Clustering as a Data Mining Technique in Health Hazards of High levels of Fluoride in Potable Water

T..Balasubramanian
Department of Computer Science,
Sri Vidya Mandir Arts and Science college, Uthangarai(PO),
Krishnagiri(Dt), Tamilnadu, India.

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

Abstract— **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 underground 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 gives meaningful decision making to this socio-economic real world health hazard. [2]**

Keywords-Data mining, Fluoride affected people, Clustering, K-means, Skeletal.

I. Introduction

A. Data Mining

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[1], 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. [5][8].

B. 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. [10]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 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.

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 [9]. By analyzing all these we came to know that the increased fluoride level in ground water creates dental, skeletal and neuro problems. In this analysis we focus only on skeletal 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, Tamil Nadu Water And Drainage Board (TWAD) suggested that the level of fluoride content in drinking water should not exceed 1.5 mg/L.[7]

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

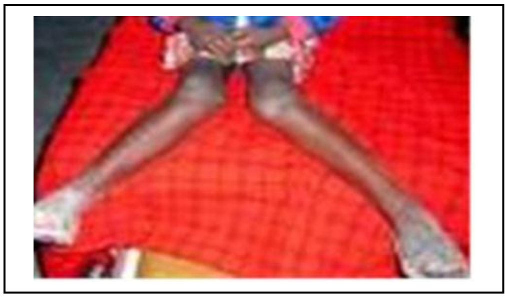


Figure 1. Skeketal Osteoroposis by Fluoride

TABLE 1. CLASSIFICATION OF SYMPTOMS OF DISEASES

Neck pain	**Joint pain**	**Body Pain**	**Foot Neck Pain**	**Class**
Low	Low	--	--	Mild Skeletal
Low	Low	Low	--	Mild Skeletal
Low	Low	Low	Low	Mile to Moderate Skeletal
Low	Medium	Low	Medium	Moderate Skeletal
Low	Medium	Low	High	Moderate Skeletal
Low	Medium	Medium	-Medium	Osteoporosis

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 questionnaires prepared in our mother tongue, Tamil since the people in and around the district are illiterate.

Total data collected from Villages and Panchayats

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

Based on the medical practitioner's advice, while classifying the data, the degrees of symptoms are placed in several compartments as follows:

Mild Skeletal Victims

Moderate Skeletal Victims

Osteoporosis Victims

With the following classification,

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

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

Those who are found with more than two medium symptoms are grouped as osteoporosis victims of skeletal disease.

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 [1]. 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.

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)[12].

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.

E. Clustering in WEKA

The classification is based on supervised algorithms. This algorithm is 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 [6].

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. [11]

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 skeletal fluorosis at different skeletal disease levels and to cluster the different water sources using by the people which are causes for skeletal fluorosis in krishnagiri district.

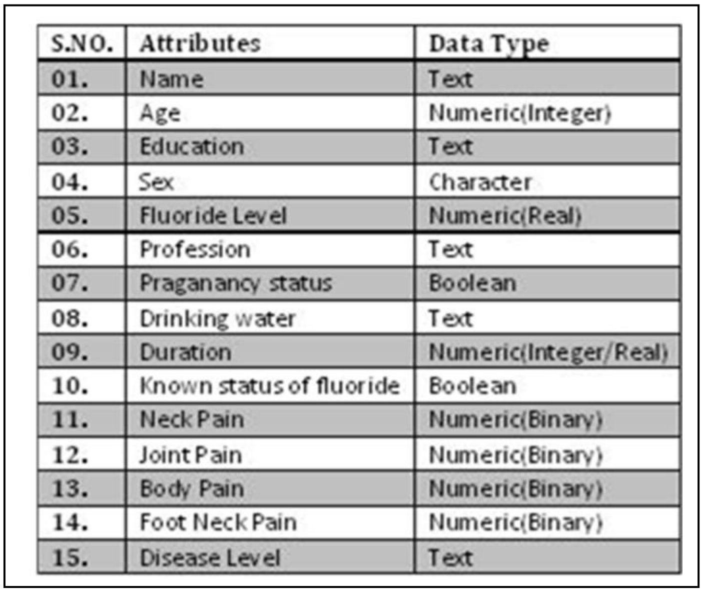
III. Discussion And Result

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.[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.

