A REVIEW ON TUMOR TYPES AND DATA MINING METHODS FOR CLUSTER CLASSIFICATION IN BIO-MEDICAL DOMAIN

S. Subash Chandra Bose[1], T. Christopher[2]

ABSTRACT

Emerging information about biological differences in tumors may different in outcomes for some people are different from others, uncertain incredible growth of tissue in living organs is difficult to identify types and vulnerability of tumor easily in high dimensional dataset.To perform effective and perfect diagnosis and treatment of cancer, identifying and classifying cancer types accurately is essential. This paper presents data mining algorithms and methods, to diagnosis tumor types by taking the features from the tumor classification on cancer data sets, and maximizes the associated data and minimizes the disassociation between the clusters; this survey provides a comprehensive overview for the classification of tumor.

Keywords: Data mining, tumor clustering, gene expression, cluster ensemble, cancer.

1. INTRODUCTION

Data mining can be performed to extracting patterns from knowledge implicitly stored in large datasets and focus on their feasibility, usefulness, effectiveness and scalability in the process of knowledge discovery, data's can be performed normally preprocessing through data cleaning, integration, data selection, and data transformation and prepared for mining task. Most popular data mining techniques has been developed and they using in data mining is classification, clustering and associations

Cluster analysis is one of the major data analysis method widely used for many practical applications in various domains, such as medical and biological applications .cluster is a processes of finding a group of object and the objects in the group will be similar or related to one another and different from or unrelated to the objects in the other groups, a good cluster will produce high quality clusters with high intra-cluster similarity and low inter-cluster similarity, quality of the cluster depends on the similarity measure used by the methods and its implementation and its ability to discover or all of hidden patterns.

When the dimensionality increases usually, only a small number of dimensions are relevant to the certain clusters, data in the irrelevant dimensions may produce noise, when dimensionality increases cluster analysis becomes meaningless. Hence, attribute reduction or dimensionality reduction is essential in data-preprocessing for cluster analysis of datasets having large number of features/attributes, reduced dimensionality of the data falls into two categories.

[1] Research Scholar, PG and Research Department of Computer Science, Government Arts College, Udumalpet, Tamilnadu, India.

[2] Assistant Professor, PG & Research Department of Computer Science, Government Arts College, Coimbatore, India.

Researchers provides different kinds of tumor clustering approaches based on single clustering algorithm to assign samples to the corresponding classes, such as self-organization map, hierarchical clustering model based clustering, non-negative matrix factorization analysis, distance-based clustering, evolutionary clustering, Traditional K-means algorithm for low dimensional data.

Dimensionality reduction, Feature Selection (FS) and Feature Reduction (FR) Feature selection algorithm aims at finding out subset of the most representative features according to some objective function discrete space. Feature extraction/Feature reduction an algorithm aims to extract features by projecting the original high-dimensional data into a lower-dimensional space through algebraic transformations, it finds an optimal solution to a problem, but computation complexity is more comparative to feature selection algorithm and proposed a method to apply PCA on original data set, dataset is transformed to possible uncorrelated variables, which reduced in size. Before applying PCA dataset needs to be normalized. The resulting dataset obtained from the application of PCA applied to K-means clustering algorithm, this framework is able to give better clustering with reduced complexity and also provides better accuracy and efficiency for high dimensional datasets.

2. Tumor Types

Different body tissue types give rise to different tumors, both benign and malignant. Table 1.Tumor List (Tissue Types) shows the different kinds of tumors each of the tissue types and it's vulnerable.

Two types of Cancers:

1) Benign: (usually curable)

Benign tumor generally harmless, it does not spread to other parts.

2) Malignant: (cancerous-growth)

It may spread to other parts of the body and they sometimes recur after they were removed.

3. Characteristics of Tumor

Tumor (neoplasm) is an abnormal new growth of cells usually grows rapidly than normal cells, and it will continue to grow if fail to treatment and damage adjacent structures.

Primary: Arise in the pancreas itself, a large gland behind the stomach located in the abdomen. It is a part of digestive system and produces important enzymes and hormones that help down to breakdown the food.

Metastatic: Cancers arise in the other organs and later spread to the pancreas.

