ISSN: 0976-3031

Available Online at http://www.recentscientific.com

CODEN: IJRSFP (USA)

International Journal of Recent Scientific Research
Vol. 10, Issue, 04(E), pp. 31928-31932, April, 2019

International Journal of Recent Scientific Research

DOI: 10.24327/IJRSR

Research Article

A SURVEY ON VISUAL DATAMINING TECHNOLOGIES OF MEDICAL DIAGNOSIS

Harishbabu. Kalidasu and Gudipati Murali

Acharya Nagarjuna University, Guntur, Andhra Pradesh, India

DOI: http://dx.doi.org/10.24327/ijrsr.2019.1004.3365

ARTICLE INFO

Article History:
Received 13th January, 2019
Received in revised form 11th February, 2019
Accepted 8th March, 2019
Published online 28th April, 2019

Key Words:

Data Mining, classification, prediction, pattern recognition

ABSTRACT

Data mining (DM) is the method of ascertaining purposeful correlation, patterns, and trends by transferring through huge data, using recognition technologies. DM accentuates on making and testing algorithms that can support the process of classification, prediction, and pattern recognition. This process makes use of computer models acquired from existing data (previous data) with finite human interaction. The idea is to amplify precision and lessen human biases by means of using automatic pre-programmed methods. As an outcome, a solid and well-grounded functional data mining algorithms can be built to categorize objects or foresee new cases of diseases.

INTRODUCTION

Data Mining

Frawley (2012) details DM as the nontrivial extraction of implicit, formerly unknown, and possible helpful particulars from data. Baxt (1990) defines DM as the course of automating information that has been found out. Moxon (2012) states that data mining is the process of finding out significant new correlation, patterns and trends by relocating through huge sums of data, using pattern recognition technologies as well as statistical and mathematical techniques. Han and Kamber (2012), disputes that DM techniques can be hold to be illustrative (summarize data and to highlight their interesting properties) or predictive (build models to forecast future behaviours).

Machine Learning (ML)

ML is a scientific discipline responsible for recognizing complex patterns and making intelligent decisions based on data. Emphasizing on making and testing algorithms, ML can assist the process of classification, prediction, and pattern recognition using computer modules. ML provides limited human involvement and uses the automatic pre-programmed methods that reduce human biases. The process of proposing the algorithm and its functionality to classify objects or predict new cases are to be built on solid and reliable data, Mitchell, 1977. The database contains a collection of instances (records or case). Each instance used by ML algorithms is formatted using same set of fields (features, attributes, inputs, or variables). When the instances contain the correct output (class label) then the learning process is called the supervised learning. Whilst the process of ML without knowing the class label of instances is called unsupervised learning. (Ozgur, 2004), clustering is a common unsupervised learning method (some clustering models are for both). The goal of clustering is to describe data. However, classification and regression are predictive methods. This research will focus on supervised machine learning.

Classification

Classification is the method of learning the target function that maps between a set of features (inputs) and a predefined class labels (output) i.e. it places data in lone groups that fits in to a common class, inferring the defining characteristics of a certain group done by Regression algorithms which make an effort to map input to domain values. For example, a regressor can forecast some but not all goods sales by bearing in mind the goods characteristics. Simultaneously, classifiers can map the input space into pre-defined classes. As a result, a classifier can foresee a new case of patient whether benign (healthy) or malignant (suffer from a certain disease).Kotsiantis *et al*, 2007, gives a detailed account of supervised ML as the look for algorithms that reason from outwardly supplied instances to

Corresponding author:* **Harishbabu. Kalidasu
Acharya Nagarjuna University, Guntur, Andhra Pradesh, India

produce general hypotheses; the general hypotheses are then used to create predictions about upcoming instances. The objective of supervised learning is to construct a succinct model of the distribution of class labels with respect to predictor features. The resulting classifier is then used to designate class labels to the testing instances where the values of the predictor features are known, and the value class label is unknown

