International Journal of Intelligent Computing and Information Science

DATA CLEANINGTOOL: USAGEOFFUZZYROUGHSETTHEORY AS MACHINE LEARNINGPRE-PROCESSING

B. I. Hameed A.. A. Elfetouh M. Abu_Elkheir

Information System Department Information Technology Department

Faculty of Computers and Information, Mansoura University-Egypt

basharibh78@gmial.com elfetouh@mans.edu.eg mfahmy78@mans.edu.eg

Abstract: *Real-world data is often incomplete, inconsistent, and/or lacking in certain behaviors or trends, and is likely to contain many errors. Data preprocessing is a crucial phase in the data mining process that involves techniques toresolve such issues. Feature selection is a popular data preprocessing procedure that is focused on omitting attributes from decision systems while still maintain the ability of those decision systems to distinguish different decision classes. A popular way to evaluate attribute subsets with respect to this criterion is based on the notion of dependency degree. In this paper, we conduct an experimental study using the generalized classical rough set framework for data-based attribute selection and reduction, based on the notion of fuzzy decision reducts to evaluate the viability of using Fuzzy rough subset feature. Experimental results shows that, general optimization can be achieved under average accuracy reduction, ±10.7 %, against high reduction rate over attributesranging from 36% to 97% and over instances from 1.7% to 44%.*

Keywords:*Fuzzy Rough, Data preprocessing, Data Mining*

1. Introduction

Every day is 86400 seconds, every second in computing environment costs large amount of data to be kept as a transactions for people everywhere; home, internet, organizations, etc. This rate makes a treasure of data to be mined and leads the future with smart decisions. Such decisions depend on mining history and extract implicit information and training to predict the future. So these data should be correct, integrated and consistent for prediction and generate helpful information for decision making process. Completeness is a hard goal can't easily achieve, so that working on refining data mistakes is opportunity for making a try to make data ready for prediction use and generate a perfect information for decision makers. Because of such explosive growth of government, business, and scientific transactions there were traditional, manual approaches to data analysis and a new optimized generation of techniques are proposed by many researchers as intelligent and automated knowledge discovery strategies in data [1, 2].

Data Pre-processing is a process which is used to optimize data quality to make data error-free during warehousing and mining process i.e. quality of data needs to be improved by using the data cleaning techniques. Existing data cleaning techniques are used to identify record duplicates, missing values, record and field similarities, and duplicate elimination [3]. Data quality issues are often multifaceted

and complex, and it is crucial for information management departments to build applications that support the goal of achieving high-quality data within an organization [4].

As data cleaning includes several types of cleaning for different types of data mistakes such as data record duplication, missing values, inconsistency, data integrity...etc. For different types of data cleaning there are many different types can be used to overcome the problem. All of efforts which is spent to clean data focus on reducing time and increase accuracy of the generated data. Because of data cleaning is a preprocess of knowledge discovery process, the generated report must be generated on the time for gaining it's value and being effective at the decision making level of any organization hierarchy. Different types are used to solve noise of the data, most recent strategies are powered by soft computing techniques including genetic algorithms, genetic programming, fuzzy sets, rough sets and much more rather than introducing hybrid strategies such as Fuzzy Rough, Neru-Rough,…….etc. In these techniques data tables are considered as a decision table or after transformation of data there is one that is constructed from different sources and need to go through mining process for extract useful information. In order to work with soft computing techniques especially fuzzy sets, rough sets or hybrid of them, a decision table is considered as input for the mining process, i.e. input of the problem. Model is constructed after both feature selection and instance reduction are made up respectively as cleaning preprocess activity. So, Fuzzy, Rough or Hybrid are considered as a tool for cleaning when perform feature selection and instance reduction. In Feature selection, all of features are examined against each other and the decision itself to make ranking, preceding of each feature and classify each is dependent/core and independent/reduct so that weak, invalid, redundant and useless attribute is eliminated without affecting the information gain provided by the decision table. Feature selection can be classified into two different style wrapper and filter style.Wrapper style, conjectures the precision of features subset based on using a statistical re-sampling mechanism (such as cross validation) by actual machine learning algorithm. Whereas instance reduction, reduce total amount of duplicated record, ambiguity instance and timely waste instance from data set for accelerating model built up process via training process without affecting consistency downward at all and increasing up the model accuracy in some sensitive matter. In filter style, executes any induction algorithm solely to filter out any undesirable features before induction launches. Filter methods typically get benefit of all the training data when selecting a subset of features.Filters have proven to be much faster than wrapper that is late to carry out induction based on frequencies calculations. Another method eliminates features whose information content relative to decision class is still the same using the remaining features [5-6].

The main goals of this paper is to measure and evaluation of cleaning data using most recent hybrid soft computing, Fuzzy Rough strategy as data cleaning tool on two different major perspective, attributes and instance with the missing values and miss classified. In section 2, basic information of two soft computing techniques are used and explained in a simple manner to show pros and cons of different strategy as impact on data cleaning. In Section 3, short term survey of recent researcher's activities and publications over data cleaning and related applications. In section 4,description and discussion of Fuzzy rough techniques and how can be applied and used as data cleaning tools. In section 5,experimental results, two different data sets are used to evaluate the performance explaining the impact of using Fuzzy Rough as data cleaning tool. Then in section 6, the work is concluded with final effective words of how Fuzzy Rough strategy is a data cleaning tool.