Table 1.Tumor List (Tissue Types)

Tissue	Benign Tumors	Malignant Tumors
Adult fibrous tissue	Fibroma	Fibrosarcoma
Embryonic (myxomatous) fibrous tissue	Myxoma	Myxosarcoma
Fat	Lipoma	Liposarcoma
Cartilage	Chondroma	Chondrosarcoma
Bone	Osteoma	Osteosarcoma
Notochord	—	Chordoma
Connective tissue, probably fibrous	Fibrous histiocytoma	Malignant fibrous histiocytoma
Blood vessels	Hemangioma, hemangiopericytoma	Hemangiosarcoma, angiosarcoma
Lymph vessels	Lymphangioma	Lymphangiosarcoma
Mesothelium	—	Mesothelioma
Hematopoietic cells	"Preleukemias", "myeloproliferative disorders"	Leukemia, of various types; aleukemic leukemia
Lymphoid tissue	Plasmacytosis	Plasmacytoma; multiple myeloma; Hodgkin lymphoma and Non-Hodgkin lymphoma
Smooth muscle	Leiomyoma	Leiomyosarcoma
Striated muscle	Rhabdomyoma	Rhabdomyosarcoma
Lymphoid tissue	Plasmacytosis	Plasmacytoma; multiple myeloma; Hodgkin lymphoma and Non-Hodgkin lymphoma
Smooth muscle	Leiomyoma	Leiomyosarcoma
Striated muscle	Rhabdomyoma	Rhabdomyosarcoma
Stratified squamous	Papilloma Seborrheic keratosis and some skin adnexal tumors	Squamous cell carcinoma; epidermoid carcinoma and some malignant skin adnexal tumors
Glandular epithelium 1. Liver 2. Kidney 3. Bile duct	Adenoma Hepatic adenoma Renal tubular adenoma Bile duct adenoma	Adenocarcinoma Hepatoma: hepatocellular carcinoma Renal cell carcinoma; hypernephroma Cholangiocarcinoma
Transitional epithelium	Transitional cell papilloma	Transitional cell carcinoma
Placenta	Hydatidiform mole	Choriocarcinoma
Testis	—	Seminoma; embryonal cell carcinoma

Glial cells (of several types)	—	Glioma, grades I-III, anaplastic; glioblastoma multiforme (grade IV)
Nerve cells	— — Ganglioneuroma	Neuroblastoma Medulloblastoma —
Meninges	Meningioma	Malignant meningioma
Nerve sheath	Schwannoma, neurilemmoma Neurofibroma	Malignant meningioma Malignant schwannoma Neurofibrosarcoma
Pituitary	Basophilic adenoma Eosinophilic adenoma Chromophobe adenoma	— — —
Parathyroid	Parathyroid adenoma	Parathyroid carcinoma
Thyroid (C cells)	C cell hyperplasia	Medullary carcinoma of thyroid
Bronchial lining (Kultschitzky cells)	—	Bronchial carcinooid; oat cell carcinoma
Adrenalmedulla Pheochromocytoma	Pheochromocytoma	Malignant Pheochromocytoma
Pancreas	Islet cell adenoma; Insulinoma; gastrinoma	Islet cell carcinoma
Stomach and intestines	Carcinoid	Malignant carcinoid
Carotid body and chemo-receptor system	Chemodectoma; paraganglioma	Malignantcarcinoid Malignant paraganglioma
Pigment-producing cells in skin, eyes, and occasional other sites	Nevus	Melanoma
Schwann cells of peripheral nervous system	Schwannoma, or neurilemmoma	Malignant schwannoma
Merkel cells in squamous epithelium (unknown function)	—	Merkel cell neoplasm (similar to oat cell)
Breast	Fibro adenoma	Cyst sarcoma phylloides
Renal anlage	—	Wilms tumor

4. Related Research Work

In [1] general PAM (Partitioning Around Medoids) algorithm selects an initial center medoids and replaced non selected medoids in data set, and it improves sum of dissimilarities of data points of nearest medoids.PAM is powerful and robust than K-means(centroids). Fuzzy c-means (FCM) is also a clustering based outlier detection technique, for high dimensional data reduction Expectation-Maximation (EM) algorithm is used. Multiple runs of clusters performed with agglomerative clustering algorithm and the results are aggregated, results produced by ensemble approach is better than single cluster algorithms. Final clustering obtains by re-clustering the consensus matrix and spectral clustering algorithm chosen consensus function and (Spectral Clustering) is applied to the components to obtain final results.PAM is similar to K-Means, both are partition algorithms, both break the dataset into groups(clusters).both work to minimize the dissimilarity.

