ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

IJESRT

INTERNATIONAL JOURNAL OF ENGINEERING SCIENCES & RESEARCH TECHNOLOGY

A Study on Different Feature Level Learning and Prediction Approaches

Shikha[*1], Shikha Khera[2]
[*1] Student, M.Tech Department of Computer Sc. & App., SPGOI Rohtak, Haryana, India
[2] Asstt. Professor, Department of Computer Sc. & App., SPGOI Rohtak, Haryana, India
shikha.shikha43@gmail.com

Abstract

Mining is about to extract the important knowledge over the available dataset. There are number of functionalities available to extract such information over the dataset. One of such approach is classification approach. Dataset Learning is about to categorize the available dataset so that effective knowledge extraction will be done. Learning approaches perform the feature level analysis over the dataset and reduced the processing dataset size so that the performance of the overall mining process will be reduced. In this paper, some of the effective learning approaches are discussed such as Decision Tree, KNN and Bayes Network. The paper has also presented the learning model of disease prediction for agricultural data also.

Keywords : Disease Prediction, Learning Algorithm, Bayes, KNN.

Introduction

When the data is in raw form and to extract the meaningful information over this dataset, some mining operations are required to implement. These mining algorithms include the data filteration, encoding, feature extraction, segmentation, classification, prediction etc. These all operations itself represent a research area of data mining and each stage improves the effectiveness of available data and allow to extract the meaningful information from the dataset. These operational stages are used in various applications individually or in combination. In most of the applications, one or more data mining operations are collected to take the decision regarding the information extraction and to perform the effective data operations. These operations depend on the application type, data type and based on user requirement. The basic data mining operations are given in figure 1.

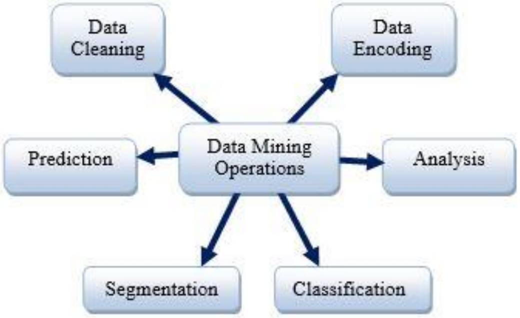

Figure 1 : Data Mining Operations

Data Cleaning is about to perform the data filtration. When the data is collected from primary source, it can have number of impurities. These impurities are defined in terms of blank data values or the missing values. In such case, to provide the effective data solution, it is required to remove these impurities over the dataset in earlier stage. The impurities can be attribute level or record level. Once the impurities are removed, the clean dataset is obtained for processing. The stage of data processing is to perform the data encoding. Encoding is about to convert the data is required form. Some applications are data type specific such as they requires data in nominal form in which data conversion to that form is required.

Another associated operation with data mining is data segmentation. Segmentation is about to extract the valuable features over the dataset. The features can be identified in terms of particular recordset or the values set. Segmentation is about to reduce the dataset size and so that the processing efficiency and accuracy will be improved.

Data classification is one of the most required tasks to extract information from Data. It is used in different contexts to perform the object or pattern recognition as well as to perform the categorization of the objects based on information analysis. It is actually defined in a hybrid scenario that itself covers the concept of object categorization, object recognition as well as enable the object search.

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

Classification is about to characterize the Data under the visual part analysis with view analysis. There are number of application areas where the classification plays an important role. These application areas include the disease classification in medical Data, , object classification in real time Data etc.

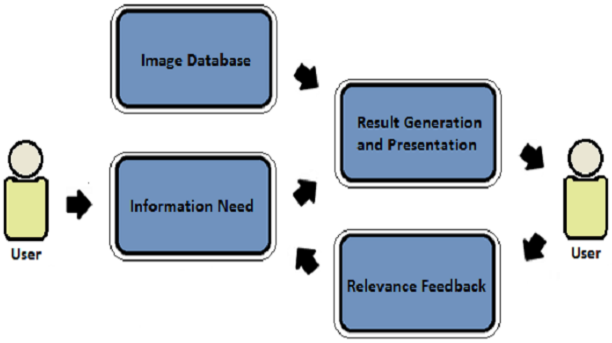

Figure 2 : Information Retrieval Process

Data classification is actually to define a tag or the annotation to the Data based on the feature based analysis. The classification process is applied on a set of arbitrary Data collected from any primary or secondary source. These Data belong to specific domain such as medical Data, geographic Data, handwritten characters, biometric Data etc. The classification procedure is divided in two broader approaches called supervised classification and unsupervised classification.

