Applied Soft Computing 12 (2012) 999–1005

Contents lists available at SciVerse ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

A new weighted rough set framework based classification for Egyptian NeoNatal Jaundice

Hala S. Own [a,*], Ajith Abraham [b,c]

[a] Department of Solar and Space Research, National Research Institute of Astronomy and Geophysics, El-Marsad Street, P.O. Box 11421, Helwan, Egypt
[b] Machine Intelligence Research Labs (MIR Labs), Scientific Network for Innovation and Research Excellence, Washington 98071, USA
[c] VSB - Technical University of Ostrava, 17, listopadu 15/2172, Ostrava - Poruba, Czech Republic

ARTICLE INFO

Article history:
Received 10 January 2010
Received in revised form 3 February 2011
Accepted 27 November 2011
Available online 16 December 2011

Keywords:
Weighted rough set
Neonatal jaundice
Class imbalance learning
Rule importance

ABSTRACT

Prediction of diseases would help physicians to make informal decision regarding the type of treatment. Jaundice is the most common condition that requires medical attention in newborn babies. Although most newborns develop some degree of jaundice, a high level bilirubin puts a newborn at risk of bilirubin encephalopathy and kernicterus, which are rare but still occur in Egypt. This paper presents a new weighted rough set framework for early intervention and prevention of neurological dysfunction and kernicterus that are catastrophic sequels of neonatal jaundice. The obtained results illustrate that the weighted rough set can provide significantly more accurate and reliable predictive accuracy than well known algorithms such as weighted SVM and decision tree considering the fact that physicians do not have any estimation about probability of jaundice appearance.

1. Introduction and motivation

Neonatal jaundice, although a normal transitional phenomenon in most infants, can occasionally become more pronounced. Blood group incompatibilities (e.g., Rh, ABO) may increase bilirubin production through increased hemolysis. Historically, Rh isoimmunization was an important cause of severe jaundice, often resulting in the development of kernicterus. Although this condition has become relatively rare in industrialized countries following the use of Rh prophylaxis in Rh-negative women, Rh isoimmunization remains common in developing countries [34]. About 60% of term babies, and 80% of preterm babies, develop jaundice in their first week of life. 10% of breastfed babies are still jaundiced at 1 month of age. Rapid differentiation between the majority of babies with jaundice who have no underlying disease (physiological jaundice) and those with pathological causes is important to detect the underlying disease and to prevent adverse sequelae such as bilirubin encephalopathy and kernicterus [1].

The increased bilirubin cause the infant's skin and whiteness of the eyes (sclera) to look yellow. Kernicterus, or bilirubin encephalopathy, is a condition caused by bilirubin toxicity to the basal ganglia and various brainstem nuclei. In the acute phase, severely jaundiced infants become lethargic, hypotonic and suck poorly. If the hyperbilirubinemia is not treated, the infant becomes hypertonic and may develop a fever and a high-pitched cry. The hypertonia is manifested by backward arching of the neck (retrocollis) and trunk (opisthotonus). Surviving infants usually develop a severe form of athetoid cerebral palsy, hearing loss, dental dysplasia, paralysis of upward gaze and, less often, intellectual and other handicaps [1].

Medical databases have accumulated large quantities of information about patients and their medical conditions. Relationships and patterns within these data could provide new medical knowledge [2–5]. Analysis of medical data is often concerned with treatment of incomplete knowledge, with management of inconsistent pieces of information and with manipulation of various levels of representation of data. Over the past two decades, several traditional multivariate statistical classification approaches, such as the linear discriminate analysis, the quadratic discriminate analysis, and the logic analysis, have been developed to address the classification problem. More advanced and intelligent techniques have been used in medical data analysis such as neural network, Bayesian classifier, genetic algorithms [6–8], fuzzy theory, and rough set. Other approaches like case-based reasoning and decision trees [9,10] are also widely used to solve data analysis problems.

All these techniques haves its own properties and features including their ability of finding important rules and information that could be useful for the medical field domain. Each technique contributes a distinct methodology for addressing problems in its domain.

Rough set theory [11–13] is a fairly new intelligent technique that has been applied to the medical domain, and is used for the discovery of data dependencies, evaluates the importance of

* Corresponding author. Tel.: +20 1001615337.
E-mail addresses: halaown@gmail.com (H.S. Own), ajith.abraham@ieee.org (A. Abraham).

1568-4946/$ – see front matter © 2011 Elsevier B.V. All rights reserved.
doi:10.1016/j.asoc.2011.11.025

attributes, discovers the patterns of data, reduces all redundant objects and attributes, and seeks the minimum subset of attributes. Moreover, it is being used for the extraction of rules from databases. One advantage of the rough set is the creation of readable "if-then" rules. Such rules have a potential to reveal new patterns in the data material. Rough set based method assumes that the target classes share similar prior probabilities. However, in many real world problems, especially in medical applications almost all the examples are labeled as one class, while far fewer examples are labeled as other classes. When the class distribution of a data set is skewed, the traditional rough set based method is biased to the majority class, and therefore performs poorly in recognition of the minority class because the priori knowledge of class distribution is not taken into account. Recently, there are few research works introducing a prior knowledge about samples into rough set. Hu et al. [17] assigned with each sample x its probability $P(x)$. Stefanowski et al. [18] proposed the removing and filtering to enhance the performance of rough when the class distribution of a data set is skewed. Liu et al. [14] introduced weights to represent the class imbalance problem. They select the significant attributes by design a weighted attribute reduction algorithm based on Guiasu weighted entropy. Finally the weighted rule extracted by introducing a weighted heuristic strategy into LEM2 algorithm.

