Proc. of the Intl. Conf. on Advances in Computer Science and Electronics Engineering — CSEE 2013

ISBN: 978-981-07-5461-7

Clusters of Ayurvedic Medicines Using Improved K-means Algorithm

Kajal C. Agrawal Meghana Nagori

Abstract - Clustering analysis method is one of the main analytical methods in data mining, the method of clustering algorithm will influence the clustering results directly. This paper discusses the importance of Ayurveda, standard k-means and improved k-means clustering algorithm and its implementation to form clusters of ayurvedic medicine for four disease. The improved method avoids computing the distance of each data object to the cluster centers repeatedly.

Keywords - clustering; k-means algorithm; Updated KMeans Algorithm; distance.

I. INTRODUCTION

Ayurveda . The Oldest Science of Life We are all aware of the scientific theories behind the evolution of Human life on this earth. It is an incredible development, but what is more amazing is the development and evolution of the knowledge and means to preserve human life. Development and growth of such a body of knowledge in India is referred to as Ayurveda, which was synonymous with the growth and evolution of Indian civilization and culture. We can find historical evidence of Ayurveda in the ancient books of wisdom known as the '*Vedas'*. The science of Ayurveda. Ayurveda is, perhaps, the oldest science of life, a system of diet, healing and health maintenance that is deeply spiritual in origin. Ayurveda is believed to have been around for more than 6000 years. Ayurveda is more than just a medical system. It is, in fact, a Science of Life!

What is Ayurveda? What is Ayurveda? Ayurveda is a Sanskrit word derived from two words '*Ayur',* which means life; and 'Veda', which means knowledge. Ayurveda is a medicinal science that deals with the life cycle of human being and the awareness of his life on the earth and the reason for their existence on earth. The study of Ayurveda includes herbal medicine, dietetics, body work, surgery, psychology and spirituality. According to Ayurveda, a living creature is composed of the mind, the body and the soul. It is the compound of these three elements that constitutes the science of life. Ayurveda teaches us to understand our body, our particular nature, and our individual mixture of elements at a deep physical, mental and emotional level. With this knowledge, we are able to identify activities, conditions, herbs and foods that either keep us healthy and in balance, or make us ill and throw us out of balance.

Basic Ayurvedic Concepts The BodyLife in Ayurveda is conceived as the union of body, senses, mind and soul. The living man is a conglomeration of three doshas . basic physical energies (*Vata, Pitta* & *Kapha*), seven basic tissues (plasma, blood, muscle, fat, bone, marrow and reproductive fluid) and the waste products of the body such as faeces, urine and sweat. Thus, the total body matrix comprises of the doshas, the tissues and the waste products of the body. The growth and decay of this body matrix and its constituents revolve around the food that gets processed into doshas, tissues and wastes. Ingestion, digestion, absorption, assimilation, and metabolism of food have interplay in health and disease, which are significantly affected by psychological mechanisms as well as by bio- fire (*Agni*). According

Mrs. Kajal C. Agrawal(Malegaonkar).
Dept. of Computer Science & Engineering.
Jawaharlal Nehru Engineering College, Aurangabad,
Maharashtra, India.

Mrs. Meghana Nagori.
Dept. of Computer Science & Engineering.
Government Engineering College, Aurangabad,
Maharashtra, India.

Proc. of the Intl. Conf. on Advances in Computer Science and Electronics Engineering — CSEE 2013

ISBN: 978-981-07-5461-7

to *Ayurveda* all objects in the universe including human body are composed of five basic elements namely, earth, water, fire, air and vacuum (ether).According to Ayurveda, understanding the three doshas is the basis to health and healing. The concept of Vata - Pitta - Kapha is unique to Ayurveda and is very difficult to translate into Western terms. *Vata* is the subtle energy associated with movement . composed of Space and Air.Vata governs all movements within the body. In balance, Vata promotes creativity and flexibility. Out of balance, Vata produces fear and anxiety. *Pitta i*s the energy of assimilation and transformation. Pitta expresses as the body.s metabolic system . made up of Fire and Water. If you have a *Pitta* constitution, you have a very alert and focused mind. In balance, *Pitta* promotes understanding and intelligence. Out of balance, *Pitta* arouses anger, hatred and jealousy. *Kapha* is the energy that forms the body.s structure . bones, muscles, tendons . and provides the glue that holds the cells together, formed from Earth and Water. In balance, *Kapha* is expressed as love, calmness and forgiveness. Out of balance, it leads to attachment, greed, and envy.

Health and SicknessAs per Ayurveda - Health or sickness depends on the presence or absence of a balanced state of the total body matrix including the balance between its different constituents.

