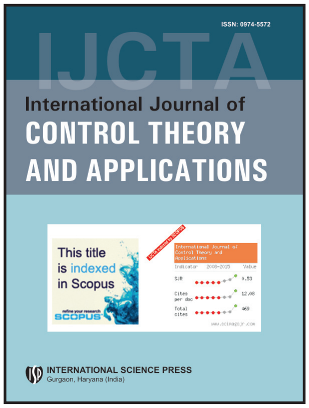
ISSN: 0974-5572

IJCTA

International Journal of
CONTROL THEORY
AND APPLICATIONS

This title is indexed in Scopus

INTERNATIONAL SCIENCE PRESS
Gurgaon, Haryana (India)

International Journal of Control Theory and Applications

ISSN : 0974-5572

© International Science Press

Volume 10 • Number 25 • 2017

Approaches to Partition Thyroid Data using Clustering Algorithms

R. Priya Anand[a], J. Jebathangam[b] and R. Bhuvana[c]

[a]Assosiate Professor, Department of MCA, Vels University, Chennai. Email: priyaa.research@gmail.com
[b]Assistant Professor, Department of Information Technology, Vels University, Chennai. Email: jeba81research@gmail.com
[c]Assistant professor, Department of CS, A.M. Jain College, Chennai. Email: bhuvanavr1981@yahoo.co.in

Abstract: Data Mining is defined as the process of extracting information from huge sets of data. In other words, we can say that data mining is extracting knowledge from data to obtain knowledge discovery, query language, classification and prediction, decision tree induction, cluster analysis, and how to mine the Web. Data Mining for Healthcare Management (DMHM) has been influential in detecting patterns of diagnosis, decisions and treatments in healthcare. Data mining has aided in many aspects of healthcare management including disease diagnosis, decision-making for treatments, medical fraud prevention and detection, fault detection of medical devices, healthcare quality improvement strategies. Healthcare system becomes very imperative to develop an automated tool that is capable of identifying and disseminating relevant healthcare information. This paper intends to provide the sample implementation of various clustering techniques used in thyroid data to improve the design of clustering methods for further enhancement.

Keywords: Data Mining, Health care Management, Clustering Methods, Thyroid Data.

1. INTRODUCTION

Characteristics of health care data, including issues of data availability and representation models, can make health care data mining applications challenging and interesting. The major challenges in health care data mining are huge volume of data, regular update, inconsistent data representation, poor integration, noise, number of variables and missing or incomplete data [1].

One popular approach that is frequently being applied in the healthcare industry and which is quite efficient in analyzing data is Data Mining. Today, Data Mining plays a vital role in widely used to understand marketing patterns, customer behavior, examine patient's data, and detect fraud, etc. Data Mining can be applied to different tasks associated to decision-making. Data Mining can potentially improve organizational processes and systems in hospitals, advance medical methods and therapies, providing better patient relationship management practices and getting better ways of working within the healthcare organization.

Dimensionality reduction and evaluation is an important task for decision making system. Dimensionality reduction improves the performance of the system and reduces the complexity of the system. Feature selection is broadly classified into two categories which are subset selection method and ranking methods. Subset selection method returns a subset of the original set of the features which are believed to be most important for classification. Ranking methods sort the attributes according to their role in the classification task. They use correlation based evaluation using rank method for dimensionality reduction [2].

Another factor is that the huge amounts of data generated by healthcare transactions are too complex and voluminous to be processed and analyzed by traditional methods. Data mining can improve decision-making by discovering patterns and trends in large amounts of complex data [3].

2. THYROID DISEASE

The thyroid is an endocrine which looks like butterfly and it is to be found in the lower front part of the neck. Thyroid is responsible for producing the thyroid hormones. These hormones are released into the blood and passed to tissues via blood. Thyroid hormones govern the body functions such as energy usage, controlling the body temperature and keeping organs working as believed [Wikipedia]. The two main thyroid hormones formed by the thyroid gland are triiodothyronine and levothyroxine, which are called T3 and T4 respectively. Thyroid hormone creation which is too less than the normal level is called hypothyroidism and too much thyroid hormone production (hyperthyroidism) generally causes thyroid function abnormalities. Decision of thyroid diagnosis is made after getting clinical tests, including the amount of thyroxin and triiodothyronine hormones and thyroid stimulating hormone (TSH) [Senthil Kumar]. Thyroid hormones are represented through TRH, TSH, T3, T4 as shown in the following diagram (Figure 1).

