ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 1, January 2015

A Survey on Feature Extraction Techniques

N. Elavarasan[1], Dr. K.Mani[2]

Research Scholar, Dept of Computer Science, Nehru Memorial College, Puthanampatti, Trichy, India[1]

Associate Professor, Dept of Computer Science, Nehru Memorial College, Puthanampatti, Trichy, India[2]

ABSTRACT: Data Mining (DM) technique is able to process the high volume of data. The data mining applications contain dataset with high dimensionality. Due to this high dimensionality, the performance of the machine learning algorithms get degraded and this problem is resolved using a technique called Dimensionality Reduction (DR). DR is an essential preprocessing technique in DM to reduce the high dimensionality. Feature Extraction is one of the important techniques in DR to extract the most important features. The goal of this survey is to provide a comprehensive review of various feature extraction approaches to improve the classification accuracy. This paper gives an over view of various feature extraction techniques which are used to the budding researchers.

KEYWORDS: Data Mining, Dimensionality Reduction, Feature Selection

I. INTRODUCTION

Data Mining (DM) is the process of finding the previously unknown information from the large amount of databases. The other terms carrying a similar meaning to data mining, is knowledge mining from databases, knowledge extraction, data or pattern analysis, data archeology, and data dredging. It predicts future trends, behaviors and knowledge-driven decision. To discover useful knowledge, the data should be preprocessed. Data preprocessing is an important technique in data mining to rectify the erroneous data present in the dataset [1]. The dataset contains high dimension of data, due to this the performance of the data mining algorithm gets degraded. The high dimensionality issue is resolved using an important technique called Dimensionality Reduction(DR).

II. DIMENSIONALITY REDUCTION

In many real-world applications, numerous features are used in an attempt to ensure accurate classification. If all those features are used to build up classifiers, then they operate in high dimensions, and the learning process becomes complex, which leads to high classification error. Hence, there is a need to reduce the dimensionality of the feature space before classification. The main objective of DR is to transform the high dimensional data samples into the low dimensional space such that the intrinsic information contained in the data is preserved. Once the dimensionality gets reduced, it helps to improve the robustness of the classifier and it reduces the computational complexity [2]. According to the adopted strategy dimensionality reduction techniques are divided into feature selection and feature extraction [3].

III. FEATURE EXTRACTION

Feature extraction technique is used to extracts a subset of new features from the original feature set by means of some functional mapping by keeping as much information in the data as possible. The following methods are commonly used for the feature extraction [4].

A. Principal Component Analysis

Principal Component Analysis (PCA) is the most popular statistical method. This method extracts a lower dimensional space by analyzing the covariance structure of multivariate statistical observations [3]. The computation of the PCA transformation matrix *S* is given as

$$S = \left(\sum_{i=1}^{n} (Y_i - m)(Y_i - m)^T \right) \qquad (1)$$

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 1, January 2015

where
n is the number of instances
Y_i is the *i-th* instance
m is the mean vector of the input data

B. Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) technique mainly projects the high-dimensional data into lower dimensional space. LDA aims to maximize the between-class distance and minimize the within-class distance in the dimensionality-reduced space [3]. The LDA is computed as

$$f(X) = trace\left((X^T S_w X)^{-1}(X^T S_b X)\right) \quad (2)$$

where
S_b is the *between-class matrix*
S_w is the *within-class matrix*

$$S_b = \frac{1}{n}\sum_{i=1}^{m} k_i\,(c_i - c)(c_i - c)^T$$

$$S_w = \frac{1}{n}\sum_{i=1}^{m}\sum_{x \in X_i} (x - c_i)(x - c_i)^T$$

where,

X_i is the index set of i^{th} class
c_i is the mean vector of i^{th} class.

IV. RELATED WORKS ON FEATURE EXTRACTION TECHNIQUES

In this section some essential contributions are listed to improve the classification accuracy by extracting the most relevant set of new features.

Tomasz Kajdanowicz et al. [5] developed a new method for feature extraction. In this method the new features are calculated by combining the network structure information and the class label. This method is able to extract the important features and show small improvement in the classification accuracy.

Mohammad et al. [6] studied the existing feature extraction methods and they found that the existing methods are unable to find the important features. Hence they developed a multi-level method which has used to extract the most important features. In this method they used binary *n-gram* method in the first level and in the second level a statistical method was applied to extract the most relevant features.

Pechenizkiy et al. [7] considered three different eigen based feature extraction methods. Among them they suggested one important method to extract the best features. And also they designed a decision support system using the suggested feature extraction method to find the improvement in the classification accuracy.

