Proceedings of the International Conference on
Pattern Recognition, Informatics and Medical Engineering , March 21-23, 2012

An Analysis on the Impact of Fluoride in Human Health (Dental) using Clustering Data mining Technique

T.Balasubramanian
Department of Computer Science,
Sri Vidya Mandir Arts and Science College,
Uthangarai(PO), Krishnagiri(Dt), Tamilnadu, India.
balaeswar123@gmail.com

R.Umarani
Department of Computer Science,
Sri Saradha College for Women,
Salem, Tamilnadu, India.
umainweb@gmail.com

Abstract: Data Mining is the process of extracting information from large data sets through using algorithms and Techniques drawn from the field of Statistics, Machine Learning and Data Base Management Systems. Traditional data analysis methods often involve manual work and interpretation of data which is slow, expensive and highly subjective Data Mining, popularly called as knowledge discovery in large data, enables firms and organizations to make calculated decisions by assembling, accumulating, analyzing and accessing corporate data. It uses variety of tools like query and reporting tools, analytical processing tools, and Decision Support System. [1][2]

This article explores data mining techniques in health care. In particular, it discusses data mining and its application in areas where people are affected severely by using the under- ground drinking water which consist of high levels of fluoride in Krishnagiri District, Tamil Nadu State, India. This paper identifies the risk factors associated with the high level of fluoride content in water, using clustering algorithms and finds meaningful hidden patterns which give meaningful decision making to this socio-economic real world health hazard.

Keywords: Data mining, Fluoride affected people, Clustering algorithms, K-Means.

I. INTRODUCTION

A. Fluoride as a Health Hazard

Fluoride ion in drinking water ingestion is useful for Bone and Teeth development, but excessive ingestion causes a disease known as fluorosis. The prevalence of fluorosis is mainly due to the consumption of more fluoride through drinking water. Different forms of fluoride exposure are of importance and have shown to affect the body's Fluoride content and thus increasing the risks of Fluoride-prone diseases. [4]

Fluorosis was considered to be a problem related to Teeth only. But it now has turned up to be a serious health hazard. It seriously affects Bones and problems like Joint pain, Muscular Pain, etc. which are well known their manifestations. It not only affects the body of a person but also renders them socially and culturally crippled. [4]

The goal of this paper by using the clustering algorithms as a tool of data mining technique is to find out the volume of people affected by the high fluoride content of potable water.

FIG 1 : MODERATE FLUOROSIS DENTAL

II. MATERIALS AND METHODS

A. Literature Survey of The Problem

To understand the health hazards of fluoride content on living beings, discussions were made with medical practitioners and specialists like General Dental, Neuro surgeons and Ortho specialists. We have also gathered details about the impact of high fluoride content water from World Wide Web [4]. By analyzing all these we came to know that the increased fluoride level in ground water create dental, skeletal and neuro problems. **In this analysis we focus only on Dental hazards by high fluoride level in drinking water**. Level of fluoride content in water in different regions of Krishnagiri District was obtained from Water Analyst . Based on the recommendations of WHO which released a water table, Tamilnadu Water And Drainage Board (TWAD)

978-1-4673-1039-0/12/$31.00 ©2012 IEEE

Suggested level of fluoride content in drinking water should not exceed 1.5 mg/L.[5].

The Water table also shows the contents of minerals and associated health hazards. We found out that Krishnagiri District of Tamilnadu in India is most affected by fluoride level in water by naturally surrounded hills in the District.

TWAD have analyzed the sample ground potable water from various regions of Krishnagiri District and maintained a table of High level fluoride (1.6mg/L to 2.4mg/L) contaminated ground drinking water of panchayats and villages list in this District. We conclude that in Krishnagiri District, many people in the villages and panchayats are severely affected by ground potable water. So we decided to make a survey and found out the combination of diseases which are possibly affected mostly by high fluoride content in water.

B. Data Preparation

Based on the information from various physicians and water analyst, we have prepared questionnaires to get raw data from the various fluoride impacted villages and panchayats, having fluoride level in water from 1.6mg/L to 2.4mg/L. People of different age groups with different ailments were interviewed with the help of questionnaire prepared in our mother tongue, Tamil since the people in and around the district are not up to the level of understanding other languages.

