Studying the Impact of Handling the Missing Values on the Dataset On the Efficiency of Data Mining Techniques

Bushra M. Hussan ***Ghaida al-Suhal***
College of Science ***College of Engineering***
Amal Hameed Khaleel
College of Science

Abstract

Medical data has potential information for extracting hidden patterns in the data sets.

Classification is form of data analysis that can used to extract models describing important data classes or to predict future data trend. Such analysis can help providing us with a better understanding of the large data.

The diagnosis of a medical from symptoms is one example of classification tasks, in which the classes could be either the various disease states or the possible therapies.

Data cleaning and normalization may improve the accuracy and efficiency of mining algorithms.

In this paper we use two data mining techniques (neural network and decision tree) on a known diabetic dataset to predict the future from the given attributes, and notice the impact of handling the missing value in the dataset at the results.

Key words : Classification, Neural Network, Decision Tree, Normalization.

1. Introduction

With the widespread use of medical information systems that include databases, which have recently featured explosive growth in their sizes, the physicians and medical researchers are faced with a problem of making use of the stored data since the traditional manual data analysis has become insufficient (Sumathi and Sivanandam 2006). Therefore, there is a clear need for semi-automatic methods for extracting knowledge from data. This need has led to the emergence

of a field called data mining and knowledge discovery (KDD) (Ghosh and Jain 2005).

Medical data mining can be used to help in predicting a future patient behavior and to improve treatment programs and help control costs and improve the efficiency of patient care (Sumathi and Sivanandam 2006), since it has great potential for exploring the hidden patterns in the data sets of the medical domain. These patterns can be utilized for fast and better clinical decision making (Devi and Khemchandani 2010).

Medical diagnosis is regarded as an important yet complicated task that needs to be executed accurately and efficiently (Patil and Kumaraswamy 2009). In general, results of a patient classification or prediction task are true only with a certain probability, Therefore, any prognostic system cannot predict always the correct future state but may just give early warnings for the treating physician (Brause 2001).

The real-life data mining applications are attractive since they provide data miners with varied set of problems, time and again. The working on diabetic disease patients databases is one kind of a real-life application. Since the detection of a disease from several factors or symptoms is a multi-layered problem. Therefore, we use the knowledge and experience of several specialists collected in databases towards assisting the diagnosis process of disease (Patil and Kumaraswamy 2009).

2. Knowledge Discovery (KDD) and Data Mining(DM)

To understand the term '*Data Mining'* it is useful to look at the literal translation of the word: *to mine* in English means *to extract*. The verb usually refers to mining operations that extract from the Earth her hidden, precious resources (Giudici 2003).

According to the Gartner Group, "*Data mining* is the process of discovering meaningful new correlations, patterns and trends by sifting through large amounts of data stored in repositories, using pattern recognition technologies as well as statistical and mathematical techniques" (Larose 2005).

Data mining refers to extracting or "mining" knowledge from large amounts of data. The term is actually a misnomer. Remember that the mining of gold from rocks or sand is referred to as gold mining rather than rock or sand mining. thus, data mining should have been more appropriately named "knowledge mining from data," which is unfortunately somewhat long. "*Knowledge - mining*" a shorter term, may not reflect the emphasis on mining from large amounts of data.

Many people treat data mining as a synonym for another popularly used term, Knowledge Discovery from Data , or KDD (Han and Kamber 2006) but now use the term KDD to denote the overall process of extracting high-level knowledge from low-level data [Sumathi and Sivanandam 2006].

Knowledge Discovery has been defined as the 'non-trivial extraction of implicit, previously unknown and potentially useful information from data (Bramer 2007). Data mining is a part of a larger process referred to as Knowledge Discovery in Database (KDD). Knowledge discovery as a process consists of an iterative sequence of the following steps (Bramer 2007, Venugopal, Srinivasa and Patnaik 2009):

Step 1: Data Cleaning .

Step 2: Data Integration .

