A Novel-weighted Rough Set-based Meta Learning for Ozone Day Prediction

Hala S. Own[1], Ajith Abraham[2]

[1]Department of Solar and Space Research, National Research Institute of Astronomy and Geophysics, El-Marsad Street, P. O. Box 11421 Helwan, Egypt. hala@cs.ku.edu.kw

[2]Machine Intelligence Research Labs (MIR Labs), Scientific Network for Innovation and Research Excellence, WA, USA, ajith.abraham@ieee.org

Abstract: Nowadays, classifier combination methodsreceives great attention from machine learning researchers. It is a powerful tool to improve the accuracy of classifiers. This approach has become increasingly interesting, especially for real-world problems, which are often characterized by their imbalanced nature. The unbalanced distribution of data leads to poor performance of most of the conventional machine learning techniques. In this paper, we propose a novel weighted rough set as a Meta classifier framework for 14 classifiers to find the smallest and optimal ensemble, which maximize the overall ensemble accuracy. We propose a new entropy-based method to compute the weight of each classifier. Each classifier assigns a weight based on its contribution to classification accuracy. Thanks to the powerful reduct technique in rough set, this guarantees high diversity of the produced reduct ensembles. The higher diversity between the core classifiers has a positive impact on the performance of minority class as well as on the overall system performance. Experimental results with ozone dataset demonstrate the advantages of weighted rough set Meta classifier framework over the well-known Meta classifiers like Bagging, boosting and random forest as well as any individual classifiers.

Keywords: Weighted Rough Set; real world web service; class imbalance learning; entropy

1 Introduction

Practical experience has indicated that hybrid intelligence techniques might be helpful to solve some of the challenging real-world problems. In an hybrid intelligence system, a synergistic combination of multiple techniques is used to build an efficient solution to deal with a particular problem. One field of the hybrid intelligent approaches that has recently become a topic for researchers is Meta learning.

Meta learning refers to employing a set of base predictors for a given classification task and then fuse the output information using a fusion technique. Meta learning approach can be found under different names in literature such as decision combination [9], mixture of experts [10], classifier ensembles [5], classifier fusion [16] consensus aggregation [7], hybrid methods [8] and more.

The main purpose of Meta learning techniques is to improve the performance of a single classifier. Different classifiers usually make different predictions on the same sample of data. This is due to their diversity and many research works illustrated that the sets of misclassified samples from different classifiers would not necessary overlap [28]. This observation motivated the idea of using multiple sets of classifiers.

The techniques used to develop Meta learning (multi classifiers) can be divided into two categories: classifiers disturbance and sample disturbance. The first approach utilizes the instability of the base classifiers. This is applied to classifiers which are very sensitive to the initialization parameters like neural networks, random forests, and decision trees. The second approach even trains the classifier with different sample subsets or to train classifiers in different feature subspaces. Bagging works by resampling the original training data set of size M to produce N bootstrap training data sets of size M. Each of the bootstrapped training data sets are then used to train the classifier. Boosting on the other hand generates a series of base models. Each model is learned from a weighted training set whose weights are determined but the classification error of the preceding model [28]. The Adaboost was introduced by Freund et al. [14] and is based on weighting the data instead of randomly sampling it by putting more weight on the misclassified examples and smaller weights on the correctly classified examples.

Staking [21] is concerned with combining multiple classifiers generated from using different learning algorithms on a single dataset. This task is performed through different phases. In the first phase, the learning is performed individually for each classifier to output a new data set. In the second phase, a meta-level classifier is learned that combines the outputs of each individual classifier.

ECOC is a technique thatmanipulates output labels of the classes [12]. In the ECOC method, a discrete decomposition matrix (code matrix) is first defined for the problem at hand. Then this problem is decomposed into a number of binary sub-problems according to the content of the code matrix. After training binary classifiers on these sub-problems and testing them on any incoming test sample, a binary output vector is created. The final label is assigned to the class with the shortest distance between this vector and the code words.

Jin and Liu proposed a novel method for heterogeneous data [17]. The classifier system divided heterogeneous data into homogeneous subsets of similar sizes in order to generate reliable and accurate classification models. They proposed a novel algorithm, HISS, which allows for data overlapping between different

clusters (strata) and promises size-balanced clusters. The partitioning method was shown to perform well with heterogeneous data classification.

