Outlier Detection in Cancer Infected Cells by Random Forest Approach and Distance-Based Outliers

S. Mohamad Haja Sherif[1] C. Imthyaz Sheriff[2]
[1]P.G Student [2]Assistant Professor
[1,2]Department of Computer Science & Engineering
[1,2]B.S.AbdurRahman University, Chennai

Abstract— In the recent past, large scale databases and files have grown beyond the capabilities and capacities of commercial database management systems. Parallel processing is very much essential to process a massive volume of data in a timely manner. Field of healthcare and medical industry is always in need for newer ways of analysing and making senses of the data available with them. Cancer related data analytics is always pushing it frontiers in deploying effective and efficient ways of analysing cancer related data. This paper deals with usage of random forest approach and distance based outlier's detection for cancer detection in large datasets.

Key words: Biomedical Text Mining Phases and Tasks, Cluster Analysis Based on Frequent Pattern

I. INTRODUCTION

Data Scientists will typically use wide range of technologies such as graphical analysis, Data Mining and Statistical according to the problem being tackled. It is necessary in the future analysis tools will most likely work seamlessly across technologies, unclear the underlying storage technology. In spite of avoiding the potential confusion between the physical representation of the and its inherent meaning (semantics) and data at rest or in-flight (syntax), for the remainder of this will use the terms 'weakly typed' and 'strongly typed' data rather than overloaded terms such as structured, unstructured or semi-structured. By strongly typed, to mean data that is tabular in nature and can be readily placed into a set of weakly typed data and relational tables refers to data where this is not the case. It helps to distinguish between the way data is encoded (e.g. XML) from the contents of the data (e.g. free text versus well-structured address data).

The objective of clustering is to perform grouping of input data into sets in such a way that a similaritymeasure is high for objects in the same cluster, and low where. "Clustering algorithms are attractive for the task of class identification The clustering algorithms developed for a large database can be divided into three prominent categories: solidity based, partitioning and hierarchical. clustering is the process of organizing data into groups within which the elements are homogeneous in some way . As an unsupervised learning technique mainly for discovering natural groups or underlying structure of a given dataset, clustering has been an active research subject in various fields including, pattern recognition, machine learning, image analysis statistical analysis and data mining. Big Data applies to information that can't be processed or analyzed using traditional tools or processes. Increasingly, organizations today are facing enormous number of Big Data challenges.

II. RELATED WORKS

A combination of AdaBoost and random forests algorithms for constructing a breast cancer survivability prediction model [1] . Random forests used as a weak learner during the boosting process to reduce over-fitting problems to improve stability and accuracy by ping sun et.al [2]. The creation of a health specific geo-demographic classification sys for a particular city .The proposed classification system is particularly designed for the public health domain by combining most reliable health data and other sources accounting for the main determinants of health.A novel random forest is used in clustering algorithm for generating clusters, which has several advantages over the k-means algorithm Ping sun et.al[3]. The investigation color classification based on random forest approach.Their random forest approach has also used IHLS color space for raw pixel based skin detection.it have evaluated random forest based skin detection and compared it to the SVM, AdaBoost, Bayesian network, Multilayer Perceptron, Naïve Bayes and RBF network[4] by Rehanullah Khan et.al. A hybrid of random forest and multivariate adaptive regression splines algorithms for building a breast cancer survivability prediction model.random forest approach is used to perform a preliminary screening of variables and to receive a important rank.The new dataset is extracted from initial WDBC dataset according to top-k important predictors and its input into the MARS procedure, inwhich it is responsible for building interpretable models for predicting breast cancer survivability by Dengju Yao et.al

III. BIOMEDICAL TEXT MINING PHASES AND TASKS

The goal of text mining is to derive implicit knowledge that hides in unstructured text and present it in an exact form.

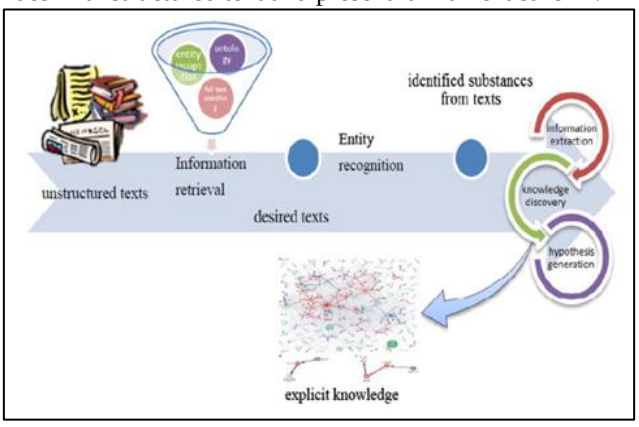

Fig. 1.1: Tasks in Bio medical Text Mining

This generally has four phases: hypothesis generation, information extraction, information retrieval and knowledge discovery. Information retrieval systems aim to get desired text on a certain topic; information extraction systems are