Step 3: Data Selection.

Step 4: Data Transformation.

Step 5: Data Mining.

Step 6: Pattern Evaluation.

Step 7: Knowledge Presentation.

Steps 1 to 4 are different forms of data preprocessing, where the data are prepared for mining. The data mining step may interact with the user or a knowledge base(Han and Kamber 2006) .

Data mining is the core part of the knowledge discovery in database (KDD) Process.

The goal of the data mining and knowledge discovery process is to develop a set of processing steps that should be followed by practitioners when conducting data mining projects (Cios and Moore 2002).

In general data mining tasks can be broadly classified into two categories: *descriptive* data mining and *predictive* data mining. *Descriptive* data mining describes the data in a concise and summary fashion and gives interesting general properties of the data whereas *predictive* data mining attempts to predict the behavior of the data from a set of previously built data models (Venugopal, Srinivasa and Patnaik 2009).

Data mining adopted its techniques from many research areas, including statistics, machine learning, database systems, neural networks, rough sets, and visualization.

1- **Statistical Approaches.**
2- **Machine Learning Approaches.**
3- **Rough Set.**
4- **Visual exploration**.

Three observations emerge from this four-step process are a listed as follows (Sumathi and Sivanandam 2006):

1. Mining is only one step in the overall process.
2. The process is not linear but involves a variety of feedback loops.
3. Visualization plays an important role in the various steps.

3. Disease Classifications (Diabetes)

Diabetes occurs because the amount of glucose in the blood is too high because the body cannot produce or properly use insulin (Huang, McCullagh, Black, Harper 2004). This is because pancreas does not produce any insulin, or not enough, to help glucose enter the body cells - or the insulin that is produced does not work properly. Insulin is the hormone produced by the pancreas that

allows glucose to enter the body's cells, where it's used as fuel for energy .

There are two main types of diabetes as follows:

I) **Type I Diabetes**, when no insulin is produced at all because the insulin-producing cells in the pancreas have been destroyed. This type of diabetes is always treated with insulin injections.

II) **Type II Diabetes**, previously called non-insulin-dependent diabetes mellitus (NIDDM) (Sumathy, Mythili , Kumar and Nadu 2010). Type II diabetes is when the body either does not produce enough insulin, or the insulin which produces does not work as well as it should (insulin resistance). This type of diabetes is treated with healthy balanced diet (Jayalakshmi and Santhakumaran 2010, Sumathy, Mythili , Kumar and Nadu 2010).

4. Preparing the Data for Classification

The following preprocessing steps may be applied to the data to help improving the accuracy, efficiency, and scalability of the classification process:

1- **Data Cleaning**: This refers to the preprocessing of data in order to remove or reduce *noise* and the treatment of *missing values*. Although most classification algorithms have some mechanisms for handling noisy or missing data, this step can help to reduce confusion during learning (Szajnar and Setlak 2010).

2- **Data Integration:** This is needed to combine data from multiple sources like databases, data cubes, flat files etc. The integration can be done in terms of metadata, correlation analysis, detecting data conflict, and resolving semantic heterogeneity (Larose 2005).

3- **Data Transformation** : The format of data in the repositories may not be suitable for processing. So, the format of the data should be transformed to suitable one for a particular task. This is done for smoothing, aggregation, generalization, normalization, and attribute construction (Larose 2005).

4- **Data Reduction:** To have reduced representation of data sets that is necessary for efficient mining this step is used. This data reduction is done in terms of dimensionality reduction, data compression ,etc (Larose 2005).

4.1 Missing Values

In many real-world datasets, data values are not recorded for all attributes. This can happen simply because there are some

attributes that are not applicable for some instances (Bramer 2007).