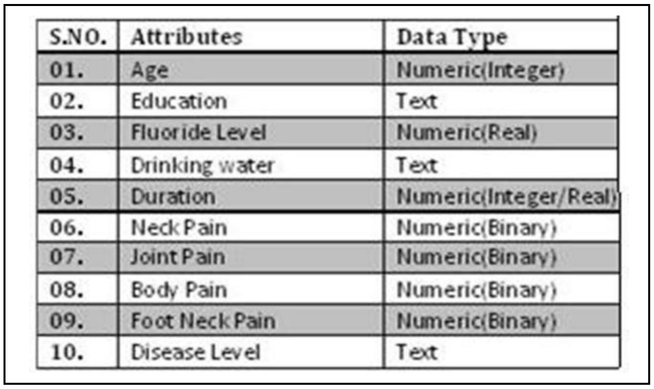
The records of data base consist of 15 attributes, from which 10 attributes were selected based on attribute selection in explorer mode of WEKA 3.6.4. (Fig 2)



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.	Praganancy status	Boolean
08.	Drinking water	Text
09.	Duration	Numeric(Integer/Real)
10.	Known status of fluoride	Boolean
11.	Neck Pain	Numeric(Binary)
12.	Joint Pain	Numeric(Binary)
13.	Body Pain	Numeric(Binary)
14.	Foot Neck Pain	Numeric(Binary)
15.	Disease Level	Text

TABLE 2. CLASSIFICATION OF ATTRIBUTES

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.	Age	Numeric(Integer)
02.	Education	Text
03.	Fluoride Level	Numeric(Real)
04.	Drinking water	Text
05.	Duration	Numeric(Integer/Real)
06.	Neck Pain	Numeric(Binary)
07.	Joint Pain	Numeric(Binary)
08.	Body Pain	Numeric(Binary)
09.	Foot Neck Pain	Numeric(Binary)
10.	Disease Level	Text

TABLE 3. SELECTED ATTRIBUTES FOR ANALYSIS

B. K-Means Metho

The k-Means algorithm takes the input parameter, k, and partitions a set of n objects into k clusters so that the resulting intracluster similarity is high but the intercluster similarity is low. Cluster similarity is measured in regard to the mean value of the objects in a cluster, which can be viewed 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] [3] [4]

$$E=\sum_{I=1}^{K}\sum_{P\in C_i}|p-m_i|^2$$

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 mi is the mean of cluster Ci . 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.

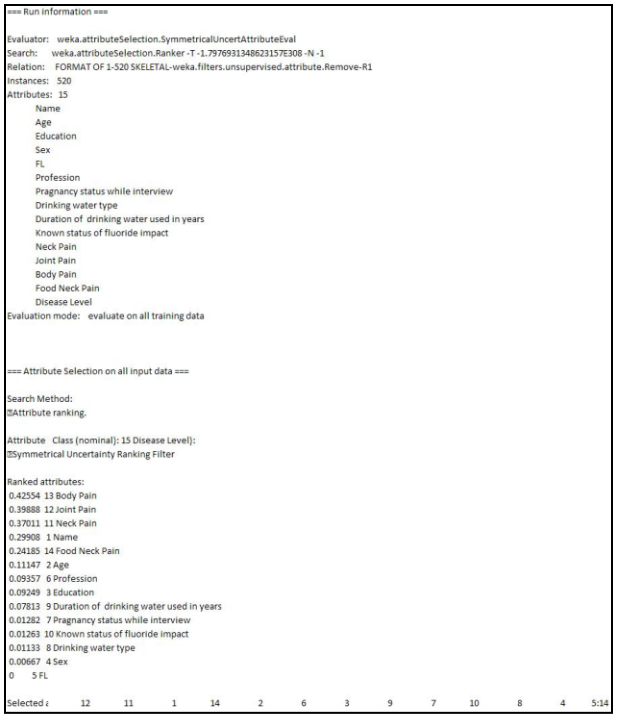
Method:

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

(2) (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;

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

(4) until no change;



```
=== Run information ===

Evaluator:   weka.attributeSelection.SymmetricalUncertAttributeEval
Search:      weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation:    FORMAT OF 1-520 SKELETAL-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
             Neck Pain
             Joint Pain
             Body Pain
             Food Neck Pain
             Disease Level
Evaluation mode:    evaluate on all training data



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

Search Method:
⊠Attribute ranking.

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

Ranked attributes:
 0.42554  13 Body Pain
 0.39888  12 Joint Pain
 0.37011  11 Neck Pain
 0.29908   1 Name
 0.24185  14 Food Neck Pain
 0.11147   2 Age
 0.09357   6 Profession
 0.09249   3 Education
 0.07813   9 Duration of  drinking water used in years
 0.01282   7 Pragnancy status while interview
 0.01263  10 Known status of fluoride impact
 0.01133   8 Drinking water type
 0.00667   4 Sex
 0        5 FL

Selected ε     12     11     1     14     2     6     3     9     7     10     8     4     5:14
```

Figure 2. Attribute selection in WEKA Explorer

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 which 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 is returned by the clustering process.