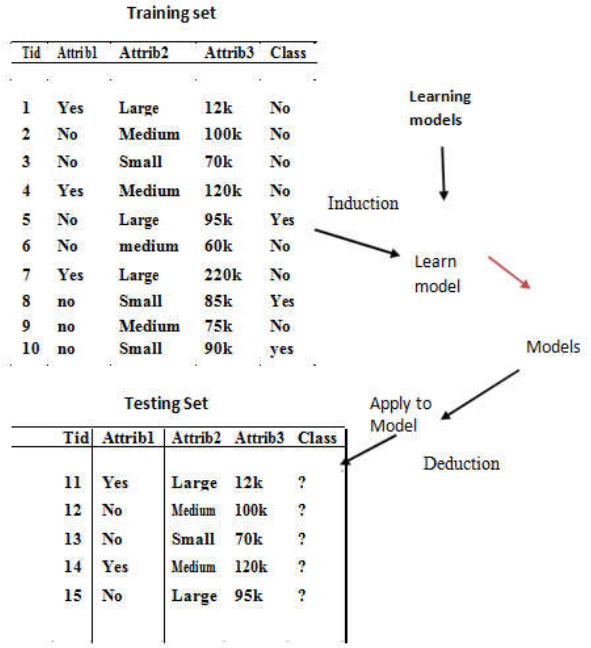

Figure 1.1 Basicapproach to building classification model

(Source: Review of Classification Techniques by Kotsiantis, 2007).

A Classification problem can be solved using the following illustration under;

Presume the goal is to categorize some objects j = 1, 2,…..m into ***k***predefined classes, where***k***stand for the number of classes, i.e. if the aim of classification is to diagnose a patient whether or not suffering from cancer then the value of ***k*** will be **2**corresponding to either benign ormalignant. The database (available data) can be structured as

Table 1.1 The confusion matrix for classifier c(y) on matrix X that contains 160 records

	PREDICTED	
	B	M
Benign (B)	70-71	15-16
Malignat (M)	5-6	90-91

The Error Rate (Er) of algorithm is computed as the total number of wrongly classified samples divided by the total number of records in the matrix X. In the example above,

Er = (15 + 5) / 160 = 0.125. Classification accuracy of the model can be calculated as Accu = 1 − *Er* 0.875……(2.1)

K-Nearest Neighbours Algorithm (K-Nn)

K-NN is an instance based machine learning algorithm that classifies feature space based on the closest training cases. *K-NN* finds the *k* closest instances to a predefined instance and makes a decision its class label by making out the most frequent class label among the training data that have the minimum distance between the query instance and training instances.

The distance metric determines the distance i.e. it minimizes the distance between similar instances and maximizes the distance between different instances. Larose (2013) provides an illustration for *k*-NN functioning as shown in the subsequent pseudo-code to describe this metric distance. Euclidean and Manhattan methods are some amongst quite a lot of of the approaches that are used for distance determination.

Procedure Testing Dataset K NN learner
For each testing instance
{
Find the k most nearest instances of the training set according to a distance metric (Euclidean distance or Manhattan distance)
Resulting Class = most frequent class label of the k nearest instances
}

Figure 1.2 Procedure for K Nn learner

Advantages of K-NN

i. It has a very efficient pattern recognition method and can be easily carried out
ii. Simple to use
iii. Strong against noisy data
iv. Can be used for large and small datasets
v. Suitable for linear and nonlinear functions
vi. Has the ability to add additional instances with no need to train the data set
vii. Its weight is used to measure features significance
viii. Missing values can be easily imputed using *k-NN*
ix. Has excellent flexibility (nonparametric model except the value of *k*) Disadvantages of using k-NN

It requires that the distance between the query instance and all other instances calculated

i. It requires the use of a huge memory
ii. It is not useful for multidimensional dataset because of high error rate
iii. It has the option of using many distance functions which may lead to different accuracy level

Anfis Structure

Adaptive Neural Fuzzy Inference System (ANFIS) has the advantages of Nearest Neighbourhood and Fuzzy Inference System (FIS) by joining rules of expert knowledge domains and it is ability to learn. The 3 major components which constitute FIS and it includes the rule base of selection fuzzy rules; database that explains the function members and a mechanism is a way of inferring a reasonable output.

Rule 1: if $\boldsymbol{x}$ is $\mathbf{A_1}$ and $\boldsymbol{y}$ is $\mathbf{B_1}$, then $f_1 = p_{1x} + q_{1x} + r_1$. ……(2.12)

Rule 1: if $\boldsymbol{x}$ is $\mathbf{A_1}$ and $\boldsymbol{y}$ is $\mathbf{B_2}$, then $f_2 = p_{2x} + q_{2x} + r_2$. ……(2.13)

For clear understanding of the above rules described as follows:

Letting the function members of fuzzy sets should be: A_i, B_i =1, 2, to $\mu A_i \mu B_i$