Diet and Ayurveda In Ayurveda, regulation of diet as therapy has great importance. This is because it considers human body as the product of food. An individual.s mental and spiritual development as well as his temperament is influenced by the quality of food consumed by the individual. Food in human body is transformed first into chyle or *Rasa* and then successive processes involve its conversion into blood, muscle, fat, bone, bone marrow, reproductive elements, and *ojas.* Thus, food is basic to all the metabolic transformations and life activities. Lack of nutrients in food or improper transformation of food leads to a variety of conditions of disease in the human body. It must be emphasized that Ayurveda is not a substitute for Western allopathic medicine. There are many instances when the disease process and acute conditions can best be treated with drugs or surgery. Ayurveda can be used in conjunction with Western medicine to make a person stronger and less likely to be afflicted with disease and/or to rebuild the body after being treated with drugs or surgery. The significance of Ayurveda has increased over a period of time as the practitioners of Ayurveda have become professional practitioners and much trusted from the beginning of 20th century and more and more people have started believing in the science of Ayurveda

Clustering is a way that classifies the raw data reasonably and searches the hidden patterns that may exist in datasets [7]. It is a process of grouping data objects into disjointed clusters so that the data's in the same cluster are similar, yet data's belonging to different cluster differ. The demand for organizing the sharp increasing data's and learning valuable information from data, which makes clustering techniques are widely applied in many application areas such as artificial intelligence, biology, customer relationship management, data compression, data mining, information retrieval, image processing, machine learning, marketing, medicine, pattern recognition, psychology, statistics [3] and so on.

K-means is a numerical, unsupervised, non-deterministic, iterative method. It is simple and very fast, so in many practical applications, the method is proved to be a very effective way that can produce good cluster in results. But it is very suitable for producing globular clusters. Several attempts were made by researchers to improve efficiency of the k-means algorithms [5].

In the literature [3], there is an improved k-means algorithm based on weights. This is a new partitioning clustering algorithm, which can handle the data's of numerical attribute, and it also can handle the data's of symbol attribute. Meanwhile, this method reduces the impact of isolated points and the

Proc. of the Intl. Conf. on Advances in Computer Science and Electronics Engineering — CSEE 2013

ISBN: 978-981-07-5461-7

"no

However, this method has No improvement on the complexity of time. This method may produce more accurate clustering results than the standard k-means algorithm, but this method does not have any improvements on the executive time and the time complexity of algorithm. This algorithm can generate the same clustering results as that of the standard k-means algorithm,This algorithm is superior to the standard k-means method on running time and accuracy, thus enhancing the speed of clustering and improving the time complexity of algorithm. An improved method can effectively shorten the running time.This paper proposed the clusters of medicine on diseases using the standard k-means and improved k-means algorithms.

This paper includes four parts: The second part details the kmeans algorithm and shows the shortcomings of the standard kmeans algorithm. The third part presents the improved k-means clustering algorithm, the last part of this paper describes the structure of project.

II. THE K-MEANS CLUSTERING ALGORITHM

A. *The process of k-means algorithm*

This part briefly describes the standard k-means algorithm. Kmeans is a typical clustering algorithm in data mining and which is widely used for clustering large set of data's. In 1967, MacQueen firstly proposed the k-means algorithm, it was one of the most simple, non-supervised learning algorithms, which was applied to solve the problem of the well-algorithm, this method is to classify the given date objects into k different clusters through the iterative, converging to a local minimum. So the results of generated clusters are compact and independent. The algorithm consists of two separate phases. The first phase selects k centers randomly, where the value k is fixed in advance. The next phase is to take each data object to the nearest center. Euclidean distance is generally considered to determine the distance between each data object and the cluster centers. When all the data objects are included in some clusters, the first step is completed and an early grouping is done. Recalculating the average of the early formed clusters. This iterative process

Continues repeatedly until the criterion function becomes the minimum. Supposing that the target object is x, xi indicates the average of cluster Ci, criterion function is defined as follows:

$$E=\sum_{i=1}^{k}\sum_{x\in C_i}\left|x-x_i\right|^2$$

E is the sum of the squared error of all objects in database. The distance of criterion function is Euclidean distance, which is used for determining the nearest distance between each data objects and cluster center. The Euclidean distance between one vector x=(x1 ,x2 ,…xn) and another vector y=(y1 ,y2 ,…yn), The Euclidean distance d(xi, yi) can be obtained as follow:

$$d(x_i,\ y_i)=\left[\sum_{i=1}^{n}(x_i-y_i)^2\right]^{1/2}$$

The process of k-means algorithm as follow:

Input:

Proc. of the Intl. Conf. on Advances in Computer Science and Electronics Engineering — CSEE 2013

ISBN: 978-981-07-5461-7

Number of desired clusters, k, and a database D={d1, d2,…dn} containing n data objects.

Output:

A set of k clusters

Steps:

1) Randomly select k data objects from dataset D as initial cluster centers.