The reasons of Missing Values can be summarized as follows (Bramer 2007)

- A malfunction of the equipment used to record the data
- Information that could not be obtained
- Faulty data collection instruments
- Data entry problems
- Data transmission problems

The presence of missing values in a dataset can affect the performance of a classifier constructed using that dataset as a training sample. In fact, missing data is a common problem in statistical analysis. Therefore, the rate of less than 1% missing data are generally considered trivial, 1-5% manageable, mean while, 5-15% require sophisticated methods to handle, and more than 15% may severely impact any kind of interpretation.

There are **several algorithms** for filling or replacement the missing values as follows (Larose 2005, Acu˜na and Rodriguez 2004) :

1. **Replace the missing value with some constant**: Specified by the analyst.
2. **Removal completer (Discard Instances)** : Removes all objects that have one or more missing values.
3. **Mean Completer (Replace by Most Frequent/Average Value)** : Replace missing values for numerical attributes with the mean value of all observed entries for that attribute.
4. **Conditioned Mean Completer** :

Similar to the algorithm described above, but the computations of the mean and mode values are conditioned to the decision classes.

5. **Replace the Missing values with a value generated at random** from the variable distribution observed.
6. **Combinatorial Completer** : This algorithm expands each missing value for each object into the set of possible values.
7. **Conditioned Combinatorial Completer :** It is similar to the one above but the sets of values are conditioned to decision class .
8. **KNN Imputation (*KNNI*)** : In this method the missing values of an instance are imputed considering a given number of instances that are most similar to the instance of interest.

4.2.Data Normalization

Variables tend to have ranges that vary greatly from each other. Therefore, data miners should normalize their numerical variables, to standardize the scale of effect each variable has on the results.

Normalization is particularly useful for classification algorithms involving neural networks, or distance measurements such as nearest neighbor classification and clustering. There are many methods for data normalization include **min-max**

normalization, z-score normalization and normalization by **decimal scaling** [Al Shalabi, Shaaban and Kasasbeh 2005].

In our search we used Min-max normalization performs a linear transformation on the original data and works by seeing how much the field value is greater than the minimum value *min (X)* and scaling this difference by the range .

Let *X* refer to our original field value and X^* refer to the normalized field value, a value of *X* is normalized to *X* by computing:

$$X^* = \frac{X - min(X)}{range(X)} = \frac{X - min(X)}{max(X) - min(X)} \quad (1)$$

X^* has a range from 0 to 1 (Larose 2005).

5.Classification Techniques

There are various techniques for supervised classification techniques. In this research we focus on neural networks, decision trees techniques.

5.1. Artificial Neural networks(ANN)

The artificial Neural Networks are very popular techniques in many areas such as pattern recognition, data mining and machine learning (Larose 2005). Specifically the neural networks, can be used for many purposes like descriptive and predictive data mining (Bramer 2007).

5.1.1. Artificial Neural Networks in Data Mining (Venugopal, Srinivasa and Patnaik 2009)

The characteristics that make neural networks are good for classification and prediction in data mining can be listed as follows:

- ✓ NNs have high tolerance of noisy data.
- ✓ NNs have the ability to classify patterns on which they have not been trained.
- ✓ NNs have the ability to classify patterns on which they have little knowledge of the relationships between attributes and classes.
- ✓ NNs are well-suited for continuous-valued inputs and outputs.
- ✓ Neural network algorithms are inherently parallel parallelization techniques can be used to speed up the computation process.

5.1.2. **Back Propagation Algorithm**

Artificial neural network(ANN) have been successfully used to solve classification problems in several domains, specifically the back propagation algorithm is very often the favorite to train feed forward neural networks. Back propagation algorithm is a classical domain dependent technique for supervised training. It works by measuring the output error calculating the gradient of this error, and adjusting the ANN weights and biases in the descending gradient direction (Jayalakshmi , Santhakumaran 2010).

BP neural network suffer from some problems such as local minimum, slow convergence speed and convergence instability in its training procedure, so we

use Learning rate parameter to adjustment weights and momentum parameter to overcome local minima (Venugopal, Srinivasa and Patnaik 2009).

learning algorithm (BP)

There are many factors that affect at the learning algorithm of back propagation and those used in our work are:

1. Learning Rate

The learning rate is a constant chosen to move the network weights toward a global minimum for sum of squared errors. The value of learning rate is choose between $0 < n < 1$ (Larose 2005).