2. Preliminaries

2.1. Fuzzy Sets Theory

Classical set or crisp set was defined as a collection of elements. In which, each element either belongs to the set or not, so that each element can be element with a grade as output of a function,$\mu_A(x)$, which considered as characteristic function of discrete distinct values 0 or 1; no other values are allowed [6]. This concept has compatibility with digital system, is not natural to human beings perceptions avoiding nothing between two extreme values so that membership can be enhanced and developed to be fuzzy not just 1 or 0.

Mathematically, fuzzy set has a membership function that allows various degrees of membership for the elements of a given set. If X is a collection of objects denoted by x, then a fuzzy subset Ain X is set of ordered pairs such that

$$\mathrm{A} = \{ x, \mu_{\mathrm{A}}(x) \mid x \in \mathrm{X} \} \tag{1}$$

Where$\mu_A(x)$ is the membership function of x in A [7]. The recent definition can be clarified a fuzzv set$A = \{(1,0.3), (2,0.6), (3,1.0), (4,0.8), (5,0.3)\}$.

Fuzzy relation in A is a fuzzy set in$A \times A$. For all a in$A$, R is a relation of $a$is the fuzzy set denoted as Ra that is defined as follow $Ra(x) = R(x, a)$ for all $a$in$A$. R can be characterized as reflexive, $\mathrm{R}(a, a) = 1$, and is symmetric,$R(a, y) = R(y, a)$ where a and y are elements of A. R is called a fuzzy tolerance relation if and only if the relation is reflexive and symmetric.

For fuzzy sets X and Yin$A, X \subseteq Y \iff (\forall_a \in A)\big(X(a) \leq Y(a)\big)$. If A is finite fuzzy set, the cardinality of A is denoted as $|A|$which is defined as [6]

$$|\mathrm{A}| = \sum_{x \in \mathrm{X}} \mathrm{A}(x) \tag{2}$$

A triangular norm (t-norm for short)$T$is a relation defined as increasing, commutative and associative mapping that is satisfying $T(1, x) = x$, [0,1]2 --> [0,1], for all x in [0,1]. So that, $T(0,0) = 1$ and $T(1, x) = x$ for all x in[0,1]. $T_M and T_L$can be defined as [8]

$$\begin{cases} \mathrm{T_M}(x, \mathrm{y}) = min(x, \mathrm{y}) \\ \mathrm{T_L}(x, \mathrm{y}) = max\ (0, x + \mathrm{y} - 1) \end{cases} \text{For } x, \text{y in [0,1]} \tag{3}$$

In turn, an implicatorI_M and I_L are considered as definition by

$$\begin{cases} \mathrm{I_M}(x, \mathrm{y}) = 1\ if\ x \leq \mathrm{y}\ \ and\ \mathrm{I_M}(x, \mathrm{y}) = y \\ otherwise,\ \ \mathrm{I_L}(x, \mathrm{y}) = min(1, 1 - x + \mathrm{y}) \end{cases} \text{For } x, \text{y in [0,1]} \tag{4}$$

2.2. Rough Set Theory

Basically, rough set analysis is based on using information table, (X, A), which is classified as set of universe of discourse $X = \{x_1 , \ldots , x_n\}$ and set of condition attributes denoted as $A = \{a_1 , \ldots , a_m\}$ that are both are finite, non-empty sets. Each a in A corresponds to an$X \rightarrow V_a$ mapping where $V_a$denotes the set of values of a over X [8].

For any subset of attributes denoted as $B$of A, the $B$is considered as indiscernibility relation denoted by $R_B$which is defined as

$$\mathrm{R_B} = \{(x, \mathrm{y}) \in \mathrm{X}^2 \text{ And } (\forall_{\mathrm{a}} \in \mathrm{B}(\mathrm{a}(x) = \mathrm{a}(\mathrm{y}))\} \quad (5)$$

$R_B$is an equivalence relation. $[x]_{R_B}$is considered as equivalence classes that is used to approximate concept via slicing the universe X as subset. There are two major approximations; lower and upper can be considered for every given relation $A \subseteq X$ and given indiscernibility relation RB lower approximation denoted by $RB{\downarrow}A$ can be defined as math relation as follow

$$\mathrm{R}B{\downarrow}\mathrm{A} = \{x \in \mathrm{X} \mid [x]_{\mathrm{R_B}} \subseteq \mathrm{A}\} \quad (6)$$

And upper approximation denoted by $RB{\uparrow}\ A$ can be defined as math relation as follow

$$\mathrm{R}B{\uparrow}\ \mathrm{A} = \{x \in \mathrm{X} \mid [x]_{\mathrm{R_B}} \cap \mathrm{A} \neq \emptyset\} \quad (7)$$

Positive region is considered the same as lower approximation $RB{\downarrow}A$ which is denoted by POS_B. Positive region contains the objects for which the valuesallow to predict the decision class clearly. The predictive ability about decision d of the attributes B can be measured by a math relation called degree of dependency of d on B and defined as

$$\gamma_{\mathrm{B}} = \frac{|\mathrm{POS_B}|}{|\mathrm{X}|}(\mathrm{B}) \quad (8)$$

Finally, using both recent defined approximation, upper and lower, a boundary region can be defined and measured as subtraction relation of upper and lower respectively[9]

$$\mathrm{BNR_B} = \mathrm{R}B{\uparrow}\ \mathrm{A} - \mathrm{R}B{\downarrow}\mathrm{A} \quad (9)$$

