Mining Health inspection report - A Graph based Approach

M.Jayasree[1], Dr.K.Venkataramana [2]
[1] *Student, Dept. of MCA, KMM Institute of Post Graduate Studies*
[2]*Associate Professor, Dept. of MCA, KMM Institute of Post Graduate Studies, Tirupati, A.P*

***Abstract*- The applications of the data mining in the different fields like e-business, commerce and trade has widely improved. The medical field also has lot of information but awareness is less. General health inspection is very important part of health care in many countries. Finding the persons at risk is important for providing the early warnings and takes the prevention to them. One of the major challenges that can be faced by learning classification model for risk prediction mainly depends on the unlabeled data that takes major portion of the collected dataset. Especially the unlabeled data in the dataset reveals about the people in the health examination whose health conditions can change greatly from healthy to too-ill. There is no proof for changing their health conditions. In this paper we propose a C 4.5 algorithm for risk predictions. C4.5 algorithm is used as the training algorithm to show rank of High risk cases with the decision tree. The health record dataset is clustered using the K-means clustering algorithm.**

***Index Terms*- Decision trees, Healthcare, Health Examination, prediction, unlabeled data.**

I. INTRODUCTION

Data mining is process of extracting hidden knowledge from large volumes of raw data. Data mining is used to discover knowledge out of data and presenting it in a form that is easily understood to humans. Data mining is one of the most important domains which help in management of healthcare data. It also helps to discover new trends from healthcare data collected from various hospitals. The data mining tools and techniques help in analyzing data collected from different hospitals and summarizing it into useful information. There are huge applications of data mining in healthcare sector like providing effective treatment, customer relationship management, detecting fraud etc.

HUGE amounts of Electronic Health Records (EHRs) collected over the years have provided a rich base for risk analysis and prediction [1] ,[2]. An EHR contains digitally stored healthcare information about an individual, such as observations, laboratory tests, diagnostic reports, medications, procedures, patient identifying information, and allergies [3] . A special type of EHR is the Health Examination Records (HER) from annual general health checkups. For example, governments such as Australia, U.K., and Taiwan [4] ,[5] , offer periodic geriatric health examinations as an integral part of their aged care programs. Since clinical care often has a specific problem in mind, at a point in time, only a limited and often small set of measures considered necessary are collected and stored in a person's EHR. By contrast, HERs are collected for regular surveillance and preventive purposes, covering a comprehensive set of general health measures, all collected at a point in time in a systematic way [6].

Identifying participants at risk based on their current and past HERs is important for early warning and preventive intervention. In this study we formulated the task of risk prediction as a multi-class classification problem using the Cause of Death(COD) information as labels, regarding the health-related death as the "highest risk". The goal of risk prediction is to effectively classify 1) whether a health examination participant is at risk, and if yes, 2) predict what the key associated disease category is. In other words, a good risk prediction model should be able to exclude low-risk situations and clearly identify the high-risk situations that are related to some specific diseases .A fundamental challenge is the large quantity of unlabeled data. For example, 92:6%of the 102,258 participants in our geriatric health examination dataset do not have a COD label. The semantics of such "alive" cases can vary from

generally healthy to seriously ill or anywhere in between. In other words, there is no ground truth available for the "healthy" cases.

We simply treat this set of alive cases as the negative class; it would be a highly noisy majority class. On the other hand, if we take this large alive set as genuinely unlabeled, as opposed to cases with known labels removed, it would become a multi-class learning problem with large unlabeled data. Most existing classification methods on healthcare data do not consider the issue of unlabeled data. They either have expert defined low-risk or control classes or simply treat non positive cases as negative [7], [8]. Methods that consider unlabeled data [9], [10], are generally based on Semi Supervised Learning (SSL) [11] that learns from both labeled and unlabeled data. Amongst these SSL methods, only handle large and genuinely unlabeled health data. However, unlike our scenario, both methods are designed for binary classification and have predefined negative cases. A closely related approach is Positive and Unlabeled (PU) learning [12], which can be seen as a special case of SSL with only positive labels available.

