A SURVEY ON DATA MINING TECHNIQUES FOR CLASSIFICATION OF IMAGES

[1]Preeti lata sahu, [2]Ms.Aradhana Singh, [3]Mr.K.L.Sinha
[1]M.Tech Scholar, [2]Assistant Professor ,[3]Sr. Assistant Professor,
Department of Comp. Sci. & Engg.
Chhatrapati Shivaji Institute of Technology, Durg Chhattisgarh, India
Email: [1]preetitanu5@gmail.com, [2]aradhanasingh@csitdurg.in ,[3]khomlalsinha@csitdurg.in

***Abstract*— With the improvement of internet and the accessibility of picture catching gadgets such as digital cameras, image scanner, the span of digital image accumulation is expanding quickly. This article reviews the current state of classification techniques, compares various classification techniques used to implement on the images such as Decision Tree, Artificial Neural Network, k- Nearest Neighbor, Genetic algorithm, Differential Evolution by highlighting the point of interest and weakness of each of the techniques. Finally, a discussion of the future techniques and methodologies which promise to enhance the ability of computer system to classify the image and current research challenges are pointed out in the field on classification of images.**

***Keywords*— Classification, Decision Tree, Artificial Neural Network, k- Nearest Neighbor, Genetic algorithm, Differential Evolution.**

I. Introduction

The amount of data kept in computer files and databases is developing at revolutionary rate. Data mining is defined as finding hidden information in a database [1]. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use [2]. The data mining model that is created can be either predictive or descriptive in nature.

A predictive model makes a prediction about estimation of information utilizing known results found from diverse information. It may be focused around the utilization of other verifiable information. It includes classification, regression, time series analysis and prediction. A descriptive model recognizes patterns or relationships in data. Clustering, summarization, association rules and sequential analysis are usually viewed as descriptive in nature.

Some basic data mining task are classification, clustering, sequential analysis, association rule. Classification maps data into predefined groups or classes. It is often referred to as supervised learning because the classes are determined before examining the data [1]. Clustering is like classification with the exception of that the groups are not predefined, but rather defined by the data alone. It alternatively referred to as unsupervised learning or segmentation. Sequential analysis is used to determine sequential patterns in data. These patterns are based on a time sequence of actions. These patterns are similar to associations in that data are found to be related but the relationship is based on time. Association refers to the data mining task of uncovering relationships among data. The best example of this type of application is to determine association rules. An association rule is a model that identifies specific types of data association.

There are various commercial ventures that are already using data mining on a regular basis. Some of these organizations include retail stores, hospitals, banks and insurance agencies. Many of these organizations are consolidating data mining with such things as statistics, pattern recognization and other vital tools. Data mining can be utilized to discover patterns and connections that would overall be hard to discover. This technique is well known with many businesses because it permits them to learn more about their customers and make smart marketing decisions. Some other application of data mining is loan/credit card approval; fraud detection in telecommunication, financial transaction; in medicine field, it analyzes patient disease history and finds relationship between diseases.

II. Problem Definition

The aim of the classification process is to categorize all pixels in a digital image into one of several classes. Classification includes a broad range of decision-theoretic approaches to the identification of images. All classification algorithms are based on the assumption that the image in input depicts one or more features and that each of these features belongs to one of several distinct classes. Image classification analyzes the numerical properties of various image features and organizes data into categories. Classification algorithms typically employ two phases of processing: training and testing. In the initial training phase, characteristic properties of typical image features are isolated and, based on these, a unique description of each classification category, i.e. training class, is created. In the subsequent testing phase, these feature-space partitions are used to classify image features [4].

Before classifying the images into classes, image preprocessing is necessary to be done on images. It produces a smooth approximation of the data and performs discontinuity detection. Classification of data is used to assign corresponding levels with respect to groups with homogeneous characteristics, with the aim of discriminating multiple objects from each other within the images classification will be executed on the basis of feature, such as density, entropy, texture etc. in the feature space [5].

