorders</td>
<td>Callotropis gigantean</td>
</tr>
<tr>
<td>Female-specific diseases</td>
<td>Aloe barbadensis, Ferula foetida</td>
</tr>
<tr>
<td>Dental diseases</td>
<td>Zingiber officinale, Cyperus rotundus, Terminalia chebula</td>
</tr>
<tr>
<td>The urinary system</td>
<td>Acorus calamus, Terminalia chebula, Terminalia bellica, Emblica officinalis, Piper cubeab</td>
</tr>
</tbody>
</table>

The Tanimoto coefficient (TC) between two points $a$ and $b$ represented by binary vectors with $k$ dimensions is calculated as:

$$
TC = \frac{\sum_{j=1}^{k} a_j \times b_j}{(\sum_{j=1}^{k} a_j^2 + \sum_{j=1}^{k} b_j^2 - \sum_{j=1}^{k} a_j \times b_j)}
$$

(6)

The TC is only applicable for a binary variable, and for binary variables the TC ranges from 0 to +1 (where +1 is the highest similarity).

We found 9 clusters for both Euclidean distance and Pearson correlation coefficient that contain similar efficacy class percentage greater or equal to 50%, but for Tanimoto coefficient, we found 20 clusters that contain greater or equal to 50% similar efficacy classes.

**Finding out most effective plants against diseases classes**

Sorting out the three columns file in ascending order according to the cluster membership ID facilitate us to find out the most homogeneous cluster, that is, formulas of similar efficacy are in a cluster. It was earlier mentioned that maximum number of homogeneous clusters was found for Tanimoto coefficient as distance measure. At present, we can use these clusters to find out most effective plants against corresponding diseases by looking for the high frequency plant (plant that is present in many formulas included in a particular cluster) in an individual cluster and choose that plant as an effective medicine for that disease. Table 4 shows the important plants effective against correspondent disease classes.

**Conclusion**

In this research, 294 Ayurvedic formulas and 260 plants were considered for clustering. Each formula is assigned to one or more efficacy groups depending on their relevancy. We applied hierarchical clustering to the dataset called ‘Ayurvedic formulas vs plants’ which is a $294 \times 260$ binary matrix. Based on the distance measure we divided 294 formulas into 100 clusters. Among three distance measures (Euclidean Distance, Correlation coefficient and Tanimoto Coefficient), we found Tanimoto Coefficient as the best choice for hierarchical clustering. By analyzing these clusters, we showed how to ascertain most effective plants against different diseases using hierarchical clustering. As a result, we can efficiently use these plants for treatment of various diseases.

Furthermore, we can specifically search for Ayurvedic formula that contains these plants for correspondent diseases. However, a lot of clinical studies on Ayurvedic products remain tentative due to poorly defined correlation with disease processes. Rigorous research in the field of Ayurvedic medicine is being conducted to keep pace with modern system of medicine.

**ACKNOWLEDGEMENT**

“This work was supported by Next Generation Interdisciplinary Research Project of NAIST and partially supported by the Ministry of Education, Culture, Sports, Science, and Technology of Japan (16K07223 and 17K00406), Platform Project for Supporting Drug Discovery and Life Science Research funded by Japan...
Agency for Medical Research (18am0101111) and NAIST Big Data Project.”

REFERENCES