However, all the previous work did not investigate the method of building information table. That is how to assign a weigh to each sample to balance the biased distribution of samples.

In this paper we apply class equal sample weighting to build a weighted information table. By using this method, samples belonging to majority class has smaller weight while samples in the minority class has larger weight. Then a weighted attribute reduction algorithm based on the significance of the attribute is applied to find the reduct set. Finally, a set of diagnosis rules are extracted based on a modified version of MLEM2 called as a weighted MLEM2 algorithm. This process leads towards the final goal of generating diagnosis rules from information or decision system of the Egyptian NeoNatal Jaundice database. The dataset comprises of 808 samples from newborns from January to December 2007 in Neonatal Intensive Care Unit in Cairo of Egypt to predict cases that will develop extreme hyperbilirubinemia with total serum bilirubin (TSB) level of greater than or equal to 428 μmol/L (≥25 mg/dL) for early intervention and prevention of neurological dysfunction and kernicterus that are catastrophic sequels of neonatal jaundice.

This paper is organized as follows: Section 2 gives a brief introduction to the rough sets. Section 3 discusses the proposed weighted rough set data analysis scheme in detail. The motivation and characteristics of NeoNatal Jaundice datasets are presented in Section 4. Experimental analysis and discussion of the results are described in Section 5. Finally, conclusions are presented in Section 6.

2. Rough sets: basic notation

2.1. Information system and approximation

Definition 1 *((Information system)).* Information system is a tuple (U, A), where U consists of objects and A consists of features. Every $a \in A$ corresponds to the function $a : U \rightarrow V_a$ where V_a is the value set of a. In the applications, we often distinguish between conditional features C and decision feature D, where $C \cap D = \phi$. In such cases, we define decision systems (U, C, D).

Definition 2 *((Indiscernibility relation)).* Every subset of features $B \subseteq A$ induces indiscernibility relation:

$$\mathrm{Ind}_B = \left\{(x, y) \in UXU : \forall_{a \in B} a(x) = a(y)\right\} \tag{1}$$

for every $x \in U$, where is an equivalence class $[x]_B$ in the partition of U defined by Ind_B.

Definition 3 *((Lower and upper approximation)).* In the rough sets theory, the approximation of sets is introduced to deal with inconsistency. A rough set approximates traditional sets using a pair of sets named the lower and upper approximation of the set. Given a set $B \subseteq A$, the lower and upper approximations of a set $Y \subseteq U$ are defined by, respectively:

$$\begin{aligned} \underline{B}Y &= \left\{x|[x]_B \subseteq X\right\} \\ \bar{B}Y &= \{x|[x]_B \cap X \neq \phi\} \end{aligned} \tag{2}$$

Definition 4 *((Lower approximation and positive region)).* The positive region $\mathrm{POS}_C(D)$ is defined by

$$\mathrm{POS}_C(D) = \bigcup_{X:X \in U/\mathrm{Ind}_D} \underline{C}X; \tag{3}$$

$\mathrm{POS}_C(D)$ is the set of all objects in U that can be uniquely classified by elementary sets in the partition U/Ind_D by means of C [15].

Definition 5 *((Upper approximation and negative region)).* The negative region $\mathrm{NEG}_C(D)$ is defined by

$$\mathrm{NEG}_C(D) = U - \bigcup_{X:X \in U/\mathrm{Ind}_D} \bar{C}X \tag{4}$$

that is the set of all objects can be definitely ruled out as member of X.

Definition 6 *((Boundary region)).* The boundary region is the difference between upper and lower approximations of a set X that consists of equivalence classes having one or more elements in common with X; it is given by the following formula:

$$\mathrm{BND}_B(X) = \bar{B}X - \underline{B}X \tag{5}$$

a rough set can be characterized using the accuracy of approximation as defined below

$$\alpha_B(X) = \frac{\left|\underline{B}X\right|}{\left|\bar{B}X\right|}, \tag{6}$$

where $|\bullet|$ denotes the cardinality of a set. X is definable with respect to B if $\alpha_B(X) = 1$, otherwise X is rough with respect to B.

2.2. Reduct and core

Definition 7 *((Degree of dependency)).* Given a decision system, the degree of dependency of D on C can be defined as

$$\gamma(C, D) = \frac{\left|\mathrm{POS}_C(D)\right|}{\left|U\right|}, \tag{7}$$

Definition 8 *((Reduct)).* Given a classification task related to the mapping $C \rightarrow D$. A reduct is a subset $R \subseteq C$ such that

$$\gamma(C, D) = \gamma(R, D) \tag{8}$$

and none of proper subsets of R satisfies analogous equality.

Definition 9 *((Reduct set)).* Given a classification task mapping a set of variables C to a set of labeling D, a reduct set is defined with respect to the power set $P(C)$ as the set $R \subseteq P(C)$ such that:

$$\mathrm{Red} = \left\{A \in P(C) : (A, D) = (C, D)\right\}.$$

That is, the reduct set is the set of all possible reducts of the equivalence relation denoted by C and D.

The reduct set is a minimal subset of attributes that preserves the degree of dependency of decision attributes on full condition attributes. The intersection of all the relative reduct sets is called core.

2.3. Significance of the attribute

Significance of features enables us to evaluate features by assigning a real number from the closed interval [0,1], expressing how important a feature is.