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.

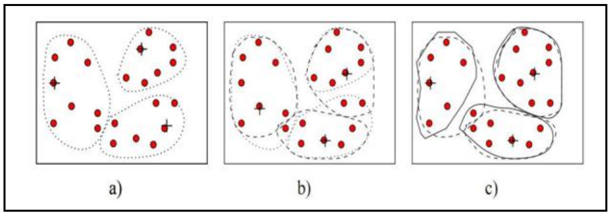
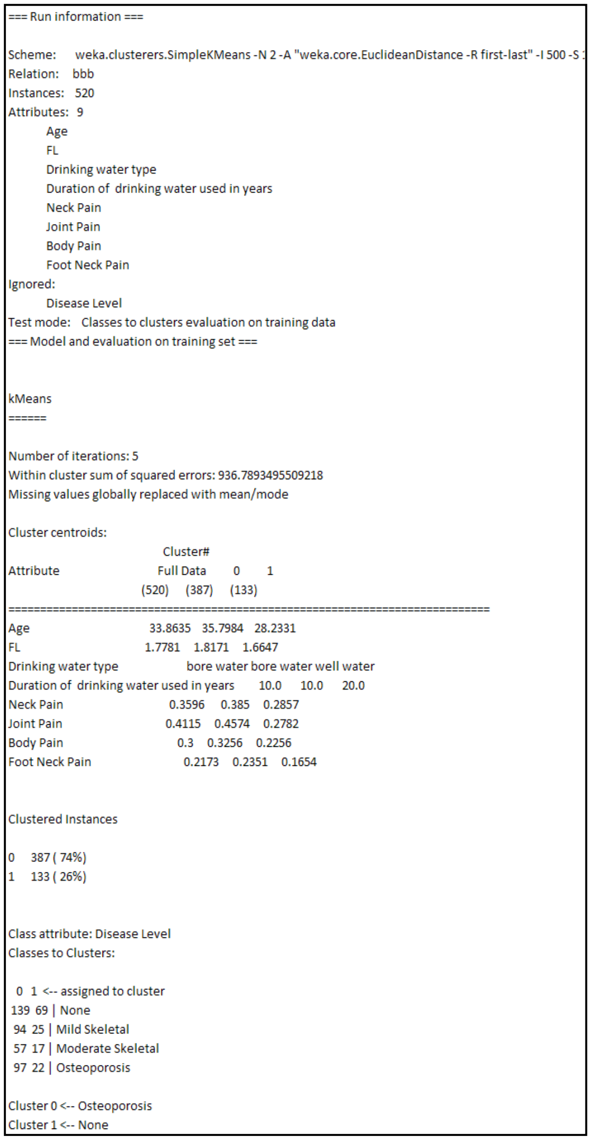


Figure 3. Clustering of a set of objects based on k-means method

It accepts the nominal data and binary sets. So our attributes selected in nominal and binary formats naturally. So there is no need of preprocessing for further process.

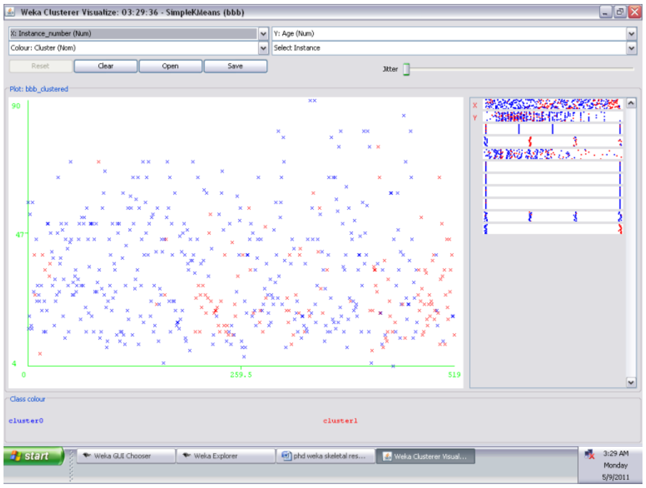
We have trained the training data by using the 10 Fold Cross Validated testing which used our trained data set as one third of the data for training and remaining for testing.