Akdemir [23] proposed the logic rule ensembles approach for unsupervised and semi-supervised cluster learning. They constructed the target variables by mapping the input variables. Each of these target variables can be used to extract several rules, and overall cluster rules are obtained from combining the rules for many target variables into an ensemble distance matrix, D(x). The clustering of the observations is accomplished by applying a distance-based hierarchical clustering algorithm to the rule-based distance matrix D(x). They use the cluster-based similarity partitioning method to combine the clusters from many rules.

Therest of the paper is organized as follows: motivation and review about imbalanced data learning are presented in section 2. Section 3 gives a brief introduction to the rough sets. Section 4 discusses the proposed weighted function. The proposed weighting rough set based meta learning is discussed in section 5 in detail. The characteristics of the Ozone data set as well as the chosen meta base classifiers are presented in Section 6. In Section 7 the proposed performance evaluation measures used in the paper are introduced. Experimental analysis and discussion of the results are described in Section 8. Finally, conclusions are presented in Section 9.

2 Imbalanced data Problem

Meta learning is one of the suggested techniques to deal with the class imbalance problem, a currently popular research area..Imbalanced data means that one of the classes has more samples than the other classes. The class with more samples is called the majority class while the other is the minority class. In many applications the minority class holds the most important information, such as in disease prediction, fraud detection, risk management, natural and physical phenomena, etc. Most classification techniques perform poorly with the minority class. There are three suggested techniques to overcome imbalanced data problems. The first is to create or modify the existing classification algorithms to deal with class imbalance problems. Data resampling is the second technique which includes over sampling or under sampling the data set to adjust the size of data set. The last approach is the feature selection, which was recently used to select a subset of features that allow the classifier to reach optimal performance [29].

The aim of modifying the algorithm is to provide adjustments on the learning algorithm (decision tree, regression, factor analysis, etc.) so it is more relevant and appropriate to imbalanced data situations. This approach is used mainly with decision tree and support vector machines (SVM); however few studies were done through this approach, since and opportunities within it are limited [27].

The main drawback of data resampling approach is that it may exclude useful information or increase the data size with artificial samples without having a real impact in the classification process, which will probably lead to the over-fitting problem [15]. The feature selection approach is applicable only for high dimension data, it selects data features that have great impact in classification of different classes, however, its performance in solving imbalanced data problems depends on the nature of the application domain.

Therefore, in imbalanced data learning, the unique optimal solution does not exist [27]. The different approaches were recently combined andapplied to SVM [26], and it had a better performance than applying separate techniques. However it is known that the SVM learning algorithm is sensitive to outliers and noise present in datasets, and it needs more work to reduce the effect of these problems.

Rough set theory [3, 6, 11] is a fairly new intelligent technique that has been applied to different domains and is used for the discovery of data dependencies, evaluates the importance of 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.

Recently, there has been a few papers introducing rough set into imbalance learning techniques. Hu et al. [20] used rough set as an ensemble model generation. They proposed attribute reduction algorithms to find a set of reduct and trained a base classifier with each reduct. Then they presented an accuracy-guided forward search and post-pruning strategy (FS-PP) to select parts of base classifiers for ensemble systems. As ensemble system is to ensemble multiple rough subspace, they called it FS-PP-EROS. On the other hand, Saha et al. [21] used rough set as ensemble combination. They combined the results of a number of individually trained classifiers to construct a decision table. Then rough set attribute reduction and rule generation processes were used to construct Meta classifiers. However, the main disadvantage of the previous algorithm is that they consider all classifiers to act on the classification performance equally likely. Moreover, it is known that the performance of the ensemble members is not uniform. Therefore when we considered an equal weight for each one this negatively affected the performance [22].