Definition 10 *((Significance)).* For any feature $a \in C$, we define its significance ξ with respect to D as follows:

$$\xi(a, C, D) = \frac{\left|\mathrm{POS}_{C\setminus\{a\}}(D)\right|}{\left|\mathrm{POS}_{C}(D)\right|}. \tag{9}$$

Based on the significance of an attribute, a heuristic attribute reduction algorithm can be designed to find a reduct by selecting an attribute with maximum significance interactively [16].

3. Weighted rough set framework

In this research, we apply class equal sample weighting to build a weighted information table. Then a weighted attribute reduction algorithm based on the significance of the attribute is applied to find the reduct set. Finally, a set of diagnosis rules are extracted based on a modified version of MLEM2 called as a weighted MLEM2 algorithm. This process leads towards the final goal of generating diagnosis rules from information or decision system of the Egyptian NeoNatal Jaundice database. Fig. 1 shows the overall steps in the proposed weighted rough sets data analysis framework. The following sections illustrate each step in detail.

3.1. Class equal sample weighting (CSW)

In the traditional rough set, all samples have equal weight without considering the distribution of samples. However, in CSW, samples belonging to majority class has smaller weight while samples in the minority class has larger weight. Algorithm 1 introduces the main steps to calculate the sample weight.

3.2. Weighted relevant attribute extraction and reduction

The basic philosophy of rough sets is to reduce the attributes in the data set based on the information content of each attribute or collection of attributes such that there is a mapping between similar objects and a corresponding decision class. In general, not all of the information contained in a data set is required; many of the attributes may be redundant in the sense that they do not directly influence which decision class a particular object belongs to. In decision tables, there often exist conditional attributes that do not provide (almost) any additional information about the objects. So, we should remove those attributes since it reduces complexity and cost of decision process [19–22].

Algorithm 1. Building weighted information table algorithm

Input: information table S
Output: weighted information table WS
for each sample $a_i \in A$
1) for each decision class $d_j \in D$
2) if $a_i \in d_j$ then
3) $w_i \leftarrow \frac{1}{(n(D)\times n(A_j))}$, where,
4) $n(D)$ is the number of decision classes and,
$n(A_j)$, is the number of samples classified as d_j

For the new weighted information table the weights generated by Algorithm 1 do not change the equivalence relation and do not change the upper and lower approximation of arbitrary subset $X \in A$. However, the introduced weights change the accuracy approximation of X [14]. This will lead to rewrite Definition 6 to introduce a weight into approximation accuracy such that:

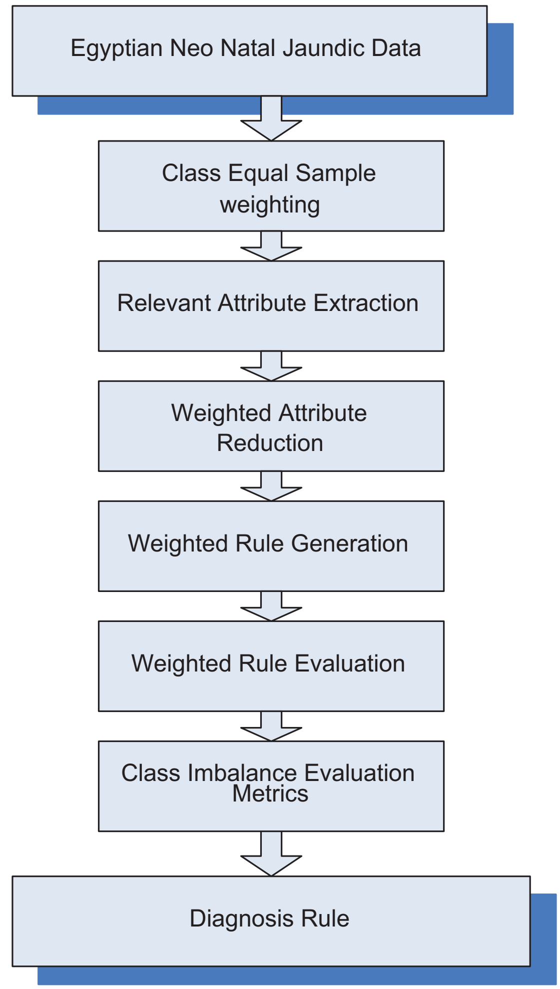

Fig. 1. Weighted rough set data analysis framework.

Definition 11 *((Weighted approximation accuracy)).* The weighted approximation accuracy can be defined as follows:

$$\alpha_B^W(X) = \frac{\left|\underline{B}X\right|_W}{\left|\bar{B}X\right|_W}, \tag{10}$$

where $\left|\underline{B}X\right|_W = \sum_{x_i \in \underline{B}X} w(x_i)$, and $\left|\bar{B}X\right|_W = \sum_{x_i \in \bar{B}X} w(x_i)$, which represents the weighted cardinality of $\underline{B}X$ and $\bar{B}X$, respectively, and $w(x_i)$ is a weight associated with the element $x_i \in X$.

Definition 12 *((Weighted degree of dependency)).* Given a weighted decision system, the weighted degree of dependency of D on C can be defined as:

$$\gamma^W(C, D) = \frac{\left|\mathrm{POS}_C(D)\right|_W}{\left|U\right|_W} \tag{11}$$

Definition 13 *((Discrimination factor)).* It measures the effect of removing the attribute $a_i \in A$ from an information table by measuring the difference between $\gamma^W(C, D)$ and $\gamma^W(C - \{a\}, D)$.

We introduce a weighted reduct and relevant attribute extraction algorithm based on the correlation factor, weighted degree of dependencies and the discrimination factors. The main steps of the reduct generation algorithm are provided below (refer to Algorithm 2).