While the unlabeled set U in a PU learning problem is similar to our alive set, its existing applications in healthcare only address binary classification problem. Nguyen et al. introduced a multi-class extension called m PUL; however, their method used a combined set of negative and unlabeled example, while in our case negative example is not available. The other key challenge of HERs is heterogeneity. It demonstrates the health examination records of Participantp1in three non-consecutive years with test items in different categories (e.g., physical tests, mental tests, etc.) and abnormal results marked black. This example shows that1) a participant may have a sequence of irregularly time-stamped longitudinal records, each of which is likely to be sparse in terms of abnormal results, and 2) test items are naturally in categories, each conveying different semantics and possibly contributing differently in risk identification. Therefore this heterogeneity should be respected in the modeling.

II. RELATED WORK

There is plenty of literature that analyzes or predicts the risk of one single disease at a time. For example, Yeh et al. [1], Shivakumar and Alby [2], and Neuvirth et al. [3] focused on diabetes analysis. The models were built for predicting the cerebrovascular disease [1]. These predictions of single diseases are formulated into the binary classification problems. However, multiple-related diseases may appear simultaneously, where binary classification cannot deal with it effectively. In this work, we focus on formulating multi label classification to resolve the multi disease risk prediction based on physical examination records.

III. PROPOSED MODEL OR ALGORITHM

Decision trees are powerful and popular tools for classification and prediction. Decision trees produce rules, which can be inferred by humans and used in knowledge system such as database. C4.5 is an algorithm for building decision trees .It is an extension of ID3 algorithm and it was designed by Quinlan .It converts the trained trees (i.e. the output of the ID3 algorithm) into sets of if-then rules. It handles discrete and continuous attributes. C4.5 is one of widely-used learning algorithms.

C4.5 algorithm builds decision trees from a set of training data using the concept of information entropy. C4.5 is also known as a statistical classifier.

- Check for base cases.
- For each element x, discover the normalized information gain from dividing on x.
  - Let x_best be the element with the highest normalized information gain.
- Create a decision node that breaks on a best.
- Repeats on the sub lists obtained by dividing on x_best, and add those nodes as children of node.

A. K-Mean Clustering

Clustering is a technique in data mining to find interesting patterns in a given dataset .The k-means algorithm is an evolutionary algorithm that gains its name from its method of operation. The algorithm clusters information into k groups, where k is considered as an input parameter. It then assigns each information's to clusters based upon the observation's proximity to the mean of the cluster. The cluster's mean is then more computed and the process begins again. The k-means algorithm is one of the simplest clustering techniques and it is commonly used in medical imaging and related fields. .K-Means algorithm is a divisive, unordered

method of defining clusters. The phases convoluted in a k-means algorithm are given consequently:

- the algorithm arbitrarily selects k points as the initial cluster centers ("means")
- Each point in the dataset is assigned to the closed cluster, based upon the Euclidean distance between each point and each cluster center.
- Each cluster center is recomputed as the average of the points in that cluster.
- Steps 2 and 3 repeat until the clusters converge. Convergence may be explained differently depending upon the performance, but it regularly explains that either no observations change clusters when steps 2 and 3 are repeated or that the changes do not make a material difference in the definition of the clusters.

The clustering is performed on preprocessed data set using the K-means algorithm with the K values so as to extract relevant data to heart attack. K-Means clustering produces a definite number of separate, non-hierarchical clusters. K-Means algorithm is a disruptive, non-hierarchical method of defining clusters.

B. C4.5 algorithm

1. Let the set of training data be S. Put all of S in a single tree node.
2. If all instances in S are in same class, then stop.
3. Split the next node by selecting an attribute A, for which there is minimum Statistical Variance. Put the split point as the Statistical Mean of the current subset of data.
4. Stop if either of the following conditions is met, otherwise

Continue with step 3:

a) If this partition divides the data into subsets that belong to a single class and no other node needs splitting.

b) If there are no remaining attributes on which the sample may be further divided.

In conventional decision tree algorithms like C4.5, the splitting will be done based on the maximum information gain concept. But here the statistical variance is used, which is defined as follows: In general, the population variance of a finite population of size N is given by equation (8.1)