2) Repeat;

3) Calculate the distance between each data object di (1 <= i<=n) and all k cluster centers cj(1<=j<=k) and assign data object di to the nearest cluster.

4) For each cluster j (1<=j<=k), recalculate the cluster center.

5) Until no changing in the center of clusters.

The k-means clustering algorithm always converges to local minimum. Before the k-means algorithm converges, calculations of distance and cluster centers are done while loops are executed a number of times, where the positive integer t is known as the number of k-means iterations. The precise value of t varies depending on the initial starting cluster centers [8]. The distribution of data points has a relationship with the new clustering center, so the computational time complexity of the kmeans algorithm is O(nkt). n is the number of all data objects, k is the number of clusters, t is the iterations of algorithm. Usually requiring k <<n and t <<n.

III. IMPROVED K-MEANS CLUSTERING ALGORITHM

The standard k-means algorithm needs to calculate the distance from the each date object to all the centers of k clusters when it executes the iteration each time, which takes up a lot of execution time especially for large-capacity databases. For the shortcomings of the above k-means algorithm, this paper presents an improved k-means method. The main idea of algorithm is to set two simple data structures to retain the labels of cluster and the distance of all the date objects to the nearest cluster during the each iteration, that can be used in next iteration, we calculate the distance between the current date object and the new cluster center, if the computed distance is smaller than or equal to the distance to the old center, the data object stays in its cluster that was assigned to in previous iteration. Therefore, there is no need to calculate the distance from this data object to the other k-1clustering centers, saving the calculative time to the k-1 cluster centers. Otherwise, we must calculate the distance from the current data object to all k cluster centers, and find the nearest cluster center and assign this point to the nearest cluster center. And then we separately record the label of nearest cluster center and the distance to its center. Because in each iteration some data points still remain in the original cluster, it means that some parts of the data points will not be calculated, saving a total time of calculating the distance, thereby enhancing the efficiency of the algorithm.

The process of the improved algorithm is described as follows:

Input:

The number of desired clusters k, and a database D= {d1, d2,…dn} containing n data objects.

Output:

A set of k clusters

Steps:

1) Randomly select k objects from dataset D as initial cluster centers.

2) Calculate the distance between each data object di (1 <=i<=n) and all k cluster centers cj (1<=j<=k) as

Proc. of the Intl. Conf. on Advances in Computer Science and Electronics Engineering — CSEE 2013

ISBN: 978-981-07-5461-7

Euclidean distance d(di , cj) and assign data object di to the nearest cluster.

3) For each data object di, find the closest center cj and assign di to cluster center j;

4) Store the label of cluster center in which data object di is and the distance of data object di to the nearest cluster and store them in array Cluster[ ] and the Dist[ ] separately. Set Cluster[i]=j, j is the label of nearest cluster. Set Dist[i]=d(di, cj), d(di, cj) is the nearest Euclidean distance to the closest center.

5) For each cluster j (1<=j<=k), recalculate the cluster center;

6) Repeat

7) For each data object di Compute its distance to the center of the present nearest cluster;

a) If this distance is less than or equal to Dist[i], the data object stays in the initial cluster;

b) Else

For every cluster center cj(1<=j<=k), compute the distance d(di, cj) of each data object to all the center, assign the data object di to the nearest center cj.

Set Cluster[i]=j;

Set Dist[i]= d(di, cj);

8) For each cluster center j(1<=j<=k), recalculate the centers;

9) Until the convergence criteria is met.

10) Output the clustering results;

The improved algorithm requires two data structure (Cluster[ ]and Dist[ ])to keep the some information in each iteration which is used in the next iteration. Array cluster[ ] is used for keep the label of the closest center while data structure Dist[ ] stores the Euclidean distance of data object to the closest center. The information in data structure allows this function to reduce the number of distance calculation required to assign each data object to the nearest cluster, and this method makes the improved k-means algorithm faster than the standard k-means algorithm.

This paper proposes an improved k-means algorithm, to obtain the initial cluster, time complexity of the improved kmeans algorithm is O(nk). Here some data points remain in the original clusters, while the others move to other clusters. If the data point retains in the original cluster, this needs O(1), else O(k). With the convergence of clustering algorithm, the number of data points moved from their cluster will reduce. If half of the data points move from their cluster, the time complexity is O(nk/2). Hence the total time complexity is O(nk). While the standard k-means clustering algorithm require O(nkt). So the proposed k-means algorithm in this paper can effectively improve the speed of clustering and reduce the computational complexity. But the improved k-means algorithm requires the pre estimated the number of clusters, k, which is the same to the standard k-means algorithm. If you want to get to the optimal solution, you must test the different value of k.