used to extract predefined types of information such as relation extraction; knowledge discovery systems help us to extract novel knowledge from text; hypothesis generation systems infer unknown biomedical facts based on text, as shown in figure 1.1. Thus, the common task of biomedical text-mining is to include named entity recognition, retrieval, knowledge discovery, relation extraction, information and hypothesis generation.

IV. DATA MINING AND BIO DATA ANALYSIS

A. Data Mining and Bio Data Analysis:

Listed below are the tasks by which Bio data can be analysed using the traditional tasks of Data Mining.

B. Data Scrubbing, Semantic Integration Along With Data Pre-Processing Of Heterogeneous Bio-Medical Databases:

In the existence of high variety of database in healthcare in which data scrubbing, semantic integration along with data pre-processing are done, such as genome databases and proteome databases, have become an important task for systematic and coordinated analysis of bio-medical databases. Data cleaning and data integration methods developed in data mining that will help the construction of data warehouses for bio-medical data analysis and the integration of bio-medical data

C. Homogeneous Search and Comparison in Biological Data:

One of the most important search problems in biological data analysis is similarity search and comparison among bio-structures and sequences. For example, gene sequences isolated from healthy and diseased tissues can be compared to identify critical differences among the two classes of genes. Initially retrieving the gene sequences from the two tissue classes, and frequently occurring patterns of each class are compared. Usually, sequences occur in the diseased samples than in the healthy samples more frequently which might indicate the genetic factors of the disease; those frequent occurring in the healthy samples might indicate mechanisms that protect the body from the disease.

D. Cluster Analysis Based On Frequent Pattern:

Most cluster analysis algorithms are based on either density or Euclidean distances. However, bio-data often consists of a lot of features which form a high dimension space, which is crucial to study differentials with scaling factors in multi-dimensional and shifting, cluster bio-data and discover pair-wise frequent patterns based on such frequent similar patterns.

E. Privacy on Preserving Mining Of Bio-Medical Data:

Although information exchange is important, hospitals and research institutes may still be reluctant to give out precious bio-medical data due to confidentiality and liability. Thus it is important to develop privacy preserving data mining methods, such as to maximally protect privacy while achieving effective data mining.

F. Data Visualization and Data Mining Of Visual:

Complex structures and sequencing patterns of genes and proteins are most effectively presented in chains, cubes, graphs and trees by various kinds of tools for visualization .Visualization and visual data mining which plays a major role in bio-medical data mining.

G. Knowledge Discovery:

Knowledge including facts, implicit or explicit, information, or descriptions, refers to the practical or theoretical understanding of a domain or a subject. Knowledge discovery is the creation of knowledge from large volumes of unstructured or structured data. The knowledge obtained may become additional value of data that can be used for further usage and discovery. Knowledge discovery is a very important part of data mining. Discovering of knowledge from biomedical text is a process with the aims to find answers for biomedical questions, such as identifying novel cancer diagnostic bio-markers or new drug targets. Knowledge discovery is compatible to integrate biomedical text with other several sources of data to generate a novel interpretive context. For example, through text mining technology together with microarray data, found out post-transcriptional control of ovarian processes as possible cause for the reproductive phenotypes and observed tumor. They also inferred that it was repetitive cycling that represented the actual link between ovarian tumor genesis and reproductive records

V. PROBLEM DEFINITION

A. Problem Definition:

This system aims to leverage the potential of data mining approaches to gain insights into the area of cancer diagnosis and prediction. Cancer cells in human body tend to behave abnormally right from their onset in comparison with normal, healthier cells . data mining methodologies like clustering ,outlier analysis is well suited for identification and further analysis of malicious cells .it is possible by analysing records of past and present patients .such records can be EHR(electronic Health Records),scans, x-rays ,prescription, lab reports etc.These records provide large amount of attributes. However not all attributes are related to cancer nor relevant for further analysis. Choosing the relevant attributes from health care data for study and analysis of cancer is quiet challenging .Ability to identify onset of cancer at an earlier stage will go a long way in effective treatment and prognosis of cancer .Any system which aims to solve these issues should analyse past cancer patient data , build prediction models based on this analysis and further refine the model based on testing with currently available patient data.

