

# Clustering: An Analysis Technique in Data Mining for Health Hazards of High Levels of Fluoride in Potable Water

**Balasubramanian. T**

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

**Umarani. R**

Department of Computer Science,  
Sri Saradha College for Women (Autonomous),  
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]

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. [2]

**KEYWORDS:** Data mining, Fluoride affected people, Clustering algorithms, J48, Naive Bayes, Health hazard.

## 1. INTRODUCTION

### 1.1 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. [8]

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. are its well known manifestations. It not only affects the body of a person but also renders them socially and culturally crippled.

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



**Fig 1: Moderate Fluoride affected teeth**

## 2. MATERIALS AND METHODS

### 2.1 Literature Survey of the Problem

To understand the health hazards of fluoride content on living beings, discussions were held 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 in water from World Wide Web [8]. 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 from TWAD. Based on the recommendations of WHO which released a water table[6], the Tamil Nadu Water And Drainage Board (TWAD) suggested the normal content of fluoride in drinking water should not be above 1.5 mg/L.[6]

The water table also shows the contents of minerals and associated health hazards. We found out that Krishnagiri District of Tamil Nadu in India is most affected by fluoride level in drinking 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 have concluded that many village people of Krishnagiri District are severely affected by ground potable water. So we have decided to make a survey and to find out the combination of diseases which are possibly affected mostly by high fluoride content in drinking water.

## 2.2 Data Preparation

Based on the information from various physicians and water analyst of TWAD, we have prepared questionnaire to get raw data from too many villagers who were affected with high level fluoride in drinking water from 1.6mg/L to 2.4mg/L.[6] People of different age groups with different ailments were interviewed based on the questionnaire prepared in our mother tongue i.e. Tamil since the people in and around the district are maximum illiterate and not studied up to the level of understanding other languages.

### Total data collected from Villages

|       |           |       |
|-------|-----------|-------|
| Men   | 251 (48%) | } 520 |
| Women | 269 (52%) |       |

As per the opinion and findings of medical practitioners, while analyzing the data for classification, the degree of symptoms of diseases are placed in several compartments as under

**None**

**Mild Dental Victims**

**Moderate Dental Victims**

**Severe Dental Victims**

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 symptoms 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

From the above, the status and degree of diseases classified as under with sample table.

**Table 1: Sample classification of symptoms of diseases**

| 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           | Moderate     | --                      |
| Low        | Low         | Medium           | Medium        | Moderate     | Any two Medium Symptoms |
| Low        | Medium      | High             | High          | Severe       | Any two High Symptoms   |

## 2.3 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 [10] However, a common goal in all cases is that the objective is to find a structural representation of data by grouping (in some sense) similar data items together. A cluster has high similarity in comparison to one another but is very dissimilar to objects in other clusters.

## 2.4 Weka 3.6.4 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).

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.

## 2.5 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.

## 2.6 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.[11]

## 2.7 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.

## 3. DISCUSSION AND RESULT

### 3.1 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.[5]

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 1)

```

Evaluator: weka.attributeSelection.SymmetricalUncertAttributeEval
Search: weka.attributeSelection.Ranker -T 1.7976931348623157E308 -N 1
Relation: FORMAT OF 1-520 DENTAL-weka.filters.unsupervised.attribute.Remove-R1
Instances: 520
Attributes: 15
  Name
  Age
  Education
  Sex
  FL
  Profession
  Pragnancy status while interview
  Drinking water type
  Duration of drinking water used in years
  Known status of fluoride impact
  Tooth Pain
  Tooth Stain
  Bad Tooth Breath
  Tooth Erosion
  Disease Level
Evaluation mode: evaluate on all training data

=== Attribute Selection on all input data ===

Search Method:
  Attribute ranking.

Attribut Class (nominal): 15 Disease Level):
  Symmetrical Uncertainty Ranking Filter

Ranked attributes:
0.47774 14 Tooth Erosion
0.422 12 Tooth Stain
0.30947 13 Bad Tooth Breath
0.29332 1 Name
0.27646 11 Tooth Pain
0.07668 3 Education
0.075 9 Duration of drinking water used in years
0.05947 6 Profession
0.03682 2 Age
0.03633 5 FL
0.01854 7 Pragnancy status while interview
0.0139 8 Drinking water type
0.00794 4 Sex
0.00662 10 Known status of fluoride impact

Selectec 12 13 1 11 3 9 6 2 5 7 8 4 10:14

```

Fig 1 Attribute selection in WEKA 3.6.4 Explorer

Table 2: classification of attributes

| 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.   | Praganency 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                   |

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.

Table 3: Selected attributes for analysis

| 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)       |

### 3.2. 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 as 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 [2]

$$E = \sum_{i=1}^n \sum_{c=1}^k |p - m_c|^2$$

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  $m_c$  is the mean of cluster  $C_c$ . 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

### 3.2.1 K-Means algorithm

Input;

- = k: the number of clusters,
- = D: a data set containing n objects

Output: A set of k clusters.

Method:

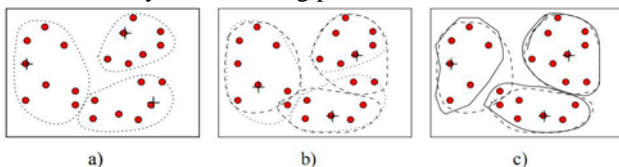
- (1) arbitrarily choose k objects from from D as the initial cluster centers;
- (2) repeat
- (3) (re)assign each object to the cluster to which the object is the most similar, based on the mean value of the objects in the cluster;
- (4) Update the cluster means, i.e., calculate the mean value of the 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 (a) 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 (a)

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 fig (b).