Now evaluate the above rule then results as:

$$Wi = \mu \ Ai \ (x) \ * \ \mu \ Bi \ (y), \ i=1,2 \qquad (2.14)$$

By evaluating the implication:

$$\text{Rule 1: } \mu \ A_1 \ (x) \quad \mu \ B_1 \ (y) \ f_1(x,y) = W_1(x,y) \ f_1(x,y) \qquad (2.15)$$

$$\text{Rule 2: } \mu \ A_2 \ (x) \quad \mu \ B_2 \ (y) \ f_2 \ (x,y) = W_2(x,y) \ f_2(x,y) \qquad (2.16)$$

By aggregating it becomes

$$f(x,y)=\frac{W1(x,y)f1(x,y)W2(x,y)f2(x,y)}{w1(x,y)+w2(x,y)} \qquad (2.17)$$

After simplifying it becomes:

$$F=\frac{W1f1-w2f2}{w1 \mid w2} \qquad (2.18)$$

Where W_1 and W_2 are FIS inference rule.

X and Y are the input & output functions and f is class represents label.

Breast Cancer Diagnosis Based On IG-ANFIS

ANFIS was first put forward by Jang in 1993. ANFIS can be effortlessly put into practice for a given input/output task. This characteristic makes it striking for various application functions. ANFIS is a grouping of two machine learning approaches; NN and Fuzzy Inference System (FIS). The ANFIS model combines the ANN and FIS tools into a "compound", meaning that there are no limitations to distinguish the respective features of ANN and FIS. ANFIS data mining technique with a pre-processing stage involving IG method intensifies breast cancer dataset's classification accuracy.

Related Work

Data Mining Applications in Medical Diagnosis

Meesad and Yen (2003) proposed a hybrid Intelligent System (HIS) which integrated the Incremental Learning Fuzzy Network (ILFN) with the linguistic knowledge representations. The linguistic rules were determined based on knowledge embedded in the trained ILFN or been took out from real experts. In addition, the method also made use of Genetic Algorithm (GA) to lessen the number of the linguistic rules to uphold high accuracy and consistency.

After being fully constructed, the system could incrementally learn new details in both numerical and linguistic forms. The proposed method was evaluated using Wisconsin Breast Cancer (WBC) Dataset. The results showed that the proposed HIS performed superior than some familiar methods.

Setiono (2006) proposed a method to take out classification rules from trained neural networks and discussed its application to breast cancer diagnosis. He made clear how the pre-processing of datasets get better accurateness of the neural network and the accuracy of the rules since some rules could be extracted from human experience, and may be incorrect. The data pre-processing involved the selection of important attributes and the removal of records with missing attribute values from Wisconsin Breast Cancer Diagnosis WBCD dataset. The rules produced by Setiono's method were more succinct and precise than those generated by previous methods stated in the literature.

Song, 2010 presented an involuntary breast cancer diagnosis, a hybrid system for diagnosing new breast cancer cases in partnership between GA and Fuzzy Neural Network. They showed that many problems having high complication and strong non-linearity with vast data to be analyzed, can use inputs reduction i.e. feature selections methods.

Arulampalam and Bouzerdoum, 2011, proposed a method for diagnosing breast cancer named Shunting Inhibitory Artificial Neural Networks (SIANNs). This was a neural network restored by human biological networks in which the neurons act together among each other's via a nonlinear mechanism called shunting inhibition. The feed forward SIANNs have been applied to numerous medical diagnosis problems and the marks were more positive than those obtained using Multilayer Perceptions (MLPs). SIANNs showed reduction in the number of inputs.

Liao, 2010, proposed a hybrid features selections method alongside with *k*-NN and support vector machine (SVM). This was used to recognize the most noteworthy genes that make obvious the highest capabilities of unfairness between sample classes. First they graded the genes in terms of their expression difference using filter method and then a clustering method based on *k*-

Principles for clustering gene expression data. SVM was applied to check the accuracy of the classification performance of candidate genes. The experimental results demonstrated the effectiveness of their method in addressing the problem.

Vijaysankari and Ramar, 2012, proposed a novel hybrid features selections method to decide on applicable features and throw away irrelevant and redundant features from the original dataset using C4.5 and Naïve Bayes classifier. The efficiency and effectiveness of the proposed method was established through extensive comparisons with other methods using real world data of high dimensionality. Experimental results on datasets make known that the algorithm improved classifier accurateness with a smaller amount error rate.

Hall and Holmes, 2003, presented a standard similarity of several attribute selection methods for oversee classification. Attributes selections is attained by cross-validation the attribute rankings in respect of a classification learner C4.5 and Naïve Bayes. The grades came to an end that features selections methods can improve the performance of some learning algorithms. The findings also concluded that Correlation based feature selection method has generated the best result among six different feature selections methods, but increasing the number of features led to a drop of performance.