3.3. Weighted rule generation and classification

The generated reducts are used to generate decision rules. The decision rule, at its left side, is a combination of values of attributes such that the set of (almost) all objects matching this combination have the decision value given at the rule's right side. The rule derived from reducts can be used to classify the data. The set of rules is referred to as a classifier and can be used to classify new and unseen data.

Algorithm 2. Weighted attribute reduction

Input: Weighted information table (WS).
Output: weighted reduct sets WR_{final}.
Let $R = \varnothing$
1) For each conditional attribute $c \in C$
Calculate $\varepsilon_{(c,D)}$, the correlation factor between $c \in C$ and the decision attributes D. If $\varepsilon_{(c,D)} > 0$
then $R = R \cup c$
end for.
2) Let P a partition of R such that $\cup P = R$ and $A \cap B = \varnothing$ if $A \in P, B \in P, A \neq B$.
3) For each
$p_i \in P, i = 1, \ldots, n$ where n is the number of sets in P
Calculate $\gamma^W(p_i, D)$.
End for.
4) Choose p_i such that $\gamma^W(p_i, D) = \max(\gamma^W(p_i, D))$.
5) Let $WR_{final} = \varnothing$
6) For each attribute $c \in p_i$ calculate its discrimination factor σ_c.
7) Let $c_{\max} = \max(\sigma_c)$
8) Let $p_i = p_i - c_{\max}$ and $WR_{\text{final}} = WR_{\text{final}} \cup c_{\max}$
9) For each $c \in p_i$ add c to $c_{\max}$ and calculate the discrimination factor of the new combination.
10) Choose c which satisfies max discrimination factor with $c_{\max}$ let $c_{\max} = c$
11) Repeat steps 8 to 10 until all attributes in p_i *are processed*

LEM2 algorithm, a rule induction algorithm used by LERS (Learning from Examples based on Rough Set theory), accepts input data sets only with symbolic attributes. MLEM2, a new algorithm, extends LEM2 capabilities by inducing rules from data with both symbolic and numerical attributes including data with missing attribute values MLEM2 accuracy is comparable with accuracy of LEM2 inducing rules from pre-discretized data sets. However, compared with other members of the LEM2 family, MLEM2 produces the smallest number of rules from the same data [23,24]. In order to introduce an imbalance learning concept into a traditional rough set, a modified version of MLEM2 is introduced. The algorithm is divided into two steps the preprocessing phase and weighted rule induction phase. The preprocessing phase describe how MLEM2 induces rules from data with numerical attributes. Rule induction in MLEM2 is conducted in the same way as in well known LEM2 rule induction algorithm, the only difference is in the kernel part of LEM2 algorithm. Since LEM2 not taking account of imbalance learning; therefore, the inner while loop will be changed to adapt the introduction of weight into rough set.

Let B be a nonempty lower or upper approximation of a concept represented by a decision-value pair (d,w). Set B depends on a set T of attribute value pairs $t = (a,v)$ if and only if:

$$\varnothing \neq [T] = \bigcap_{t \in T} [t] \subseteq B \tag{12}$$

where $[(a,v)]$ denotes the set of all examples such that for attribute a its values are v [23].

The main steps of the Rule Generation and classification algorithm are provided below (refer to Algorithm 3).

Algorithm 3. Weighted rule induction

Input: weighted reduct sets $WR_{\text{final}} = \{wr_1 \cup wr_2 \cup wr \cup \ldots wr_n\}$
Output: set of weighted rules
Phase 1 (preprocessing)
Input: weighted reduct sets $WR_{\text{final}} = \{wr_1 \cup wr_2 \cup wr \cup \ldots wr_n\}$
Output: a set of blocks represent the search space B
1) For each numerical attribute $wr = (x_1, x_2, \ldots, x_n) \in WR_{\text{final}}$ do
2) Sort the values of wr,
3) For the sorted set of wr, compute a set of cut points $Q = (q_1, q_2, \ldots, q_{n-1})$ where $q_j = \frac{x_i + x_{i+1}}{2}$
4) For each q_j define q_{j1} and q_{j2}, such that q_{j1} contains all cases for which values of the
Numerical attribute are smaller than q_j, and q_{j2} contains all cases for which values of the numerical attribute are larger than q_j.
Phase 2 (processing)
Input: a set of blocks represent the search space B
Output: set of weighted rules
1) $G := B$;
2) $WT := \Phi$;
3) *While* $G \neq \Phi$
4) *Begin*
5) $T := \Phi$;
6) $T(G) := \{t : [t] \cap G \neq \Phi\}$,
7) *While* $T = \Phi$ *or* $[T] \not\subseteq B$
8) *Begin*
9) *Select a pair* $t \in T(G)$ *such that the cardinality of* $[t] \cap G_W$ *is maximum; if a tie occurs, select a pair* $t \in T(G)$ *with the smallest cardinality of* $[t]_W$; *if another tie occurs, select first pair;*
10) $T := T \cup \{t\}$,
11) $G := [t] \cap G$
12) $T(G) := \{t : [t] \cap G \neq \Phi\}$
13) $T(G) := T(G) - T$
14) *end* $\{while\}$
15) *for each* $t \in T$ *do*
16) *if* $[T - \{t\}] \subseteq B$
17) *then* $T := T - \{t\}$
18) $WT : WT \cup \{T\}$
19) $G := B - \cup_{T \in WT} [T]$
20) *end while*
21) *for each* $T \in WT$ *do*
22) *if* $\cup_{S \in WT - \{T\}} [S] = B$ *then* $WT := WT - \left\{ T \right\}$

3.4. Rule evaluation

When rules are generated, the number of objects that generate the same rule is typically recorded. The quality of rules that are generated based on attributes included in the reduct is connected with its quality. We would be especially interested in generating rules which cover possibly largest parts of the universe. Covering the universe space with more general rules implies smaller size of a rule set. We could, therefore, use this idea in measuring the quality of a reduct. If a rule is generated more frequently across different rule sets, we can believe that this rule is more important than other rules. The rule importance measure [25] R_I is used as an evaluation to study the quality of the generated rule. It is defined by:

$$R_I = \frac{\tau_r}{\rho_r} \tag{13}$$

where τ_r is the number of times where a rule appears in the set of generated rules and ρ_r is the number of total cases.