2. Momentum

The back-propagation algorithm is made more powerful through the addition of a momentum term α [(Bramer 2007)]. The momentum factor belonging to [0, 1] (Karray and de Silva 2004).

To avoid the slow convergence of the algorithm, researchers have devised a modified weight updating algorithm in which the change of the weight of the upcoming iteration (at time t+1) is made dependent on the weight change of the current iteration (at time t).

3. Sigmoid Activation Function

The sigmoid function combines nearly linear behavior, curvilinear behavior, and nearly constant behavior, depending on the value of the input. depending on the location of x

The sigmoid function is sometimes called a squashing function, since it takes any real-valued input and returns an output bounded between zero and one.

We used this topology

The number of units in the input layer is equal to the number of attributes in the database = 8

The number of hidden layers = 1

The number of units in each hidden layer = 6

The number of units in the output layer =1 (It classifies it as diabetes or not diabetes)

The total error = 0.0000001

The learning rate =0.28

The momentum =0.8

Normalizing the input values for each attribute measured in the training tupelos will help speed up the learning phase. Typically, input values are normalized so as to fall between 0.0 and 1.0, even for categorical variables therefore we used max-min normalization method.

Neural networks can be used for classification (to predict the class label)

For classification, one output unit may be used to represent two classes (where the value 1 represents one class, and the value 0 represents the other). If there are more than two classes, then one output unit per class is used , Therefore we need to one output unit to diagnose diabetes.

70% or (500) of the samples are used for training and 30% or (268) is used for testing.

The learning algorithm of BP neural network is described as follows:

1.Create an architecture consists of three layers: input layer, one hidden layer and output layer.

2.Replace the missing data by the values calculated from hybrid k-nearest neighbor Imputation approach.

3.Normalize the input data using max-min method.

4.Initialize the network parameters.(learning rate =0.3, Momentum=0.8)

5. Initialize weights by random values.

6.Repeate

{**Feed forward stage**}

6.1. For each node in the input layer, assign the inputs.

6.2 For each node in the hidden layer calculate the output.

6.3. For each node in the output layer calculate the output.

{**Back propagation stage**}

6.4.For each output unit calculate its error.

6.5.For each hidden unit calculate its error.

{**Update weights**}

6.6. Update weight for each output unit .

6.7. . Update weight for each hidden unit.

until the maximum epochs are reached or the desired output is identified or the minimum gradient is reached.

5.2. Decision Tree Approach

Decision trees (DT) are one of the fundamental techniques used in data mining. They are tree-like structures used for classification, clustering, feature selection, and prediction (Berry and Browne2006).

A decision tree is a flowchart-like tree structure that represents the knowledge for classification , where each internal node (no leaf node) denotes a test on an attribute, each branch represents an outcome of the test, and each leaf node (or *terminal node*) holds a class label and the topmost node in a tree is the root node (Han and Kamber 2006) .

For a given record, the classification process starts from the root node. The attribute in the node is tested, and the value determines which edge is to be taken. This process is repeated until a leaf is reached. The record is then classified as the class of the leaf. Decision tree is a simple knowledge representation for a classification model, but the tree may be very complicate and difficult to interpret (Wong and Leung 2002). Decision trees accept several types of variables: nominal, ordinal, and interval. A variable can be of any type regardless of whether it serves as an input or as the target.

Decision trees are easily interpretable, amenable to graphical display, and intuitive for humans and fast and usually produce high-quality solutions

(Berry and Browne 2006).

5.2.1. The Requirements of DT algorithms

There are several requirements must be met before decision tree algorithms may be applied:

1. Decision tree algorithms represent supervised learning, and as such require preclassified target variables.

2. The training data set should be rich and varied.
3. The target attribute classes must be discrete. That is, one cannot apply decision tree analysis to a continuous target variable. Rather, the target variable must take on values that are clearly demarcated as either belonging to a particular class or not belonging (Larose 2005).

