International Journal of Science and Research (IJSR)
ISSN (Online): 2319-7064
Index Copernicus Value (2013): 6.14 | Impact Factor (2015): 6.391

Data Mining and Face Recognition

Mohammad Mohsen Ahmadinejad[1], Alireza Ahmadinejad[2], Mohammad Hadi Ahmadinejad[3]

[1]Department of Computer Science University of Kerala, India

[2]Department of Computer Islamic Azad University Kerman branch, Iran

[3]Department of Electronic, Islamic Azad University Kerman branch, Iran

Abstract: ***Data mining is definitely anemerging multidisciplinary subjectthat facilitates discovering of previously unknown correlations, patterns and trends from big amounts of data located in multiple data sources. It is a powerful new technology with good potential to help businesses produce full use of the accessible data for competitive advantages. Data mining application achievementreports have been informed in various areas among them; healthcare, Banking and finance, telecommunication and artificial intelligence. Classification is the most commonly applied data mining technique which is used in face recognition systems. A face recognition system is a dynamic topic in the field of biometrics. The human face has a principal role, which consists of complicated combination of features that allow us to communicate emotions and express our feelings. Principal Components Analysis (PCA) and Kernel Principal Components Analysis (KPCA) are techniques that have been used in face feature extraction and recognition. In this paper, data mining technique,key challenges in data mining, application of data mining is evaluated and kernel PCA algorithm is use for face recognition. To find Recall of KPCA algorithm Yela database is used to shows Kernel-PCA performance.***

Keywords: Data mining technique, key challenges in data mining, application of data, KPCA, Face Recognition.

1. Introduction

A data Mining process is sifting through data store to extract valid patterns, previously unknown and relationships that provide useful information. When these patterns are found they can more be useful to make certain decisions for development or improve of businesses. For decades, major components of data mining technology have been under development in research areas that includes artificial intelligence, machine learning and statistics.Lately, the maturation of these techniques along with high performance relational database engines and broad data integration efforts produce these technologies more efficient for current data factory environments. Data Mining uses sophisticated data analysis instruments and visualization techniques to s portion the data and evaluate the probability of more events. Data mining technology has become popular with many businesses so it allows them to learn more about their customers and make smart marketing decisions for extract more customers. Data mining techniques in health care have been applied in the diagnosis of diseases such as tuberculosis, diabetesetc.Several major data mining techniques have been developed and used in data mining projects include classification, clustering, prediction, association, and sequential patterns etc.

Support vector machines are data mining classification models, whichare used for face recognition. Support vector machines which are supervised learning models which includes associated learning algorithms that analyze data used for classification and regression analysis.Face recognition system is nowadays emerging in multiple areas such as security, protection, safety, robotic, surveillance etc. There are numerous approaches for developing an online face recognition system and also several algorithms are currently existed to identify and recognize a face from a given dataset.

The PCA has been proven as a powerful method for face recognition as its optimal compression scheme minimizes the mean squared error, which is between the original images and their reconstructions for any given level of compression [1]. PCA is mathematically defined as a linear transform that transform the data to a new coordinate system such that best variance which is use to find best Eigen-faces. A PCA algorithm computes means, covariance, variances and correlations of large data sets [2].

Traditional PCA only allows linear dimensionality reduction but, Kernel PCA allows us to generalize linear PCA to nonlinear dimensionality reduction[3]. A Kernel PCA algorithm is a nonlinear form of PCA, which is a classification method and works better in complicated spatial structure of high-dimensional feature.In this paper is evaluated data mining technique and a klpca method for face recognition.

In this paper, data mining technique, key challenges in data mining, application of data mining is evaluated and kernel PCA algorithm is use for face recognition. The performance of KPCA method obtained from Yela database shows satisfactory result.

2. Related Work

Kim *et* al.(2002) presented a Kernel PCA based face feature extraction method, Therefor they used polynomial kernel principal components to compute the product space of input pixels to generate a facial pattern. To show the effectiveness of the proposed method, an SVM method was used as the recognition with ORL database [6].