3.5. Evaluation metrics for class imbalance learning

The measures of the quality of classification are built from a confusion matrix (shown in Table 1) which records correctly and incorrectly recognized examples for each class. Accuracy is the most common evaluation metric for most traditional application. But accuracy is not suitable to evaluate imbalanced data set, since many practitioners have observed that for extremely skewed class distribution the recall of the minority class is often 0, which means that there are no classification rules generated for the minority class [26]. For this reason, additional metrics are used.

Table 1
Confusion matrix.

	Predicted class				
True class	C_1	C_2	C_3	...	C_K
	C_2	n_{11}	n_{12}	....	n_{1k}
	C_3	n_{21}	n_{22}		n_{2k}
	.	.	.	.	.
	.	.	.	.	.
	C_K	n_{k1}	n_{k2}		n_{kk}

In this study, we have used the following performance measures:

$$\mathrm{TP}_{(i)\mathrm{RATE}} = \frac{\sum_{i=1}^{k}\mathrm{TP}(i)}{\left(\sum_{i=1}^{k}\mathrm{TP}(i)+\sum_{i=1}^{k}\mathrm{FN}(i)\right)} \tag{14}$$

$$\mathrm{Recall}_{(i)} = \frac{n_{ii}}{\sum_{j=1}^{k} n_{ij}} \tag{15}$$

$$\mathrm{Precision} = \frac{n_{ii}}{\sum_{j=1}^{k} n_{ji}} \tag{16}$$

$$F_{(i)} - \mathrm{Measure} = \frac{2 \times \mathrm{Recall}_i \times \mathrm{Precision}_i}{\mathrm{Recall}_i + \mathrm{Precision}_i} \tag{17}$$

$$G\text{-mean} = \left(\prod_{i=1}^{k} R_i\right)^{1/k} \tag{18}$$

The previous measures are popular evaluation metrics for imbalance problems [27]. It is clear that neither recall nor precision is adequate by themselves. *F*-Measure is suggested in [28] to integrate these two measures as an average. It is obvious that *F*-measure will be high when both the recall and precision are high. When the performances of each class are interested, class classification performance of each class should be equally represented in the evaluation measure. Kubat et al. [29] suggested the *G*-mean as the geometric means of recall values of every class. As each recall value representing the classification performance of a specific class is equally accounted, *G*-mean is capable to measure the balanced performance among classes of a classification output [30,31].

4. Egyptian NeoNatal Jaundice data set collection

A total of 808 medical records were collected from newborns during January to December 2007 in Neonatal Intensive Care Unit in Cairo, Egypt. Retrospective data of all neonatal jaundice cases were collected from patient's files and descriptive analysis of these data was done. These data include the following: sex, gestational age, postnatal age, and weight at day of presentation, the day of onset of jaundice after delivery and duration of stay in hospital.

The ratio of 474 male to 334 female was 1.4:1. There were 643 full terms (79.58%), 53 near terms (6.55%) and 113 preterms (13.98%). The mean postnatal age of patients on admission was 5.75 ± 4 days (ranging from 1 to 20 days except one case diagnosed as Crigler–Najjar syndrome, was admitted to NICU at 60 days old). The median age of onset of jaundice was 3 days with the interquartile range (IQR) of one day. The mean weight of patients was 2658.6 ± 710 g (ranging from 740 to 4900 g). The mean duration of stay is of 7.21 ± 8.72 days (ranging from 1 to 86 days).

The total and direct bilirubin levels were measured several times for the studied patients with the detection of peak of total bilirubin and the day on which the peak occurred. The peak of total bilirubin ranged from 6.5 to 65.5 mg/dl with a mean value of 24.55 ± 9.16 at mean age 6.2 ± 3.58 days (ranging from 1 to 33 days). The median peak of total bilirubin was 23 mg/dl with median age for the peak was the 5 days. Among 808 studied cases, a peak of total bilirubin was reported in files of 781 cases.

Table 2
Neonatal jaundice data.

Field	Description
Sex	Male or female
Age/day	Postnatal age per day on admission
gest. Age	Gestational age [either F = full term ($\geq$ 38 weeks), N = near term (35–37 weeks) or P = preterm ($\leq$34 weeks)]
Wt/g	Weight per gram on admission
Onset of J at day	Postnatal age of patient per day in which onset of jaundice was occurred
Days of adm.	Days of admission in hospital (duration of stay in hospital)
Peak of T bil	Peak of total bilirubin level (mg/dl)
bil peak at day	Postnatal age of patient per day in which total bilirubin peak was recorded
T bil d of presentation	Total bilirubin level (mg/dl) at day of presentation.
D bil d of presentation	Direct bilirubin level (mg/dl) at day of presentation.
T bil 24h later	Total bilirubin level (mg/dl) after 24 h after presentation
D bil 24h later	Direct bilirubin level (mg/dl) after 24 h after presentation
T bil after 2day	Total bilirubin level (mg/dl) after 2 days after presentation
D bil after 2day	Direct bilirubin level (mg/dl) after 2 days after presentation
T bil before disc	Total bilirubin level (mg/dl) before discharge from hospital or death
D bil before disc	Direct bilirubin level (mg/dl) before discharge from hospital or death
Pattern	According to type of jaundice, patients are classified into three pattern along their duration of stay: (1) patients with indirect hyperbilirubinemia, (2) patients with indirect hyperbilirubinemia then changed into direct hyperbilirubinemia and (3) patients with direct hyperbilirubinemia