5.2.3.Learning algorithm of (DT)

Input: the records in the dataset each one with 8 attributes.

Output: decision rules.

Steps:

1.Initialize the empty tree, the root is the current node;

2.Repeat

2.1. Calculate the entropies for each attribute;

2.2. Calculate the profit for each attribute;

2.3.Choose the maximum profit;

2.4 Check the test for the current node;

2.5.If the node is final then affect to a class otherwise select a test and create the under tree.

Until obtaining a decision tree.

6. Discussion and Conclusion

Our goal is to observe the impact of data mining technique on diabetic dataset, in this study we have implemented two algorithms of classification, neural network and decision trees on the PIMA indian diabetic dataset(which has 768 record each with 9 numeric variables(number of pregnant, OGTT plasma glucose, diastolic blood pressure, triceps skin fold thickness, serum insulin, BMI, diabetes pedigree function, age, diabetes onset(0, 1)) before and after handling missing values by hybrid k-nearest neighbor imputation method. also the goal is to use the first 8 variables to predict the 9th attribute value The performance of the algorithms is evaluated based on the accuracy and the Area Under Curve(AUC):s.

Table (1) shows Some scenarios of training with different setup.

We noticed from previous training scenarios that momentum has little effect on accuracy so we kept momentum 0.8, and tried with several smaller learning rates. Note that lower learning rate like 0.28 improves the accuracy but takes much time than a higher learning rate like 0.6 which make the NN learn faster. when increase the number of iterations and the learning rate and momentum rate is constant then the accuracy increase but no large.

Table (2) represent the Performance of the approaches on the dataset with missing values.

Table (3) represent Performance of the approaches on the dataset without missing value.

Figure(1)and Figure(3) represent the difference of the areas under the curve for the values of the dataset by using NN before and after handling missing values.

Figure(2) and Figure(4) represent the difference of the areas under the curve for the values of the dataset by decision tree before and after handling missing values.

Table (1)

Training	Training Instances	Iterations	Learning Rate	Momentum	Stopping Criteria	Sqr Error Convergence	Accuracy (%)
1	500	10000	0.3	0.8	yes	0	70 %
2	500	10000	0.3	0.9	yes	10^{-4} or when recognized class	70 %
3	500	20000	0.3	0.8	No	10^{-5} or when recognized class	60 %
4	500	30000	0.3	0.8	Yes	10^{-5} or when recognized class	60 %
5	500	100000	0.3	0.8	Yes	10^{-5} or when recognized class	65 %
6	500	100000	0.6	0.8	Yes	10^{-5} or when recognized class	65 %
7	500	100000	0.3	0.8	Yes	10^{-6} or when recognized class	75%
8	500	100000	0.28	0.8	Yes	10^{-6} or when recognized class	77%
9	500	1000000	0.1	0.8	Yes	10^{-6} or when recognized class	77.6 %
10	500	50000	0.3	0.6	Yes	10^{-6} or when recognized class	75 %
11	500	100000	0.3	0.9	Yes	10^{-6} or when recognized class	75.7 %
12	500	100000	0.6	0.5	Yes	10^{-6} or when recognized class	75.7 %

Table (2)

Methods / Performance	Neural network	Decision tree
Precision	**0.79365**	**0.71429**
Recall (Sensitivity)	**0.80214**	**0.808389**
Specificity	**0.51852**	**0.46535**
Accuracy	**71.64179**	**67.91045**
F_measure	**0.79787**	**0.75843**