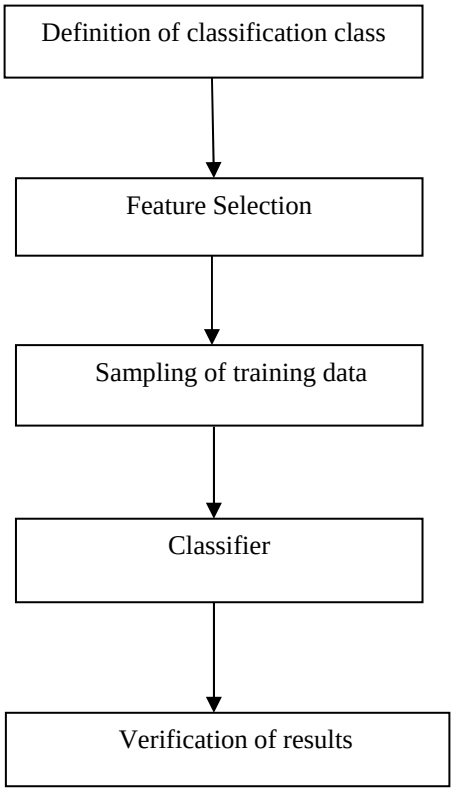

Fig.1 Procedure of Classification

III. Proposed Methodology

There are different techniques of classification of images:

A. Decision Tree
B. Artificial Neural Network
C. k-Nearest Neighbor
D. Genetic Algorithm
E. Bayesian Theorem

a. Decision Tree

Decision tree builds classification or regression models in the form of a tree structure. It breaks down a dataset into smaller and smaller subsets while at the same time an associated decision tree is incrementally developed. The final result is a tree with decision nodes and leaf nodes. A decision node has two or more branches. Leaf node represents a classification or decision. The topmost decision node in a tree which corresponds to the best predictor called root node. Decision trees can handle both categorical and numerical data [6].

b. Artificial Neural Network

Artificial neural networks were initially developed according to the elementary principle of the operation of the (human) neural system. Since then, a very large variety of networks have been constructed. All are composed of units (neurons), and connections between them, which together determine the behaviour of the network. This network consists of three or more neuron layers: one input layer, one output layer and at least one hidden layer. In most cases, a network with only one hidden layer is used to restrict calculation time, especially when the results obtained are satisfactory. All the neurons of each layer (except the neurons of the last one) are connected by an axon to each neuron of the next layer [12].

c . k- Nearest Neighbour

k-nearest neighbors is a simple algorithm that stores all available cases and classifies new cases based on a similarity measure (e.g., distance functions). The idea behind this method is to build a classification method using no assumption about the form of the function, $y=f(x_1,x_2,x_3,...x_p)$ that relates the dependent variable, y, to the dependent variables $x_1,x_2,x_3,...,x_p$. The only assumption we make is that it is a "smooth" function. This is a non parametric method because it does not involve estimation of parameters in an assumed function form such as the linear form that we encountered in linear regression [13]. k-nearest neighbor has been used in statistical estimation and pattern recognition.

d. **Genetic Algorithm**

Algorithm is started with a set of solutions called population. Solutions from one population are taken and used to form a new population. This is motivated by a hope, that the new population will be better than the old one. Solutions which are selected to form new solution are selected according to their fitness - the more suitable they are the more chances they have to reproduce. This is repeated until some condition is satisfied. Genetic Algorithms (GAs) are adaptive heuristic search algorithm based on the evolutionary ideas of natural selection and genetics.

e. **Bayesian Theorem**

Bayesian classifiers are the statistical classifiers. They are able to predict class membership probabilities such as the probability that a given tuple belongs to a particular class. Naïve Bayes classification is based on the Bayes rule. Classification is made by combining the impact that the different attributes have on the prediction to be made. The approach is called naïve because it assumes the independence between the various attribute values [1].

Explaining all of these algorithms is beyond the scope of this paper, there is short introductory information about some of these classification algorithms are shown in table 1.

[Table (1): Different classification techniques]

Methods	Fundamental	Work of Classifier	Pros	Cons	Application
Genetic algorithm	• algorithm is started with a set of solutions called population. Solutions from one population are taken and used to form a new population. This is motivated by a hope, that the new population will be better than the old one. Solutions which are selected to form new solution are selected according to their fitness.	• classifier is a set of rule based system suited to rule discovery algorithm. The rule must lend themselves to process that extract and recombine "building blocks" from currently useful rules to form new rules, and rules must interact simply and in a highly parallel fashion.	• Solve problem with multiple solution. • Easily transferred to existing simulation and models.	• Certain optimization problem cannot be solved by means of genetic algorithm. • There is no absolute assurance that a genetic algorithm will find a global optimum.	• airlines Revenue Management, Audio watermark insertion/ detection, RNA structure prediction.
Artificial Neural Network	• initially developed according to the elementary principle of the operation of the neural system.	• Composed of units and connection between them, which together determine the behavior of the network.	• parallel processing • network learning	• no structured methodology available in artificial neural network. • greater computation burden.	• Airline security control, OCR system, sales forecasting, target marketing, prediction of stock price index.