Another effective approach associated with classification process is prediction approach. Prediction is about to identify the accurate decision from the dataset. There are number of such approaches for prediction analysis based on classification approaches. The prediction is about forecasting the future aspect of dataset so that the information extraction over the dataset will be generated and identified. There are number of such approaches to identify the disease over the dataset accurately.

After implementing any of the mining operation, the final work is to perform the analysis over the dataset. There are number of analysis approaches used to predict the results over the dataset. The analysis parameters are generally statistical and used to identify the error rate or the accuracy.

In this paper, A study on different learning approaches on real time datasets is defined. In this section, a stage wise exploration of data mining process and the associated processing terms is defined. In section II, the work defined by the earlier researchers is discussed. In section III, different classification approaches are defined and explored based on which the work is carried on. In section IV, the conclusions obtained from the paper are discussed.

Related Work

In this section, the exploration of the data mining operations used by different researchers is discussed. These operations includes the filteration, segmentation, classification and prediction approaches. K. S. Chen[1] has presented a data cleaning approach to improve the classification over the dataset. Author defined a statistical approach for data classification called Kalman Filter approach for dataset analysis and to classify the record. The Kalman filter approach actually analyze the performance measure over the dataset and defined work on filtered data. Author obtained the satisfactory results in reasonable time. Another work defined by A.T.Shuen[2] on data leaning approach by using the region fusion and learning approaches. Author defined the work on satellite dataset. Author compared this dataset with existing block dataset and performs the effective categorization of data. A work on data similarity analysis and segmentation was defined N.Abbadeni[3] in year 2003. Author defined a content based analysis approach over the dataset to identify the data similarity and to resolve the fundamental issues regarding dataset processing.

Author improved the information retrieval approach over the dataset by using the concept of coefficient analysis based similarity analysis. Author defined data database formation under the similarity analysis so that the weighted value analysis under the hierarchical data construction and organization are defined by the author.

Author defined an effective model to separate the homogenous and heterogeneous data elements under auth regressive analysis. The presented model by the author was based on statistical analysis so that the effective data reterival and perceptual data cleaning is performed by the user. Author defined the directional analysis on dataset features so that effective classification and prediction rate will be obtained[4].

Kentaro Toyama[5] has defined a system level analysis on dataset values to perform the data extraction on geographical digital data. Author defined the work to different data structures and different types of data to handle the location tag problem. Author work on location based data optimization so that location oriented information will be retrieved effectively. Another work on data classification for MAP algorithm was performed by L. Yuan[6]. Author defined a GMM based approach for data distribution and parametric density distribution under likelihood parameter analysis. Author defined a statistical approach for iterative analysis of data values so that effective computations will be drawn. Lan Gao[6] defined a fuzzy rule based data analysis approach satellite data. Author presented a C Means clustering inspired approach for data separation and division to the clusters. Author performed the data extraction and transition so that the effective dataset conversion and cluster adaptation will be performed. Another work on supervisied leaning based approach was presented by R.Lorenzo[7]. Author presented a three layered neural approach for data classification and recognition. Author used the unsupervised learning approach for data distingution and recognition. Author identified the similar and dissimilar data elements over the dataset so that effective data organization and distribution will be done. Q.Wu[8] has presented an effective knowledge discovery approach for intelligent systems. Author combined the data mining and machine learning approaches under the neuron model for intelligent systems so that the effective fuzzy rules will be identified over the dataset. Another work on data transformation and evoluation for satellite data considered by Peterson[9]. Author presented a training and testing effective dataset organization and partitioning approach for data correction based on reasoning analysis. Author presented the data classification under temporal vector analysis so that time based decisions will be taken from the dataset.

F. Chen[10] has presented a data reasoning and case based information classification approach is suggested. Author presented a statistical and machine learning based approach for similar data identification and to divide the dissimilar data in separate data groups. Author work on ancillary information to perform data division so that effective data encoding and coversion will be obtained.

From the literature review, it is found that the researchers main focus is on data learning approaches and classification so that effective data distribution and recognition will be performed.