Belal and Al Daoud (2005) proposed an algorithm to initiate clusters. The proposed algorithm based on how to find a group of medians which are extracted from dimension with variance of maximum. The algorithm applied on different data sets and achieve good results.

Saeys, 2007 reconsidered the significance of feature selection approach in a set of familiar bioinformatics applications. They concentrated into, the two large input dimensionality, and the tiny sample sizes. The results proved that the features selection applications are primary in dealing with high dimensional applications.

Khan and Ahmed (2004) proposed algorithm computes cluster centers for k-means clustering. This algorithm based on 2 observations that some patterns are similar to each other. That's why the same cluster members irrespective to initial cluster centers. The cluster centers calculated using the above methodology founded to be very close to respective cluster centers. The procedure is applicable to cluster algorithms for continuing data. The results shown as improved solutions, consistent solutions of proposed algorithm.

Hamerly and Elkan, 2002 investigated the behaviour of k-means standard clustering algorithm: Fuzzy k means, Gaussian expectation maximization and other new two variants of k-harmonic means. We show that algorithms behave differently from one another on simple dimensionality low datasets and other image segmentation tasks.

Donald Rubin, 2008 classified the missing feature values from the literature into three types: Missing completely at random, missing at random, and missing not at random.

Missing Completely At Random (MCAR)

MCAR makes out how the missing values happened. Here the probability that a feature value is missing is not related to the feature value or to the value of any other features in the dataset e.g. data may be lost because equipment malfunctioned, the weather was dreadful and could not documentation the observation for a certain day, people got sick, or the data were not entered rightly.

Missing At Random (MAR)

MAR is the case when the existence of missing feature value does not be dependent on the feature value itself and may well depend on other features values in the dataset e.g. a depressed person is more likely not to report income just due to depression.

Howell David 2010 (2009) explains that the most popular methods for dealing with missing feature values are omitting instances, imputation, and expectation maximization. All these methods can be applied in conjunction with any classifier that operates on complete data. These methods are

Omitting Instances

In this method, any record of data that contains missing features values is deleted from the data set. After omitting instances that contain missing features values, classification process run on the remaining instances. The main drawback of this method is discarding important information in some cases. This is not a common method. However, it could be used in cases of a small amount of missing data.

Features Imputation

This is a well-known method for constructing missing features values in the datasets for learning purposes. The imputation method can be divided into two major types: single imputation and multiple imputations. In single imputation, 27 the missing features values are substituted by the correspondence features values according to certain rules such as the features values means, mode, median, or algorithm e.g. the mean imputation calculate the mean of feature f in the dataset that contain values which is then used to fill the features f that has missing values. The scenario for constructing missing features values in multiple imputations is similar to the scenario for single imputation. However, the multiple imputation use more than one value to fill missing features values in the dataset, such as mean of observed feature values, the mode of observed feature values, and regression method. However multiple imputations approach has a number of drawbacks include the computational cost being higher than in single imputation. However, the classification performance (accuracy) is higher than single imputation.

Expectation Maximization (EM)

Expectation Maximization is of the most effective methods for handling missing data. To perform Expectation Maximization; the mean, variance, and covariance are estimated from instances whose data is complete, Moss S and Hancoock E, 2009. Expectation Maximization uses maximum likelihood procedures to estimate regression equations to calculate the relationships between variables.

The Proposed Approach

IG-ANFIS

IG-ANFIS data mining method for Cancer Diagnoses has been used. The approach makes use of the advantages of ANFIS and IG method. The output of IG became the input for ANFIS.

Treating Missing Feature values

The approach for building missing feature values was based on iterative nearest neighbours' and distance metrics. This approach employed weighted k-nearest neighbours' algorithm and propagated the classification accuracy to a certain threshold. Classification accuracy in the constructed dataset was computationally differentiated with original dataset which consists of some missing feature values.

Missing feature values that matters but still missing makes a problem for researchers in data mining applications. Handling unknown attributes values with the most appropriate values is a general unease in data mining and knowledge discovery. It was significant to Construct missing values most supervised and unsupervised data mining they affect the eminence of learning and performance of classification algorithm.

Training ANFIS Model

The method to train ANFIS is the hybrid learning algorithm. This algorithm makes use of the gradient descent method and Least Square Estimate (LSE). Each cycle of the hybrid learning consists of a forward pass and a backward pass. In the forward pass the signal travels forward until Layer 4 and the resulting parameters are recognized using the LSE method. In the backward pass the errors are propagated backward and the premise parameters are well-run by gradient descent.