Veerabhadrappa and Lalitha Rangarajan [8] designed a hybrid method to extract the features. In this method they used multi-level process to extract the important features. In the first level they used statistical method to extract the best features and in the second level they analysed the quality of the individual features which are extracted in the first level. Finally, based on the features quality measure the best features are extracted.

Suganya et al. [9] developed a new algorithm to extract the relevant features and to improve the classification accuracy. In this algorithm they adopted a new approach called clustering based feature extraction. The most relevant features are extracted using the supervised clustering algorithm. The algorithm uses a probability density function as a measure.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 1, January 2015

Hence, the algorithm efficiently extracts the most important features and shows a slight improvement in the classification accuracy.

Hua-Yan Wang et al. [10] designed a new approach to extract the efficient features for the compositional data. In this approach they introduced a family of DR projections that preserve all relevant constraints, and then found the optimal projection that maximizes the estimated Dirichlet precision on projected data. This approach extracts the efficient features by reducing its dimension and also improves the classification accuracy.

Hoang Vu Nguyen et al. [11] analyzed the existing feature extraction technique and they found many disadvantages to extract the most efficient features. Hence, they developed a new method called Dimensionality Reduction/Feature Extraction for OUTlier Detection (DROUT). In this method they mainly concentrated to outlier data and to extract the most relevant features. Further weighted adjusted scatter matrix is used to extract the efficient features. This measure is mainly used to detect the outlier and extract the best features.

Nojun Kwak[12] developed a new method to extract the best features by analyzing existing feature extraction methods like Principal Component Analysis (PCA), Linear Discriminant Analysis (LDA), and Independent Component Analysis (ICA) etc. These three methods searches a set of linear combinations of the original features, whose mutual information with the output class was maximized. The mutual information between the extracted features and the output class was calculated by using the probability density estimation based on the Parzen window method. A greedy algorithm using the gradient descent method was used to determine the new features. Hence, this method efficiently extracts a new set of new features and the classification accuracy was improved.

Many data mining applications generated the dataset with the small sample size and it is very difficult to perform the analysis. Hence, Sitanshu Sekhar Sahu et al. designed a new hybrid feature extraction method. In this method they used F-score to extract the most efficient features with a small amount of data [13]

Muhammad Fahad Shinwari et al. [14] have studied the various feature extraction methods and found that the methods are unable to extract the most relevant features and decided to develop a new framework which helps to extract the important features. In this framework they used two important statistical measures namely, Linear Discriminant Analysis and Cross-correlation to extract the features. This framework efficiently extracts the new set of features. The obtained new set of features shows the improved the classification accuracy.

Gladis Pushpa Rathi and Palani [15] used the most common feature extraction technique to extract the features. In their research they used PCA and LDA to extract the most relevant features. This set of newly obtained features is applied to a Support Vector Machine (SVM) classifier and it shows improved classification accuracy.

Sandya et al. [16] developed a new feature extraction method using fuzzy logic. In this method the fuzzy system generates a fuzzy score. This score is used to extract the most relevant features. They found that this method extract the efficient features and shows the better classification accuracy.

V. CONCLUSION

In this paper, a survey is carried out to extract the new set of features efficiently. Many feature extraction algorithms proposed by different researchers are discussed and the issues present in the existing algorithm were identified. Hence, the future work is to overcome the issues and to propose a new feature extraction algorithm which will extract the new set of features and to improve the classification accuracy.

REFERENCES

1. Ranshul Chaudhary, Prabhdeep Singh, Rajiv Mahajan, *A Survey on Data Mining Techniques*, International Journal of Advanced Research in Computer and Communication Engineering, Volume 3, Issue 1, 2014, pp. 5002-5003.
2. Alireza Sarveniazi, *An Actual Survey of Dimensionality Reduction*, American Journal of Computational Mathematics, Volume 4, 2014, pp. 55-72.