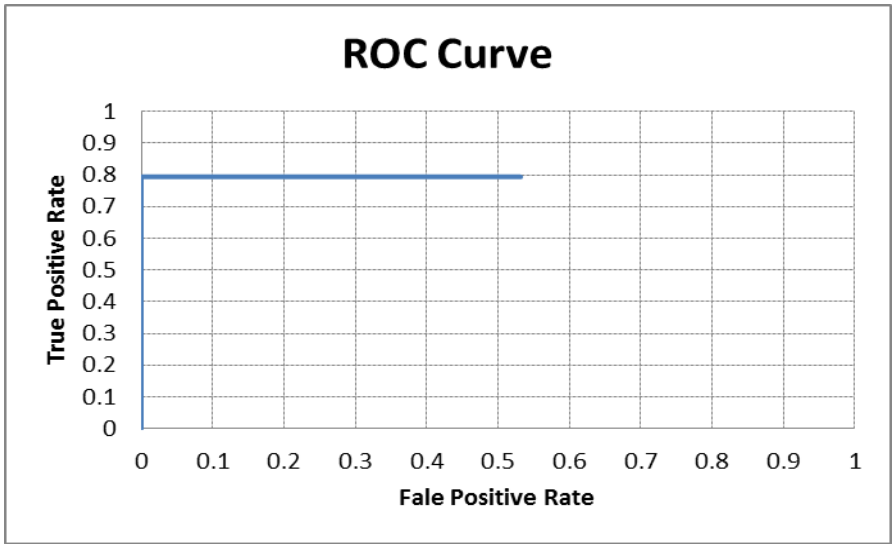

Figure (1)

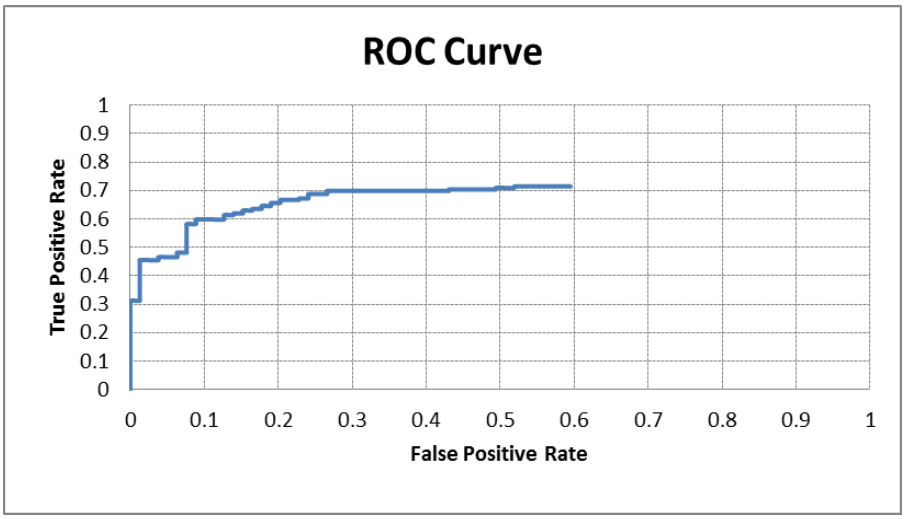

Figure (2)

Table (3)

Methods / Performance	Neural network	Decision tree
Precision	**0.962963**	**0.8783**
Recall (Sensitivity)	**0.77447**	**0.8783**
Specificity	**0.78788**	**0.70886**
Accuracy	**77.61194**	**82.83582**
F_measure	**0.85849**	**0.8783**