A decision system $(X, A \cup \{d\})$ is a derived version of information system as defined by rough sets. Decision table is used for classification purposes, in which $d(d \notin A)$ is considered attribute called decision attribute based on the values v_k that d was assigned. Decision attribute classifies the instance objects available in the universe of discourse into finite set relative to finite set of the valuesV_d, so that X is partitioned into a number of decision classesX_k where k is the distinct values of V_d [8, 9].
For R_A and R_B are indiscernibility relation where $B \subseteq A$ is called a decision reduct if $PSO_B = PSO_A$ i.e, B preserves the decision making power of A, and if it cannot be further reduced i.e., there exists no proper subset B of B such that $PSO_B = PSO_A$[8].
For Example, next table is decision table with attribute set A = {Headache, temp} and decision d = {Flu} whereas universe of discourse U={u_1,.., u_8} assuming that B is a relation defined as B = {u | Flu(u) = no}.
U/R = {{u1}, {u2}, {u3}, {u4}, {u5, u7}, {u6, u8}}
B = {u1, u4, u5, u8}
$PSO_B = RB{\downarrow}$A = {u1, u4}
$RB{\uparrow}\ A$ = {u1, u4, u5, u6, u7, u8}
$NEG_B = U - R\mathrm{B} \uparrow A$ = {u2, u3}
$BNR_B = RB{\uparrow}\ A\ - RB{\downarrow}A$ = {u5, u6, u7, u8}

Table 1: Sample Decision Table

U	Headache	Temp	Flu
U_1	Y	Normal	No
U_2	Y	High	Yes

U_3	Y	V-High	Yes
U_4	N	Normal	No
U_5	N	High	No
U_6	N	V-High	Yes
U_7	N	High	Yes
U_8	N	V-High	No

3. Related Work

Due to the big growth of data and file, many data types and different applications of organization leads to increasable styles of data mining process including major large number of styles for data cleaning behavior that is so related to application area of mining domain. Many researchers' works on mining preprocessing in different fields using various strategy of building mining model. Here some common recent works were survived in short comparison. Joaquin et al proposed a hybrid model. Basically, it is a steady-state GA for IS where, every time a fixed number of evaluations has beenspent, an RST based FS procedure is applied to modify the features considered during the search. Therefore, at any time onlya single feature subset will be used in the whole search procedure. As the search progresses, this subset will be upgraded andadapted, to fit with the best subset of instances found. Its main steps: Initialization (Step 1), feature selection procedure (Step 4), Instance Selection procedure (Step 5), and Output (Step 7). The rest of the operations (Steps 2, 3 and 6) control whether each of theformer procedures should be carried out [10].

Acharjyaetalproposedan association rule prediction model that consists of preprocess and post process. They processed the data after data cleaning by using rough set on fuzzy approximation space and ordering rules. Based on the classification obtained in preprocess, Bayesian classification is used in post process to predict the missing association of attribute values. The main advantage of this model is that, it works for both literature and numerical data [11].

The Quick Reduct Algorithm is an efficient algorithm for finding reducts [12]. This is widely used is several soft computing implementations using Rough Sets. Algorithm attempts to calculate a reduct without exhaustively generating all possible subsets. It startsoff with an empty set and adds in turn, one at a time, those attributes that result in the greatest increase in the rough setdependency metric, until this produces its maximum possible value for the dataset.

Sharma and Kumar proposed a method which first reduces the size of the database to be searched, and the remainder reduces the number of candidates. Before giving pruning rule about equivalent candidates, they restrict the term equivalent candidate XΦY to a canonical form such that X is always the candidate that is generated earlier than candidate Y. For example, for a relational schema R={A, B, C, D}, canonical form corresponds to alphabetical order. So AΦB and BC→D are in canonical form, but DΦB and CB→D are not. The Purpose is to discover all functional dependencies in a dataset [13].

Elgamal et alproposed system with extensibility as the central idea, and it enables users to customize the data cleaning operations to meet their needs, rather than try to adapt to the rules set forth by the system. Each step of the framework is well suited for the different purposes. Some of the data cleaning techniques are suited for the particular work of the data cleaning process. In addition, the framework offersthe user interaction by selecting the suitable algorithm. The framework start with Removing unimportant characters such as (special characters, title or salutation, ordinal forms and common words),then Expand abbreviations using Reference table, Check the type of row (j) if it a numeric type convert string into number, sort number and put into LOG table,Else if it alphabet type select first

character of every word, sort the characters in alphabetic order, combine them together to obtain the alphabetic token and put into LOG table, Else ifitisalphanumeric type split alphanumeric to numeric and alphabetic, combine numeric together and alphabetic together, sort the components, put numeric first then put alphabetic to formulate the components, and put into LOG table [14].

4.1.Fuzzy-Rough Sets as Data Cleaning Tools

In the proposed application of using data reduction strategy both feature and instance as cleaning tools is based on searching the initial space of attributes given as a header of dataset characteristic attributes and their values as a dataset instances. The process of cleaning data using soft computing sets is divided into several subsequent steps start by searching the given space towards to the filtering the attribute as feature subset selection process going to minimize the number of trained instance via instance selection provided that dealing with missing value as air push to enhance the classifier accuracy which is built upon training phase using the reduct data.

First, Various heuristic search strategies can be used for searching the problem space but the most important strategy with a reasoning time are hill climbing and Best First [15]. Every search strategy was trailed with the feature selection mechanism which can be correlation based feature selection (CFS), fuzzy rough feature selection or other. In these study, only both Fuzzy Rough set Feature Selection (FRFS) and Correlation based Feature Selection are considered and discussed how can be applied as a data cleaning tools.

