AN INVESTIGATION STUDY ON FEATURE EXTRACTION TECHNIQUES ON HIGH DIMENSIONAL DATA

[1]NITHYA C, [2]SARAVANAN V
[1]Ph.D SCHOLAR, [2]ASSOCIATE PROFESSOR and HEAD
PG DEPARTMENT OF IT, HINDUSTHAN COLLEGE OF ARTS AND SCIENCE,
COIMBATORE, INDIA.

ABSTRACT: Data mining is the process of discovering interesting knowledge patterns from large amount of data stored in database. It is an essential process where the intelligent techniques (i.e., machine learning, artificial intelligence, etc) are used to extract the data patterns (i.e., features). The aim of data mining process is to extract the useful information from dataset and transform it into understandable structure for future use. Feature extraction is the process of extracting the relevant features from large database for dimensionality reduction. Many conventional feature extraction techniques were used to project the data in low dimensional space to highlight the property of scattered data. But, the feature extraction time and feature extraction accuracy performance was not improved for high dimensional dataset. Our research objective is to extract the feature from high dimensional data with higher feature extraction accuracy and minimal space complexity by studying the existing issues.

Keywords: Data mining, feature extraction, dimensionality, scattered data, high dimensional data

1. INTRODUCTION

Data mining techniques has the ability to provide the set of useful rules for performing the tasks. Feature extraction is an essential process for addressing the machine learning problems. Features extraction is an essential one for the implementation of decision support system as it identifies abnormal one through selecting the essential features. The feature extraction techniques like Principle Component Analysis (PCA) and Linear Discriminant Analysis (LDA) and their variations are employed in pattern recognition, computer vision and data mining. PCA and LDA was used in many fields for face recognition, palmprint recognition and gene expression data classification. The feature extraction techniques aimed on global structure for dimensionality reduction. Feature extraction is used to encode the high dimensional data into low dimensional space. The feature extraction results are enhanced by constructing set of application-dependent features called feature engineering. Feature engineering is the process of employing the domain knowledge of data to create the features for efficient machine learning algorithms performance. Feature engineering is primary one for machine learning application.

This paper is organized as follows: Section II discusses existing feature extraction techniques on high dimensional data, Section III shows the study and analysis of the existing feature extraction techniques in data mining, Section IV explains the possible comparison between them using dataset. Section V portrays discussion and limitation on existing feature extraction techniques with future direction. Section VI concludes the paper.

2. LITERATURE SURVEY

An ensemble scheme was introduced in [1] for cancer diagnosis and classification with three stages. A hybrid filter based feature selection method was used for selecting the genes. The selected genes were mapped through PSO-dICA method which is a modification of dICA. The mapped features were classified using SVM classifier. But, the prediction accuracy was not increased using ensemble scheme. Class-dependent Locality Preserving Projections method was designed in [2] within-class multimodality. Class-dependent Locality Preserving Projections evaluated every class independently and generated particular projection for each one of them. The query pattern was designed by means of every class output and classified depending on class that fits pattern. But, the feature extraction was not carried out in efficient manner. The lightweight open information extraction (OIE) tool was introduced in [3] to obtain correct knowledge triples resulting in seeding of ontological knowledgebase. But, the error rate was not minimized using lightweight open information extraction tool.

An Orthogonal Feature Extraction (OFE) model with feature ranking techniques was introduced in [4] depending on higher cancer prediction accuracy. The designed model was introduced with feature extraction methods through 5-fold cross-validation. But, the feature extraction time was not minimized through choosing the independent vectors from top ranked attributes iteratively. Two algorithms termed feature selection and feature extraction were introduced in [5] with diverse statistical properties for dimensionality reduction. Multi-view spectral embedding algorithms were developed. A random survival forests identified the local neighborhood relations from right censored survival data. However, the space complexity remained unaddressed.