The total bilirubin level was measured at day of presentation, then after 24 h later, then after 2 days, afterwards before discharge or death. The mean values were 23.1 ± 9.87 (ranging from 2.1 to 65.5 mg/dl), 19.85 ± 6.76 (ranging from 4.9 to 49.5 mg/dl), 16.09 ± 5.84 (ranging from 3.1 to 51.3 mg/dl) and 12.34 ± 6 (ranging from 0.74 to 51.7 mg/dl), respectively. The direct bilirubin level was measured at the same time with total bilirubin. The mean values were 1.55 ± 3.4 (ranging from 0.02 to 38 mg/dl), 1.58 ± 3.21 (ranging from 0.03 to 25.36 mg/dl), 1.65 ± 3.35 (ranging from 0.01 to 24.8 mg/dl) and 1.22 ± 2.64 (ranging from 0.01 to 25.36 mg/dl), respectively.

These data are presented for prediction of the risk of neonatal jaundice and extreme hyperbilirubinemia of newborns. The data set of 808 records, 16 predictor variables and 1 target variable, was constructed. The target variable 'pattern' has three possible values "1" (indirect hyperbilirubinaem), "2" (changed from indirect to direct hyperbilirubinaemia) and "3" (direct hyperbilirubinaemia). Table 2 shows the predictive attributes and their description, used in our work.

Table 3 describes the class distribution within the data set. As shown in Table 3, class B and C are two small classes which posses only 6% and 3% samples, respectively.

5. Experimental analysis and discussion

Two experiments were conducted to evaluate the weighted rough set framework, for all of the following experiments 60% split was used for training and the remaining 40% was used for

Table 3
The 808 data set class distributions.

Index	Class name	Class size	Class distribution
A	Indirect hyperbilirubinaem	737	91%
B	Changed from indirect to direct hyperbilirubi-naemia	46	6%
C	Direct hyperbiliru-binaemia	25	3%

Table 4
The 525 data set class distributions.

Index	Class name	Class size	Class distribution
A	Indirect hyperbilirubinaem	265	82%
B	Changed from indirect to direct hyperbilirubinaemia	41	13%
C	Direct hyperbilirubinaemia	19	6%

testing. The first experiment was conducted using the 808 records. The second experiment was conducted using 325 records of cases which had extreme hyperbilirubinemia the class distribution of this data set shown in Table 4. The computations of rules have been done only on training set. The results of computations of rules were applied to the classification of objects from the tested dataset. We can see from Tables 3 and 4 that the class distribution of each data set is imbalanced. The ratio of majority class A to class B and C is 16.02% and 29.48%, respectively, in Table 3. Moreover, in Table 4 the ratio was 6.46% and 13.94%.

By applying the introduced reduct generation algorithm (refer to Algorithm 2) we compute the weighted dependency degree and the classification quality for each attribute. We reach the minimal number of reducts that contains a combination of attributes which has the same discrimination factor. The final generated reduct sets, which are used to generate the list of rules for the classification, are:

{“D bil d of presentation”, “D bil 24 h later”, “T bil after 2 day”,
“T bil before disc”, “D bil before disc”}

The entire attributes generated are all core attributes. Using the minimal reduct set, a set of rules is generated. Part of these rules is listed in Table 5.

As a result of our new approach, we observed very good rules like R1 and R2 the first one “R1” is true for 147 cases with rule importance = 0.89. As well as the second rule “R2” which has support as 134 cases with rule importance = 0.76.

In order to evaluate the performance of the proposed framework, comparative experiments are conducted. Weighted SVM and

Table 5
Part of generated rules.

Rule No.	Rule
R1	(D bil 24h later = [0.1,8.645])&(“D bil before disc = [0.012.195]”) ⇒ (Pattern = {A[147]})
R2	(“T bil after 2day = [4.59,26.0]”)&(“D bil before disc = [0.01,2.195]”) ⇒ (Pattern = {A[134]})
R3	(“D bil d of presentation” = “[0.02,11.9]”)&(“T bil before disc” = “[1.1,16.75]”)&(“D bil before disc” = “[2.195,5.045]”) ⇒ (Pattern = {B[10]})
R4	(“D bil d of presentation = [0.02,11.9]”)&(“T bil after 2day = [4.59,26.0]”)&(“D bil before disc = [5.045,25.36]”) ⇒ (Pattern = {B[10]})
R5	(“T bil before disc = [16.75,51.7]”)&(“D bil before disc=[0.01,2.195]”) ⇒ (Pattern = {A[30]})
R6	(“D bil 24h later = [0.1,8.645]”)&(“T bil after 2day = [4.59,26.0]”)&(“T bil before disc = [16.75,51.7]”) ⇒ (Pattern = {A[19]})
R7	(“D bil d of presentation = [11.9,38] ⇒ (Pattern = {C[7]})

decision tree (C4.5) were employed as comparative methods with our proposed framework. LIBSVM [32] was used to implement weighted SVM for classification problem. LIBSVM provides an efficient parameter selection tool using cross validation via parallel grid search under the kernel of the radial basis function type. Decision tree [10] is formalism for expressing mappings from attribute values to classes and consists of tests or attribute nodes linked to two or more sub trees and leafs or decision nodes labeled with a class which means the decision. Because of the very simple representation of accumulated knowledge, they also give us the explanation of the decision, and that is essential in medical applications. Weka package [33] was utilized to construct a decision tree.