Learnining Techniques

To extract the valuable information from the dataset, it is required to extract the dataset features. The process of feature extraction and feature analysis based on parametric analysis is required to improve the effectiveness of different data oriented tasks. These tasks include the data prediction, recognition etc. While working will learn algorithm, some authenticated dataset is required with feature value analysis. The learning of data values based on input data analysis and mapping is shown in figure 3.

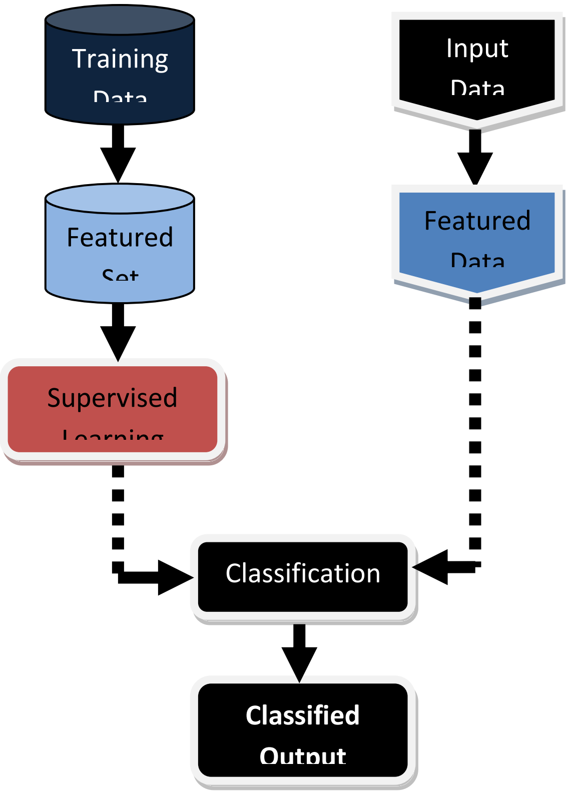

Figure 3 : Classification Model

The effectiveness of classification or recognition process is based on the learning algorithm. There are number of learning algorithms already available to process on dataset. The figure is showing the basic flow of recognition and classification process. As shown in figure, this model requires, the authenticated dataset with on which the preprocessing stage is implemented to perform the data cleaning and data encoding. Just after the encoding process, the learning process is defined in the work. In this work, a rule based learnining model is shown. After the effective learning process, the next stage is to perform the recognition or classification based on the learning feature capabilities. The performance analysis and data value analysis are the basic stages of learning model. Once the learning is performed, the next work is to perform the data classification or recognition based on the analysis. In this paper, some of the effective learning and classification approaches are discussed. Some most effective approaches defined here are Navie Bayes approach, KNN and Decision Tree based approach.

ISSN: 2277-9655
Scientific Journal Impact Factor: 3.449
(ISRA), Impact Factor: 1.852

A) Naive Bayes Classifier

This classification algorithm is based on the data estimation and prediction under the density value analysis. Author defined a knowledge analysis and data extraction scheme for class definition and extraction model. Author defined N value analysis under feature class analysis and class identification. The rule is here defined under the probabilistic value analysis. Author presented feature vector based analysis and relative feature analysis so that effective classification process will be obtained. Author defined class formation approach under feature value analysis given as under

$$P(C_j|X_1,X_2..X_n) = \frac{P(C_j)\,P(X_1,X_2..X_n|C_j)}{P\,(X_1,X_2..X_n)}$$

Here P(Cj) is the prior probability of class Cj, $P(X_1,X_2..X_n|C_j)$ represents the conditional probability of feature vector respective to the class.

The process of Naive Bayes classification combines the model with some decision rule. The decision rule here defines the hypothesis to obtain the most probabilistic value. Based on which, the maximum a posterior mapping based decision rule can be formed. This classification process can be performed to obtain the most effective matched Data.

B) K-Nearest Neighbor

This is one of the simplest classification technique that based on the computation process to achieve the accurate result. In most of the classification implementation, it gives more accurate results. This method is processed on the feature vector and perform the distance based analysis between the input object and the training objects. Here k represents the number of classes in the training set. Now when the distance analysis is performed, based on the best-fit analysis the nearest neighbor is identified. The featured classes are differentiated based on the analytical distance based decision vector. Here figure 4 is showing the phenomenon of KNN process. This approach is the prediction approach that use some metric based measure for distance analysis. One of such measure is the Euclidean distance based analysis.