Jointly Sparse Discriminant Analysis (JSDA) was designed in [6] to determine the essential factors in breast cancer and removed key features for increasing the accuracy during diagnosis and prediction. JSDA started sparse regular term to criterion. A convergent iterative algorithm was designed to address the optimization issue. Though the diagnosing accuracy was increased, feature extraction time was not minimized using JSDA. A new dimension reduction termed p-norm singular value decomposition (PSVD) was designed in [7] for recognizing the low-rank approximation matrix to biomolecular data. Schatten p-norm used regularization function in optimization model. For improving the PSVD, K-means clustering method was used for tumor clustering depending on low-rank approximation matrix. But, the prediction accuracy was not improved using p-norm singular value decomposition method.

A data mining based approach was introduced in [8] for breast cancer classification and diagnosis. The designed approach evaluated with two breast cancer datasets and identified the best techniques in predicting benign/malignant lesions as well as breast density classification. But, the prediction accuracy was not improved using data mining based approach. The cancer registries used Record Linkage algorithm in [9] for linking new cancer records with existing ones. Though algorithm had many linking process, there was few percentage of records that failed to link mechanically. But, the feature selection time was not reduced to the required level using Record Linkage algorithm.

3. FEATURE EXTRACTION TECHNIQUES

In machine learning, pattern recognition and image processing, feature extraction process is used to extract the relevant features from large dataset. Feature extraction process initiates from the collection of measured data and makes derived features as informative and non-redundant through subsequent learning and generalization steps. Feature extraction is mainly used for dimensionality reduction. When the input data of an algorithm is large to process and redundant, then it is transformed into reduced set of features. The determination of initial subset features is termed as feature selection. The selected features have relevant information from input data where the desired task are performed by minimum representation rather than complete initial data. The drawback of feature selection is information loss during the selection of relevant features. In addition, it is difficult to determine the optimal number of essential features. Dimensionality is minimized after selecting the representation set of features from original feature set in feature extraction process.

Feature extraction comprises minimal amount of resources describing large dataset. The large number of data analysis needs large amount of memory and computation power. In addition, it cause classification algorithm to overfit to training samples and simplify inadequately to new samples. Feature extraction is an essential process for constructing the combinations of variables for classification with better accuracy.

3.1 Ensemble Feature Selection and Modified Discriminant Independent Component Analysis for Microarray Data Classification

An ensemble scheme is designed for cancer diagnosis and classification in three stages as described in figure 1. Initially, hybrid filter based feature selection method used modified Bayesian logistic regression (BLogReg), Ttest and Fisher ratio for selecting the genes. Secondly, the selected genes are mapped through PSO-dICA method. Finally, the mapped features are classified by using SVM classifier. For increasing the effectiveness of PSO-dICA method, traditional microarray data like Colon, Lung cancer, DLBCL, SRBCT, Leukemia-ALL and Prostate Tumor datasets are employed.

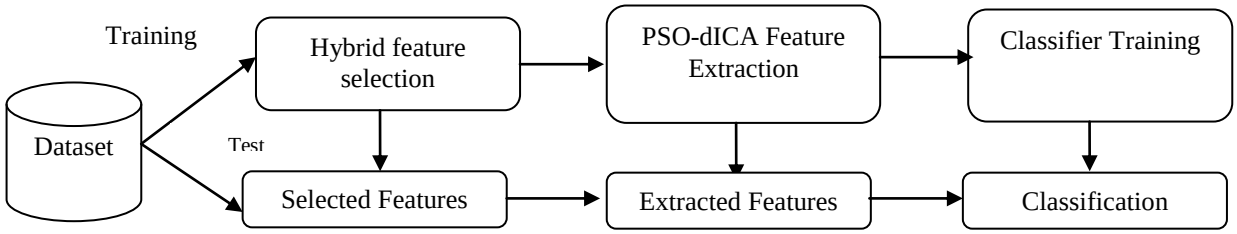

Figure 1 Ensemble Scheme Architecture

3.1.1 Hybrid feature selection

In feature selection, a subset of relevant features is chosen from many features for classification or improves efficiency when using all features. Hybrid scheme joins blogreg, Ttest and Fisher methods to select the features in first stage where the features are ranked depending on three methods. The sum of rankings of three techniques is considered as final ranking. The features with best hybridized rank are chosen as top ranked features.