Respecting to classification performance in Table 6; performance on class A reported by *F*-measure is approximately the same in the three methods since it is the majority class. By applying our method performances of classes B and C are significantly better than the other two methods with *F*-measure and also the overall *G*-mean. The *F*-measure of class B using decision tree is better than that of weighted SMV. But weighted SVM achieves best *F*-measure than decision tree for class C.

As illustrated in Table 7, the classification quality of class B and class C was improved by more than 11% when compared with the other two methods but the proposed method still outstand the other two methods by more than 19%. Also like the 808 data set, the *F*-measure of class B using decision tree is better than that of weighted SMV. But weighted SVM achieves best *F*-measure than decision tree for class C.

As reported in Table 7, there are only two misclassified cases; the experts conclude that the reason of misclassification is different for each case. The first case patient number 24, she is a case of indirect hyperbilirubinemia along her duration of stay (class A) but an error

Table 6
Confusion matrix and performance measures for 808 data set.

Predicted		A	B	C	Measure	TPRATE	Recall	Precision	*F*-Measure	*G*-Mean
Actual	A	263	0	0	Proposed method	1	1.00	1.00	1.00	0.98
	B	0	16	1		1	0.94	1.00	0.97	
	C	0	0	7		0.88	1.00	0.88	0.93	
Actual	A	293	1	0	Decision tree	0.99	0.99	0.99	0.99	0.74
	B	4	12	2		0.66	0.66	0.7	0.68	
	C	0	4	7		0.63	0.63	0.63	0.63	
Actual	A	293	1	0	LIBSVM	0.99	0.99	0.97	0.98	0.73
	B	0	13	5		0.722	0.72	0.76	0.74	
	C	2	3	6		0.54	0.54	1	0.7	

Table 7
Performance measure of 325 data set.

Predicted					Measure	TPRATE	Recall	Precision	*F*-Measure	*G*-Mean
		A	B	C						
Actual	A	103	0	1	Proposed method	0.99	0.99	0.99	0.99	0.98
	B	1	16	0		1	0.94	1.00	0.97	
	C	0	0	9		0.9	1.00	0.90	0.95	
Actual	A	103	2	0	Decision tree	0.98	0.98	0.98	0.98	0.80
	B	1	13	1		0.87	0.86	0.72	0.788	
	C	1	3	6		0.6	0.6	0.85	0.706	
Actual	A	104	1	0	LIBSVM	0.99	0.99	0.95	0.97	0.86
	B	2	13	0		0.86	0.86	0.86	0.86	
	C	3	1	6		0.6	0.75	0.6	0.99	

occurred and this case classified as class C due to high levels of direct bilirubin at day of presentation and 24 h later (approximately 18.8% and 19% of total bilirubin levels, respectively).

6. Conclusions

Classical rough set rough just works in nominal domain and treats each class as equal weight. Moreover, Classification of data with imbalanced class distribution has posted a significant drawback of the performance attainable by most standard classifier learning algorithm. In this paper, we have proposed a weighted rough set framework for early intervention and prevention of neurological dysfunction and kernicterus that are catastrophic sequels of neonatal jaundice. The proposed framework tackles the class imbalance problem with multiple classes.

In our collected data set, the ratio of majority class to the other minority classes is 16.02% and 29.48%, respectively.

The performance of the proposed framework on minority classes increases significantly reported by recall value, *F*-measure and *G*-mean compared with weighted SVM and decision trees.

References

[1] http://pediatrics.aappublications.org/cgi/content/full/108/3/763 (accessed on May 14, 2010).

[2] N. Zhong, A. Skowron, Rough sets based knowledge discovery process, International Journal of Applied Mathematics and Computer Science 11 (3) (2001) 603–619.

[3] S. Carlin, J. Komorowski, A. Ohrn, Rough set analysis of medical datasets: a case of patients with suspected acute appendicitis, in: Proceedings of Workshop on Intelligent Data Analysis in Medicine and Pharmacology (ECAI 98), Brighton, UK, 1998, pp. 18–28.

[4] K. Cios, W. Pedrycz, R. Swiniarski, Data Mining Methods for Knowledge Discovery, Kluwer Academic, Boston, MA, USA, 1998.

[5] N. Lavrajc, E. Keravnou, B. Zupan, Intelligent Data Analysis in Medicine and Pharmacology, Kluwer Academic, Boston, MA, USA, 1997.

[6] G. Beligiannis, I. Hatzilygeroudis, C. Koutsojannis, J. Prentzas, A GA driven intelligent system for medical diagnosis, knowledge-based intelligent information and engineering systems, Lecture Notes in Computer Science 4251 (2006) 968–975.

[7] A.V. Adamopoulos, P.A. Anninos, S.D. Likothanassis, G.N. Beligiannis, L.V. Skarlas, E.N. Demiris, D. Papadopoulos, Evolutionary self-adaptive multimodel prediction algorithms of the fetal magnetocardiogram, in: Proceedings of the 14th International Conference on Digital Signal Processing (DSP 2002), 2002, pp. 1149–1152.