Gan*et al.* (2005)in their research, presented the advantages of PCA, and an improved method. Gan*et al.* did research on the normalization of within class average face image. They compared with traditional PCA method, and their results showed more acceptable method to process samples with

different class and same class. This showed that a higher correct recognition rate can be acquired, and then a better efficiency can be achieved [7].

Timotius*et al.* (2010) presented, that KPCA method is utilized to extract features from the input images, SVM method is applied to classify the input images. They compared the performance of this face recognition method to other commonly used methods that is shown the combination of KPCA and SVM achieves a higher performance as compared SVM,and the combination of kernel principal component analysis (K-PCA)[8].

Vieira, MatheusAlves, et al (2012) in their research develop a methodology for contributing in the automation of sugarcane mapping over big areas, with time series of remotely sensed imagery. They have combined two major method which includes object based image analysis and Data Mining. Object based image analysis used to represent the knowledge, which required mapping sugarcane, and data mining is usedto generate the knowledge model. Data mining algorithm applied to generates decision trees (DT) from pass prepared training set. After training, the DT was used to the Landsat time series, So generating the desired thematic map with sugarcane ready to harvest. The classification accuracy is calculated over a set of 500 points, which not previously used during the training stage. The statistics indicated that the classification is shown an overall accuracy of 94% and for a Kappa coefficient of 0.87. Their Results show that the combination of OBIA and DM techniques is very effecfive and promising for the sugarcane classification process [9].

El Traboulsi et al.(2016) in their research survey a semi supervised discriminant embedding which is the semi-supervised extension of Local Discriminant Embedding (LDE). Therefore, since this type of methods is in general dealing with high dimensional data, the small sample size issue very often happens. So this problem happens when the number of available samples is less than the sample dimension. Thus The classic solution to this issue is to reduce the size of dimension of the original data thus, the reduced number of features is less than the number of samples. They mention This can be achieved by applying Principle Component Analysis (PCA). So a SDE needs a dimensionality reduction or an explicit matrix regularization, with the shortcomings both methods may suffer from. In their paper, they propose an exponential version of SDE to overcoming the SSS issue,Sothe latter emphasizes the discrimination property by enlarging distances between samples which belong to different classes. So experiments result has made on seven benchmark datasets, whichshow the superiority of our method over SDE and number of state of theart semi-supervised embedding technique [10].

3. Data Mining Techniques

Several data mining techniques is developed and used in data mining, which include association, classification, clustering, prediction and sequential patterns etc.

3.1 Classification

One of the most commonly applied data mining technique is Classification, which employs a set of pre-classifiedsample to develop a model that could classify the population of records at large that this approach frequently employs decision tree or neural network based classification algorithms.

A data classification ia a process whichis involves to learning and classification.Thetraining data are analyzed by classification algorithm in learning and in classification test data are used to estimate the accuracy of the classification rules. Soif the accuracy is acceptable the rules can be applied to the new data tuples.

The classifier training algorithm uses pre-classified examples to determine the set of parameters required for proper discrimination then encodes these parameters into a model called a classifier. A classification normally uses prediction rules to express knowledge.Thus, Prediction rules that expressed in the form of IF-THEN rules, where the antecedent consists of a conjunction of conditions and the rule consequent predicts a certain predictions attribute value for an item that satisfies the antecedent Types of Classification Models.

3.1.1Different classification models

a. Neural Networks
b. Support Vector Machines
c. Bayesian Classifiers
d. Classification based on Associations
e. Decision Trees

3.2 Clustering

Clustering is a data mining technique of collection set of data objects into numerous groups or clusters thus objects within the cluster have high similaritybecause are very dissimilar to objects in the other clusters. Dissimilarities and similarities have assessed on basis of the feature values explaining the objects. Clustering algorithms areused to categorize data, organize data for data compression and model construction, for recognize and detection of outliers etc. Thus, common approach for all clustering techniques is to find clusters Centre, which will represent each cluster. Cluster Centremay signify with input vector may tell thatcluster vectorsfit in with to by calculating a similarity metric in input vector and all cluster Centre and determining which cluster is nearest or most similar one.So a cluster analysis has used as astandalone data-mining tool to gain insight into the data distribution, or as a preprocessing stage for different data mining algorithms running on the detected clusters. Many clustering algorithms have developed and categorized from a few aspects such as dividing methods, density based methods, hierarchical methods and grid-based methods.