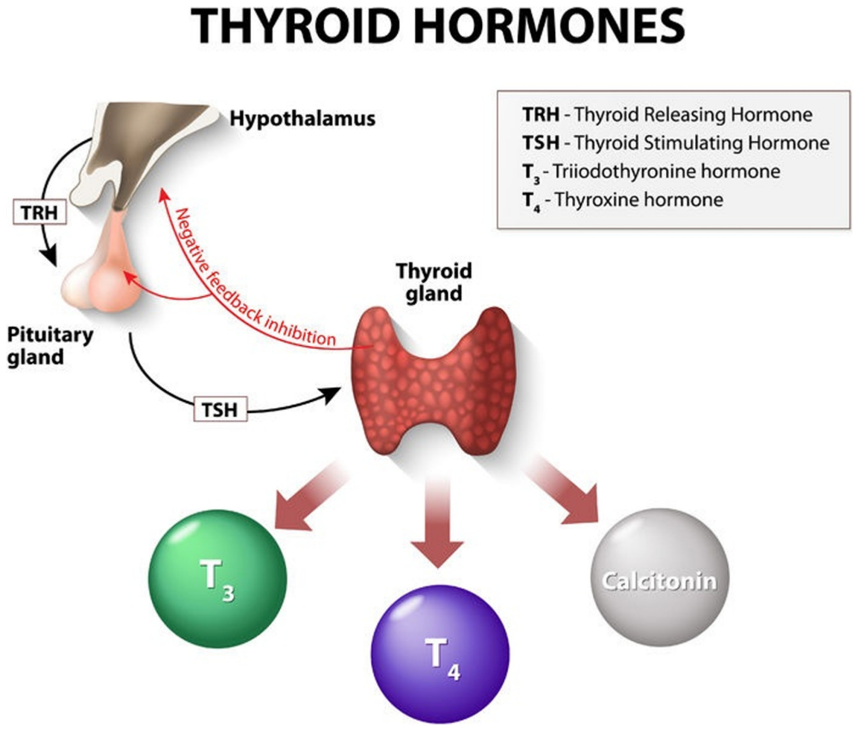
THYROID HORMONES
Hypothalamus
TRH - Thyroid Releasing Hormone
TSH - Thyroid Stimulating Hormone
T_3 - Triiodothyronine hormone
T_4 - Thyroxine hormone
TRH
Negative feedback inhibition
Thyroid gland
Pituitary gland
TSH
T_3
T_4
Calcitonin

Figure 1

3. CLUSTERING DATA FACES NEW CHALLENGES

- Information overload – Advances in medical equipments collective with high computing ability is increasing the amount of data collected and stored in health care industry.

- Knowledge discovery and retrieval of information from such huge databases is challenging and are prohibitively costliest one.
- Too many disease markers (attributes or dimensions) obtainable for decision making and are heterogeneous in nature.

The high awareness for quality care among public and increased life expectancy is increasing the demand for quality health services. But with hackneyed and tired physicians, stressful work conditions, etc., misdiagnosis and imprecise treatment solutions normally occur.

Clustering Algorithms

"Clustering" is a set of such clusters that usually contains all objects in the data set. Additionally, it also informs the relationship of the clusters with each other, for example a chain hierarchy of clusters put inside or embedded in each other.

- **Filtered Clustering:**

  A filter adds a new nominal attribute that represents the clusters assigned to every instance by specified clustering algorithm. Either the clustering algorithm is built with the first batch of data or ones specifications are serialized for clustered model file to be used instead.