In [2] K-means clustering algorithm used, often does not work well or high dimensional data, to improve efficiency apply Principal Component Analysis (PCA) on original dataset to get reduced dataset contain possible uncorrelated variables, and reduced dataset is applied to K-means algorithm to determine precise no of clusters, and additionally they used Euclidian distance to maximum among all the data objects to make algorithm more effective and efficient.

In [3] Shieng, Zhang Changshui, Zang Xuegong present Principal component analysis (PCA), and Fisher analysis (FA) and another name is linear Fisher discriminant analysis. PCA is used to reduce the dimensional of the gene expression data through orthogonal transformation, and FA is used to handle the interval-scaled attributes.PCA is helpful in eliminate the correlations and remove the noise.

In [4] Knowledge based cluster ensemble approach (KCE) Performance of KCE is use in the several data set in the UCI repository. And it's compared with single cluster approaches K-means, Spectral Clustering (SC) Hierarchical Clustering, Self-Organization Map (SOM), and Partition around medoids (PAM).Single clustering algorithms perform average accuracy value, while knowledge based cluster approach improving robustness and stability of single clustering algorithms and KCE increasing the accuracy the approach.

In [5] Random double clustering based framework (RDCCE), selects a basic clustering algorithm, performing cluster on future dimension and generated subset to the center of clusters then new dataset is generated. Next RDCCE adopts selected clustering algorithm to partition the new sample into several groups to obtain clustering solution.

In [6] Consensus matrix is constructed with set of clustering solutions, finally RDCCE uses normalized cut algorithm to obtain final result. Mostly gene selection removes large number of

immaterial gene and improves the classification accuracy, and most of cancer classification problem are derived from the biological data sets.

In [7] Chun-Hou Zheng,De-Shuang Huang, Lei Zhang,and Xiang-Zhen Kong, presents Independent Component Analysis (ICA) for dimension reduction and reduces the noice, and it removes linear correlations as well as higher order dependencies in dataset, It aims to transformed coefficients mutually independent, and several algorithm has been proposed to implement ICA, such as FastICA and JADE Whereas PCA (principal Component Analysis) developed for separation of independent sources from their linear mixtures, it used to decorrelate the gene expression dataset. Clustering with Nonnegative Matrix Factorization (NMF) algorithm reduces the dimensionality of the gene dataset and it is efficient method to identify distinct molecular pattern.

In [8] Zhiwen Yu and Hantao Chen have analyzed three semi-supervised clustering ensemble frameworks. Feature based semi-supervised clustering ensemble framework (S-SSCE), double selection based semi-supervised clustering ensemble framework DS-SSCE and modified double selection semi-supervised clustering ensemble framework MDS-SSCE to perform tumor clustering on bio-molecular data [8]. This approach applied to perform tumor clustering from the bio-molecular data under the clustering ensemble framework and consider multiple clustering solution selection strategies at the same time.Parwise constraints based K-means clustering algorithm (PC-Means) is adopted to estimate cancer samples. Feature selection for low dimensional spaces.

Fuzzy rough set theory they introduced gain ratio and proposed to an attribute selection algorithm based on gain ratio in fuzzy rough set theory [9], and it used in tumor classification problem and experiments on tumor data sets in gene expression are conducted to evaluate the reduction and classification results compare to fuzzy rough model based on gain and crisp rough set methods.

In [10] Zhiwen Yu et al developed an Fuzzy cluster ensemble framework is done in Bio-molecular Data.HFCEF-first applies the Affinity Propagation (AP) algorithm to perform clustering on the sample dimension, it randomly selects one sample from each cluster which is served as the base sample, and obtains a set of basic samples. HFCEF repeats the above process B times, and generates B base sample sets. Here fuzzy membership function is adopted to capture the relationship between the samples in the original dataset P and the base samples and obtain a set o fuzzy matrix. Finally consensus unction is designed to summarize the fuzzy matrices and obtain the final results. Normalized cut algorithm (Ncut), used to maximize the association within the cluster and minimize the disassociation between the clusters.

