International Journal Of Mechanical Engineering And Information Technology
An Official Publication Of IGM Publication

Performance Analysis of K++ and Apriori Algorithm in Terms of their Effectiveness Against Various Diseases

Authors

Anisha Dahiya[1], Kamal Saluja[2]
Ganga Technical Campus, Soldha, Bahadurgadh
Email: *anniedahiya28@gmail.com kamalsalu@gmail.com*

Abstract
Heart disease is a general term that means that the heart is not working normally. Babies can be born with heart disease. This is called congenital heart disease. If people get heart disease later, it is called acquired heart disease. Most heart disease is acquired. Globally, heart diseases are the number one cause of death. About 80% of deaths occurred in low- and middle income countries. If current trends are allowed to continue, by 2030 an estimated 23.6 million people will die from cardiovascular disease (mainly from heart attacks and strokes). About 25 per cent of deaths in the age group of 25- 69 years occur because of heart diseases. If all age groups are included, heart diseases account for about 19 per cent of all deaths. It is the leading cause of death among males as well as females. The proportion of deaths caused by heart disease is the highest in south India (25 per cent) and lowest - 12 per cent - in the central region of India.[1] The healthcare industry gathers enormous amounts of heart disease data which, unfortunately, are not "mined" to discover hidden information for effective decision making. However, there is a lack of effective analysis tools to discover hidden relationships and trends in data. In this paper we are studying the techniques like K Means and Apriori for the elimination of manual tasks and easier extraction of data directly from electronic records, transferring onto secure electronic system of medical records which will save lives and decrease the cost of the healthcare services.
***Keywords**: k means Clustering algorithm, Apriori algorithm, Patient record, Heart disease*

INTRODUCTION

According to the World Health Organization heart disease is the first leading cause of death in high and low income countries and occurs almost equally in men and women [2]. By the year 2030, about 76% of the deaths in the world will be due to non-communicable diseases (NCDs) [3]. Cardiovascular diseases (CVDs), also on the rise, comprise a major portion of non communicable diseases. In 2010, of all projected worldwide deaths, 23 million are expected to be because of cardiovascular diseases. In fact, CVDs would be the single largest cause of death in the world accounting for more than a third of all deaths [4]. For CVDs specifically, in 2005, the age standardized mortality rate for developing nations like India, China, and Brazil was between 300-450 per 100,000, whereas it was around 100-200 per 100,000 for developed countries like USA and Japan [5]. According to a recent study by the

Registrar General of India (RGI) and the Indian Council of Medical Research (ICMR), about 25 percent of deaths in the age group of 25- 69 years occur because of heart diseases. The core functionalities of data mining are applying various methods and algorithms in order to discover and extract patterns of stored data [6]. From the last two decades data mining and knowledge discovery applications have got a rich focus due to its significance in decision making and it has become an essential component in various organizations. The field of data mining have been prospered and posed into new areas of human life with various integrations and advancements in the fields of Statistics, Databases, Machine Learning, Pattern Reorganization and healthcare.

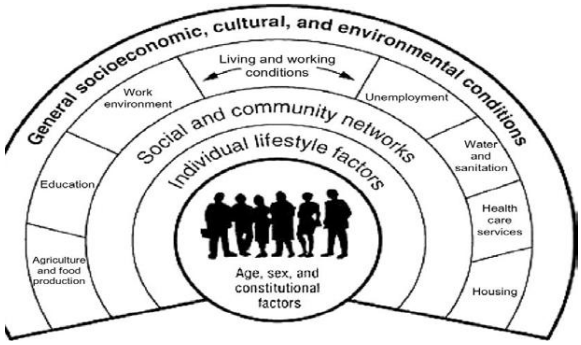

Figure 1. Factors Responsible for Diseases [7]