3.1.2 Improved PSO-dICA

dICA computes the separating matrix 'W' where the independent variables are attained by gradient of cost function. Gradient method falls in local optima. Gradient-based techniques failed to attain high level of accuracy due to the significant complexity involved in Independent Component Analysis. Discriminant Independent Component Analysis depending on PSO is designed. PSO fitness function increases the Negentropy and Fisher ration simultaneously. The optimization problem is to identify the values of particle with higher fitness function. The position of particle presents the separating matrix 'W'. The step is easier than calculation of derivation in Gradient method.

3.1.3 Microarray Classification Framework

After the selecting the features (i.e., genes) from large dataset (i.e., microarray data) by hybrid method, the selected genes are converted using PSO-dICA feature extraction method. Finally, the extracted features are given to the SVM with Gaussian kernel classifier for classification. This in turn helps to improve the classification accuracy.

3.2 Class-wise Feature Extraction Technique for Multimodal Data

Feature extraction is an essential process in machine learning problems. The feature extraction process encodes the high dimensional data into low dimensional space. Many feature extraction methods project data to low dimensional space to emphasize the property of scattered data. Class-dependent Locality Preserving Projections method operates in within-class multimodality scenarios. Class-dependent Locality Preserving Projection method calculates the every class separately and generates projection for all of them. The designed method examines a query pattern by output of each class and categorized depending on class that fits the pattern in better way.

A supervised method termed Class-dependent Locality Preserving Projections (CdLPP) is introduced for feature extraction. The designed method considered the local neighborhood structure of patterns per class to protect the original class structure. CdLPP manages the scenarios with intra-class multimodality. The success in protection of within-class structure is an essential for conservation of original class multimodality. CdLPP is used separately for every class to identify the reduced space that increases the preservation of original class structure. The strategy is enhanced by class-wise negative scatter matrix that are computed by all training patterns which does not belong to the class under evaluation. CdLPP manages the intra-class multimodal situations. CdLPP aimed on preserving intraclass multimodal structure of original data. CdLPP increases the dispersion between positive and negative classes through negative scatter matrix.

3.3 Lightweight predicate extraction for patient-level cancer information and ontology development

A lightweight open information extraction (OIE) tool is introduced to derive the accurate knowledge triples that result in seeding of an ontological knowledgebase. A custom application is developed with information extraction software library to perform the tasks for producing the knowledge triples from textual sources. The open information extraction called ClausIE is carried out from patient health information similar to cancer. ClausIE is a lightweight information extraction library that generates knowledge triples in subject, predicates, objects or representations. A semi-formal method is designed for helping the subject matter experts to initialize the growth of ontology from textual sources and a front-end user tool allow learner subject matter experts to use ClausIE. A formalized evaluation is used to guide the subject matter experts for accessing the results.

It is essential to find out the whether the extracted knowledge of patient health information is favorable for ontology serialization with ClausIE. ClausIE presents many output options to improve the output results. An appropriate extraction pattern within ClausIE presented as an accurate export of triples to facilitate ontology growth for patient-level knowledge of cancer information. In addition, it is essential to find out the particular extraction of propositions that contribute to the accurate tuple information with original source and exact decomposition of tuples. With decomposition, it studies whether the n-ary representations by ClausIE increases the decomposition of the tuples without impediments.

4. COMPARISON OF FEATURE EXTRACTION TECHNIQUES & SUGGESTIONS

The performance of existing feature extraction technique is compared using OASIS dataset and Epileptic Seizure Recognition Data Set with respect to number of instances taken to perform experiment. OASIS dataset comprises 9 features, namely MR session, subject, gender, age, etc with 416 instances. Epileptic Seizure Recognition Data Set comprises 179 features with 11500 instances. The attribute characteristics are integer and real. The dataset is used for classification and clustering tasks. Various parameters are used for efficient feature extraction from high dimensional data. The parameters such as feature extraction accuracy, feature extraction time and space complexity is compared for three existing feature extraction techniques, namely Ensemble scheme, Class-dependent Locality Preserving Projections (CdLPP) method and Lightweight Open Information Extraction (OIE) tool.