In this paper we propose a hybrid approach combining algorithm modification and feature selections to solve the class imbalance problem. A modification of classical rough set theory is proposed by introducing a new weighting function. We apply a weighted entropy-based function to build a weighted Meta information table. By using this method, samples (classifiers) are weighted by its local contrast entropy of the training set. After building our weighted Meta information table, a weighted rough set reduction technique, which was proposed in our previous work [22], is applied to find the core base classifiers. This step will guarantee high diversity between ensembles. The higher diversity between the core classifiers has a positive impact on the performance of minority class as well as in overall system

performance [28]. Finally, a set of classifications rules are extracted based on a modified version of MLEM2 called a weighted MLEM2 algorithm [22]. We apply our scheme to an ozone data set, a highly imbalanced dataset.

The generated rules will be able to classify the minority class (ozone day) with high accuracy.

3 Rough Sets: Basic Notation

3.1 Information System and Approximation

Definition 1 (Information System) An 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 $a : U \to V_a$, where Va is the value set of a. In the applications, we often distinguish between conditional features, C, and decision features, D, where C ∩ D = ϕ . In such cases, we define decision systems (U, C, D).

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

$$Ind_B = \{(x, y) \in UXU : \forall_{a \in B} a(x) = a(y)\} \tag{1}$$

for every x $\in$ U, $[x]_B$ is an equivalence class in the partitioning 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,

$$\underline{B}Y = \{x \mid [x]_B \subseteq X\} \tag{2}$$

$$\overline{B}Y = \{x \mid [x]_B \cap X \neq \phi\}$$

Definition 4 (Lower Approximation and Positive Region) The positive region $POS_C(D)$ is defined by

$$POS_C(D) = \bigcup_{X : X \in U / \overline{Ind}_D} \underline{C}X \quad ; \tag{3}$$

$POS_C(D)$ is the set of all objects in U that can be uniquely classified by elementary sets in the partition U/IndD by means of C [15].

Definition 5 (Upper Approximation and Negative Region) The negative region NEGC(D) is defined by

$$NEG_C(D) = U - \bigcup_{X:X \in U/Indo} \overline{C}X \tag{4}$$

that is the set of all objects that can be definitely ruled out as a 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:

$$BND_B(X) = \overline{B}X - \underline{B}X \tag{5}$$

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

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

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.

3.2 Reduct and Core

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

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

Definition 8 (Reduct) Given a classification task related to the mapping $C \vec{\square} D$, A reduct is a subset $R \subseteq C$ such that

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

and none of the 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

$\Re = \{A \subseteq P(C) : (A, D) = (C, D)\}$. 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 the core.

3.3 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:

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

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 [11].

4 The Proposed Weighting Function

Several weighting functions have been introduced in ensemble learning. The most primitive one is simple voting. In the simple voting, the final decision is taken according to the number of votes given by the individual classifiers. Unfortunately, Matan [13] verified that in some cases, simple voting might perform worse than any of the members of combined classifiers. Therefore, several weighting voting methods were proposed to tackle this problem [2, 3, 5]. In this approach the decision of each classifier is multiplied by a weight to reflect its individual confidence in the decision.

In this paper, we introduce an entropy-based method to compute the weight of each classifier. We define the Local Contrast Entropy (LCE) function which is based on the relationship between each classifier and the overall entropy. We were motivated by the fact that if the classifier has a higher local contrast entropy it means that it makes a significant contribution to the classification accuracy. The

fundamental concept of the proposed technique is to reward the individual classifier a weight according to its local contrast entropy.

Entropy is widely used for the measuring of local information content or uncertainty and the information content in a probability distribution [1]. The entropy function is calculated by the following formula:

$$H = -\sum_{i=1}^{N} P_i \log P_i \tag{10}$$

To take into account the classification accuracy of each classifier in classifying a minority class,let D= {D1, D2, ….., DN} be the set of N classifiers,where D is considered to be a set of individual variables.

Each classifier Di assigns an input feature vector x to one of the possible classes C.

We can define the local contrast entropyas follows:

$$L(¿¿i) = \frac{D}{\sum_{j=1}^{N} D_j(x)} \tag{11}$$

¿.

where , Di(x) is the classification accuracy of the classifier Di in classifying the minority class.

Therefore our idea is to assign for each classifier a weight equal to its local contrast entropy:

$$w(D_i) = \frac{D_i}{\sum_{j=1}^{N} D_j} \tag{12}$$

This weight represents its ability to correctly predict the minority class.