K-means algorithm

K-means algorithm is applied for dealing with medical database for clustering. To increase the efficiency of mining process, some pre-processing need to be done to the data. It is a process of semi-automatically analyzing large databases to find pattern that are valid, novel, useful and understandable. Clustering techniques have a wide use and importance nowadays. This importance tends to increase as the amount of data grows and the processing power of the computers increases. Clustering applications are used extensively in various fields such as artificial intelligence, pattern recognition, economics, ecology, psychiatry and marketing. The main purpose of clustering techniques is to partition ate a set of entities into different groups, called clusters. These groups may be consistent in terms of similarity of its members. As the name suggests, the representative-based clustering techniques uses some form of representation for each cluster. Thus, every group has a member that represents it. The motivation to use such clustering techniques is the fact that, besides reducing the cost of the algorithm, the use of representatives makes the process easier to understand. There are many decisions that have to be made in order to use the strategy of representative-based clustering.[8] For example, there is an obvious trade-off between the number of clusters and the internal cohesion of them. If there are few clusters, the internal cohesion tends to be small. Otherwise, a large number of clusters makes them very close, so that there is little difference between adjacent groups. Another decision is whether the clusters should be mutually exclusive or not, that is, if an entity can co-exist in more than one cluster at the same time K-Means is a simple learning algorithm for clustering analysis. The goal of K-Means algorithm is to find the best division of n entities in k groups, so that the total distance between the group's members and its corresponding centroid, representative of the group, is minimized. The aim of this study is to test the K-means algorithm which is one of data mining techniques at medical data bases. Cluster analysis is one of the major data analysis which helps to identify the natural grouping in a set of data item. Clustering is operation or process of partitioning a given set of objects into disjoint cluster.

There are several motivations for clustering are as under [9]-

1-A good clustering has predictive power. Since cluster labels are meaningful which lead to more

efficient description of given data, and will help to choose better actions.

2- Second, clusters can be a useful aid to communication because they allow lossy compression. In lossy compression, the aim is to convey in as few bits as possible a reasonable reproduction of a picture; one way to do this is to divide the image into N small in an alphabet of K image-templates, then we send a close fit to the image by sending the list of labels of the matching template.

3-Failures of the cluster may highlight interesting objects that deserve special attention.

4- Clustering algorithms may serve as models of learning process in neural systems.

K-means clustering is one of the most accepted and well known clustering techniques because of its simplicity and good behavior in many applications. For identifying the attributes that will be used in the clustering and these attributes are apparent clustering attributes for heart disease patients. Initial centroid selection is an important matter in K-means clustering and strongly affects its results[7] .The generation of initial centroids are based on actual data points using inlier method, outlier method, range method, random attribute method, and random row method .Here we deals with inlier method.

In generating the initial K centroids using the inlier method the following equations are used:

$Ci = Min (X) – i$ where $0 \leq i \leq k$ (1)

$Cj = Min (Y) – j$ where $0 \leq j \leq k$ (2)

Where C (ci, cj) is the initial centroid and min (X) and min (Y) is the minimum value of attribute X and Y respectively. K represents the number of clusters.

The k-means algorithm is a simple iterative method to partition the given dataset into User-specified number of clusters, k. This function depends on fuzzy logic in its work it assumes many centers, and then finds the smallest distance from the points to these centers, then rearrange these points as clusters. The distance of each cluster from fixed center is less than the distance from other centers. The algorithm operates on a set of d dimensional vectors. The algorithm is initialized by picking k points as the initial k cluster representative. Techniques for selecting these initial seeds include sampling at random from the dataset, setting them as the solution of clustering a small subset of the data or perturbing the global mean of the data.

Strength of k-Means cluster algorithm:

- Relatively efficient O (knt) where k-number of clusters, n-number of objects, t-number of iteration.
- Easy to implement and understand.
- Objects automatically assigns to clusters.
- Often terminate at local optimum

Limitations of k-Means cluster algorithm:

- User need to provide input k as number of clusters.(need to specify k)
- Different initial k objects may produce different clustering results.
- Unable to handle noisy data and outlier.
- Not suitable for non-convex shapes.
- Does not apply directly to categorical data[10]