ISSN(Online): 2320-9801
ISSN (Print) : 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Vol. 3, Issue 1, January 2015

3. Daniel Engel, Lars Hüttenberger, Bernd Hamann, *A Survey of Dimension Reduction Methods for High-dimensional Data Analysis and Visualization*, LNCS Springer, 2014, pp. 1-16.
4. Khalid, Samina, Khalil Tehmina, Nasreen Shamila, *A Survey of Feature Selection and Feature Extraction Techniques in Machine Learning*, IEEE Science and Information Conference, 2014, pp. 372-378.
5. Tomasz Kajdanowicz, Przemysław Kazienko, Piotr Doskocz, *Label-Dependent Feature Extraction in Social Networks for Node Classification*, Lecture notes in computer science (Springer), volume 6430, 2010, pp.89-102.
6. Mohammad M. Masud, Latifur Khan, Bhavani Thuraisingham, *A scalable multi-level feature extraction technique to detect malicious executable*, Lecture Notes in Computer Science (Springer),Volume 10, 2008, pp. 33-45.
7. M. Pechenizkiy, S. Puuronen, A. Tsymbal, *Feature Extraction for Classification in the Data Mining Process*, International Journal "Information Theories & Applications", Volume 10, 2008, pp. 271-278.
8. Veerabhadrappa,Lalitha Rangarajan, *Multi-Level Dimensionality Reduction Methods Using Feature Selection and Feature Extraction*, International Journal of Artificial Intelligence & Applications, Volume 1, Number 4, 2010, pp. 54-68.
9. Suganya.D, Kowshika.A, *Enhanced Mining of High Dimensional Data using Efficient Clustering Algorithm*, International Journal of Advanced Research in Computer Science and Software Engineering,Volume 3, Issue 12, 2013, pp. 1094-1096.
10. Hua-Yan Wang,Qiang Yang,Hong Qin, Hongbin Zha, Dirichlet *Component Analysis: Feature Extraction for Compositional Data*,International Conference on Machine Learning,2008,pp.20-28.
11. Hoang Vu Nguyen, Vivekanand Gopalkrishnan, *Feature Extraction for Outlier Detection in High-Dimensional Spaces*, Journal of Machine Learning Research, Volume 10, Issue 2, 2010, pp. 252-262.
12. Nojun Kwak, *Feature Extraction based on Direct Calculation of Mutual Information*,International Journal of Pattern Recognition and Artificial Intelligence,Volume 21, Number 7, 2007, PP. 1213-1231.
13. Sitanshu Sekhar Sahu, Ganapati Panda, Ramchandra Barik,*A Hybrid Method of Feature Extraction for Tumor Classification Using Microarray Gene Expression Data*, International Journal of Computer Science & Informatics, Volume 1, Issue 1,2011,pp. 22-26.
14. Muhammad Fahad Shinwari, Naveed Ahmed, Hassan Humayun, Ihsan ul Haq, Sajjad Haider and Atiq ul Anam, *Classification Algorithm for Feature Extraction using Linear Discriminant Analysis and Cross-correlation on ECG Signals*, International Journal of Advanced Science and Technology, Volume 48, 2012, pp. 149-161.
15. V.P.Gladis Pushpa Rathi,Dr.S.Palani,*A Novel Approach For Feature Extraction And Selection on MRI Images for Brain Tumor Classification*, International Journal of Computer Science and Information Technology, Volume 2, Issue 1, 2012, pp. 225-234.
16. Sandya H. B., Hemanth Kumar P. , Himanshi Bhudiraja, Susham K. Rao, *Fuzzy Rule Based Feature Extraction and Classification*, International Journal of Soft Computing and Engineering, Volume 3, Issue 2, 2013, pp. 42-47.

BIOGRAPHY

N. Elavarasan is working as an Assistant Professor in the Department of Computer Applications, Thanthai Hans Roever College, Perambalur, Tamil Nadu, India since 2000.He has 14 years of experience in teaching. After did his MCA, he got his M.Phil (Computer Science) in Madurai Kamaraj University, Madurai,Tamil Nadu, MBA Algappa University,Karaikudi, Tamil Nadu and MA(English) Algappa University,Karaikudi, Tamil Nadu.He has authored books on "Web Metrics"and "Advanced Visual Basics 6.0". He is Currently pursuing doctor of philosophy Programme at Nehru Memorial College (Autonomous), Puthanampatti and his current area of research is Data mining. He has published research papers in national and international journals.

Dr. K. Mani is working as an Associate Professor in the Department of Computer Science, Nehru Memorial College, Puthanampatti, Tamil Nadu since 1989. After did his MCA, he got his Graduation in Operations Research from Operational Research Society of India, Kolkatta and obtained his MTech in Advanced Information Technology from Bharathidasan University, Trichy, Tamil Nadu. He completed his Ph. D degree from Bharathidasan University relating to enhancing security and optimizing the run time in cryptographic algorithms. His current research area include cryptography, data mining and coding theory. He has published a number of research papers in national and international journals and conferences.