Further data set can be numeric or categorical. An inherent geometric property of numeric data is exploited to naturally define distance function between data points. Whereas categorical data maybe derived from either quantitative or

qualitative data where observations are directly observed from counts

3.2.1 Data Clustering Techniques

a. K-Means Clustering
b. Hierarchical Clustering
c. DBSCAN Clustering
d. OPTICS
e. STING

4. Association Rule Mining

Association rule learning is reallya well-known and effectively researched method for finding interesting relations between variables in big databases. Association rule learning is designed to identify powerful rules found in databases using various measures of interestingness that is based on the concept of powerfulrules.A typical and popularexample of association rule mining is Market Basket Analysis. Therefore, the issue is to generate all association rules, which may have help and confidence greater than the user specified minimum support and minimum confidence.

5. Key challenges in data mining

It is apparent that data mining is an emerging and strong technology to manage a wide range of challenges.

Application of the existing data mining techniques and algorithms has confrontedproblemsdue to of inadequacies. Significantproblemscontainthe requirement to scale up for large dimensional data and high speed streams.Sothe contamination in sequential and time series data, mining multi-agent data and distributed data mining, privacy and security of data, mining complex knowledge from heterogeneous and complex data, interpretation of results that is including visualization or any other methods.Therefore, the knowledge can be simplyrecognized and directly usable by humans, applying algorithms designed for small datasets when dealing with big data sets, etc.

On the interpretation of resultsthat delayed discovery of even highly valuable knowledge invalidates the usefulness of the discovered knowledge.

6. Most outstanding challenges in data mining

6.1 Sensitive and Private Data

Data mining is now an important technology for getting knowledge from big and diverse quantities of data. Therefor there has been growing concern which use of this technology is violating individual privacy. Several popular data mining applications exist that deal with sensitive datasuch as financial records and people medical.

If the main collection of data is not appealing as it puts their privacy into risk and some certain cases the data might belong to different, perhaps competitive, organizations thatlooks to exchange knowledge without the exchange of raw private data.

Lots of research has moved out on how to deal with this chalange. Suhad Abu Shehab and Al-Hamami [13] designed an application that provides protection for privacy and knowledge in data mining.Thereforin this application, privacy protection of individuals is achieved by adding a white Gaussian noise to selected columns in a database to be mined for knowledge protection and an unauthorized user is done by encrypting the result of data mining before it appears to the unauthorized users by using Rijndaelmethod.

6.2 Distributed data and operations

The shift towards intrinsically distributed complex issue solving environments is prompting a selection of new data mining research and development problems [14]. As reported in [12], [14] and [15], the data which is stored in distributed computing environments on heterogeneous platforms become impossible to bring to a centralized place because of both technical and organizational reasons. Consequently, development of methods, tools, and services is need that facilitates the mining of distributed data [14].

As for distributed operations, more data mining operations and algorithms will be required on the grid and to facilitate seamless integration of resources into distributed data mining systems for complex problem solving, novel algorithms, tools, grid services and IT infrastructure required to be developed.

6.3 Data Quality

The quality of data is a very important factor when considering solutions for managing data in support of Performance of organizations. Quality is common that when acquiring and entering data, simple and complex errors can be committed. Thus the errors in a large database may be due to a number of factors among them missing attribute values and corrupted vales.

To eliminate the errors, data analysis methods and data cleaning methods which can handle noise are required, as well as outlier mining methods for the discovery and analysis of exceptional cases. Thecommon data cleaning tasks include data acquisition and metadata, filling in missing values, identifying outliers and smooth out noisy data and correcting inconsistent data, unifying date formats, converting nominal to numeric values.

7. Application of Data Mining

Data mining is used in a wide range of fields including Health/healthcare/InsuranceTelecommunication, Corporate surveillance, Bioinformatics, Text Mining and Web Mining, Banking and Finance, Bibliomining, Targeted Marketing and Customer Segmentation , etc.