The Apriori algorithm

Apriori is an algorithm for frequent item set mining and association rule learning over transactional databases. It proceeds by identifying the frequent individual items in the database and extending them to larger and larger item sets as long as those item sets appear sufficiently often in the database. The frequent item sets determined by Apriori can be used to determine association

rules which highlight general trends in the database: this has applications in domains such as market basket analysis. One of the most popular data mining approaches is to find frequent item sets from a transaction dataset and derive association rules. Finding frequent itemsets (item sets with frequency larger than or equal to a user specified minimum support) is not trivial because of its combinatorial explosion. Once frequent itemsets are obtained, it is straightforward to generate association rules with confidence larger than or equal to a user specified minimum confidence. Apriori is a seminal algorithm for finding frequent itemsets using candidate generation. It is characterized as a level-wise complete search algorithm using anti-monotonicity of itemsets, "if an itemset is not frequent, any of its superset is never frequent". By convention, Apriori assumes that items within a transaction or itemset are sorted in lexicographic order. Let the set of frequent itemsets of size k be Fk and their candidates be Ck. Apriori first scans the database and searches for frequent itemsets of size 1 by accumulating the count for each item and collecting those that satisfy the minimum support requirement. It then iterates on the following three steps and extracts all the frequent itemsets.[11]

1. Generate Ck+1, candidates of frequent itemsets of size k +1, from frequent itemsets of size k.
2. Scan the database and calculate the support of each candidate of frequent itemsets.
3. Add those itemsets that satisfies the minimum support requirement to Fk+1.

Function Apriori generates Ck+1 from Fk in the following two step process:

1. Join step: Generate RK+1, the initial candidates of frequent itemsets of size k + 1 by taking the union of the two frequent itemsets of size k, Pk and Qk that have the first k−1 elements in common.

RK+1 = Pk ∪ Qk = {item1 , item2, . . . , itemk−1, itemk , itemk' }
Pk = {item1, item2. . . itemk−1, itemk }
Qk = {item1, item2. . . itemk−1, itemk' }
Where, item1 < item2 < · … < itemk < itemk'_ .

2. Prune step: Check if all the itemsets of size k in Rk+1 are frequent and generate Ck+1 by removing those that do not pass this requirement from Rk+1. This is because any subset of size k of Ck+1 that is not frequent cannot be a subset of a frequent itemset of size k + 1. Function subset finds all the candidates of the frequent item sets included in transaction t. Apriori, then, calculates frequency only for those candidates generated this way by scanning the database. It is evident that Apriori scans the database at most kmax+1 times when the maximum size of frequent itemsets is set at kmax.

Apriori Property [12]

A frequent itemset can be a defined as a subset of frequent itemset i.e., if {PQ} is a frequent itemset, both {P} and {Q} should be a frequent itemset.
1. Iteratively discover frequent itemsets with cardinality from 1 to k (k-itemset).
2. Use the frequent itemsets to produce association rules. Join Step: Ck is generated by joining Lk-1with itself Prune Step: Any (k-1)-itemset that is not frequent cannot be a subset of a frequent k-itemset
Initialize: K: = 1, C1 = all the 1- item sets; read the database to count the support of C1 to determine L1. L1 := {frequent 1- item sets};
k: =2; //k represents the pass number//
While (Lk-1 ≠) do
begin
Ck: = gen_candidate_itemsets with the given Lk-1 Prune (Ck) for all candidates in Ck do count the number of transactions of at least k length that are common in each item Ck
Lk := All candidates in Ck with minimum support;

k := k + 1;
end

Key Attributes:

Patientid – Patient's identification number

Input attributes:

1. Sex (value 1: Male; value 0: Female)
2. Chest Pain Type (value 1: typical type 1 angina, value2: typical type angina, value 3: non-angina pain; value 4: asymptomatic)
3. Fasting Blood Sugar (value 1: > 120 mg/dl; value 0: < 120 mg/dl)
4. Restecg – resting electrographic results (value 0: normal; value 1: 1 having ST-T wave abnormality; value2: showing probable or definite left ventricular hypertrophy)
5. Slope – the slope of the peak exercise ST segment (value1: unsloping; value 2: flat; value 3: down sloping)
6. Exang – exercise induced angina (value 1: yes; value 0: no)
7. CA – number of major vessels colored by fluoroscopy (value 0 – 3)
8. Trest Blood Pressure (mm Hg on admission to the hospital)
9. Thal (value 3: normal; value 6: fixed defect; value7: reversible defect)
10. Cholesterol (mg/dl)
11. Oldpeak – ST depression induced by exercise relative to rest
12. Thalach – maximum heart rate achieved

Excluding these data records it also includes personal questions about the patient. They are smoking, overweight, alcoholic, daily fast food habit and going through regular exercise.