Total data collected from Villages and Panchayats

Men 251 (48%)
Women 269 (52%) **520**

Based on the medical practitioners advise, while classifying the data , the degree of symptoms are placed in several compartments as follows:

None
Mild Dental Victims
Moderate Dental Victims
Severe Dental Victims

With the following classification.

No symptoms found grouped as none.

Those who are found with one to three low symptoms are grouped as Mild victims of dental disease.

Those who are found with four low symptoms or one to three medium and one high symptom are grouped as Moderate victims of dental disease.

Those who are found with more than two high symptoms are grouped as severe victims of dental diseases

C. Clustering as the Data Mining application

Clustering is one of the central concepts in the field of unsupervised data analysis, it is also a very controversial issue, and the very meaning of the concept “clustering” may vary a great deal between different scientific disciplines [6]. However, a common goal in all cases is that the objective is to find a structural representation of data by grouping (in some.

Tooth Pain	Tooth Stain	Bad Tooth Breath	Tooth Erosion	Dental Class	Remark
No	No	No	No	None	--
Low	Low	Low	--	Mild	Any three Low Symptom
Low	Low	Low	Low	Mode-rate	--
Low	Low	Medium	Medi-um	Mode-rate	Any two Medium Symptoms
Low	Medium	High	High	Severe	Any two High Symptoms

TABLE 1 CLASSIFICATION OF SYMPTOMS OF DISEASES

sense) similar data items together. A cluster has high similarity in comparison to one another but is very dissimilar to objects in other clusters

D. Weka as a data miner tool

In this paper we have used WEKA (to find interesting patterns in the selected dataset)**,** a Data Mining tool for clustering techniques.. The selected software is able to provide the required data mining functions and methodologies. The suitable data format for WEKA data mining software are MS Excel and ARFF formats respectively. Scalability-Maximum number of columns and rows the software can efficiently handle. However, in the selected data set, the number of columns and the number of records were reduced. WEKA is developed at the University of Waikato in New Zealand. “WEKA” stands for the Waikato Environment of Knowledge Analysis. The system is written in Java, an object-oriented programming language that is widely available for all major computer platforms, and WEKA has been tested under Linux, Windows, and Macintosh operating systems. Java allows us to provide a uniform interface to many different learning algorithms, along with methods for pre and post processing and for evaluating the result of learning schemes on any given dataset. WEKA expects the data to be fed into be in ARFF format (Attribution Relation File Format). [8]

WEKA has two primary modes: experiment mode and exploration mode .The exploration mode allows easy access to all of WEKA’s data preprocessing, learning, data processing, attribute selection and data visualization modules in an environment that encourages initial exploration of data. The experiment mode allows larger-scale experiments to be run with results stored in a database for retrieval and analysis.[9]

E. Clustering in WEKA

The basic classification is based on supervised algorithms. Algorithms are applicable for the input data. The process of grouping a set of physical or abstract objects into classes of similar objects is called clustering.. The Cluster tab is also supported which shows the list of machine learning tools. These tools in general operate on a clustering algorithm and run it multiple times to manipulating algorithm parameters or input data weight to increase the accuracy of the classifier. Two learning performance evaluators are included with WEKA.

The first simply splits a dataset into training and test data, while the second performs cross-validation using folds. Evaluation is usually described by the accuracy. The run information is also displayed, for quick inspection of how well a cluster works.

F. Experimental Setup

The data mining method used to build the model is cluster. The data analysis is processed using WEKA data mining tool for exploratory data analysis, machine learning and statistical learning algorithms. The training data set consists of 520 instances with 15 different attributes. The instances in the dataset are representing the results of different types of testing to predict the accuracy of fluoride affected persons. According to the attributes the dataset is divided into two parts that is 70% of the data are used for training and 30% are used for testing.[9]

G. Learning Algorithm

This paper consists of an unsupervised machine learning algorithm for clustering derived from the WEKA data mining tool. Which include:

- K-Means

The above clustering model was used to cluster the group of people who are affected by Dental fluorosis at different dental disease levels and to cluster the different water sources using by the people which are causes for Dental fluorosis in Krishnagiri District.