5 Weighted Rough Set Based Meta Classifier

In this paper, we train a set of different classifiers D= {D1, D2, ….., DN} on an ozone dataset. We will divide the data into three sets, atraining, testing and validation set. The training set used to train each classifier to build the classification model for each classifier. Then, the generated model is then tested using the testing set. The output of each classifier Dj on sample xi is di(xi).

Now we will define the new weighted meta decision table:

Definition (Weighted Meta Decision Table) The weighted meta decision table is a tuple (U, D, Dec), where U consists of objects and D consists of features. Every d $\in$ D corresponds to the function $d : U \to V_d$ where Vd is the value set of d. Dec is the decision feature, where Dec ∩ D =φ.

In our proposed approach the objects are a set of trained classifiers. Each classifier generates an instance in the meta decision table containing the prediction made by the classifier, as conditional feature D, and the class label as decision feature Dec.

Informally speaking, in the meta decision table the columns represents the classifier name and the rows represents the Ozone instance data in the validation set. The values represents the prediction of the corresponding classifier which reflects its correctness in the classification process. The next step in the proposed approach is to form a weighted meta decision table. As shown in Table 1. The entry in information table U is defined as:

$Ui_{,}j= w(Dj)$ if training sample xi is classified correctly by base classifier Dj.

$Ui_{,}j= 0$ otherwise.

Where w (Dj) calculated by equation 12. The decision class in this table represents the actual class for the Ozone day data.

Table 1

Weighted meta Decision table

ID	BFTREE	J48	RD	RPFTREE	DT	PART	MLP	RBFN	SMO	BAYES	NAIV	IBK	LWL	LAZYSTAR	OZONE
x1	0.909	0.895	0.934	0.920	0.929	0.890	0.891	0.937	0.937	0.692	0.621	0.883	0.937	0.891	NOT OZONE
x2	0.909	0.895	0.934	0.920	0.929	0.890	0.891	0.937	0.937	0.692	0.000	0.883	0.937	0.891	NOT OZONE
x3	0.909	0.895	0.934	0.920	0.929	0.890	0.891	0.937	0.937	0.692	0.621	0.883	0.937	0.891	NOT OZONE
x4	0.000	0.000	0.000	0.000	0.000	0.890	0.000	0.000	0.000	0.692	0.621	0.000	0.000	0.000	OZONE
x5	0.909	0.895	0.934	0.920	0.929	0.890	0.891	0.937	0.937	0.692	0.621	0.883	0.937	0.891	NOT OZONE
x6	0.909	0.895	0.934	0.000	0.929	0.890	0.891	0.000	0.000	0.692	0.621	0.883	0.000	0.000	OZONE
x7	0.909	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.692	0.621	0.000	0.000	0.000	OZONE
x8	0.909	0.895	0.934	0.920	0.929	0.890	0.891	0.937	0.937	0.692	0.621	0.883	0.937	0.891	NOT OZONE
x9	0.909	0.895	0.934	0.920	0.929	0.890	0.891	0.937	0.937	0.692	0.621	0.883	0.937	0.891	NOT OZONE
x10	0.000	0.000	0.934	0.920	0.929	0.000	0.891	0.937	0.937	0.692	0.621	0.000	0.937	0.000	NOT OZONE

Once we build the decision table; the next step is to reduce the attributes in the data set based on the information content of each attribute or collection of attributes. Generally in information tables, there often exist conditional attributes that do not provide significant information for identifying the decision class. So we should remove those attributes, since it reduces complexity and cost of the decision process [11].

Our aim in this step is to find a subset of base classifiers that maximize the overall accuracy for each decision class. Our motivations in this step are the following:

Some of the base classifiers produce good classification results for one of the decision classes but not all. Producing a reduct set for each class will decrease the overall complexity since we will use only a subset of the base classifiers.

The important effect of reduct set extraction is that we will know which base classifiers are more significant for each decision class.

By extracting the reduct set, we exclude all redundant classifiers. As a consequence, we guarantee diversity. Diversity among the base classifiers is considered important when constructing a classifier ensemble.