7.1 Health/Healthcare/Insurance

Medical data mining in healthcare is considered as an important and complicated task that needs to be executed efficiently and accurately. Healthcare data mining attempts to solve real world health problems in diagnosis and treatment of diseases such as liver diseases and heart .Vikas

[18] apply different methods of classifier techniques in the diagnosis of heart disease patientswhich results show that bagging algorithm gives an accuracy of 85.03%, which makes it one of the most successful data mining techniques used in the diagnosis of heart diseases. Thus In health insurance, data mining is used in claims analysis that medical procedures are claimed together. Alsoit helps to forecast on the customers with potential to purchase new policies and those with fraudulent behavior Text Mining and Web Mining.

7.2 Text mining and web mining

Text mining is the process of searching big volumes of documents from specific keywords or key phrases. So by looking literally thousands of documents numerous relationships involving the documents may be established. Using text mining however, quicklyderives certain patterns in the comments, which may help to identify a common set of customer perceptions not captured by the other survey questions. Extension of text mining is web mining. Web mining is an exciting new field, which integrates data and text mining within a website. It improves the website with smart behavior, which includes suggesting related links or recommending new products to the consumer Finance and Banking. It is mainly used for credit fraud prediction, risk evaluation and for analyzing trends and profitability.

8. Support Vector Machines (SVM) model

Support vector machines is one models of data classification in data mining which isgiven a set of training examples, each marked example can belong to one of two categories, thus Support vector machines training algorithm builds a model that assigns new examples into one category or the other category of non-probabilistic binary linear classifier. Support vector machines model is anexplanation and representation of the examples as points in space, each example will map to new space. So the examples of the separate categories are divided by a clear gap which is as wide as possible. New examples are mapped into that same placesthen predicted to belong category based which is in side of the gap the example fall on.

If data can't labeled, The supervised learning is not possible, therefore an unsupervised learning approach is required. This unsupervised learning approach attempts to find natural clustering of the data to groups, and then map new data to these formed groups.

SVM is a discriminative classifier technically defined by a separating hyperplane. In other words, a Support Vector Machine algorithm outputs is an optimal hyperplane which categorizes new examples.

In addition to performing linear classification, A SVM technique may efficiently perform a non-linear classification with using the kernel trick, implicitly mapping inputs into larg dimensional feature spaces.Therefore, kernel methods are a class of algorithms for pattern analysis,face imageand object recognition, which isknown as the best element in support vector machine

8.1 Kernel PCA

Kernel principal component analysis is an efficient method for face recognition. Which is used to run the real world application with large scale training set. The KPCA is extended from PCA method to represent nonlinear mappings in a higher dimensional feature space.

A standard PCA only allows linear dimensionality reduction, So if we have data that has more complicated structures which may not be good represented in a linear subspace, standard PCA will not be very helpful. But, kernel PCA allows us to generalize standard PCA to nonlinear dimensionality reduction thus we can use kernel methods to simplify the computation.

The KPCA is used for the nonlinearity of face recognition problem by using a nonlinear kernel function then a dimensional reduction is performed. The images are first transformed from image space into a feature space. KPCA process diagram for face recognition is shown in fig [3].

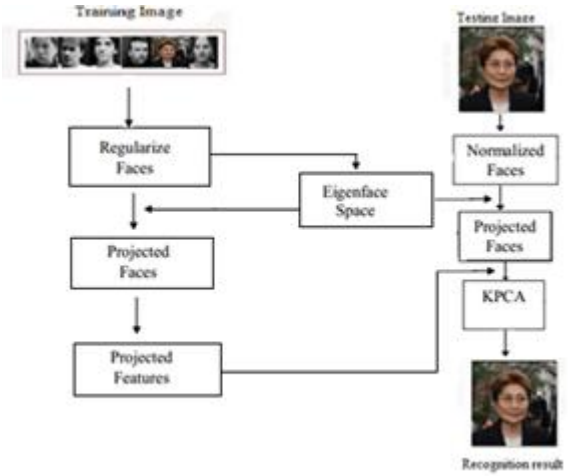

Figure 3: Kernel PCA process for face recognition

8.1.1 The Standard Kernel PCA Steps

a. Construct the kernel matrix K from the training data set which shows in equation (5).

$$K_{i,j}= K(x_i, x_j) \quad (5)$$

b. Compute the Gram matrix K' using Equation (6).

$$K'=K-1_nK-K1_n+1_nK1_n \quad (6)$$

c. Solve the vectors a_i(substitute K with K_{ak}) by using the equation (7).

$$Ka_k= \lambda_k Na_k \quad (7)$$

d. Compute the kernel principal components using equation (8).

$$y_k(x)= \phi(x)^T V_k=\sum_{i=1}^{k} a_{ki}K(x , x_i) \quad (8)$$

where $K(x, x_i)$ is given by Guassian Kernel function.