4.1 Feature Extraction Accuracy

Feature extraction accuracy is defined as the ratio of number of relevant features extracted from the total number of instances in a large database. It is measured in terms of percentage (%).

$$Feature\ Extrcation\ Accuracy = \frac{Number\ \ of\ relevant\ \ features\ \ extracted}{Total\ \ number\ \ of\ instances} \qquad (1)$$

From (1), when the feature extraction accuracy is higher, the method is said to be more efficient.

Table 1 Tabulation for Feature Extraction Accuracy

Number of Instances (Number)	Feature Extraction Accuracy (%)		
	Ensemble scheme	CdLPP method	Lightweight OIE tool
10	79	65	71
20	83	67	73
30	85	70	75
40	87	72	78
50	90	75	81
60	91	77	83
70	92	80	86
80	94	82	89
90	95	85	92
100	97	88	94

The table 1 describes feature extraction accuracy comparison takes on three methods, namely Ensemble scheme, Class-dependent Locality Preserving Projections (CdLPP) method and Lightweight Open Information Extraction (OIE) tool. From table it is clear that the Ensemble scheme has higher feature extraction accuracy than Class-dependent Locality Preserving Projections (CdLPP) method and Lightweight Open Information Extraction (OIE) tool. The graphical representation of feature extraction accuracy is described in figure 2.

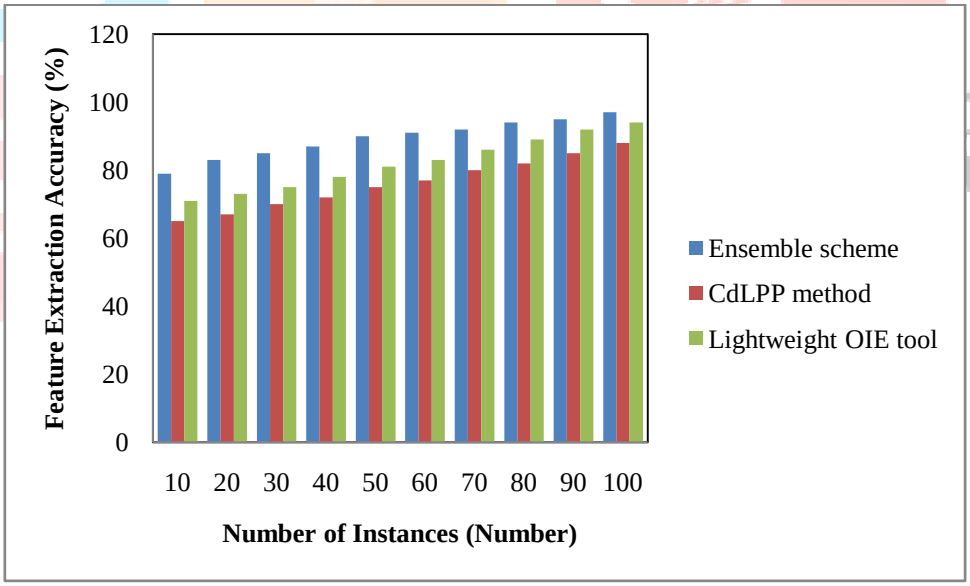

Figure 2 Measure of Feature Extraction Accuracy

Figure 2 explains the feature extraction accuracy comparison of existing feature extraction techniques. From the figure it is known that, feature extraction accuracy of ensemble scheme is comparatively higher than that of Class-dependent Locality Preserving Projections (CdLPP) method and Lightweight Open Information Extraction (OIE) tool. Research in ensemble scheme has 18% higher feature extraction accuracy than Class-dependent Locality Preserving Projections (CdLPP) method and 9% higher feature extraction accuracy than Lightweight Open Information Extraction (OIE) tool.