For the new weighted meta decision table, the weights generated do not change the equivalence relation and do not change the upper and lower approximation of arbitrary subset [22].

Finally, a set of classification rules are extracted based on a modified version of MLEM2, called the weighted MLEM2 algorithm [22]. This process leads to the final goal of generating classification rules from the information or decision system of the Ozone day database. Figure 1 shows the overall steps in the proposed Weighted Rough Set based Meta Classifier.

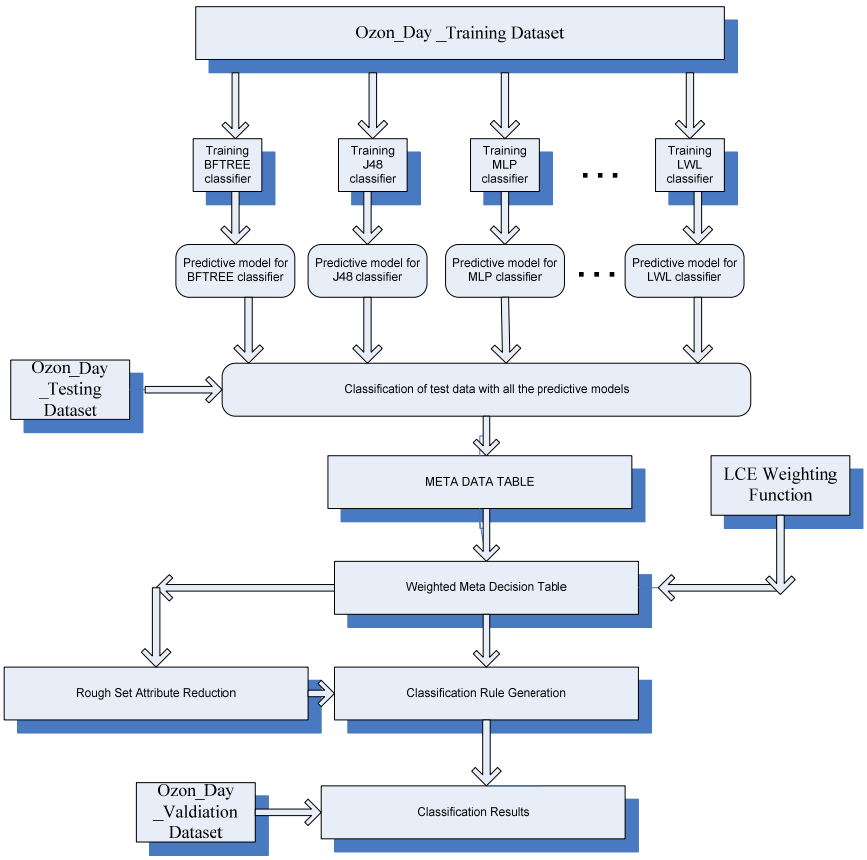

Figure 1
A Weighted Rough Set based Meta Classifier Scheme

6 Case Study: Ozone Day Prediction

6.1 Dataset

In this work, we use the dataset in the UCI machine learning repository [30] called "ozone". The number of objects is 2534 with 71 conditional attributes and a decision attribute. More information about the dataset can be found in [18]. The data set possesses very important properties that makes it a good example for imbalanced data learning:it contains high dimension of features, andit is very biased towards one of its decision classes. Table 2 describes the class distribution within the data set. As shown in Table 2, the NOZON class is small 6.3% compared with the NONOZON classes 93.7%.

Table 2
Class distribution on ozone data set

Class name	class size	class distribution
NONOZONE	2374	93.7%
OZONE	160	6.3%

With most well-known classifiers, when the class distribution of a data set is skewed, the classification method will be biased to the majority classes, and therefore will perform poorly in recognition of the minority classes because the prior knowledge of class distribution is not taken into account. Other classifiers will ignore the minority classes and will treat them as noise [26]. In this paper, we apply our proposed approach to try to overcome these well-known problems of inbalanced data.

6.2 Base Classifiers Proposed

We use 14 known classifiers:

RFTree: for building a best- first decision tree classifier.

J48: generating a pruned or unpruned C4.5 Decision Tree.

DT: DecisionTable: using simple decision table majority classifier.