III. Discussion And Result

A. Attributes selection

First of all, we have to find the correlated attributes for finding the hidden pattern for the problem stated. The WEKA data miner tool has supported many in built learning algorithms for correlated attributes. There are many filtered tools for this analysis but we have selected one among them by trial.[7]

Totally there are 520 records of data base which have been created in Excel 2007 and saved in the format of CSV (Comma Separated Value format) that converted to the WEKA accepted of ARFF by using command line premier of WEKA.

The records of data base consists of 15 attributes, from which 10 attributes were selected based on attribute selection in explorer mode of WEKA 3.6.4. (fig 2)

We have chosen Symmetrical random filter tester for attribute selection in WEKA attribute selector. It listed 14 selected attributes, but from which we have taken only 8 attributes. The other attributes are omitted for the convenience of analysis of finding impaction among peoples in the district.

S.No.	Attributes	Data Type
01.	Name	Text
02.	Age	Numeric(Integer)
03.	Education	Text
04.	Sex	Character
05.	Fluoride level	Numeric(Real)
06.	Profession	Text
07.	Pregnancy status	Boolean
08.	Drinking water	Text
09.	Duration	Numeric (Integer/Real)
10.	Known status of fluoride	Boolean
11.	Tooth Pain	Numeric (Binary)
12.	Tooth Stain	Numeric(Binary)
13.	Bad Tooth Breath	Numeric(Binary)
14.	Tooth Erosion	Numeric(Binary)
15.	Disease Level	Text

TABLE 2: CLASSIFICATION OF ATTRIBUTES

S.No.	Attributes	Data Types
1.	Age	Numeric(Integer)
2.	Fluoride Level	Numeric(Real)
3.	Drinking water	Text
4.	Duration	Numeric(Integer/Real)
5.	Tooth Pain	Numeric(Binary)
6.	Tooth Stain	Numeric(Binary)
7.	Bad Tooth Breath	Numeric(Binary)
8.	Tooth Erosion	Numeric(Binary)

TABLE 3: SELECTED ATTRIBUTES FOR ANALYSIS

B. K-Means Method

The k-Means algorithm takes the input parameter, k, and partitions a set of n objects into k clusters so that the resulting intra cluster similarity is high but the inter cluster similarity is low. Cluster similarity is measured in regard to the mean value of the objects in a cluster, which can be viewed a the cluster's *centroid or center of gravity.*

The K–Means algorithm proceeds as follows :

First , it randomly selects k of the objects, each of which initially represents a cluster mean or center. For each of the remaining objects, an object is assigned to the cluster to which it is the most similar, based on the distance between the object

and the cluster mean. It then computes the new mean for each cluster. This process iterated until the criterion function converges. Typically, the square-error criterion is used, defined as [3]

$$E=\sum_{i=1}^{K}\sum_{P\in C_i}|p-m_i|^2 \qquad (1)$$

Where E is the sum of the square error for all objects in the data set; p is the point in space representing a given object; and m_i is the mean of cluster C_i . In other words, for each object in each cluster, the distance from the object to its cluster center is squared, and the distances are summed. This criterion tries to make the resulting k clusters as compact and as separate as possible.

1) K-Means algorithm:

Input;
= k:the number of clusters,
= D:a data set containing n objects
Output: A set of k clusters.
Method:

(1) arbitrarily choose k objects from from D as the initial cluster centers;
(2) repeat
(3)

FIG 2: ATTRIBUTE SELECTION IN WEKA EXPLORER

(3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;

(4) Update the cluster means, i.e., calculate the mean value of the objects for each cluster;

(5) Until no change;

Suppose that there is a set of objects located in space as depicted in the rectangle shown in fig 3a. Let k = 3; i.e., the user would like the objects to be partitioned into three clusters.

According to the algorithm above we arbitrarily choose three objects as the three initial cluster centers, where cluster centers are marked by a "+". Each objects is distributed to a cluster based on the cluster center to which it is the nearest. Such a distribution forms encircled by dotted curves as show in fig 3a.

Next, the cluster centers are updated. That is the mean value of each cluster is recalculated based on the current objects in the cluster. Using the new cluster centers, the objects are redistributed to the clusters based on which cluster center is the nearest. Such a redistribution forms new encircled by dashed curves, as shown in fig3b.