4.2 Feature Extraction Time (FET)

Feature extraction time is defined as the amount of time taken for extracting the relevant feature from large dataset. Feature extraction time is the difference of starting time and ending time of feature extraction process. It is measured in terms of milliseconds (ms).

$$FET = Ending\ time - Starting\ time\ of\ feature\ extraction \quad (2)$$

From (2), when the feature extraction time is lesser, the method is said to be more efficient.

Table 2 Tabulation for Feature Extraction Time

Number of Instances (Number)	Feature Extraction Time (ms)		
	Ensemble scheme	CdLPP method	Lightweight OIE tool
10	36	21	45
20	39	23	47
30	41	26	51
40	43	28	54
50	45	30	56
60	47	32	59
70	50	35	62
80	52	37	65
90	53	40	67
100	56	42	71

The table 2 explains feature extraction time comparison takes on three methods, namely Ensemble scheme, Class-dependent Locality Preserving Projections (CdLPP) method and Lightweight Open Information Extraction (OIE) tool. From table it is clear that Class-dependent Locality Preserving Projections (CdLPP) method consumes lesser feature extraction time than Ensemble scheme and Lightweight Open Information Extraction (OIE) tool. The graphical representation of feature extraction time is described in figure 3.

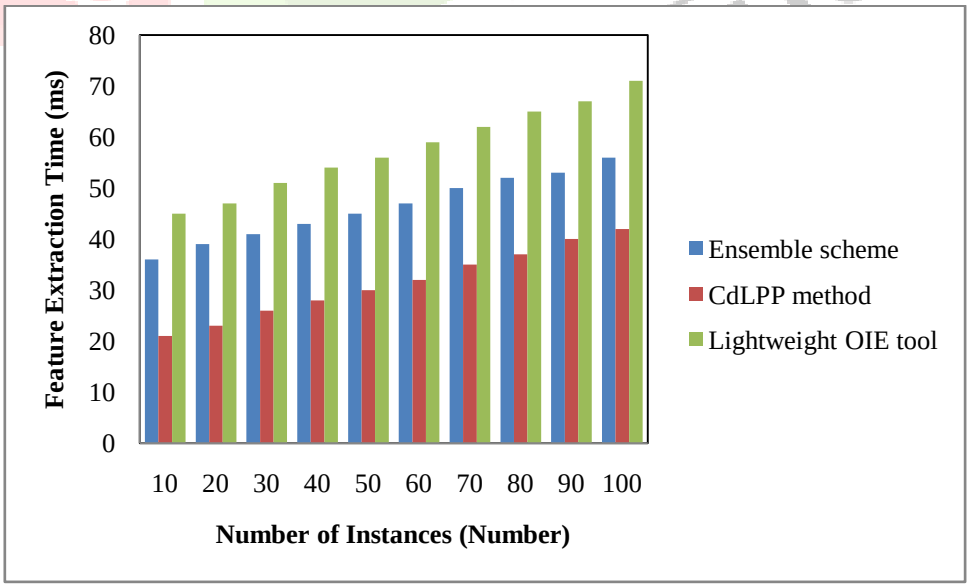

Figure 3 Measure of Feature Extraction Time

Figure 3 explains the feature extraction time comparison of existing feature extraction techniques. From the figure, feature extraction time of Class-dependent Locality Preserving Projections (CdLPP) method is comparatively lesser than ensemble scheme and Lightweight Open Information Extraction (OIE) tool. Research in Class-dependent Locality Preserving

Projections (CdLPP) method consumes 33% lesser time for feature extraction than Ensemble Scheme and 46% lesser time for feature extraction than Lightweight Open Information Extraction (OIE) tool.

4.3 Space Complexity (SC)

Space complexity is defined as the amount of memory space consumed for storing the data instances after extracting the relevant features. It is measured in terms of megabytes (MB).

$$SC = n * memory\ space\ consumed\ after\ feature\ extraction \quad (3)$$

From (3), when the space complexity is lesser, the method is said to be more efficient.