MLP: Multilayer perceptron with number of different hidden layers chosen as (attribs + classes) / 2.

RD: RandomForest: generating a forest of random trees.

REPTree : Fast Decision tree learner.

PART: using divide and conquer to generate a PART decision list.

RBFN: RBFNetwork: implementation of normalized Gaussian radial basis function network.

SMO: implementation of Sequential minimal optimization algorithm for training a support vector classifier.

BNet: BayesNet, Bayes network learner using various search algorithms and quality measures.

NV: NaiveBayes, using naïve base classifier with estimator class.

IBk: K- nearest neighbors classifier.

LWL: Locally weighted learning.

lazy.KStar: instance based classifier; it uses an entropy-based distance function.

7 Performance Evaluation Measure

The evaluation criterion is a key factor both in the assessment of the classification performance and guidance of the classifier modelling. Traditionally, the accuracy rate has been the most commonly used empirical measure. The validation is done on a validation data set to guarantee the split fairness; we use 10 fold cross-validations in all of our measures. The confusion matrix is a matrix of size CxC, and each entry in the matrix Ai,j represents the number of instances observed in class Ci, and classified in class Cj. The rows of the matrix represent the set of classes and the column represents the classification result for each class. When working with skewed data, accuracy doesn't adequately serve as a measure for the success of ensembles as it is strongly biased with the majority class. What we need is another measure to correctly distinguish between the numbers of correctly classified examples of different classes. In this study, we used the following performance measures [19]:

Recall: Theratio of the number of correctly classified instances to the number of total instances of that class.

Precision: The ratio of number of correctly classified instance of the class to the number of predicted instance to that class.

F-measure: The weighted average of the precision and recall.

The most important property of these metrics is that they can be distinguished between positive and negative classes independently. Therefore it gives us a clear view inside the classification method especially when dealing with imbalanced data as we search for a good performance in the minority class.

8 Empirical Analysis

By applying the rough set reduct generation; we computed the weighted dependency degree and the classification performance for each classifier. We reached the minimal number of reducts that contains a combination of classifiers that have the same discrimination factor. The final generated reduct sets, which are used to generate the list of rules for the classification, are {LWL, RBFN, SMO, RPFTree, NV}. This reduct set represents the best classifiers for ozone day. After producing the reduct set; a set of rules will be generated. Table 3 introduces the best rules of the 34 generated decision rules.

The overall accuracy of the proposed technique is represented in Table 4. The table shows that the proposed system achieving high accuracy rate in predicting majority and minority class.

Table 3
The most significant rule generated

Rule #	Rule
1	RBFN==(0.467048-inf)& SMO ==(0.460142-inf)& RPFTree ==(0.464681-inf)& NV =(-inf-0.610093]=>OZON==NOTOZON
2	SMO==(-inf-0.46843]&LWL==(-inf-0.46843]&RBFNETWORK==(-inf-0.46843]=>(OZON==OZON)
3	(RONDOMF==(0.467048-inf)& RPFTREE==(0.460142-inf)& DECISIONTABLE==(0.464681-inf)& NV=(-inf- 0.702379]=> OZON=NOTOZON
4	RPFTree ==(0.447711-inf)&lwl==(0.460142-inf)& NV =(0.464681-inf)& SMO ==(0.610093-inf)=> OZON==NOTOZON

Table 4
The overall accuracy

	Non-Ozone	Ozone	No. of tested objects	Accuracy
Non-Ozone	955	2	957	0.998
Ozone	0	57	57	1
True Positive Rate	1	0.97		

8.1 Comparison with Other Meta Classifier Techniques

We compare the performance of the proposed technique with well-known meta classifiers such as Adaboost, bagging, and stacking. The compression done in terms of recall, precision and accuracy. Table 5 summarizes the performance of different meta classifiers when applied to the ozone data set.

Table 5
Comparing the proposed technique with other meta classifiers techniques

Method	Recall	Precision	Accuracy
Adaboost	98%	97%	96%
Stacking	96%	95%	92%
Bagging	95%	85%	89%
Our Method	100%	99%	99%

In the next experiments, we want to investigate the performance of our technique in predicting each class. The classification performance of each classifier in terms of recall, precision, accuracy and f measure for non ozone class is presented in Table 6.