This process iterates, leading to fig 3c. The process of iteratively reassigning objects to clusters to improve the partitioning is referred to as *iterative relocation.*

Eventually, no redistribution of the objects in any cluster occurs, and so the process terminates. The resulting cluster are returned by the clustering process.

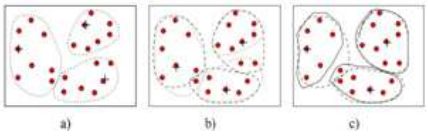

FIG 3: CLUSTERING OF A SET OF OBJECTS BASED ON K-MEANS METHOD

C. K-Means in WEKA

The learning algorithm k-Means in WEKA 3.6.4 accepts the training data base in the format of ARFF. It accepts the nominal data and binary sets. So our attributes selected in nominal and binary formats naturally. So no need of preprocessing for further process.

We have trained the training data by using the 10 Fold Cross Validated testing which used our trained data set as one third of the data for training and remaining for testing. After training and testing which gives the following results.(fig 4)

1) Euclidean distance

K-means cluster analysis supports various data types Type equation here.such as Quantitative, binary, nominal or ordinal, but do not support categorical data. Cluster analysis are based on measuring similarity between objects by computing the distance between each pair.

There are a number of methods are for computing distance in a multidimensional environment.

Distance is a well understood concept that has a number of simple properties.

1. Distance is always positive

2. Distance from point x to itself is always zero
3. Distance from point x to point y cannot be greater than the sum of the distance from x to some other point z and distance from z to y.
4. Distance from x to y is always the same as from y to x.

It is possible to assign weights to all attributes indicating their importance. There are number of distance measures such as Euclidean distance, Manhattan distance and Chebychev distance. But in this analysis Weka tool used Euclidean distance.

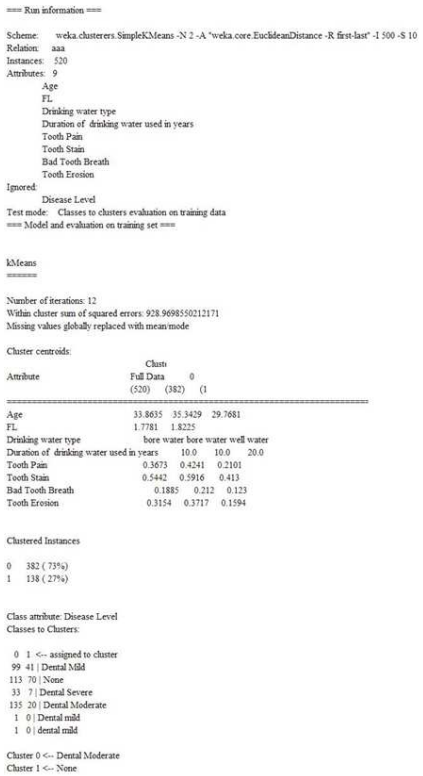

FIG 4: KMEANS IN WEKA BASED ON DISEASES SYMPTOMS

Euclidean distance of the difference vector is most commonly used to compute distances and has an intuitive appeal but the largest valued attribute may dominate the distance. It is there fore essential that the attributes are properly scaled.

Let the distance between two points x and y be D(x,y).

$$D(x,y) = (\sum(x_i-y_i)^2)^{1/2} \qquad (2)$$

2) Clustering of Disease symptoms

The collected disease symptoms such as Tooth pain, Tooth stain, Bad tooth breath and Tooth erosion as raw data, supplied to Kmeans method is being carried out in Weka using Euclidean distance method to measure cluster centroids. The result is obtained in iteration 12 after clustered. The centroid cluster points are measured based on the diseases symptoms and the water they are drinking. Based on the diseases symptoms in raw data the Kmeans clustered two main clustering units. From the confusion matrix above we came to know that the district mainly impacted by dental moderate.(Fig 4)

3) Water type in clustering

In Krishnagiri District the people are affected by fluorosis through drinking water. The source of water in this district of many type while collecting data it is categorized such as river water, well water, bore water and pond water.

With regard to that we have to find that which is main source of drinking water in Krishnagiri District of Tamil Nadu in India. With the support of Weka 3.6.4 data miner tool based on clustering we found the following is the source of flurosis drinking water.(Fig 5)

The above implementation algorithm yields results that the Krishnagiri District residing people affected by the dental disease moderately and the water source is
bore water. We came to the conclusion of this result by confusion matrix of K-means algorithm

The K-means method clustered train the data up to 100% so the error rate completely reduced. The time taken to build the algorithm relatively too small.