8.1.2 Error Rates between PCA and Kernel PCA

We have used Yale database with a dataset of 486 KPCA. The error rate of each method is calculated and is given in the table 1.

Table 1: Error rate in training and testing data

Error rate	Training data	Testing data
KPCA	7.90	13.56%

8.1.3 Performance Evaluation

The recall, precession and overall accuracy is calculated by taking 410 images from 41 different persons in 10 different poses. The result is shown in Table 2.

Table 2: compartive PCA,KPCA

KPCA	
Recall	Precision
86.58%	89.19%

9. Conclusion

In this paper, we evaluated data mining technique, key challenges in data mining, application of data mining,association rule mining and kernel PCA algorithm is use for face recognition. To find Recall of KPCA algorithm Yela database is used to shows Kernel-PCA performance.

References

[1] R. Agrawal, T. Imielinski and A. Swami, "Data Mining: A Performance Perspective", IEEE Transactions on Knowledge and Data Engineering 5(6) (1993), 914–925.
[2] Shieh, Jiann-Cherng, " The Integration System for Librarians' Bibliomining", Asia- Pacific Conference on Library & Information Education & Practice, 2009
[3] V. Uma, M. Kalaivany and G. Aghila, "Survey of Sequential Pattern Mining Algorithms and an Extension to Time Interval Based Mining Algorithm", International Journal of Advanced Research in Computer Science and Software Engineering 3(12), December -2013, pp. 1178-1183
[4] Vikas Gupta, "A survey on Data Mining: Tools, Techniques, Applications, Trends and Issues", International Journal of Scientific & Engineering Research Volume 4, Issue3, March-2013 ISSN 2229-5518.
[5] VikasChaurasia and Saurabh Pal, "Data Mining Approach to Detect Heart Diseases", International Journal of Advanced Computer Science and Information Technology (IJACSIT) Vol. 2, No. 4, 2013, Page: 56-66, ISSN: 2296-1739.
[6] Kim, Kwang In, Keechul Jung, and Hang Joon Kim. "Face recognition using kernel principal component analysis." Signal Processing Letters, IEEE 9.2 (2002): 40-42.
[7] Gan, Jun-ying, Dang-pei Zhou, and Chun-zhi Li. "A method for improved PCA in face recognition." International Journal of Information Technology 11.11 (2005): 79-85.
[8] Timotius, Ivanna K., IwanSetyawan, and Andreas A. Febrianto. "Face recognition between two person using kernel principal component analysis and support vector machines." International Journal on Electrical Engineering and Informatics 2.1 (2010): 55-63.
[9] Vieira, MatheusAlves, et al. "Object based image analysis and data mining applied to a remotely sensed Landsat time-series to map sugarcane over large areas." Remote Sensing of Environment 123 (2012): 553-562.
[10] El Traboulsi et al. "Matrix Exponential based Semi-supervised Discriminant Embedding for Image Classification." Pattern Recognition(2016).

Author Profile

MohammadMohsen Ahmadinejad, took his Post Graduation in Scientific computing in interdisciplinarySchool of scientific computing, University of Pune in 2010. He has two years of teaching experience after the Post Graduate level and four Years' experience in the research field. He is currently pursuing his Ph.D in Online face recognition in unconstrained environments at University of Kerala, India. His research interests include online methods for face recognition and methods for extract face features.

Alireza Ahmadinejad, took his Post graduation in Electronic Engineering in Islamic Azad University Kerman branch in 2014. He is currently pursuing his master's degree in computer, field of artificial intelligence in Islamic Azad University Kerman branch, Iran.

Mohammad Hadi Ahmadinejad is currently pursuing his bachelor's degree in electronic engineering in Islamic Azad University Kerman branch, Iran.