JOURNAL OF PURE AND APPLIED MICROBIOLOGY, June 2015. Vol. **9**(2), p. 1395-1400

Classification of Breast Cancer Cells using Novel DPSC Algorithm

S. Gowri[1], G.S. Anandha Mala[2] and G. Mathivanan[3]

[1]Department of Information Technology, Sathyabama University, Chennai- 600119, India.
[2]Department of Computer Science and Engineering, Easwari Engineering College, Chennai-600089, India.
[3]Dept of Information Technology, Sathyabama University, Chennai- 600119, India.

(Received: 10 April 2015; accepted: 12 May 2015)

Cells are basically in the count of millions and trillions. On examination of any kind of disease, samples of blood are collected initially from the patients and are sent to test. During the test, the data is monitored and noted continuously, especially for diseases like cancer the cells which are been classified or distinguished based on their dimensions, mass, etc, were physical appearance matter as a significant feature for the classification. For the evaluation of our proposed algorithm and the existing clustering algorithm, we utilize the dataset of the breast cancer. The most dominant concept in the domain of Data mining is the Clustering technique, which has become the most significant entail in the formulation of structure to data. The complication of data reading from a substantially large data of any sector in the ongoing circumstances, which has become the fundamental entity. The main base of this clustering technique is its proficiently structured classification. Taking the factor of structure based classification of data into consideration, this paper presents the novelty in the process of clustering, which is not derived from any traditional techniques in clustering, rather it is idealized using the concept of the most popular search engine, the Google's search engine. Google uses their Crawler technique to collect data on net and index it into its storage space. Using this crawler a clustering technique was formulated for the purpose algorithm.

Keywords: Breast Cancer, Classification; Google's Crawler; Data Mining.

The day-to-day growth of data in all the sectors is becoming extremely immense as days pass by, storing of this data has also become a strenuous job for the data control. Future vision shows a sign of substandard way of handling data, worsening the storage space. As the matter of this begin a fact for all the sector, possible means of storage was formulated, which cannot be meant as a ineffective methodology but could be indicated as a traditional method. Well know clustering techniques which are applied over document clustering are k-means, bisecting k-means, agglomerative hierarchical and so on. Though they are the most qualified techniques formulated by our forefathers of data scientists, there aren't any other novel techniques introduced thereafter independently, only techniques which are hybridized using those traditional techniques are begin proposed.

K-means the most traditional technique for clustering analysis in data mining aims to partition n observations into k clusters in which each observation belongs to the cluster with the nearest mean, serving as a prototype of the cluster. This results in a partitioning of the data space into Voronoi cells. Hierarchical Clustering is also a traditional method which builds hierarchy of clusters which falls into two types: agglomerative and divisive. Where agglomerative approach is the "bottom-up" approach of clusters and divisive approach is the "top-down" approach.

* To whom all correspondence should be addressed.
E-mail: gowri.it@sathyabamauniversity.ac.in

The approach which is been proposed in this paper is the novel approach which is independent of all the traditional approaches, idealized from the Google's crawler approach who main aspects are the "finding" and "indexing". "Finding" process fetches the data wherever it goes and the "indexing" process indexes the found data. The same is replicated in the offline document clustering process where the Agent initialized for the dataset parses through all the data in the dataset and does the "Pick" and "Place" functionality. "Pick" collects the data at which it parses and "Place" keeps the data into its respective cluster group. This process take place by classifying all the data which are relevant to the set criteria like the dimensional criteria set for the breast cancer dataset, which used for this evaluation process.

The rest of the paper is organized as follows. Section II discusses the related works. In Section III, the concept of Google's Crawling technique is discussed in detail. Section IV discusses on the proposed methodology, which includes the description of the architecture and the proposed algorithm. Section V discusses the results and analyses for the performance of the proposed system. Section VI concludes and proposes the future work that can be carried out.