[8] G.N. Beligiannis, L.V. Skarlas, S.D. Likothanassis, K.G. Perdikouri, Nonlinear model structure identification of complex biomedical data using a genetic-programming-based technique, IEEE Transactions on Instrumentation and Measurement 54 (December (6)) (2005) 2184–2190.

[9] H.S. Own, Aboul Ella Hassanien, Rough sets: a medical knowledge discovery (RS-MKD) scheme for prostate cancer data analysis, International Journal of Computer Science and Software Technology (IJCSST) 2 (2) (2009).

[10] D. Ślęzak, Various approaches to reasoning with frequency based decision reducts: a survey, in: L. Polkowski, S. Tsumoto, T.Y. Lin (Eds.), Rough Sets in Soft Computing and Knowledge Discovery: New Developments, Physica Verlag, Heidelberg, Germany, 2000.

[11] Z. Pawlak, Rough sets, International Journal of Computer and Information Science 11 (1982) 341–356.

[12] Z. Pawlak, Rough Sets Theoretical Aspect of Reasoning about Data, Kluwer Academic, Boston, MA, USA, 1991.

[13] Z. Pawlak, J. Grzymala-Busse, R. Slowinski, W. Ziarko, Rough sets, Communications of the ACM 38 (11) (1995) 89–95.

[14] J. Liu, Q. Hu, D. Yű, A weighted rough set based method developed for class imbalance learning, International Journal Of Information Science (Elsevier) 178 (2007) 1235–1256.

[15] Z. Pawlak, A. Skowron, Rudiments of rough sets, Information Sciences 177 (1) (2007) 3–27.

[16] J. Liu, Q. Hu, D. Yu, A comparative study on rough set based class imbalance learning, International Journal of Knowledge Based System (Elsevier) 21 (2008) 753–763.

[17] Q.H. Hu, D.R. Yu, Z.X. Xie, J.F. Liu, Fuzzy probabilistic approximation spaces and their information measures, IEEE Transaction on Fuzzy System 14 (2) (2006) 191–201.

[18] J. Stefanowski, S. Wilk, Rough sets for handling imbalanced data: combining filtering and rule based classifiers, Fundamental Information 72 (1) (2006) 391–479.

[19] L. Polkowski, Rough Sets: Mathematical Foundations, Physica Verlag, Heidelberg, Germany, 2003.

[20] J.A. Starzyk, N. Dale, K. Sturtz, A mathematical foundation for improved reduct generation in information systems, Knowledge and Information Systems Journal (2000) 131–147.

[21] J. Bazan, H.S. Nguyen, S.H. Nguyen, P. Synak, J. Wróblewski, Rough set algorithms in classification problem, in: L. Polkowski, S. Tsumoto, T.Y. Lin (Eds.), Rough Set Methods and Applications: New Developments in Knowledge Discovery in Information Systems, Physica Verlag, Heidelberg, Germany, 2000, pp. 49–88.

[22] G. Gediga, I. Duentsch, Statistical techniques for rough set data analysis, in: Rough Sets: New Developments, Physica Verlag, Heidelberg, Germany, 2000, pp. 545–565.

[23] W. Jerzy, Grzymala-Busse, MLEM2—discretization during rule induction, in: Proceedings of the IIPWM'2003, International Conference on Intelligent Information Processing and WEB Mining Systems, Springer-Verlag, Zakopane, Poland, June 2–5, 2003, pp. 499–508.

[24] J.W. Grzymala-Busse, MLEM2 rule induction algorithms: with and without merging intervals, Studies in Computational Intelligence (SCI) (Springer-Verlag) 118 (2008) 153–164.

[25] J. Li, N. Cercone, A rough set based model to rank the importance of association rules, in: Proceedings of the 7th International Workshop on New Directions in Rough Sets, Data Mining, and Granular-Soft Computing (RSFD-GrC'99), Springer, Berlin, Germany, Lecture Notes in Computer Science (2005) 109–118.

[26] X. Guo, Y. Yin, C. Dong, G. Yanga, G. Zhou, On the class imbalance problem, IEEE Fourth International Conference on Natural Computation 4 (2008) 192–201.

[27] V. Garcia, J.S. Sanchez, R.A. Mollineda, R. Alejo, J.M. Sotoca, The Class Imbalance Problem In Pattern Classification And Learning, http://www.lsi.us.es/redmidas/CEDI07/%5B30%5D.

[28] Y. Liu, B. Feng, G. Bai, Compact Rule Learner on Weighted Fuzzy Approximation Spaces For Class Imbalanced And Hybrid Data, RSCTC, Springer-Verlag, Berlin, 2008, pp. 262–271.

[29] M. Kubat, S. Matwin, Addressing the curse of imbalanced training sets: one sided selection, in: Proceedings of the Fourteenth International Conference on Machine Learning, Morgan Kaufmann, Nashville, TN, USA, July 8–12, 1997, pp. 186–197.

[30] A. Fernandes, M. Jose del Jesus, F. Herrera, Hierarchical fuzzy rule based classification system with genetic rule selection for imbalanced data-sets, International Journal of Approximate Reasoning 50 (2009) 561–577.

[31] Y. Sun, M.S. Kamel, Y. Wang, Boosting for learning multiple classes with imbalanced class distribution, in: Proceedings of the Sixth International Conference on Data Mining, Las Vegas, NV, USA, June 26–29, 2006.

[32] C.C. Chang, C.J. Lin, LIBSVM: A Library for Support Vector Machine, http://www.csie.ntu.edu.tw/cjlin/libsvm, 2001.

[33] http://www.cs.waikato.ac.nz/~ml/weka/.

[34] http://emedicine.medscape.com/article/974786-overview (Updated: March 24, 2010).