For both recent listed search methods, the Best First search was used in the final experiments as it gave better results in some cases. Searching a space using Best First technique starts with empty set of features as initial then generates all possible single feature growths. Next, start evaluation of the highest subset expanding in the same manner by adding single features at a time. While the expanded feature improve the subset results the search continue whereas the expanded feature has no improvements then the search drops back to the next best unexpanded subset and continues from there as a backtracking point. Under reasonable time space, Best First will explore the entire search space, so it is common to limit the number of subsets expanded that result in no improvement. Finally, the best subset found is returned when the search terminates. Hill Climbing search strategy is another mechanism which works on optimizing the returned subset search space. In Hill Climbing, search process are performed as greedy forward or backward search through the space of attribute subsets. It may start without any attributes, subset attributes or all attributes as arbitrary point in the space. Search process continues iterate after next until the modification of subset space either by adding or removing any remaining attributes results in a decrease in evaluation, at which search would be stopped. Hill Climbing can rank attributes during the process of traversing the space from one side to the other and recording the order that attributes are selected [15]. It is very clear to evaluate the search space using feature subset selector to decide which is the best and have the same level of information as possible with no ambiguity and doesn't cause data inconsistent set. Initially data set would have n features which would be ignore some of these features in the final subset so that there are (n^2) possible subsets where each must be evaluated using the trailed feature selector strategy, CFS and FRFS are described in subsequent paragraphs showing how such evaluation method is performed.

4.2. Correlation-based Feature Selection

Major feature subset feature selection algorithms are tailed with a search strategy. Correlation Feature Subset (CFS) uses a search algorithm along with a function to evaluate the merit of feature subsets. CFS perform a heuristic test process for measurement the "goodness" of feature subsets for being taken into

account the usefulness of individual features for predicting the class label along with the level of inter-correlation among them. The heuristic of CFS can be formulated as good feature subsets is the subset that contains features highly correlated with ability to be predictive for the class eliminating uncorrelated feature that are not predictive of each other [15].If the correlation between each of the components in a test and the outside variable is known, and the inter-correlation between each pair of components is given, then the correlation between a composite test consisting of the summed components and the outside variable can be predicted using the following equation where r_{zc} is the correlation between the grouped components and the outside variable, $\overline{r_{zi}}$ is the weighted average of the correlations between the components and the outside variable, $\overline{r_{ii}}$ is the average inter correlation components and k represents the number of the components that is used and summed [15].

$$\mathrm{r_{zc}} = \frac{\mathrm{k}\,(\overline{\mathrm{r_{zi}}})}{\sqrt[2]{\left(\mathrm{k} + \mathrm{k}\,(\mathrm{k}-1)(\overline{\mathrm{r_{ii}}})\right)}} \tag{10}$$

Due to Pearson's correlation coefficient, all variables have been standardized. So that the correlation between a composite and an outside variable is a function of the number of component variables including composite and magnitude of the inter-correlations among these variables, together and the outside variable. So that there are three primary conclusions can be concluded from the relation representing correlation dependency:

1. If the correlation between the components and the outside variables is high, then the correlation between the composite and the outside variables is high.
2. When inter-correlations among the components are low then the correlation between the composite and the outside variables is high.
3. During continues increase of number of components in the composite, the correlation between the composite and the outside variables increase.

CFS is considered as a filter powered by a slight modification of the previous equation for being more clarified for simplifying feature subset selection in order to prediction purposes. CFS's evaluation function can be re-written as follow where M_s is the merit of a feature subset S of k features, $\overline{r_{cf}}$is the mean correlation of feature class $f \in S$ and $\overline{r_{ff}}$ is the average of inter-correlation of feature [15].

$$\mathrm{M_s} = \frac{\mathrm{k}\,(\overline{\mathrm{r_{cf}}})}{\sqrt[2]{\left(\mathrm{k} + \mathrm{k}\,(\mathrm{k}-1)(\overline{\mathrm{r_{ff}}})\right)}} \tag{11}$$

The CFS relation is considered as a simple filter algorithm that based on ranking the feature subsets in the search space of all possible feature subsets. Testing of correlation is based on heuristic stated correlation that includes three variations each one employing a quality of the attribute measurement technique. The acceptance of a feature will depend on the extent to which it predicts classes in areas of the instance space not already predicted by other features. So redundant features should be selected out as because of being highly correlated with one or more of the remaining features. Generally, CFS doesn't develop the dependency on any modules such as discretization. Whereas measure feature subset correlation using three different quality measurement factors including CFS-UC that is based on using symmetrical uncertainty, CFS-MDL which is powered by Minimum Description Length principle for using normalized symmetric and last is CFS-relief that uses symmetrical relief. There are many different search strategies which is allowed to be used during heuristic evaluation process for correlation feature subset election; forward selection and backward elimination are the most applicable

reasonable strategies because of auto termination after five consecutive fully expanded subsets having no improvement of current. In Forward mechanism, selection process begins with no features and greedily adds one feature at a time until no possible single feature addition results in a higher evaluation. Whereas backward elimination begins with the full feature set and greedily removes one feature at a time as long as the evaluation does not degrade [15].