After training and testing this gives the following results.



```
=== Run information ===

Scheme:     weka.clusterers.SimpleKMeans -N 2 -A "weka.core.EuclideanDistance -R first-last" -I 500 -S
Relation:   bbb
Instances:  520
Attributes:  9
            Age
            FL
            Drinking water type
            Duration of  drinking water used in years
            Neck Pain
            Joint Pain
            Body Pain
            Foot Neck Pain
Ignored:
            Disease Level
Test mode:   Classes to clusters evaluation on training data
=== Model and evaluation on training set ===


kMeans
======

Number of iterations: 5
Within cluster sum of squared errors: 936.7893495509218
Missing values globally replaced with mean/mode

Cluster centroids:
                                   Cluster#
Attribute                         Full Data      0        1
                               (520)    (387)    (133)
===========================================================================
Age                              33.8635   35.7984   28.2331
FL                              1.7781    1.8171    1.6647
Drinking water type                    bore water bore water well water
Duration of  drinking water used in years      10.0      10.0      20.0
Neck Pain                           0.3596     0.385    0.2857
Joint Pain                         0.4115    0.4574    0.2782
Body Pain                             0.3    0.3256    0.2256
Foot Neck Pain                        0.2173    0.2351    0.1654


Clustered Instances

0     387 ( 74%)
1     133 ( 26%)


Class attribute: Disease Level
Classes to Clusters:

  0   1  <-- assigned to cluster
 139  69 | None
  94  25 | Mild Skeletal
  57  17 | Moderate Skeletal
  97  22 | Osteoporosis

Cluster 0 <-- Osteoporosis
Cluster 1 <-- None
```

Figure 4. K-means in weka based on diseases symptoms



Weka Clusterer Visualize: 03:29:36 - SimpleKMeans (bbb)

X: Instance_number (Num)

Y: Age (Num)

Colour: Cluster (Nom)

Select Instance

Plot: bbb_clustered

Figure 5. Disease symptoms in clusters of kmeans in weka

1) *Euclidean distance*

K-means cluster analysis supports various data types such as Quantitative, binary, nominal or ordinal, but do not support categorical data. Cluster analysis is based on measuring similarity between objects by computing the distance between each pair. There are a number of methods for computing distance in a multidimensional environment.

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

- Distance is always positive
- Distance from point x to itself is always zero
- 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.
- 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. 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 therefore essential that the attributaes are properly scaled.

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

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

2) *Clustering of Disease Symptoms*

The collected disease symptoms such as Neck pain, Joint pain, Body \pain, Foot Neck 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 skeletal osteoporosis. (Fig 3)

IV. Conclusion

Data mining applied in health care domain, by which the people get beneficial for their lives. As the analog of this research we found out that the meaningful hidden pattern from the real data set collected the people impacted in Krishnagiri district is 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 disease.

However the Primary Health hazards of fluoride are Dental and Bone diseases which disturbed their daily 000000 life. It is primary duty of the Government to providing good hygienic drinking water to the people and reduces 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. 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.

Reference

[1] Jain, M. Murty, and P. Flynn, “Data clustering: A review,” ACM Computing Surveys, vol. 31, no. 3, pp. 264–323, 1999.

[2] Jiawei Han and Micheline Kamber – Data mining concepts and Techniques. -Second Edition –Morgan Kaufmann Publishers

[3] Arun K.Pujari –Datamining Techniques – University Press.

[4] Introduction to Datamining with case studies - G.K.Gupta PHI.

[5] Berrry Mj Linoff G Data mining Techniques: for Marketing, Sales and Customer support USA.Wiley,1997.

[6] Weka3.6.4 data miner manual.

[7] Water Quality for Better Health – TWAD Released Water book.

[8] Data mining Learning models and Algorithms for medical applications – White paper - Plamena Andreeva, Maya Dimibova, Petra Radeve

[9] Elementary Fuzzy Matrix Theory and Fuzzy Models for Social Scientists - W.B.Vasantha Kandasamy (e-book :http:mit.iitm.ac.in)

[10] Professionals statement calling for an End to water Fluoridation – Conference Report (www.fluoridealert.org)

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

[12] The WEKA Data Mining Software: An Update, Peter Reutemann, Ian H. Witten, Pentaho Corporation, Department of Computer Science

Author's Profile

Dr.R.Uma Rani received her Ph.D., Degree from Periyar University, Salem in the year 2006. She is a rank holder in M.C.A., from NIT, Trichy. She has published around 40 papers in reputed journals and national and international conferences. She has received the best paper award from VIT, Vellore , Tamil Nadu in an international conference. She was the PI for MRP funded by UGC. She has acted as resource person in various national and international conferences. She is currently guiding 5 Ph.D., scholars. She has guided 20 M.Phil., scholars and currently guiding 4 M.Phil., Scholars. Her areas of interest include information security, data mining, fuzzy logic and mobile computing.

T.Balasubramanian received his M.Sc computer Science in Jamal Mohamed College, Trichy under Bharathidasan university and Mphil Degree from Periyar University. Now persuing his Ph.D research under Bharathiar University, Coimbatore. Doing research under health care domain in Datamining applications. He published 6 research papers in various National, International conferences and 4 papers in various International journals.