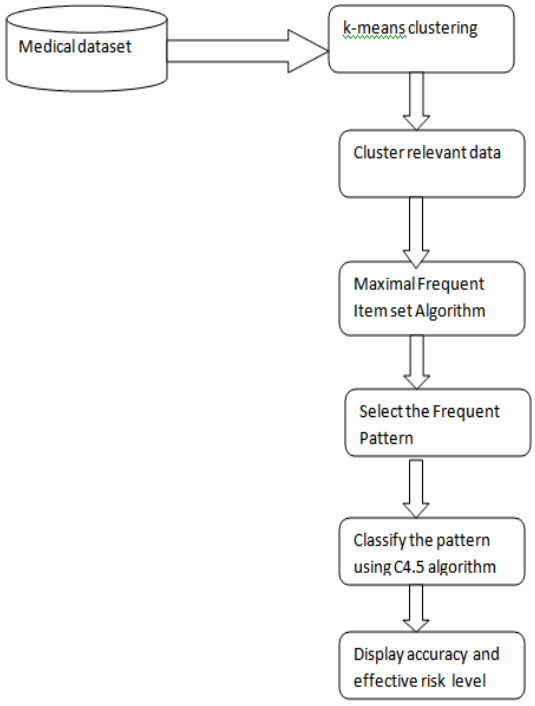
$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2 \qquad - (8.1)$$

Where μ is the population mean as given by equation (8.2):

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \qquad - (8.2)$$

Here the assumption is that, if a subset of the data is having low variance then there is a chance that they converge to a particular class in minimum number of iterations as there is minimum variation in the data for that attribute.

IV. SYSTEM ARCHITECTURE



V. CONCLUSION

Mining of the health data is a somewhat exigent due to its heterogeneity, intrinsic noise and especially it contains large amount of unlabeled data. In this paper we are proposing a new algorithm called c4.5 algorithms to face these challenges. The c4.5 algorithm based on decision tree technique and k-means clustering. In this paper we predicted the risks for the participants' based on their annual health examination.

REFERENCES

[1] M. F. Ghalwash, V. Radosavljevic, and Z. Obradovic, "Extraction of interpretable

multivariate patterns for early diagnostics," IEEE International Conference on Data Mining, pp. 201–210, 2013.

[2] T. Tran, D. Phung, W. Luo, and S. Venkatesh, "Stabilized sparse ordinal regression for medical risk stratification," Knowledge and Information Systems, pp. 1–28, Mar. 2014.

[3] M.Eichelberg,T.Aden,J.Riesmeier,A.Dogac,and G.B.Laleci,"A survey and analysis of Electronic Healthcare Record standards," ACM Computing Surveys, vol. 37, no. 4, pp. 277–315, 2005.

[4] C. Y. Wu, Y. C. Chou, N. Huang, Y. J. Chou, H. Y. Hu, and C. P. Li, "Cognitive impairment assessed at annual geriatric health examinations predicts mortality among the elderly," Preventive Medicine, vol. 67, pp. 28–34, 2014.

[5] "Health assessment for people aged 75 years and older,"http://www.health.gov.au/internet/main/publishing.nsf/ Content/mbsprimarycare mbsitem 75andolder, accessed: 201505-03.

[6] L. Krogsbøll, K. Jørgensen, C. Grønhøj Larsen, and P. Gøtzsche, "General health checks in adults for reducing morbidity and mortality from disease (Review)," Cochrane Database of Systematic Reviews, no. 10, 2012.

[7] G. J. Simon, P. J. Caraballo, T. M. Therneau, S. S. Cha, M. R. Castro, and P. W. Li, "Extending Association Rule Summarization Techniques to Assess Risk of Diabetes Mellitus," IEEE Transactions Knowledge and Data Engineering, vol. 27, no. 1, pp. 130–141, 2015.

[8] L. Chen, X. Li, S. Wang, H.-Y. Hu, N. Huang, Q. Z. Sheng, and M. Sharaf, "Mining Personal Health Index from Annual Geriatric Medical Examinations," in 2014 IEEE International Conference on Data Mining, 2014, pp. 761–766.

[9] B. Qian, X. Wang, N. Cao, H. Li, and Y.-G. Jiang, "A relative similarity based method for interactive patient risk prediction," Data Mining and Knowledge Discovery, vol. 29, no. 4, pp. 1070–1093, 2015.

[10] J. Kim and H. Shin, "Breast cancer survivability prediction using labeled,unlabeled,andpseudo-labeledpatientdata,"Journalofthe American Medical Informatics Association : JAMIA, vol. 20, no. 4, pp. 613–618, 2013.

[11] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using Gaussian fields and harmonic functions," International Conference on Machine Learning, vol. 20, no. 2, pp. 912–919, 2003.

[12] B. Liu, W. S. Lee, P. S. Yu, and X. Li, "Partially Supervised Classification of Text Documents," in International Conference on Machine Learning, 2002, pp. 387–394.