Related Works

Clustering the field of medical has become the most enthusiastic entity where clustering in such field as a lot of variants and attributes for formation of clusters take place on enormous basis. The Study of few proposed clustering algorithms for clustering of medical data are stated in this section.

An important application of microarray data in functional genomics is to classify samples according to their gene expression profiles such as to classify cancer versus normal samples or to classify different types or subtypes of cancer. One of the major tasks with gene expression data is to find co-regulated gene groups whose collective expression is strongly associated with sample categories. In this regard, a gene clustering algorithm is proposed by Maji and Das (Maji, P. and Das, C., 2012) to group genes from microarray data. It directly incorporates the information of sample categories in the grouping process for finding groups of co-regulated genes with strong association to the sample categories, yielding a supervised gene clustering algorithm. The average expression of the genes from each cluster acts as its representative. Some significant representatives are taken to form the reduced feature set to build the classifiers for cancer classification. The mutual information is used to compute both gene-gene redundancy and gene-class relevance. The performance of the proposed method, along with a comparison with existing methods, is studied on six cancer microarray data sets using the predictive accuracy of naive Bayes classifier, K-nearest neighbor rule, and support vector machine. An important finding is that the proposed algorithm is shown to be effective for identifying biologically significant gene clusters with excellent predictive capability.

Cancer subtype information is critically important for understanding tumor heterogeneity. Existing methods to identify cancer subtypes have primarily focused on utilizing generic clustering algorithms (such as hierarchical clustering) to identify subtypes based on gene expression data. The network-level interaction among genes, which is key to understanding the molecular perturbations in cancer, has been rarely considered during the clustering process. The motivation of Yiyi Liu et al (Yiyi Liu, Quanquan Gu, Jack P Hou, Jiawei Han and Jian Ma, 2014) was to develop a method that effectively incorporates molecular interaction networks into the clustering process to improve cancer subtype identification.

Coming to novelty in clustering algorithms independently, a few research works are been presented below which show the routine hybridization of traditional clustering algorithms and haven't shown a novel algorithm.

A hybrid unsupervised learning algorithm, articulated as rough-fuzzy c-means, is propounded by Pradipta and Sankar (Pradipta Maji and Sankar K. Pal., 2007). It consists of a sensible incorporation of the ethics of rough sets and fuzzy sets. While the idea of lower and upper approximations of rough sets deal with improbability, ambiguity, and incompleteness in class definition, the membership function of fuzzy sets facilitates proficient management of overlying divisions. The concept of crisp lower bound and fuzzy boundary of a class, introduced in rough-fuzzy c-means, facilitates proficient choice of

cluster models. Numerous quantitative indices are initiated based on rough sets for calculating the performance of the propound c-means algorithm. The efficacies of the algorithm, along with a comparison with other algorithms were demonstrated on a set of real life data sets.

A hybrid clustering algorithm based on K-mean and K-harmonic mean (KHM) was described by Ravi (Ravindra Jain, 2012). The proposed algorithm was tested on five different datasets. The research was focused on fast and accuracy in clustering. The performance was compared with the traditional K-means & KHM algorithm. The result obtained from proposed hybrid algorithm was found to be much better than the traditional algorithm.

Google's Crawler

"Crawler" is a generic term for any program. This Crawler is used mainly for the purpose if automation in discovering and scanning websites, this process is accomplished by following links from one webpage to another. The well known Google's Search engine's crawler is the Googlebot. (Sergey Brin and Lawrence Page, 1998), (http://en.wikipedia.org/wiki/Larry_Page)

This web crawling robot of Google finds and retrieves pages which are on the web (www) and hands them off to the Google indexer. A little spider scurrying across the strands of cyberspace could replicate the Google's Googlebot, but in reality Googlebot doesn't traverse the web at all. It functions more like a web browser in a PC. It sends request to the web servers for a web page after which it downloads the entire page, and hands it over to Google's indexer. The Figure 1 below shows the Google's Crawler Architecture Diagram.