In [11] Matlab function is capable of counting and assigning the pixels in the each frame to the

respective, ROI (region of interest), colored pixels in the frames assigned as true positive (TP) from false positive and noncolored pixels in frames assigned as true negatives (TN) or false negatives (FN).

In [13] and [14] various classifying methods are carried out by researchers for tumor classification, finally they choose most efficient intelligent classification technique to recognize normal and abnormal brain images. Support vector machines (SVM) is used on high dimensional histograms to improve image quality, remove noise, here PCA is used to reduce the future set. And SVM mainly used two classifiers, linear non-linear boundaries.

In [15] they used fuzzy c-means for; Brain tumor segmentation mainly used using two image processing methods. Threshold-based and Region-based(region-based mostly used in two dimensional image segmentation),FCM divides the group of data into two or more clusters, on the basis of the distance between the cluster and the data point, and it gives best result for overlapped data set and its best than k-means algorithm.

In [16] Markov clustering algorithm used for clustering, method is fast and scalable unsupervised clustering method for graphs, and it is a hierarchical netlist.The input of the clustering process is a netlist of logic blocks and their inter connections.

[17] Uses graph clustering algorithms and embed the input graph into Euclidean space by eigendecomposing matrix and then cluster the embedding using a geometric clustering algorithm.

5. Comparision of Single Clustering Over Multiple Clustering on Review

Comparing [2] & [4] Previously many single clustering approaches are used for tumor clustering in many practical applications in emerging areas like bio medical domain. Feature selection PAM (Partitioning Around Medoids), fuzzy c-means (FCM), expectation-Maximation (EM), Principal component analysis(PCA), Knowledge based cluster ensemble approach (KCE) and double clustering based framework (RDCCE) have used for better results, when the dimensionality increases again computational complexity is very high for large datasets, feature reduction and feature selection algorithms is applied over the dataset and provides the subeset for the original data to reduce the disassociation for feature selection and reduce distance calculation between data points to reduce the time complexity, for calculation distance increases for large datasets and its fail to discover the relevant clusters in high-dimensional data.

In [8] &[10] To improve the efficiency over high-dimensional data, Clustering ensemble framework is used in resent researches and it integrate multiple clustering solutions to increase accuracy, robust and stable to provide effective results. Multiple clustering is used to reduce the disassociation between the clusters and increase the association

between the clusters, by combining two or more techniques to improve the accuracy.

Pair-wise is constraints used in single clustering algorithm, to improve efficiency knowledge based and hybrid cluster ensemble classifier, and dataset generates the pair-wise result then algorithm transforms pair-wise constraints for best results.

Comparing single and multiple clustering techniques, multiple clustering solutions give better results over high dimensional datasets.

6. Conclusion and Future

This paper presents a tumor clustering based on Bio-molecular data, and compared with many traditional single clustering approaches for low dimensional datasets and it perform average accuracy value, robustness noise and it increases computational complexity, to perform efficient clustering on bio-molecular datasets, clustering ensemble framework and multiple clustering solution selection will be for future cancer classification using gene expression data, and it will provide more systematical approach in different tumor types, and still we have to improve few more techniques to achieve the goal of cancer classification.

7. References

[1] Matthew C.Clark, Lawrence O.Hall, Dmirty B.Goldgof, Robert Velthuizen, F.Reed Murtagh, and Martin S.Silbiger, *"Automatic Tumor Segmentation Using Knowledge-Based Techniques"*, IEEE Transactions on Medical Imaging,Vol.17,pp.0278-0062, Apr 1998.

[2] Rajashree Dash, Debahuti Mishra, Amiya Kumar Rath, Milu Acharya, *"A Hybridized K-means Clustering Approach for High Dimensional dataset"* International Journal of Engineering, Science and Technology, Vol.2, pp.59-66, 2010.

[3] Weng Shieng, Zhang Changshui, Zang Xuegong, *"PCA-FA: Applying Supervised Learning to Analyze Gene Expression Data"* Tsinghua Science and Technology, Vol.9, pp.428-434, 2004.

[4] Zhiwen Yu, Hau-San Wongb, Jane You,Qinmin Yang, Hongying Liao, et al., *"Knowledge Based Cluster Ensemble for Cancer Discovery From Bimolecular Data"*, IEEE Transaction on Nano BioScience, Vol.10, pp.1536-1241, June 2011.