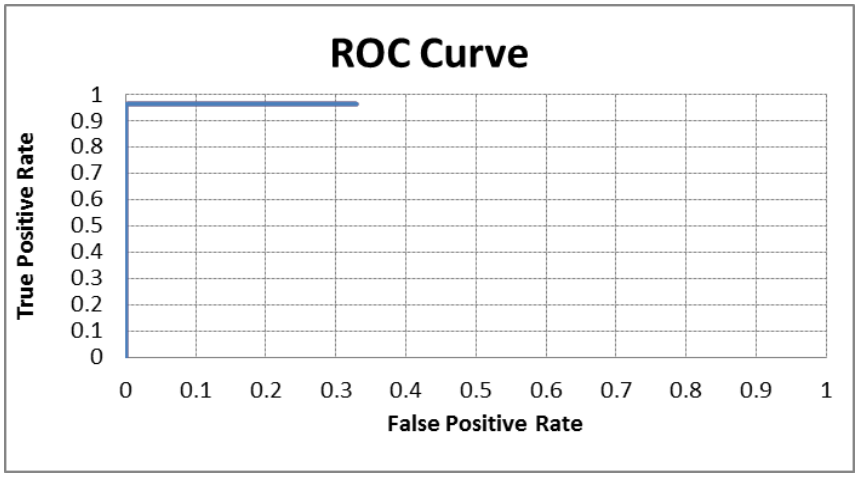

Figure (3)

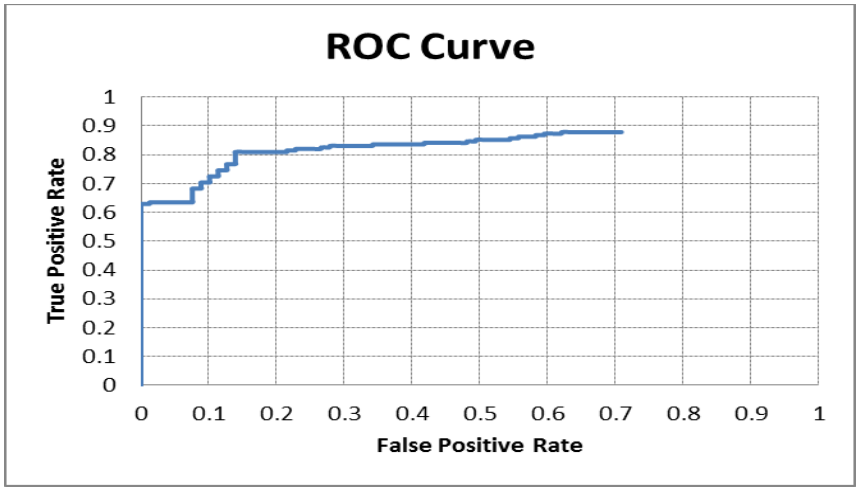

Figure (4)

References :

A.Ghosh and L.C. Jain, "***Evolutionary Computation in Data Mining***", published by Berlin Heidelberg, Germany, 2005.

D. T. Larose, "***Discovering Knowledge in Data***", published by John Wiley & Sons, United States of America, 2005.

D. T. Larose, "***Discovering Knowledge in Data***", published by John Wiley & Sons, United States of America, 2005.

E. dgar Acu˜na and C. Rodriguez , "**The treatment of missing values and its effect in the classifier accuracy**", PP. (639-648), 2004.

F. Karray and C. de Silva , "***Soft Computing and Intelligent Systems Design***", published by Pearson Education Limited, England, 2004.

J. Han and M. Kamber, "***Data Mining: Concepts and Techniques***", Second Edition, published by Elsevier, United States of America, 2006.

K.R.Venugopal, K.G. Srinivasa, and L.M. Patnaik, "***Soft Computing for Data Mining Applications***", published by Berlin Heidelberg, German, 2009.

K. J. Cios and G. W. Moore, "**Uniqueness of Medical Data Mining**", In proceeding of Artificial Intelligence in Medicine journal, Vol. (26), PP. (1-24), 2002.

K.R.Venugopal, K.G. Srinivasa, and L.M. Patnaik, "***Soft Computing for Data Mining Applications***", published by Berlin Heidelberg, German, 2009.

L. Al Shalabi, Z. Shaaban, and B. Kasasbeh, "**Data Mining: A Preprocessing Engine**", In proceeding of Journal of Computer Science Vol. (2), No. (9), PP. (735-739), 2006.