This process iterates, leading to fig (c). 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 2 Clustering of a set of objects based on k-means method**

## 3.3 K-Means in WEKA

### 3.3.1 Disease symptoms in clustering

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. The collected 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.

### 3.3.2 Performance Analysis

The database was implemented through the Weka 3.6.4 data miner tool in exploration mode using kmeans clustering algorithm. The test was based on the Cluster to class value. The cluster model was performed on disease level class attribute. Totally it was clustered 520 records, Here the cluster algorithm performed on all instances successfully ie it clustered 100% in database. All attributes are clustered on the mean value of each using Euclidean distance measure. Final result obtained in 12<sup>th</sup> iteration. Initially two seeds were set, as cluster, so all attributes are bounded with in only two clusters based on class values, let them 0 and 1. In each iteration the class values are changed as any of two clusters based on mean value of attribute and finally the two classes Dental Moderate and None were final clusters. As per the Confusion Matrix, in Dental Moderate cluster had 382 records and None cluster had 138 records including true positive and false positive values.

=== Run information ===

```
Scheme: weka.clusterers.SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S 10
Relation: aaa
Instances: 520
Attributes: 9
```

```
Age
FL
Drinking water type
Duration of drinking water used in years
Tooth Pain
Tooth Stain
Bad Tooth Breath
Tooth Erosion
```

Ignored:

```
Disease Level
```

Test mode: Classes to clusters evaluation on training data

=== Model and evaluation on training set ===

kMeans

=====

Number of iterations: 12

Within cluster sum of squared errors: 928.9698550212171

Missing values globally replaced with mean/mode

Cluster centroids:

| Attribute                                | Clusti                           |         |         |
|--|----------------------------------|---------|---------|
|  | Full Data                        | 0       | 1       |
|  | (520)                            | (382)   | (1)     |
| Age                                      | 33.8635                          | 35.3429 | 29.7681 |
| FL                                       | 1.7781                           | 1.8225  |         |
| Drinking water type                      | bore water bore water well water |         |         |
| Duration of drinking water used in years | 10.0                             | 10.0    | 20.0    |
| Tooth Pain                               | 0.3673                           | 0.4241  | 0.2101  |
| Tooth Stain                              | 0.5442                           | 0.5916  | 0.413   |
| Bad Tooth Breath                         | 0.1885                           | 0.212   | 0.123   |
| Tooth Erosion                            | 0.3154                           | 0.3717  | 0.1594  |

Clustered Instances

```
0 382 ( 73%)
1 138 ( 27%)
```

Class attribute: Disease Level

Classes to Clusters:

```
0 1 <- assigned to cluster
99 41 | Dental Mild
113 70 | None
33 7 | Dental Severe
135 20 | Dental Moderate
1 0 | Dental mild
1 0 | dental mild
```

Cluster 0 <- Dental Moderate

Cluster 1 <- None

**Fig 3 Kmeans in weka 3.6.4 based on diseases symptoms**



### 3.3.3 Result

In exact 135 records were originally clustered in Dental Moderate cluster and 70 records were originally clustered in none cluster. From the confusion matrix above we can come to the conclusion that the Krishnagiri District is impacted Dental Moderate disease state (fig 3).

### 4. CONCLUSION.

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 in 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 District came 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] Berry Mj Linoff G, "Data mining Techniques: for Marketing, Sales and Customer support USA", Wiley, 1997.
- [5] Weka3.6.4 data miner manual. 2010.
- [6] Water Quality for Better Health – *TWAD Released Water book*. Published IEC, TWAD, Chennai. mail: twadboard@vsnl.in
- [7] Plamena Andreeva, Maya Dimibova and Petra Radeve, "Data mining Learning models and Algorithms for medical applications – White paper", page no.44, 2004.
- [8] Professionals statement calling for an End to water Fluoridation – *Conference Report NRC Review, 2006.* ([www.fluoridealert.org](http://www.fluoridealert.org))
- [9] "Analysis of Liver Disorder Using Data mining algorithms", Global Journal of computer science and Technology, 1.10 issue 14 (ver1.0) November 2010 page 48 to 52.
- [10] Peter Reutemann, Ian H. Witten, "The WEKA Data Mining Software: An Update - White paper", Pentaho Corporation. SIGKDD Explorations Volume 11, issue 1 pages 10 to 18.
- [11] T.Balasubramanian and R.Umarani "An Analysis on the Impact of Fluoride in Human Health (Dental) using Clustering Data mining Technique " Proceedings of International Conference on IEEE PRIME 2012, held at Periyar University, on Mar21-23.

### AUTHORS PROFILE



T. Balasubramanian (Corresponding author) received his M. Sc Computer Science Degree from Jamal Mohamed College, Trichy affiliated with Bharathidasan University and M. Phil Degree from Periyar University. Now pursuing his Part time Ph. D research in Bharathiar University, Coimbatore. Now he is working As

Asst. Professor, Department of Computer Science in Sri Vidya Mandir Arts and Science College, Uthangarai, Krishnagiri Dt. His research area is of Data Mining application Techniques. He has published 9 research papers in various National, International conferences and 6 paper in various international journals.

Dr. R. Uma Rani has completed her M.C.A. from NIT, Trichy in 1989.



She did her M.Phil from Mother Teresa University, Kodaikanal. She received here Ph.D from Periyar University, Salem in 2006. Her area of interest includes Information Security, Data mining and Mobile communications. She has published about 50 papers in National and International conferences. She is also working as Associate Professor in Department of Computer

Science, Sri Sarada College for women, Salem. She has published 35 papers in International and National journals