Googlebot actually consists of many computers which keep requesting and fetching pages which are done in milliseconds of time unlike the PC's web browser. Googlebot can request thousands of different pages concurrently. Consequently, to avoid overwhelming of web servers, or crowding requests from human users, Googlebot deliberately makes requests of each individual web server slowly than its capability. As you can see in the Diagram the agents are present in fetcher, who would fetch the Web pages from the internet which is over the air. For the purpose of reference to the agent, the URL is tagged along with the agent in order to fetch the page and return the page.

The proposed system would comprise of the similar agent conceptualization which would be carried out in offline process for Clustering of documents which are given as the input from a large dataset.

Proposed Methodology

The Detail overview of the proposed Clustering Algorithm is given in this section. A Detailed description of the architecture and the algorithm formulated are stated.

Architecture Overview

To make over the idealization form the Google's Crawler process this propounded Clustering Technique brings in novelty by building an architecture which has the replication of the

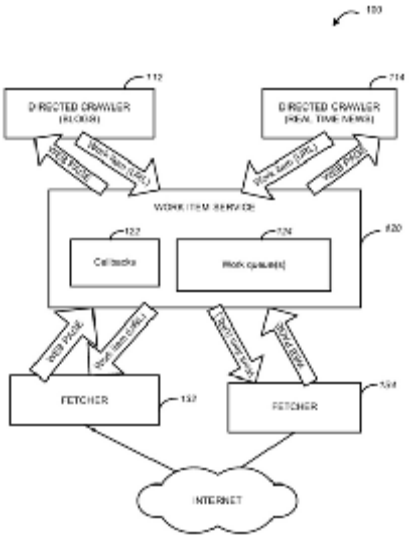

Fig. 1. Crawler Architecture

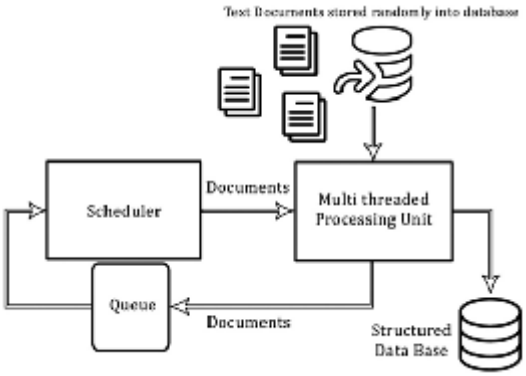

Fig. 2. Proposed Cluster Algorithm Architecture

processing of the Googlebot.

The Agent which is initialized carries a buffer bucket in which it collects the keyword and there weights on parsing through a file. The fetched content is then stored as one of the attributes in document form for the files. The Documents created for the files are then referred for the purpose of the agent to place the file in the rite cluster in which it fits to according to the content of the document.

Algorithm Description

The algorithm designed for the proposed system is formulated using 3 module functions, the three functions are Energy Checker, Agent's Action and Agent Creation, the description of each function is given below:

The Energy Checker function tests and de-energizes the energy of the agent which is crawling about the dataset as it forms a document of the file; this functionality is attained utilizing the looping concept, which would then call upon the action function of the agent according to the energy left.

The Agent's Action function has the pick and placing operation, where the agent fetches the words from the files and inserts into the document and once the agent gets to the "end of file" position the agent places the document into the respective path directory by checking the weight count of the word in the document formed.

The Agent Creation if the 1st function which has the process of creating the agent and initializing its buffer bucket, energy, max-point of parsing through files, frontier which is the end of file edge line and so on.