Table 3 Tabulation for Space Complexity

Number of Data (Number)	Space Complexity (MB)		
	Ensemble scheme	CdLPP method	Lightweight OIE tool
10	50	61	42
20	54	63	45
30	56	67	48
40	58	71	51
50	62	74	55
60	65	77	57
70	68	80	60
80	71	82	63
90	73	86	66
100	76	91	69

The table 3 explains space complexity comparison takes on three methods, namely Ensemble scheme, Class-dependent Locality Preserving Projections (CdLPP) method and Lightweight Open Information Extraction (OIE) tool. From table it is clear that Lightweight Open Information Extraction (OIE) tool consumes lesser memory space than Ensemble scheme and Class-dependent Locality Preserving Projections (CdLPP) method. The graphical representation of space complexity is described in figure 4.

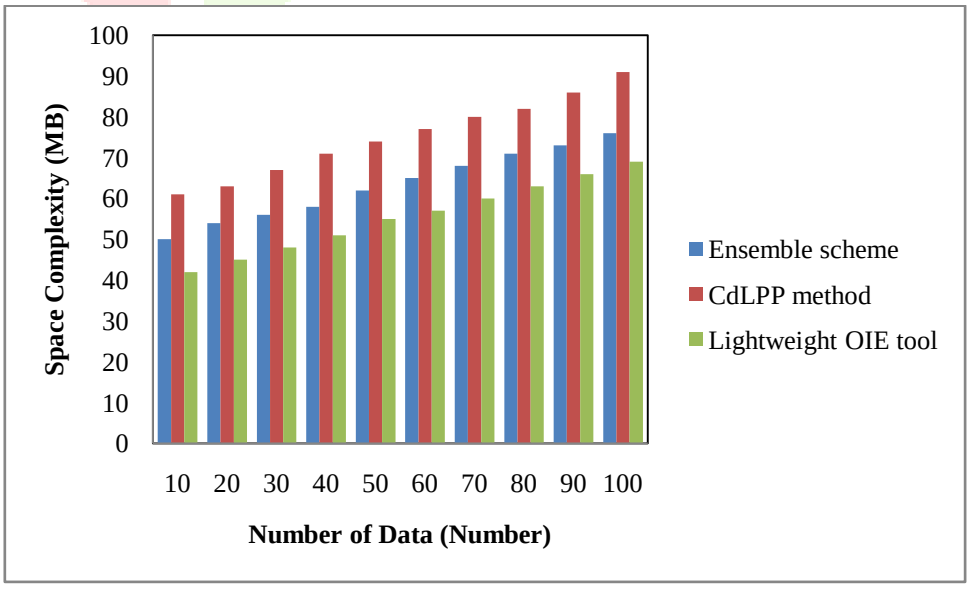

Figure 4 Measure of Feature Extraction Time

Figure 4 describes the space complexity comparison of existing feature extraction techniques. From the figure, space complexity of Lightweight Open Information Extraction (OIE) tool is lesser than ensemble scheme and Class-dependent Locality Preserving Projections (CdLPP) method. Research in Lightweight Open Information Extraction (OIE) tool consumes 12% lesser space than Ensemble Scheme and 26% lesser space than Class-dependent Locality Preserving Projections (CdLPP) method.

5. DISCUSSION ON LIMITATION OF DIFFERENT FEATURE EXTRACTION TECHNIQUES

An ensemble scheme was designed for cancer diagnosis and classification with three stages. A hybrid filter based feature selection method using modified Bayesian logistic regression (BLogReg), Ttest and Fisher ratio was applied for selecting genes. The selected genes were mapped through PSO-dICA method. The mapped features are classified by SVM classifier. The ensemble scheme extracts the informative features from Microarray data. By identifying least correlated features and discriminative features, the decision accuracy is enhanced. However, sampling methods was not used to remove data scarcity and imbalance problem in microarray data for classification problem

Class-dependent Locality Preserving Projections method was operated within-class multimodality. Class-dependent Locality Preserving Projections separated the class and created projection. The method examined the query pattern by output of each class and classified depending on class that better fits the pattern. However, the feature extraction was not carried out in efficient manner. In addition, verification of multimodality in multivariate datasets was an open and hard problem.