B. Existing System;

Outlier detection in the large set of data there are different algorithms to find the outliers. The approaches and techniques which are used before is based on depth ,density and distance etc. Data analytics in cancer predominantly focus on analysing past patient records to determine the relationship between various parameters such as id, clump thickness, cell size, cell shape, marginal adhesion, epithelial cell, bare nuclei ,bland chromatic and nucleoli the type of cancer .some systems focus on symptomatic analysis and prediction of cancer based on available health care data .outlier also deployed in detection of cancer infringement in human body through analysis of scanned images, oncology

based laboratory test. There is an inherent need to route the learning's of such systems into the actual diagnostic and treatment solutions in healthcare. This is quiet challenging in existing systems as they lack in proper feedback mechanisms which can quickly translate the knowledge attained out of analysis into actionable inputs and deliver to appropriate healthcare systems and health care professionals.

C. Proposed System:

The proposed system aims to evaluate the health care datasets (past and present) of cancer patients. Such evaluation will be directed towards the hypothesis that there is strong correlation between the columns such as marginal adhesion, epithelial cell, id, clump thickness, cell size, cell shape, bare nuclei, bland chromatic, nucleoli and cancer (type of cancer, stage of cancer) .clustering and outlier analysis will form a bed rock of this system as they are well suited in identifying abnormalmalicious cells from normal ones. Inorder to do effectively along with existing data sets have to be analysed in order to find the existing patterns, correlation among various data attributes and cancer .Knowledge of such patterns and correlation will be of immense value in improvising outlier detection algorithms .This system will be based on such pattern generation, clustering and outlier analysis.

D. Algorithms:

Algorithm for the distance based outlier detection Import the dataset file to continue the process by declaring an object for the dataset file.cluster centres and ids are created.The distances between objects and cluster centres are been calculated. In order to find the largest distances and the outliers as a result plot the outliers and cluster centres.

A popular method of finding these outliers is to use the distance to an example's k nearest neighbors as a measure of abnormal. However, algorithms for finding distance-based outliers have moderate scaling properties, making it knotty to apply them to large datasets typically available in security domains

1) *Procedure: Find Outliers*

Input:

- k, the number of nearest infected cell;
- n, the number of outliers to return;
- D, a set of examples in random order.

Output:

O, a set of outliers.

Let maxdist(x, Y) return the maximum distance between x and an example in Y .

Let Closest(x, Y , k) return the k closest examples in Y to x.

begin

1) c + 0 // set the cutoff for pruning to 0
2) O +0 // initialize to the empty set
3) while B + get-next-block(D) { // load a block of examples from D
4) Neighbors(b) +0 for all b in B
5) for each d in D {
6) for each b in B, b =←d {
7) if |Neighbors(b)| <k or distance(b,d) <maxdist(Neighbors(b),b) {
8) Neighbors(b) + Closest(b,Neighbors(b) U d, k)
9) if score(Neighbors(b),b) <c {
10) remove b from B
11) }}}}
12) O + Top(B U O,n) // keep only the top n outliers
13) c + min(score(o)) for all o in O // the cutoff is the score of the weakest outlier
14) }

return O

Random forest is very popular and powerful techniques in the pattern recognition and machine learning problem such as skewed and high- dimensional classification. RF is used to construct a collection of individual decision tree classifiers which utilized regression trees and the classification algorithms the regression tree algorithm is also called as CART which generates a binary tree in a rule based method via splitting the node by yes or no answers that are provided by predictors in which the process named as binary recursive partitioning . The aim of the CART is to maximize the difference of heterogeneity, however, in the real world data sets the over fitting problem that causes the classifier to have a high error of prediction in the unseen data set often encounters. To maximize the class clearly within the two resulting subsets at each step the rule is generated. in order to measures the impurity of a data partition or set of training instances given index is used in the CART

Fig. 2 Input: L: training sample k: number of input instance to be used at each of the tree G: number of generated trees in random forest.

```
1) E is empty
2) for b=1 to G
3) Lb = bootstrapSample(L)
4) Cb = BuildRandomTreeClassifiers(Lb,k)
5) N=N U{Cb}
6) next b
7) return N
```

Fig. 2: Random forests algorithm

Several research studies applied by the random forests algorithm to construct the decision trees. However, the random forests classifier is weak in high noise data that could cause an over fitting problem which reduced the accuracy of models in the unseen data set (test set). Moreover, it suffers from the difficulty of tree growing without pruning

VI. Conclusion

Random forest based approach in combination with distance based outlier detection was used for cancer cell detection and stage predication. These approaches have potential to be deployed in large scale databases without any perceivable loss of efficiency.