Algorithm

```
Initialize Agent:
Computation (initial_point, agent[x]) {
agent[x].energy •! Max_point;
agent[x].DSize •! MAX_BUFFER;
agent[x].frontier •! eof(initial_point);
insert(agent[x], population);
while (visited < MAX_PAGES) {
foreach parallel agent (population) {
Action (agent[x])
agent[x].energy •! agent[x].energy + newenergy;
if (agent[x].frontier > agent[x].fsize) {
dequeue_words(agent[x].frontier,
word_count(D[x]));
}
Energy (agent[x])
}
}
}

Agent's action:
Action (agent[x])
{
foreach active agent:
        D[x] •! Create_new_document()
Pick_keywordsFilepath(agent[x])
        E(agent[x]) •! E(agent[x]) – cost(D) +
relevancy(D)
        Energy (agent[x])
}

Pick_keywordsFilepath(agent[x])
{
word := pick_and_dequeue(agent[x].frontier);
D[x] := fetch(words);
}

Agent's energy E loop:
Energy (agent[x], population){
foreach alive agent:
        get_process (data[..], agent[x])
        Action (agent[x]) // fetch_words and
find_path
        update_energy_environment_state
(agent[x])
        if(E(agent[x]) > Max_point)
                child_agents •!
create.newagents(agent[i], agent[i+1])
                insert(child_agents,
population);
        elseif (E(agent[x])< 0)
                delete(agent[x])
replenish environmental resources
}
```

RESULTS AND ANALYSIS

For the results and decision we take an algorithm for the purpose of comparison with the proposed clustering technique. The algorithm chosen for comparison is the most famous algorithm which is the Hierarchical Clustering Algorithm.

The Dataset used for the purpose generating the clustering results is a medical which is a standard bench mark dataset generated by Dr. William H. Wolberg et al. in the year 2013, this

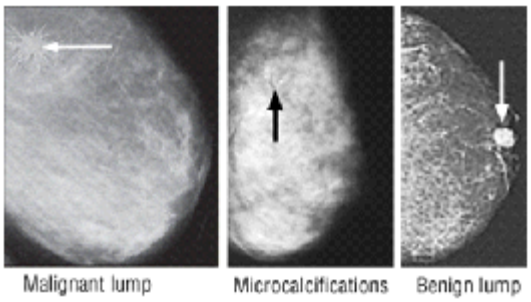

Fig. 3. Malignant and Benign Lump

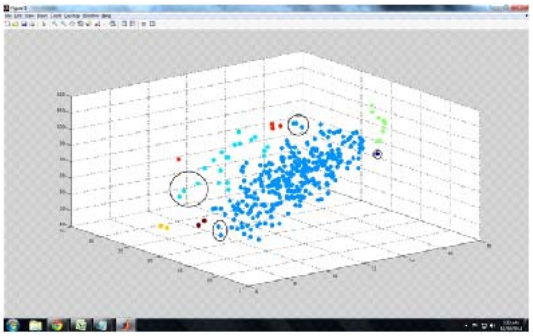

Fig. 4. Cluster Formation Graph of Hierarchical Clustering Algorithm

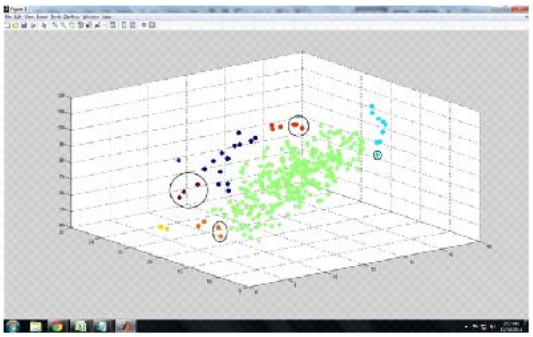

Fig. 5. Cluster Formation Graph of Proposed DPSC Algorithm

dataset comprises of the Breast Cancer Wisconsin (Diagnostic) Dataset. The total number of attributes and instances are 569 and 32 respectively. It is a multivariate characteristic type dataset with real attribute characters.