IV. IMPLEMENTATION

A. We considered four disease.

B. Prepared the database according to that disease.

C. Graphical Representation.

We are using Weka 3.7.7 release to demonstrate the clustering of data using Improvised KMeans Clustering Algorithm. For this purpose we are reprogramming Weka 3.7.7 and adding an Improved K-Means Clustering algorithm to its libraries. That will result in representation of clustering by both Simple KMeans

ISBN: 978-981-07-5461-7

Algorithm & Improved KMeans Algorithm. We are reprogramming this tool using Java 6 programming in Netbeans 6.0.1.

D. Backend.

We are using ARFF& MS Access databases for this project.

IV. DATABASE

Database Design

A. Disease

Field	Type
Sr_no	Number
D_Name	Text

B. Kalp (Medicine)

Field	Type
Sr_no	Number
K_Name	Text

C. Herbs

Field	Type
Sr_no	Number
H_Name	Text

D. Disease to Kalp Relational Table

Field	Type
Sr_no	Number
Disease	Number
Kalp	Number

E. Kalp to Herbs Relational Table

Field	Type
Sr_no	Number
Kalp	Number
Disease	Number

REFERENCES

[1] Yuan F, Meng Z. H, Zhang H. X and Dong C. R, "A New Algorithm to Get the Initial Centroids," Proc. of the 3rd International Conference on Machine Learning and Cybernetics, pp. 26–29, August 2004.

[2] Sun Jigui, Liu Jie, Zhao Lianyu, "Clustering algorithms Research",Journal of Software ,Vol 19,No 1, pp.48-61,January 2008.

[3] Sun Shibao, Qin Keyun,"Research on Modified k-means Data Cluster Algorithm"I. S. Jacobs and C. P. Bean, "Fine particles, thin films and exchange anisotropy," Computer Engineering, vol.33, No.13, pp.200–201,July 2007.

[4] Merz C and Murphy P, UCI Repository of Machine Learning Databases, Available: ftp://ftp.ics.uci.edu/pub/machine-learning-databases

[5] Fahim A M,Salem A M,Torkey F A, "An efficient enhanced k-means clustering algorithm" Journal of Zhejiang University Science A, Vol.10, pp:1626-1633,July 2006.

[6] Zhao YC, Song J. GDILC: A grid-based density isoline clustering algorithm. In: Zhong YX, Cui S, Yang Y, eds. Proc. of the Internet Conf. on Info-Net. Beijing: IEEE Press, 2001. 140−145. http://ieeexplore.ieee.org/iel5/7719/21161/00982709.pdf

[7] Huang Z, "Extensions to the k-means algorithm for clustering large data sets with categorical values," Data Mining and Knowledge Discovery, Vol.2, pp:283–304, 1998.

[8] K.A.Abdul Nazeer, M.P.Sebastian, "Improving the Accuracy and Efficiency of the k-means Clustering Algorithm",Proceeding of the World Congress on Engineering, vol 1,london, July 2009.

[9] Fred ALN, Leitão JMN. Partitional vs hierarchical clustering using a minimum grammar complexity approach. In: Proc. of the SSPR & SPR 2000. LNCS 1876, 2000. 193−202. http://www.sigmod.org/dblp/db/conf/sspr/sspr2000.htm

[10] Gelbard R, Spiegler I. Hempel's raven paradox: A positive approach to cluster analysis. Computers and Operations Research, 2000,27(4):305−320.

[11] Huang Z. A fast clustering algorithm to cluster very large categorical data sets in data mining. In: Proc. of the SIGMOD Workshop on Research Issues on Data Mining and Knowledge Discovery. Tucson, 1997. 146−151.http://www.informatik.uni-trier.de/~ley/db/conf/sigmod/sigmod97.html

[12] Ding C, He X. K-Nearest-Neighbor in data clustering: Incorporating local information into global optimization. In: Proc. of the ACM Symp. On Applied Computing. Nicosia: ACM Press, 2004. 584−589.http://www.acm.org/conferences/sac/sac2004/

[13] Hinneburg A,Keim D.An efficient approach to clustering in large multimedia databases with noise.In:Agrawal R,Stolorz PE,Piatetsky- Shapiro G,eds.Proc.of the 4th Int'l Conf.on Knowledge

ISBN: 978-981-07-5461-7

Discovery and Data Mining(KDD'98).New York:AAAIPress,1998.58~65.
[14] Zhang T,Ramakrishnan R,Livny M.BIRCH:An efficient data clustering method for very large databases.In:Jagadish HV,Mumick IS,eds.Proc.of the 1996 ACM SIGMOD Int'l Conf.on Management of Data.Montreal:ACM Press,1996.103~114.
[15] Birant D, Kut A. ST-DBSCAN: An algorithm for clustering spatialtemporal data. Data & Knowledge Engineering, 2007,60(1): 208-221. 67
[16]www.ayurbest.com/cms/media/Newpapper_article_on_Ayurveda.pdf[17]www.ayukalp.com