Lightweight open information extraction (OIE) tool was introduced to obtain accurate knowledge triples for seeding of ontological knowledge base. A custom application with information extraction software library facilitated the tasks for creating the knowledge triples from textual source. OIE tool delivers an effective and accessible method to the development ontologies. But, the methodology was to be implemented for assessing the extracted information transformation to OWL/RDF triples.

5.1 Future Direction

The future direction of the feature extraction technique can be carried out using machine learning techniques for increasing the feature extraction accuracy and minimizing the feature extraction time.

6. CONCLUSION

A comparison of different existing feature extraction techniques on high dimensional data is studied. From the study, it is observed that the existing techniques consumes large amount of time for extracting features and accuracy was not improved. The survival review shows that the existing Class-dependent Locality Preserving Projections failed to perform verification of multimodality in multivariate datasets in easy manner. In addition, sampling methods was not employed for reducing the data scarcity and imbalance problem in microarray data for classification. The wide range of experiments on existing methods computes performance of the many feature extraction techniques with its limitations. Finally, from the result, the research work can be carried out using machine learning techniques for enhancing feature extraction accuracy and minimizing the feature extraction time.

REFERENCES

[1] Maryam Mollaee and Mohammad Hossein Moattar, “A novel feature extraction approach based on ensemble feature selection and modified discriminant independent component analysis for microarray data classification”, Biocybernetics and Biomedical Engineering, Elsevier, Volume 36, Issue 3, 2016, Pages 521-529

[2] Elias R. SilvaJr, George D. C. Cavalcanti, TsangIng Ren, “Class-wise feature extraction technique for multimodal data”, Neurocomputing, Elsevier, Volume 214, November 2016, Pages 1001-1010

[3] Muhammad Amith, Hsing-Yi Song, Yaoyun Zhang, Hua Xu and Cui Tao, "Lightweight predicate extraction for patient-level cancer information and ontology development", BMC Medical Informatics and Decision Making, Springer, Volume 17, Issue 73, 2017, Pages 23-34

[4] Hao Jiang, Wai-Ki Ching and Wenpin Hou "On Orthogonal Feature Extraction Model with Applications in Medical Prognosis", Applied Mathematical Modelling, Elsevier, Volume 40, Issues 19–20, October 2016, Pages 8766-8776

[5] Sebastian Polsterl, Sailesh Conjeti Nassir Navab and Amin Katouzian, "Survival analysis for high-dimensional, heterogeneous medical data: exploring feature extraction as an alternative to feature selection", Artificial Intelligence in Medicine, Elsevier, Volume 72, September 2016, Pages 1-11

[6] Heng Kong, Zhihui Lai, Xu Wang and Feng Liu, "Breast cancer discriminant feature analysis for diagnosis via jointly sparse learning", Neurocomputing, Elsevier, Volume 177, 2016, Pages198–205

[7] Xiang-Zhen Kong, Jin-Xing Liu, Chun-Hou Zheng, Mi-Xiao Hou and Juan Wang, "Robust and Efficient Biomolecular Clustering of Tumor Based on p-Norm Singular Value Decomposition", IEEE Transactions on NanoBioscience, Volume 16, Issue 5, July 2017, Pages 341 – 348

[8] Joana Diz, Goreti Marreiros, and Alberto Freitas, "Applying Data Mining Techniques to Improve Breast Cancer Diagnosis", Journal of Medical System, Springer, Volume 40, Issue 9, September 2016, Pages 1-7

[9] Yannik Siegert, Xiaoyi Jiang, Volker Krieg and Sebastian Bartholomaus, "Classification-Based Record Linkage with Pseudonymized Data for Epidemiological Cancer Registries", IEEE Transactions on Multimedia, Volume 18, Issue 10, October 2016, Pages 1929 – 1941