Table (1): Different classification techniques (continue)]

Methods	Fundamental	Work of Classifier	Pros	Cons	Application
Decision Tree	• powerful, straight forward and easy classification algorithm. • represented by rule if then else condition to classify the data items.	• Recursively partition a dataset of records using depth first approach until all the data item belong to a particular class are identified.	• construction does not require any domain knowledge. • Handle high dimension data. • Implement in parallel and series fashion.	• output attribute must be categorical. • limited to one output attribute.	• in decision making system, teaching, research area.
Bayesian Network	• powerful probabilistic representation. • graphical model. • Also called belief network.	• This classification learns from training data the conditional probability of each attribute. A_i given the class label C. • Classification is then done by applying Bayes rule to compute the probability of C given the particular instances of $A_{i...}A_n$ and then predicting the class with highest posterior probability.	• simplify the computation. • exhibit high accuracy and speed when applied to large database.	• the assumption made in class conditional independence. • lack of available probability data.	• in computational biology and bioinformatic, medicine, document classification, information retrieval, semantic search, image processing, data fusion, etc.
k-Nearest Neighbor	• is one of the best known distance base algorithm • is considered as statistical learning algorithm.	• When given an unknown sample, a k-nearest neighbor classifier searches the pattern space for the k training sample. • It is lazy learning algorithm.	• analytically traceable. • uses local information which can yield highly adaptive behavior. • implement in parallel and simple.	• large storage requirement. • highly susceptible to the curse of dimensionality • slow in classifying test tuple that are closest to the unknown sample.	• in pattern recognition, image databases, internet marketing, cluster analysis, etc.

IV. CONCLUSION

Data mining offers guarantee to uncover hidden patterns within large amounts of data. These hidden patterns can potentially be used to predict future behavior. The availability of new data mining algorithms, however, should be met with caution. First of all, these techniques are only as good as the data that has been collected. Good data is the first requirement for good data exploration. Assuming good data is available, the next step is to choose the most appropriate technique to mine the data. However, there are tradeoffs to consider when choosing the appropriate data mining technique to be used in a certain application. There are definite differences in the types of problems that are conductive to each technique. The "best" model is often found by trial and error: trying different technologies and algorithms. Often times, the data analyst should compare or even combine available techniques in order to obtain the best possible results. In this paper, we studied and summarized the different kinds of classification techniques and their pros and cons.

References

[1] Data Mining Introductory and Advanced Topics, Margaret H. Dunhum

[2] http://en.wikipedia.org/wiki/Data_mining

[3] An Introduction to Data Mining, prof. S. Sudarshan, CSE Dept, IIT Bombay

[4] http://homepages.inf.ed.ac.uk/rbf/HIPR2/classify.htm

[5] http://www.jars1974.net/pdf/12_Chapter11.pdf

[6] http://www.saedsayad.com/decision_tree.htm

[7] http://en.wikipedia.org/wiki/Decision_tree_learning

[8] Rokach, Lior; Maimon, O. (2008). *Data mining with decision trees: theory and applications*. World Scientific Pub Co Inc. ISBN 978-9812771711

[9] http://en.wikipedia.org/wiki/Artificial_neural_network#Real-life_applications

[10] A Compréhensive Study of Artificial Neural Networks, Vidushi Sharma, Sachin Rai, Anurag Dev, Volume 2, Issue 10, October 2012.

[11] Classifier Systems and Genetic Algorithms, L.B. Booker, D.E. Goldberg and J.H. Holland

[12] Artificial Neural Network as a classification method in the behavioural science, David Reby, Sovan Lek, Ioannis Dimopoulos, Jean Joachim, Jacques Lauga, Ste´phane Aulagnier, 1996.

[13] Insight into Data Mining, K.P.Soman, Shyam Diwakar,V.Ajay.