[5] Zhiwen Yu, Hantoe Chen, Jane You,Liu ellow, Hau-San Wong, Guqiang Han, Le Li, et al., *"Adaptive Fuzzy Consensus Clustering Framework or Analysis of Cancer Data"*, IEEE/ACM Transactions on Computational Biology and Bioinformatics, pp.1545-5963, 2013.

[6] Shu-Lin Wang, Yi-Hai Zhu, Wei Jia, and De-Shaung Huang, et al., *"Robust Classification Method of Tumor Subtype by Using Correlation Filters"*, IEEE/ACM

Transactions on Computational Biology and Bioinformatics, Vol.9, No.2, Mar/April 2012.

[7] Chun-Hou Zheng,De-Shuang Huang, Lei Zhang, Xiang-Zhen Kong, et al., *"Tumor Clustering Using Nonnegative Matrix Factorization with Gene Selection"*, IEEE Transactions on Information Technology in Biomedicine, Vol.13, No.4, pp.1089-7771, July 2009.

[8] Zhiwen Yu, Hantao Chen, Jane YouHau-San Wong , Jiming Liu, and Guoqiang Han Le Li, et al., *"Double Selection Based Semi-Supervised Clustering Ensemble for Tumor Clustering from Gene Expression Profiles"* IEEE Transactions on Computational Biology and Bioinformatics, Vol 11.No.4 , July/August 2013.

[9] Jianhua Dai, Qing Xu, et al., *"Attribute Selection based on Information Gain Ratio in Fuzzy Rough set Theory with Applications to Tumor Classification"*, Elsevier,Vol.13, pp.211-221, 2012.

[10] Zhiwen Yu,Hantao Chen Jane You, Guoqiang Han Le Li, et al., *"Hybrid Fuzzy Cluster Ensemble Framework For Tumor Clustering from Bio-molecular Data"*, IEEE Transactions on Computational Biology and Bioinformatics, Vol.13, 2013.

[11] Giulia Soloperto, Francesco Conversano, Antonio Ggreco, Ernesto Casciaro, Roberto Franchini, and Sergio Casciaro, et al., *"Advanced Spectral Analysis for Real-Time Automatic Echo graphic Tissue-Typing of Simulated Tumor Masses at Different Compression Stages"*, IEEE Transactions on Ultrasonic's, Ferroelectrics, and Frequency Control, Vol.59, December 2012.

[12] Visakh.R, Lakshmipathi.B, et al., *"Constraint based Cluster Ensemble to Detect Outliers in Medical Datasets"*, International Journal of Computer Applications, Vol.86, pp.133-135, 2012.

[13] C.Logeswaran, P.Bharathi, M.Gowthami, et al., *"Brain Tumor Detection Using Hybrid Techniques and Support Vector Machine"*, International Journal of Advances Research in Computer Science and Software Engineering, Vol.5, 2015.

[14] Chinnu.A, et al., *"MRI Brain Tumor Classification Using SVM and Histogram Based Image Segmentation"*, International Journal of Computer Science and Technologies, Vol.6, pp.1505-1508, 2015.

[15] Jin Liu,Min Li,Jianxin Wang,angxiang Wu, Tianming Liu,Yi Pan,et al., *"A Survey of MRI-Based Brain Tumor Segmentation Methods"*, Tsinghuh Science and Technology, Vol.19, pp.578-595,2014.

[16] Dai Hui,Zhou Qiang, Bian Jinian *"Markov Clustering-Based placement Algorithm or Hierarchical FPGAs"*, Tsinghuh Science and Technology,Vol.16,pp.62-68,2011.

[17] Scott White, Padhraic Smyth, et al., *"A Spectral Clustering Approach to Finding Communities in Graphs"*, University of California, 2011.

Authors' Biography

S. Subash Chandra Bose is currently pursing Ph.D Computer Science in Govt. Arts and Science College, Udumalpet. His area of interest is data mining. He has attended various national and international conferences. He has published 3 papers in various journals.

Dr.T.Christopher is presently working as Assistant Professor of Computer Science,at Post Graduate and Research Department of Computer Science, Government Arts College (Autonomous), Coimbatore - 641 018, Tamilnadu, India (affiliated to Bharathiar university, Coimbatore-641046).He has published 30 research papers in International/National Journals; His area of interest include knowledge miningand Network security. He has to credit 25years of teaching and research experience.