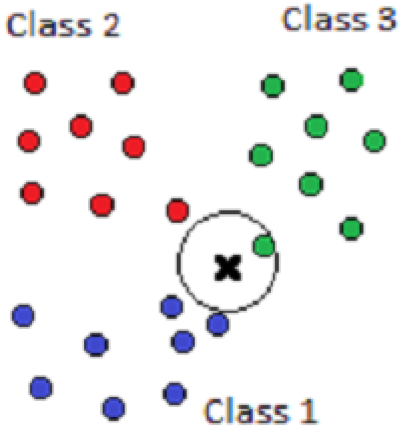

Figure 4 : KNN Classification Process

C) Decision Tree

The decision tree approach is the recursive process approach that performs the conditional analysis on the intermediate nodes as well as leaf nodes to perform the classification process. The process starts from the leaf node and moves towards the root to perform the classification. This process is implemented on subsequent nodes till the classification is not performed. This classification process includes the space division so that the attribute based sub-space division will be performed. Each sub space group will represent the separte class. It is the decision tree process called Random Forest classification. Here the feature vector is defined at the leaf node to perform the classification and random variables are been used for the analysis.

Conclusion

In this paper, a study based representation of different classification approaches is defined. The paper includes the study of the classification process along with model exploration. The work also includes the exploration of different supervised learning approaches that are effective to perform the Data classification.

References

[1] K. S. Chen,"Filtering Effects on Polarimetric SAR I m g e Classification", 0-7803-3836-7/97@ 1997 IEEE, 1997

[2] A. T. Shuen Ho,"Improving SAR Data classification In Tropical Region Through Fusion With SPOT Data", 0-7803-4403-0/98@1998 IEEE, 1998

[3] N. Abbadeni,"Content Representation and Similarity Matching for Texture based Data Retrieval", MIR'03, November 7, 2003,

Berkeley, California, USA 1581137788/03/00011 p 63-70, 2003

[4] *K. Toyama,"Geographic Location Tags on Digital Data",MM'03, November 2-8, 2003, Berkeley, California, USA. 1-58113-722-2/03/0011 pp 156-166, 2003*

[5] *L. Yuan,"SAR Data Classification Based on MAP via the EM Algorithm", Proceedings of the 6th World Congress on Intelligent Control and Automation, June 21 - 23, 2006, Dalian, China 1-4244-0332-4/06©2006 IEEE, 2006*

[6] *L. Gao," A New Fuzzy Unsupervised Classification Method for SAR Data", 1-4244-0605-6/06@2006 IEEE, 2006*

[7] *R. Lorenzo,"A New Unsupervised Neural Network for Pattern Recognition with Spiking Neurons", 2006 International Joint Conference on Neural Networks Sheraton Vancouver Wall Centre Hotel, Vancouver, BC, Canada July 16-21, 2006 0-7803-9490-9/06©2006 IEEE, 2006*

[8] *Q. Wu," Knowledge Representation and Learning Mechanism Based on Networks of Spiking Neurons", 2006 IEEE International Conference on Systems, Man, and Cybernetics October 8-11, 2006, Taipei, Taiwan 1-4244-0100-3/06@2006 IEEE, 2006*

[9] *M. R. Peterson,"A Satellite Data Set for the Evolution of Data Transforms for Defense Applications",GECCO'07, July 7-11, 2007, London, England, United Kingdom 978-1-59593-698-1/07/0007 pp 2901-2906, 2007*

[10] *F. Chen,"SAR Data Classification Using Case-based Reasoning Method", 1-4244-1212-9/07©2007 IEEE, 2007*

[11] *X.She,"The Boosting Algorithm with Application to Polarimetric SAR Data Classification", 1-4244-1188-2/07 ©2007 IEEE, 2007*

[12] *G. Chang,"Polarimetric SAR Data Classification Based on the Degree of Polarization and Co-polarized Phase-Difference Statistics", Asia-Pacific Microwave Conference 2007 1-4244-0749-4/07@2007 IEEE, 2007*

[13] *V.V.Chamundeeswari," Unsupervised Land Cover Classification of SAR Data by Contour Tracing", 1-4244-1212-9/07 ©2007 IEEE, 2007*

[14] *M. Oster,"A Spike-Based Saccadic Recognition System", 1-4244-0921-7/07 © 2007 IEEE, 200*

[15] *Y.Oh,"Classification Of Polarimetric Sar Data Using The Degree Of Polarization And The Co-Polarized Phase Difference", IGARSS 2008 978-1-4244-2808-3/08©2008 IEEE, 2008*