- **Hierarchical clustering:**

  In data mining and statistics, hierarchical clustering which is referred as hierarchical cluster analysis or HCA, is a method of cluster analysis which build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types:

  1. **Agglomerative:** This is a "bottom up" approach: each study starts in its own cluster, and pairs of clusters are merged as one move up the hierarchy.
  2. **Divisive:** This is a "top down" approach: all comments start in one cluster, and splits are performed recursively as one move down the hierarchy.

  In general, the merges and splits are determined in a greedy manner. The outcome of hierarchical clustering is usually presented in the form of dendrogram.

- **Density based clustering:**

  To discover clusters with arbitrary shape, density based clustering style have been developed hence typically regard clusters as dense region of objects in the data space that are separated by regions of low density. The restrictions mentioned above can be overcome by using a new approach, which is based on density for deciding which cluster of each element will be in DBSCAN, which stands for density-based algorithm for discovering clusters in large spatial databases with noise.

- **EM clustering:**

  Expectation–Maximization (**EM**) **algorithm** is an iterative method for result maximum likelihood or maximum a posteriori (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables. The EM iteration alternates between performing an expectation (E) step, which creates a function for the expectation of the log-likelihood evaluated using the current estimate for the parameters, and a maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the E step. These parameter-estimates are then used to determine the distribution of the latent variables in the next E step.

4. WEKA TOOL

Weka is a set of open source ML algorithms applied for pre-processing, classifiers, clustering, and association rule. Weka is formed by researchers at the University of Waikato in New Zealand. It is a Java based tool used in the field of data mining. It uses flat text files to explain the data. It can work with a wide variety of data files including its own “arff” format and C4.5 file formats. The advantages of Weka are:

- Free availability & Portability.
- Fully implemented in the Java programming language and thus runs on almost any modern computing platforms (Windows, Mac OS X and Linux).
- Comprehensive collection of data preprocessing and modeling techniques.
- Supports standard data mining tasks: data preprocessing, clustering, classification, regression, visualization, and feature selection.
- Easy to use GUI and Provides access to SQL databases.
- Using Java Database Connectivity and can process the result returned by a database query.

5. RESULTS AND DISCUSSION

For this study, the data set used is hypothyroid.arff and is implemented in Weka tool 3.6. The attributes shortlisted are as listed below.

The selected attributes are:

1. Age
2. Sex
3. Thyroxin
4. Query on thyroxin
5. On antithyroid medication
6. Sick
7. Pregnant
8. Thyroid surgery
9. I131 treatment
10. Query hypothyroid
11. Query hyperthyroid
12. Lithium
13. Goiter
14. Tumor

6. PERFORMANCE ANALYSIS

Based on the implementation, the performance of each clustering algorithms were evaluated and is shown in the Table 1.

Table 1
Shows the Performance of Clustering Algorithm

Clustering algorithm	*Execution time-in seconds*
EM	190.8 Seconds
Filtered Clusterers	0.06 Seconds
Hierarchical Clusterer	42.81 Seconds
Make Density Based Clusterer	0.09 Seconds