4.3.Fuzzy Rough Set Theory

Fuzzy Rough set theory is a researcher's works towards hybrid fuzzy set with the concept of approximation rough. Fuzzifying the formulas for lower and upper approximation of rough set theory is the main focus of every hybrid model of fuzzy rough set theory. Chris Cornelis et. al [8]. Have proposed a concept of generalized fuzzy approximation including upper and lower. Every rough set A can be generalized to be a fuzzy set in$X$, that in turn, allow different objects belongs to varied concept with a membership degree rather than evaluating objects' indiscernibility as a fuzzy tolerance relation R to estimate closeness of every object. For a relation set A and fuzzy tolerance relation$R$, the upper and lower approximation of A can be adopted and formulated as [17]:

$$(\mathrm{R}\downarrow \mathrm{A})(\mathrm{y}) = inf_{x\in \mathrm{X}}\, \mathrm{I}\big(\mathrm{R}(x,\mathrm{y}),\mathrm{A}(x)\big), \tag{12}$$

$$(\mathrm{R}\uparrow \mathrm{A})(\mathrm{y}) = \sup_{x\in \mathrm{X}}\, \mathrm{T}\big(\mathrm{R}(x,\mathrm{y}),\mathrm{A}(x)\big) \tag{13}$$

for all x and y in X where I is implicator and T is a t-norm.
In turn of refining the approximation definition, approximate equality between two objects can be estimated by computation of a parameterized relationR_a, where a is a numeric attribute with range$l(a)$, where$x$ and y are elements in X then

$$\mathrm{R}_a(x,\mathrm{y}) = \max\,(0,\, \min\,(1, \beta - \alpha \tfrac{|a(x)-a(\mathrm{y})|}{\mathrm{l}(a)})) \tag{14}$$

Where α and $\beta$are parameters such that ($\alpha \geq \beta \geq 1$) that were used to determine the granularity ofR_a. Obviously, discernibility, or distance, of two objects x and y can be computed as the complement of their closeness: $1 - R_a(x,y)$. Whereas a is nominal attribute and x and y are elements in X then discerning objects$R_a(x,y)$ can be defined as

$$\mathrm{R}_a(x,\mathrm{y}) = \begin{cases} 1, & a(x) = a(\mathrm{y}) \\ 0, & \text{Otherwise} \end{cases} \tag{15}$$

Generally, for any subset $B$of A, the fuzzy B-indecernibility relation by $R_B(x,y) = \min_{a\in B} R_a(x,y)$. Due to the stated definition, R_B is a fuzzy tolerance relation even attributes are nominal where discretization can be applied. In [14], an extended framework were proposed to deal with numeric attributes in more concrete manner based on evaluating feature subset as increasing [0-1] valued ratio of discernibility relative to the decision attribute. In turn, B relation measurement can be denoted as a fuzzy decision reduct denoted by Fuzzy M-decision reduct that can be defined as $degree\alpha and for all B \subset B,\ M(B) < \alpha if M be monotonic P(A) \rightarrow [0,1] mapping B \subseteq A$and$and\ 0 < \alpha \leq 1$.
In order to minimize and finding out a reduct of decision table, a measure can be defined as [17]:

$$\gamma_{\mathrm{B}} = \frac{|\mathrm{POS_B}|}{|\mathrm{POS_A}|} \text{ or } \delta_{\mathrm{B}} = \frac{min_{x\in \mathrm{X}}\, \mathrm{POS_B}(x)}{min_{x\in \mathrm{X}}\, \mathrm{POS_A}(x)} \tag{16}$$

which are used to evaluate the precision of decisions reduct B of given decision table with set A as decision provided that $B \subseteq A$ and POS_B is a fuzzy set in X for a fuzzy B-indiscernibility relation such that $POS_B(y) = \left(\bigcup_{vk \in V_d} R_B \downarrow X_k\right)(y)$ where y in U. Using fuzzy positive region allows gradual membership values by defining increasing [0, 1]-valued measure to obtain fuzzy decision reducts [8].

5. Experimental Results

5.1. Data Set(s), Performance Measure and Runtime Environment

In 1997, H. Altay Guveniret.al. have created a dataset officially to determine the type of arrhythmia from the ECG recordings which is called "Cardiac Arrhythmia Database"; short name is **arrhythmia**. In which they were concerning of distinguishing between the presence and absence of cardiac arrhythmia besides classification to be one of 16 groups. The 16 different groups are labeled as classes through 01 to 16 where class 01 refers to 'normal' ECG whereas classes 02 to 15 refers to different classes of arrhythmia and last class 16 refers to the rest of unclassified ones. On other hand, in 1998, R.S. Michalskiet.al. were responsible for collecting and formulating a dataset called "Large Soybean Database", short name is **soybean**, which includes 19 classes. First 15 classes are considered as their prior work. Their dataset were powered by 35 categorical attributes, some nominal and some ordered. The folklore seems to be that the last four classes are unjustified by the data since they have so few examples. The values for attributes are encoded numerically, with the first value encoded as "0" the second as "1" and so forth. An unknown values is encoded as "?" where "dna" is used to denote does not apply. These two datasets were used during analysis preprocess using different search strategies including best first and Hill Climbing with a tailed feature subset selection evaluator to minimize and find out reduct with same level of decision reduction or more optimally. Analysis process were performed using "WEKA", Waikato Environment for Knowledge Analysis, machine learning tool which is considered as the most powerful open source machine learning and data mining tool. WEKA is supported and developed under supervision and management of the University of Waikato. In the following table, two dataset which are available online on [18], were concluded with most characteristics describing each in concrete factors. Experimental were performed using Java runtime environment version 8 update 25 as major perquisites for WEKA 3.7.2. The experimental were done using a computer system having the following specification of hardware and software respectively, Intel Core i-5 second edition, 4GB RAM and Windows 7 professional 64 bit.