M. Bramer, "***Principles of Data Mining***", published by London Limited, England, 2007.

M. Berry and M. Browne, "***Lecture Notes in Data Mining***", published by World Scientific, United States of America, 2006.

M. L. Wong and K. S. Leung , "***Data Mining Using Grammar Based Genetic Programming and Applications***", published by Kluwer Academic, United States of America, 2002.

P. Giudici, "***Applied Data Mining***", published by John Wiley & Sons, England, 2003.

R. Devi and V. Khemchandani, "**Application of Data Mining Techniques For Diabetic Dataset**", Proceedings of the 4th National Conference, Computing For Nation Development, Bharati Vidyapeeth's Institute of Computer Applications and Management, New Delhi, February 25 – 26, 2010.

R. W. Brause, "**Medical Analysis and Diagnosis by Neural Networks**", In proceeding of Lecture Notes in Computer Science , Vol. (2199), PP.(1-13), 2001.

S. Sumathi and S.N. Sivanandam, "***Introduction to Data Mining and its Applications***", published by Berlin Heidelberg, German, 2006.

S.B.Patil and Y.S.Kumaraswamy, "**Intelligent and Effective Heart Attack Prediction System UsingData Mining and Artificial Neural Network**", In proceeding of European Journal of Scientific Research, Vol. (31), No. (4), PP. (642-656), 2009.

Sumathy, Mythili , P. Kumar, and T. Nadu , "**Diagnosis of Diabetes Mellitus based on Risk Factors**", In proceeding of International Journal of Computer Applications, Vol. (10), No. (4), PP. (1-4), 2010.

T.Jayalakshmi and A. Santhakumaran, "**Improved Gradient Descent Back Propagation Neural Networks for Diagnoses of Type II Diabetes Mellitus**", In proceeding of Global Journal of Computer Science and Technology, Vol. (9), No. (5), PP. (94-97), 2010.

T.Jayalakshmi and A. Santhakumaran, "**Improved Gradient Descent Back Propagation Neural Networks for Diagnoses of Type II Diabetes Mellitus**", In proceeding of Global Journal of Computer Science and Technology, Vol. (9), No. (5), PP. (94-97), 2010.

Y.Huang, P.McCullagh, N. Black, and R. Harper," ***Feature Selection and Classification Model Construction on Type 2 Diabetic Patient's Data*** ", published by Berlin Heidelberg, German, 2004 .

W. Szajnar and G. Setlak, "**A Concept of Design Process of Intelligent System Supper Ting Diabetes Diagnostics**", In proceeding of Methods and Instruments of Artificial Intelligence, PP. (168-178), 2010.

دراسة اثر معالجة القيم المحذوفة في قاعدة البيانات على كفاءة تقنيات تنقيب البيانات

بشرى محمد حسن
كلية العلوم
قسم علوم الحاسبات

غيداء عبدالرزاق
كلية الهندسة
قسم هندسة الحاسبات

امل حميد خليل
كلية العلوم
قسم علوم الحاسبات

البيانات الطبية تمتلك معلومات كامنة لاستخلاص انماط مخبأة في قواعد البيانات.

التصنيف هو شكل من تحليل البيانات لاستخلاص انماط لوصف اصناف البيانات المهمـــــة او لتخمين اتجاه البيانات المستقبلي. بعض التحليلات تساعد بتزويدنا بفهم جيد للبيانات الكبيرة التشخيص الطبي من الاعراض، هو احد الامثلة من مهمات التصنيف التي تكون فيها الاصناف الناتجـــة اما حالات المرض المختلفة او الحالة الممكنة.

تنظيف البيانات وتسويتها يحسن الدقة والكفاءة في خوارزميات التنقيب.

في هذا البحث تم استخدام تقنيتين (الشبكات العصبية وشجرة اتخاذ القرار) على قاعدة بيانات عالمية معروفة لمرض السكر لتخمين المستقبل من الصفات المعطاة وملاحظة اثر معالجة القيم المحذوفة بقاعـــدة البيانات على النتائج.