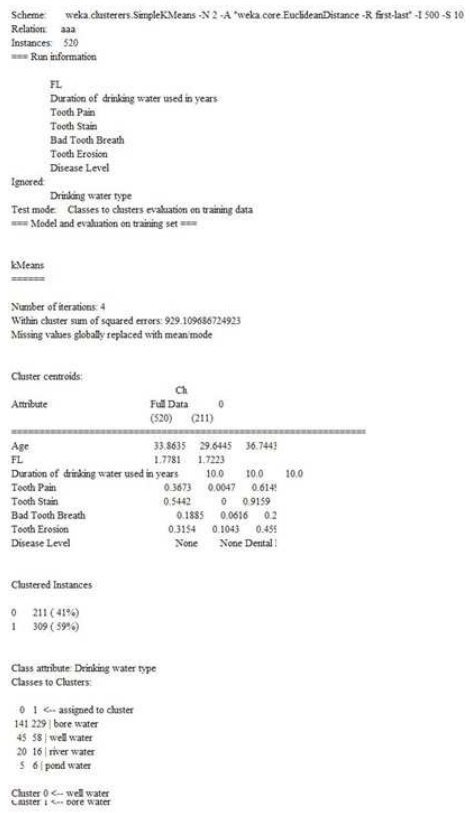

FIG 5: KMEANS IN WEKA ON WATER TYPE

IV. CONCLUSION.

The K-means algorithm was implemented using Weka 3.6.4 data miner. It clustered into two major clusters units with the class variables. The clusters were varied in zigzag manner, slightly with each iterations and finally in the 12th iteration and finally maximum of its attributes belongs to Dental moderate class. Based on this we can conclude that the Krishnagiri District impacted with moderate dental disease.

Data mining applied in health care domain, by which the people get beneficial for their lives. As the analog of this research found the meaningful hidden pattern that from the real data set collected the people impacted Krishnagiri District by drinking high fluoride content of potable water. By which we can easily know that the people do not get awareness among themselves about the fluoride impaction. If it continues in this way, it may lead to some primary health hazards like Kidney failure, mental disability, Thyroid deficiency and Heart diseases.

However the Primary Health hazards of fluoride are Dental and Bone diseases which disturbed their daily meager life. It is primary duty of the Government to providing good hygienic drinking water to the people and reduce the fluoride content potable water with the latest technologies and creating awareness among the people in some way like medical camps and taking documentary films. If continues in this way after 10 to 20 years there may be the possibilities of Severe Dental impaction among people in Krishnagiri District. Through this research the problem of fluoride in Krishnagiri come to light. It is a big social relevant problem. Pharmaceutical industries also can identify the location to develop their business by providing good medicine among people with service motto.

REFERENCES

[1] Jiawei Han and Micheline Kamber, "Data mining concepts and Techniques",Second Edition, Morgan Kaufmann Publishers second edition,2008.

[2] Arun K.Pujari, "Datamining Techniques", *University Press, First edition, fourteenth reprint,* 2009.

[3] G.K.Gupta, "Introduction to Datamining with case studies", PHI. 2009

[4] Professionals statement calling for an End to water Fluoridation – c*onference Report NRC Review,2006. (www.fluoridealert.org)*

[5] Water Quality for Better Health – *TWAD Released Waterbook.* Published IEC, TWAD, Chennai. mail:twadboard@vsnl.in,2009.

[6] A.Jain, M. Murty, and P. Flynn, 'Data clustering: A review," *ACM Computing Surveys*, vol. 31, no. 3, pp. 264–323, 1999.

[7] Weka3.6.4 data miner manual 2010.

[8] "Analysis of Liver Disorder Using Data mining algorithms", Global Journal of computer science and Technology, l.10 issue 14 (ver1.0) November 2010, pp. 48 -52.

[9] Peter Reutemann, Ian H. Witten," The WEKA Data Mining Software: AnUpdate - White paper", Pentaho Corporation. SIGKDD Explorations Volume 11, issue 1 pp. 10 - 18, 2005.