Table 2 Dataset(s) Principles Features

Names of data sets	Soybean	Arrhythmia
No. of attributes	36	280
No. of instances	683	452
No. of attributes that contain nominal	36	76
No. of attributes that contain numeric	0	204
No. of distinct class	19	13
No. of attributes that contain missing	34	5

Experimental were performed based on selection different dataset with ranging interval of their principles feature to be more reasonable about the derived results. In the experiments there are two major search strategies that are used to write down results; best first and hill climbing rather than others during running which preserve same derivations. Every dataset during experiment were evaluated using different performance measure includes decision reduct accuracy, consistency, total number of feature in attribute subset, total number of instance selected and missing value management. Because of data

mining applications are the most interested applications in comprehensible results of data preprocessing techniques and feature subset reduction for prediction purpose or any other machine language fields, it is clear that naive Bayes/ perceptron/ back propagation or any other as a data classifier should be required in order to maximize predictive performance [19]. There are twelve of scenarios that are applied during execution of the experiments including using two subset feature selection evaluator and two reasonable search strategies; best first and hill climbing as mentioned before. The twelve different scenarios are applied on two different datasets. Next table abbreviated the twelve different scenarios that were used for experimenting soft computing Fuzzy Rough selection as data preprocessing tool.

Table 3 Abbreviation of Experimental results

Scenario #	Abbreviation
Scenario [1] "SC1"	Perform attribute subset selection using Fuzzy Rough tailed with a Best First search strategy.
Scenario [2] "SC2"	Perform attribute subset selection using Fuzzy Rough tailed with a Hill Climbing search strategy.
Scenario [3] "SC3"	Perform attribute subset selection using Correlation dependency relation tailed with a Best First search strategy.
Scenario [4] "SC4"	Perform attribute subset selection using Correlation dependency relation tailed with a Hill Climbing search strategy.
Scenario [5] "SC5"	Scenario [1] subsequent with instance reduction based on Fuzzy Rough Entropy model.
Scenario [6] "SC6"	Scenario [2] subsequent with instance reduction based on Fuzzy Rough Entropy model.
Scenario [7] "SC7"	Scenario [3] subsequent with instance reduction based on Fuzzy Rough Entropy model.
Scenario [8] "SC8"	Scenario [4] subsequent with instance reduction based on Fuzzy Rough Entropy model.
Scenario [9] "SC9"	Scenario [5] subsequent with Replace Missing Values of the reduced dataset with the means.
Scenario [10] "SC10"	Scenario [6] subsequent with Replace Missing Values of the reduced dataset with the means.
Scenario [11] "SC11"	Scenario [7] subsequent with Replace Missing Values of the reduced dataset with the means.
Scenario [12] "SC12"	Scenario [8] subsequent with Replace Missing Values of the reduced dataset with the means.

5.2. Experimental Results & Discussion

Using twelve different scenarios listed in table 3, data cleaning process were evaluated using two datasets over different factors of performance metrics. Accuracy is the most powerful factor which is used to measure the quality of predictor and classifier based on the original decision reduction knowledge against three different subsequent level of reduction process for decision information table. Table 4, 5 and 6 are used to list the experimental analysis of accuracy of multi-layer Perceptron classifier that is so simple neural network learning algorithm. Missing values have a concrete face of two enumerated either correlated values or separate values. For correlated values, the best way to deal with unknowns depends on their meaning in the domain where commonly treating it as a separate value is the best approach. While separate values that are truly representing missing information a more sophisticated scheme should be used with missing entries.

5.2.1. Accuracy

In table 4, the classifier is initialized with acceptable rate of Soybean dataset, 94.4 and reasonable accuracy of arrhythmia 73.5. Going through different scenarios the accuracy rate was shacked in wide

range upward and downward especially SC2 and SC4 were accuracy was recorded as 89, 47.5 and 94.4, 73.5 for Soybean and Arrhythmia respectively. However, accuracy were down for Arrhythmia in SC6 and SC8, they are raised in some manner to be increased against down of Soybean accuracy at the same scenario, see table 5.

Table 4
Accuracy after Feature Selection

Scenario	Soybean	arrhythmia
Original	94.4	73.5
SC1	89	47.5
SC2	89	47.5
SC3	93.9	71.1
SC4	94.4	73.5

Table 5
Accuracy after Instance Reduction

Scenario	Soybean	Arrhythmia
SC5	90.7	93.8
SC6	90.7	93.8
SC7	94.2	71.4
SC8	91.9	59.4

Table 6
Accuracy after Replace Missing Values

Scenario	Soybean	arrhythmia
SC9	90.2	93.8
SC10	90.2	93.8
SC11	94.4	74
SC12	92.2	59.4

fields in the same possible manner for the nominal were done. Table 6, shows that Fuzzy Rough computing is powerful tools for data cleaning especially for datasets that contains nominal data fields which are subject for fuzzification under supervision of rough approximation rather than numeric fields that have challenge in such process. It is noticeable from SC9 and SC10 of Arrhythmia to see that, Fuzzy Rough subset selection as data cleaning have great impact factor of classifier rate and accuracy rather than correlation subset that minimize and reduce classification rate in bad manner, see table 2. Opposite to pros, cons are existed for Fuzzy Rough as data cleaning tools that is hard to discretize numerical