Attributes Description:

The two main Attributes of the Cancer Dataset are:

- ID number
- Diagnosis (M = malignant, B = benign)

The ten real-valued features computed for each nucleus of the Breast Cancer Cell are:

- Radius (The radius of the nucleus was calculated using the mean of distances from center to points on the perimeter)
- Texture (The Nucleus Texture is Computed with the help of the standard deviation of gray-scale values)
- Perimeter
- Area Accumulated
- Smoothness (local variation in radius lengths)
- Compactness $(perimeter^2 / area - 1.0)$
- Concavity (severity of concave portions of the contour)
- Concave points (number of concave portions of the contour)
- Symmetry
- Fractal Dimension ("coastline approximation" - 1)

Figure 3 shows the microscopic image of the Malignant and Benign Lump, were the structure of the formed lump vary in shape.

For the Experimental results of the proposed technique and the clustering technique used for comparison purpose is only the part of dataset which comprises of the attribute value obtain during the benign diagnosis which is the starting stage of the Breast Cancer. Here the clustering is based on the following few attributes which is the Radius, Area and Perimeter of the nucleus. Taking these attributes as the base parameters for the clustering techniques, the grouping is done in such a way that the nucleus of the similar dimensions would be grouped up into one cluster. Each Cluster which is formed by the Algorithms should have the nucleus of the same dimensional characteristics.

The Graphical representation of the clusters formed from the input Dataset by the hierarchical clustering and the proposed clustering technique is shown in below figure Figure 4 and Figure 5 respectively. The Output was generated using MATLAB tool, version R2009b. As show in the respective figures the clusters formed by the hierarchical clustering technique shows a little uneven in proportions of the cluster formation which even included the nuclei with dissimilar characteristics, while the proposed technique shows a quite closely precise cut of the cluster

formation where only one or two nucleolus which did not have a room to form a cluster had to combine with the cluster which are closest amongst the clusters availed.

CONCLUSION

As a known fact upon the current clustering techniques, they are mostly derived from the traditional clustering techniques which seem to be origin specific and derive more operational features from the existence rather than overall novelty. Consequently this novel clustering technique is been formulated, which is idealized using the Google's Crawling Technique for their search engine. It has been proven with the results shown above in the section 5 that the proposed Technique has resulted in a precise manner than the results of the Hierarchical Clustering Technique which is also a well know foremost technique amongst all the existing traditional techniques.

As a part of future enhancement the Clustering algorithm would be made more precise over the existence time of the Agent inside the dataset. This clustering process is then integrated into a system which we have developed for the forensic department to analyze suspicion in Digital textual data which gathered from various network channels like SMS, Mails, chat rooms, etc.

ACKNOWLEDGMENTS

We thank Sathyabama University for providing us with various resources and unconditional support for carrying out this work.

REFERENCES

1. Sergey Brin and Lawrence Page, "The anatomy of a large-scale hypertextual Web search engine", *Computer Networks and ISDN Systems*, 1998; **30**; 107- 117.
2. Pradipta Maji and Sankar K. Pal., "RFCM: A Hybrid Clustering Algorithm Using Rough and Fuzzy Sets", *Fundamenta Informaticae*, 2007; **80**, pp.475-496.
3. Maji, P. and Das, C.,"Relevant and Significant Supervised Gene Clusters for Microarray Cancer Classification", *IEEE Transactions on NanoBioscience*, **11(2)** , April 2012, pp. 161 - 168.

4] Ravindra Jain, "A Hybrid Clustering Algorithm for Data Mining", *CCSEA, SEA, CLOUD, DKMP, CS & IT 05*, 2012, pp.387-393.

5. Yiyi Liu, Quanquan Gu, Jack P Hou, Jiawei Han and Jian Ma, "A network-assisted co-clustering algorithm to discover cancer subtypes based on gene expression", *BMC Bioinformatics*, Feb 2014
6. uci,http://archive.ics.uci.edu/ml/machine-learning-databases/breast-cancer-wisconsin/wdbc.data
7. http://en.wikipedia.org/wiki/Larry_Page