References

[1] JareeThongkam, GuandongXu and Yanchun Zhang on " AdaBoost Algorithm with Random Forests for Predicting Breast Cancer Survivability" of IEEE International Conference 2008.
[2] Ping Sun, Irena Begaj, Iris Fermin and Jim McManus on "Creating Health Typologies with

Random Forest Clustering" of IEEE International Conference 2010.

[3] Dengju Yao, Jing Yang, Xiaojuan Zhan on "Predicting Breast Cancer Survivability Using Random Forest and Multivariate Adaptive Regression Splines" of International Conference on Electronic & Mechanical Engineering and Information Technology 2011.

[4] Rehanullah Khan, Allan Hanbury, Julian Stoettingeron"SKIN DETECTION: A RANDOM FOREST APPROACH"of Proceedings of 2010 IEEE 17th International Conference on Image Processing September 26-29, 2010.

[5] Xindong Wu, Xingquan Zhu, Gong-Qing Wu and Wei Ding, "Data Mining with Big Data", of the IEEE Transactions on Knowledge and Data Engineering, Vol. 26,No. 1, January 2014.

[6] Kamalpreet Singh, Ravinder Kaur "Hadoop: Addressing Challenges of Big Data "proceedings of IEEE International Congress on Big Data , pp. 686-689 ,Feb 2014.

[7] Reza Mokhtari and Michael Stumm " BigKernel High Performance CPU-GPU Communication Pipelining for Big Data-style Applications" proceedings of IEEE 28th International Parallel & Distributed Processing Conference, pp. 819-828, May 2014.

[8] RamanaNagavelli, Dr.C.V.GuruRao "Degree of Disease Possibility (DDP): A mining based statistical measuring approach for disease prediction in health care data mining" proceeding on IEEE International Conference on Recent Advances and Innovations in Engineering, May 2014.

[9] Chin-Ho Lin, Liang-Cheng Huang, Chih-Ho Liu, Han-Fang Cheng, I-Jen Chiang "Temporal Event Tracing on Big Healthcare Data Analytics" of IEEE International Congress on Big Data,pp. 281-287, July 2014

[10] Carson Kai-Sang Leung,Richard Kyle MacKinnon ,Fan Jiang "Reducing the Search Space for Big Data Mining for Interesting Patterns from Uncertain Data " proceeding on IEEE International Congress on Big Data, pp. 315-322 ,July2014.

[11] HaoLin, Shuo Yang , Samuel P. Midkiff "RABID : A Distributed Parallel R for Large Datasets" proceedings of IEEE International Congress on Big Data ,pp. 725-732, October 2014.

[12] Victor Levy on "A Predictive Tool for Nonattendance at a Specialty Clinic An Application of Multivariate Probabilistic Big Data Analytics", of IEEE International Conference,Vol 27,October 2013.

[13] Aziz Nasridinov Young-Ho Park "Visual Analytics for Big Data using R " proceedings of IEEE Third International Conference on Cloud and Green Computing ,Pp. 564-567,October 2013 .

[14] Subham Khanna, Sonali Agarwal on "An Integrated Approach towards the prediction of Likelihood of Diabetes" in the International Conference on Machine Intelligence, Vol.62 ,Dec 2013.

[15] Jiaqi Zhao , Jie Tao , Lizhe Wang , Rajiv Ranjan , and Joanna Kołodziej "Using Traditional Data Analysis Algorithms to Detect Access Patterns for Big Data Processing " proceeding of IEEE International Conference on High Performance Computing and Communications , pp. 1097-1104 ,November 2013.

[16] Rezwan Ahmed, George Karypis"Algorithms for Mining the Evolution of Conserved Relational States in Dynamic Networks", Proceedings of of the IEEE International Conference on Data Mining,pp 1-10, December 2011

[17] Le Zhou ,Junjie Li ,ZhiyongZhong ,Joshua Zhexue Huang ,Jin Chang ,ShengzhongFeng "Balanced Parallel FP-Growth with MapReduce " proceeding on IEEE Conference, pp.243-246 , Nov 2010.

[18] AnjanaGosain, AmitKumar"Analysis of Health Care Data Using Different Data Mining Techniques", pp. 1-6, July 2009.