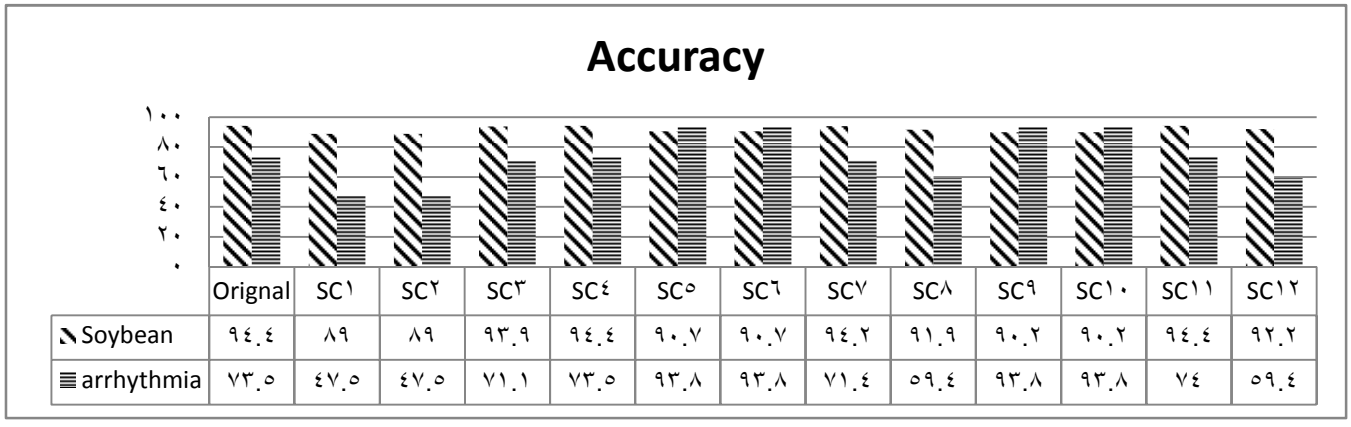

Figure 1. Accuracy of Different level of data cleaning

, compromise the overall In Figure 1, accuracy of back propagation classifier for original datasets going further for preprocessing and cleaning feature subset then followed by instance reduction and finally remove noisy dataset with replacing missing data by the mean and ignore meaning less unknown values. **.2.2. Reduction Rate** Reduction rate is considered as the complement of the percentage of removed elements divided by the total number of element before elimination process regardless element is instance or attribute in the dataset.Based on main characteristic of the used datasets which are listed in table 2 and having deep look at table 7 and 8, Fuzzy Rough subset selection tool can be considered as a perfect reduction tools. It would be able to minimize the dataset size over its two different dimensions; attributes and instances, because of its ability to minimize the decision reduct as optimal as possible with least error rate. From table 7 and 8, also, Fuzzy Rough reduction rate would be affected by the missing value rate as reverse relation. When missing

values is minimized as possible, the reduction rate would be maximized as possible because of missing values leads to ambiguity that reduces the elimination process over the two dimension of any dataset. Figure 2effects of attribute reduction and instance selection.

Table 7
Reduction rate after feature selection

Scenario	Soybean	arrhythmia
Original	0	0
SC1	0.555556	0.971429
SC2	0.555556	0.971429
SC3	0.361111	0.903571
SC4	0	0

Table 8
Reduction Rate after instance reduction

Scenario	Soybean	arrhythmia
Original	0	0
SC5	0.013177	0.442478
SC6	0.013177	0.442478
SC7	0.01757	0.059735
SC8	0.013177	0.002212

Figure 2. Reduction rate (Attribute Selection vs. Instance Reduction)

5.2.3. Consistency

Every dataset is considered as decision table that can be defined as knowledge representation system called K= (U, A) as KRS. Every dataset is comprised of two subsets of attributes, called condition and decision attributes. In dataset, decision table, for every (a) there is a rule denoted as dx (a) = a (x) where x is used to refer a label of the decision rule dx, dx|C is the restriction of dx or conditions of dx and dx|D is the restriction of dx or decision of dx. The decision rule dx is consistent in the given dataset if for every instance x and y, dx|C = dy|C implies dx|D = dy|D otherwise the decision rule is inconsistent. In turn the dataset is said to be consistent if all its decision rules are consistent; otherwise it is inconsistent [9]. Because of the Fuzzy Rough subset evaluation strategy is powered by the principles of the rough set theory, using Fuzzy Rough represents add on to the consistency factor of any dataset that would be analyzed using it. Table 9,10and 11 represents the gained consistency of using Fuzzy Rough evaluator as soon as starting analysis reaching the top of consistency level after preprocess weak, noisy, fault or misclassified instance and missing values per tuples. Figure 3, is used to display the powerful of using Fuzzy Rough subset selector as data cleaning tool and its consistency impact.

Table 9
Consistency after Feature Selection

Scenario	Soybean	arrhythmia
Original	99.7	100
SC1	99.7	100
SC2	99.7	100
SC3	99.1	100
SC4	99.7	100