The process is recurring until it attains the least possible error or a predefined threshold. In other words; the total parameter set is divided into three: S = set of total parameters, S_1 = set of

premise (nonlinear) parameters, S_2 = set of consequent (linear) parameters. So, ANFIS uses a two pass learning algorithm: where S_1 is unmodified and S_2 is computed using a LSE algorithm. In Backward Pass, S_2 is unmodified and S_1 is computed using a gradient descent algorithm such as back propagation. So, the hybrid learning algorithm uses a grouping of steepest descent and least squares to become accustomed the parameters in the adaptive network. The effortless process followed by ANFIS is:

Forward pass: present the input vector - calculate the node outputs layer by layer - repeat for all data A and y formed - identify parameters in S_2 using least squares - compute the error measure for every training pair.

CONCLUSION

This chapter has offered a background study of the important data mining technologies used in the existing research. These problems posed by the current techniques have been recognized. The problem were missing feature values and how to process them, huge features and attributes and how to select the most beneficial ones, taking out accurate diagnostic markers that can predict the early onset of the disease and monitoring of various stages of the disease. IG-ANFIS approach reduced the number of features to the optimal using the IG and the output was then fed as input to ANFIS.

References

1. Frawley, W.J., Piatetsky-Shapiro, G & Matheus, C.J., (1992). Knowledge discovery In Databases: An Overview. *AI Magazine, 13(3*), 57.
2. Baxt, W.G. (1990). Use Of An Artificial Neural Network For Data Analysis In Clinical Decision-Making: The diagnosis Of Acute Coronary Occlusion. *Neural Computation,* 2(4), 480-489.
3. Moxon. B. (1996). Defining Data Mining: The Hows and Whys of Data Mining and How it Differs from Other Analytical Techniques. DBMS Data Warehouse Supplement, Miller Freeman. *Inc., San Francisco,CA*.
4. Ozgur, A. (004). *Supervised and unsupervised machine learning techniques for text documentcategorization* (Doctoral dissertation, Bogazici University).
5. Setiono, R, (2006). Generating Concise and Accurate Classification Rules for Breast Cancer Dignosis, *Artificial Intelligence in Medicine,,* 18(3), 205-219.
6. Song, H, Seun, L., Dongwon, K and Gwitae, P. (2010). New methodology of computer aided diagnostic system on breast cancer, *in Proceedings of the Second international conference on Advances in NeuralNetworks* – volume Part III. Chongquing: Springer – Verlag, 780-789.
7. Singapore cancer registry factsheet, (2012). Most Frequent Cancers in Men and Women, Retrieved from
8. http://globocan.iarc.fr/factsheets/populations/factsheet.asp?uno=900.
9. Vijayasankari, S. And Ramar, K (2012). Enhancing Classifier Performance Via Hybrid Feature Selection and Numeric Class Handling – A Comparative study. *International Journal of Computer Applications,* 41(17), 30-36.
10. Widrow, B and Hoff., M (1989). Adaptive Switching Circuits, in *WESCON Conference Record*. 709-717.
11. Moss S and Hancock E.R.: rationale underpinning expectationandmaximization.http://www.academia.edu/1786470/Teacher_beliefs_about_feedback_within
12. Howell David 2010, Missing feature values http://www.uvm.edu/~dhowell/StatPages/Missing_Data/Missing.html.
13. Meesad, P. and Yen, G. (2003). Combined numerical and linguistic knowledge representation and its application to medical diagnosis. *Component and SystemsDiagnostics, Prognostics, and Health Management* II. 4733,. 98-109.
14. Rubin, D.B., (1976). Inference and missing data., *Biometrika,* 63 (3), 581-592.
15. Saeys, Y., Inza,L and P. Larranaga, A. (2007). Review of feature selection techniques in bioinformatics. *Bioinformatics,* 23 (19), 2507-2517.
16. Hall, M., & Holmes, G. (2003). Benchmarking attribute selection techniques for discrete classdata mining. Knowledge and Data Engineering, IEEE Transactions on, 15(6), 1437-1447.

How to cite this article:

Harishbabu. Kalidasu and Gudipati Murali., 2019, A Survey on Visual Datamining Technologies of Medical Diagnosis. *Int J Recent Sci Res.* 10(04), pp. 31928-31932. DOI: http://dx.doi.org/10.24327/ijrsr.2019.1004.3365

****