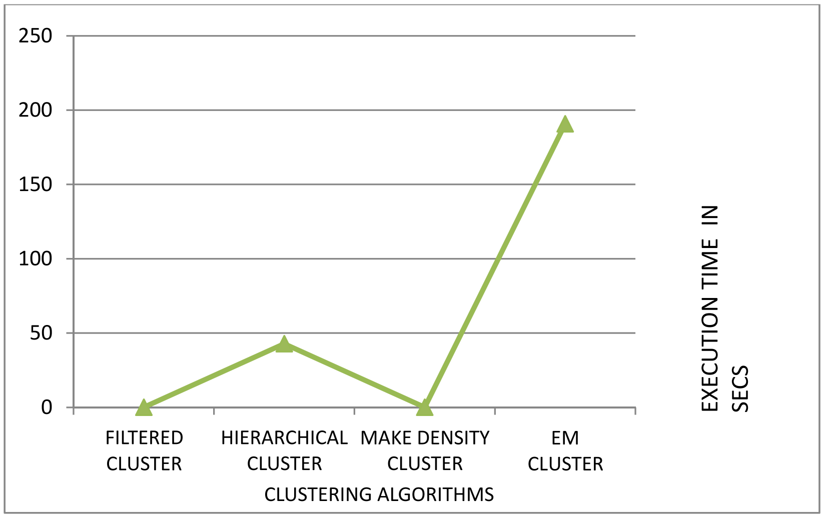

Figure 2: Clustering Algorithm in Execution Time

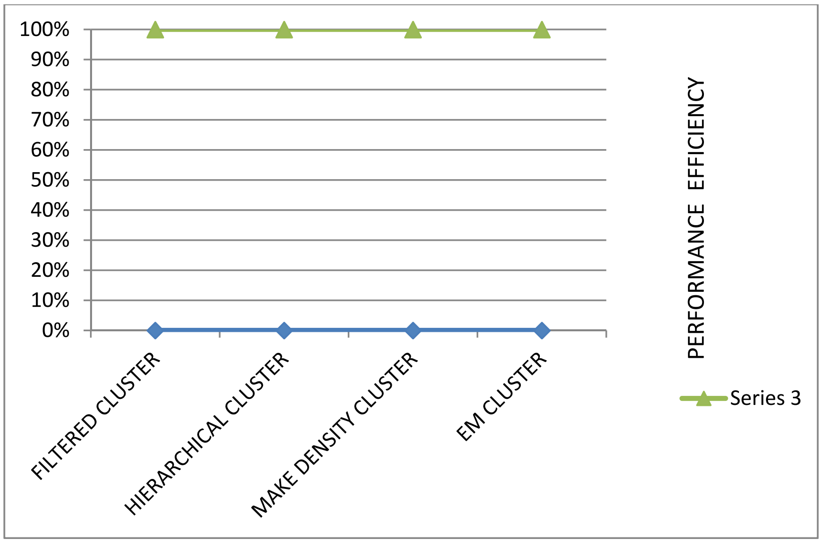

Figure 3: Clustering Algorithm in Performance Efficiency

7. CONCLUSION

Performance of the clustering method is measured by the percentage of the incorrectly classified instances. As the percentage of the incorrectly classified attribute is low, performance of the clustering is good. Filtered clustering gives better performance compared to Density based clustering. Also this algorithm's result is independent of number of cluster, while hierarchical cluster result is highly dependent on the number of cluster. In future, other parameters also can be taken into consideration which can facilitate doctors for efficient decision making process.

REFERENCES

[1] Fatemeh Hosseinkhah, Hassan Ashktorab, Ranjit Veen, M. Mehdi Owrang, “Challenges in Data Mining on Medical Databases”, IGI global, pp. 502.

[2] D. Senthilkumar, N. Sheelarani and S. Paulraj, “Classification of Multi-dimensional Thyroid Dataset Using Data Mining Techniques: Comparison Study”, Advances in Natural and Applied Sciences, 9(6) Special 2015, Pages: 24-28.

[3] P. Kalyani, ”Approaches to Partition Medical Data using Clustering Algorithms”, International Journal of Computer Applications (0975 – 8887) Volume 49– No. 23, July 2012.

[4] Vishal Shrivastava, 2Prem narayan Arya, ”A Study of Various Clustering Algorithms on Retail Sales Data”, Volume 1, No. 2, September – October 2012 International Journal of Computing, Communications and Networking Available Online at http://warse.org/pdfs/ijccn04122012.pdf.

[5] S. Revathi and T. Nalini, “Performance Comparison of Various Clustering Algorithm”, International Journal of Advanced Research in Computer Science and Software Engineering Research Paper Available online at: www.ijarcsse.com.

[6] Papillary, Follicular, Medullary, and Anaplastic Thyroid Cancers, “Incidence and Types of Thyroid Cancer”.

[7] N.S. Nithya, Dr. K. Duraiswamy, P. Gomathy, “A Survey on Clustering Techniques in Medical Diagnosis”, International Journal of Computer Science Trends and Technology (IJCST) – Volume 1, Issue 2, Nov-Dec 2013.