Table 10
Consistency after Instance Reduction

Scenario	Soybean	Arrhythmia
SC5	100	100
SC6	100	100
SC7	100	100
SC8	100	100

Table 11
Consistency after Replace Missing Values

Scenario	Soybean	arrhythmia
SC9	100	100
SC10	100	100
SC11	100	100
SC12	100	100

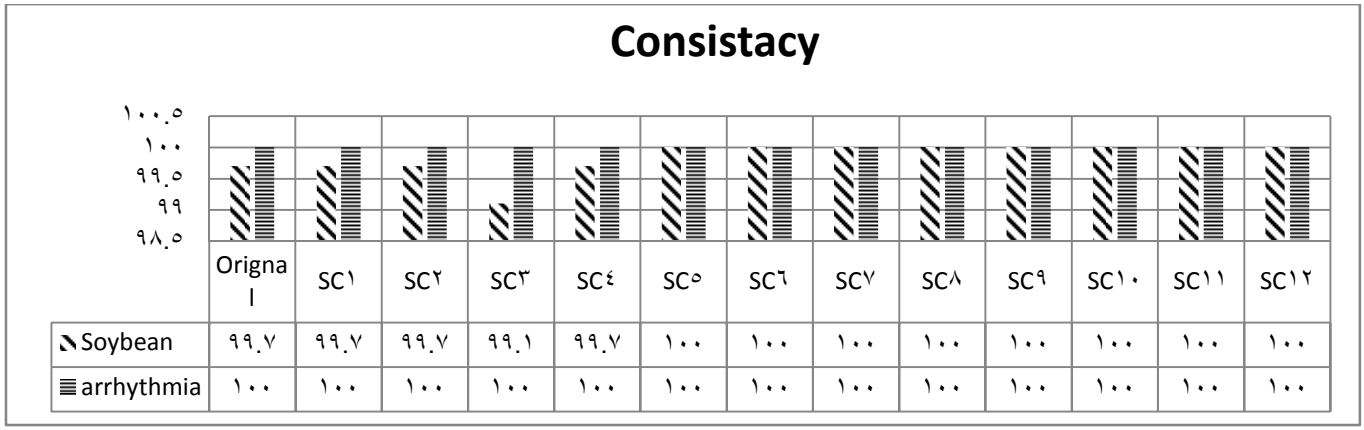

Figure3. Consistency of dataset over different level of data cleaning

6. Conclusion

In the proposed study, experimental results were used to induct how fuzzy rough feature subset evaluator can be used as data cleaning tools in preprocessing phase of different machine learning applications. Because of texture of Fuzzy rough of being mathematical model, it is considered as a powerful tool of data cleaning especially when the dataset were classified as numerical attribute set rather than benefit of use dataset with nominal attribute set under ability of being discretizing. The consistency of dataset is considered under control during analysis especially missing values ratio decreased and gradually guaranteed while using fuzzy rough concept.

References

1. F. Elgamal, N.A. Mosa, N.A. Amasha, "Application of Framework for Data Cleaning to Handle Noisy Data in Data Warehouse", International Journal of Soft Computing and Engineering, Vol. 3, pp. 226-231, 2014
2. MagdiKamel, "Data Warehousing and Mining", IGI Global, 2009. online URL:http://www.igi-global.com/chapter/data-preparation-data-mining/10872, last visit : Dec 2014
3. Israr Ahmed and Abdul Aziz," Dynamic Approach for Data Scrubbing Process", International Journal on Computer Science and Engineering Vol. 02, No. 02, pp 416-423, 2010.
4. Jason D. Van Hulse, TaghiM. Khoshgoftaar, Haiying Huang, "The pairwise attribute noise detection
5. M. Beirlaen and C. StraSsEr, Non-monotonic reasoning with normative conflicts in multi-agent deontic logic, Journal of Logic and Computation, 2013
6. Zadeh, L.: Fuzzy sets. Information and Control 8 (1965) 338–353
7. T. Terano, K. Asai, M. Sugeno, Fuzzy Systems Theory and its Applications, 1992

8. Chris Cornelis, Germ´anHurtadoMart´ın, Richard Jensen, and Dominik ´Sl¸ezak, "Feature Selection with Fuzzy Decision Reducts", Rough Sets and Knowledge Technology Lecture Notes in Computer Science, Vol. 5009, pp 284-291, 2008
9. Rough sets, Theoretical aspects of reasoning about data, 1991
10. Joaquín D., Chris C., Salvador G., and Francisco H., "Enhancing evolutionary instance selection algorithms by means of fuzzy rough set based feature selection",Informatics and Computer Science Intelligent Systems Applications Elsevier, Vol. 186, Pp. 73- 92, 2011.
11. D. P. Acharjya, Debasrita Roy, and Md. A. Rahaman, "Prediction of Missing Associations Using Rough Computing and Bayesian Classification ", International Journal of Intelligent Systems and Applications,Vol. 11, Pp. 1-13, 2012.
12. K., Anitha, and P., Venkatesan,"FEATURE SELECTION BY ROUGH QUICK REDUCT ALGORITHM", International Journal of Innovative Research in Science, Engineering and Technology Vol. 2, 2013.
13. P., Sharma, and V., Kumar "Data Dependencies Mining In Database by Removing Equivalent Attributes", International Journal of Scientific Research in Computer Science & Engineering. Vol. 1, Pp. 7- 11. 2013
14. A., F.,Elgamal, N., A.,Mosa, and N., A.,Amasha,"Application of Framework for Data Cleaning to Handle Noisy Data in Data Warehouse", International Journal of Soft Computing and Engineering (IJSCE), Vol. 3, Pp. 226 - 231 2014.
15. M. A. Hall, "Correlation-based Feature Subset Selection for Machine Learning", thesis , Hamilton, New Zealand, 1998
16. Radzikowska, A., Kerre, E.: A comparative study of fuzzy rough sets. Fuzzy Sets and Systems, Vol. 126, pp. 137–156, 2002
17. Jensen, R., Shen, Q.: Fuzzy-rough sets assisted attribute selection. IEEE Transactions on Fuzzy Systems, Vol. 15, pp. 73–89, 2007
18. Dataset online url: http://repository.seasr.org/Datasets/UCI/arff/, last visit on Oct 2014.
19. P. Langley and S. Sage, "Induction of selective Bayesian classifiers", In Proceedings of the Tenth Conference on Uncertainty in Artificial Intelligence, Seattle, W.A, Morgan Kaufmann, 1994