Limitations of Apriori Algorithm

- Needs several iterations of the data
- Uses a uniform minimum support threshold
- Difficulties to find rarely occuring events
- Alternative methods (other than appriori) can address this by using a non-uniform minimum support thresold
- Some competing alternative approaches focus on partition and sampling.

IMPLEMENTATION DETAILS

By using the techniques, Apriori and K-means the heart and various other disease prediction is performed on the patient databases. Apriori algorithm is used for finding the frequent itemsets from candidate itemset. And weights are assigned on the given itemsets to perform the prediction. The rule discovery of classification improves the predictive correctness of the classification system. In K-means clustering, clusters are generated from the datasets through inlier method. Then prediction is performed under basis of decision tree, these two methods are effective when evaluating the results and it can be used in real time.

PERFORMANCE EVALUATION

The effectiveness of models was tested. The purpose was to determine which gave the highest percentage of correct predictions for diagnosing patients disease. Apriori and K-means by undertaking patient records, After analyzing the result, and accuracy of two methods, It shows that K-Means had better accuracy when compared to Apriori.

CONCLUSION

The implementation paper provides a comparison study of two techniques; K-means clustering algorithm and Apriori algorithm over heart disease prediction system. Apriori algorithm Rule M uses Weighted Support and Confidence Framework to extract Association rule from data warehouse and takes a training data set and generates a small set of

rules to organize future data. The k-means clustering is the technique to cluster the attributes from the patient record. The decision tree with K-means clustering can enhance the classifier's performance in diagnosing heart disease. The initial centroid selection technique among the K-means clustering techniques is used here because it can provide better performance over heart disease prediction. The experiments shows that K-means algorithm makes the system more accurate and efficient when compared Apriori algorithm.

REFERENCES

1. World health organization report from http://www.who.int/mediacentre/factsheets/fs310.pdf.

2. Preventing Chronic Disease: A Vital Investment. World Health Organization Global Report.

3. Global Burden of Disease. 2004 update (2008). World Health Organization
4. Coronary Heart Diseases in India. Mark D Huffman. Center for Chronic Disease Control. http://sancd.org/uploads/pdf/factsheet_CHD.pdf
5. Fayadd, U., Piatesky -Shapiro, G., and Smyth, P. 1996. From Data Mining To Knowledge Discovery in Databases, AAAI Press / The MIT Press, Massachusetts Institute Of Technology. ISBN 0–26256097.

6. R. Kandwal, P. K. Garg and R. D. Garg, "Health GIS and HIV/AIDS studies: Perspective and retrospective", Journal of Biomedical Informatics, vol. 42, (2009), pp. 748-755.

7. K A Abdul Nazeer, S D Madhu Kumar,"Enhancing the K-means clustering algorithm by using O(nlog) heuristic method for findingbetter initial centroids", computer societyIEEE,nazeer@nitc.ac.in, madhu@nitc.ac.in,2011.

8. Cambridge University http://www.cambridge.org, 2003.

9. Huang, "Extensions to the k-Means Algorithms for Clustering Large Data Sets with Categorical Values, Data Ming and Knowledge Discovery, vol. 2, no. 3, pp. 283-304, 1998.

10. Shreve, H. Schneider, O. Soysal, "A methodology for comparing classification methods through the assessment of model stability and validity in variable selection", Decision Support Systems, Vol. 52, pp. 247–257, 2011.

11. Goswami D.N,Chaturvedi Anshu, Raghuvanshi C.S," *An Algorithm for Frequent Pattern Mining Based On Apriori*," (IJCSE) International Journal Vol. 02, No. 04, 2010, 942-947