In Table 7, the classification performance of each classifier in terms of recall precision, accuracy and f measure for ozone class is summarized.

Table 6
Classification performance for non ozone day

Measure / Classifier	RFTree	J48	DT	MLP	RD	REPTree	PART	RBFN	SMO	BNet	NV	IBk	LWL	lazy.KStar
PRECISION	0.95	0.95	0.94	0.95	0.94	0.94	0.95	0.93	0.93	0.98	0.98	0.95	0.93	0.95
RECALL	0.98	0.95	0.99	0.96	0.99	0.98	0.96	1	1	0.75	0.66	0.95	1	0.96
F-MEASURE	0.96	0.95	0.96	0.96	0.96	0.96	0.95	0.96	0.96	0.85	0.79	0.95	0.96	0.95
ACCURACY	0.98	0.96	0.99	0.96	0.99	0.98	0.96	1	1	0.75	0.66	0.95	1	0.96

Table 7
Classification performance for ozone day

Measure / Classifier	RFTree	J48	DT	MLP	RD	REPTree	PART	RBFN	SMO	BNet	NV	IBk	LWL	lazy.KStar
PRECISION	0.47	0.34	0.38	0.45	0.72	0.34	0.36	0	0	0.17	0.14	0.36	0	0.37
RECALL	0.24	0.26	0.05	0.38	0.05	0.1	0.32	0	0	0.78	0.85	0.35	0	0.31
F-MEASURE	0.32	0.29	0.08	0.42	0.09	0.15	0.34	0	0	0.29	0.25	0.35	0	0.34
ACCURACY	0.24	0.26	0.05	0.38	0.05	0.1	0.32	0	0	0.78	0.85	0.35	0	0.31

As evident from Tables 6 and 7, although the performance measures of the proposed classifiers in identifying a nonozone day are high its performance is very low with the ozone day. This is because of the imbalanceof the data, as we mentioned before. From table 8, we see that the overall accuracy of each classifier may give us a false indication about the ability of classifier to correctly identify the correct class of the instances.

Table 8
Performance of the14 individual classifiers

Measure / Classifier	RFTree	J48	DT	MLP	RD	REPTree	PART	RBFN	SMO	BNet	NV	IBk	LWL	lazy.KStar
Accuracy	93.5 %	92.1%	93. 4%	93.2 %	93.8 %	93.1 %	92.1 %	93.6%	93.6 %	75 %	67.8 %	92 %	93.6%	90.2%

8.2 Meta Learning Performance Measures

The studies have shown that the accuracy of prediction model in meta learning depend highly on the degree of diversity (difference) of the base classifiers. It has been proved to be one of the main reasons for the success of ensembles from both theoretical and empirical perspectives [28]. The larger the diversity value, the more evenly distributed the predictions are for the base classifiers, while a smaller

diversity value represents base classifiers that have more biased predictions.[25]. Therefore it is important to measure the diversity of our chosen classifiers.

8.2.1 Diversity Test

Meta learning techniques are expected to increase the prediction performance by considering the opinions from multiple classifiers. Therefore, diversity becomes an important factor. If every classifier gives the same opinion, constructing multiple classifiers becomes meaningless. Generally, larger diversity causes better recall for minority. To prove that the generation of reduct set guarantee high diversity between the different classifiers. We introduce two measures: a correlation measure between the different classifiers and a pair-wise Q-statistic.

8.2.2 Correlation Measure

We adopt the method used in [24] to calculate the correlation between two classifiers i and j. They calculated the total correlation between two classifiers as the sum of correlations of individual instances.

$$Total\ correlation(i, j) = \sum_{a} correlation(d_i, d_j) \tag{13}$$

Where di denotes the output of classifier i for the correct class on instance a , the results are shown in Table 9. We consider a correlation under 0.5 as a weak correlation (denoted as bold in the tables). As we see from the Table 9 The lowest correlated classifiers are LWL, RBFN, SMO, NV, and RPFTree. When the correlation becomes low the similarity between the classifiers is also low. In Meta learning we aim to group high diversity classifiers together to guarantee high performance predication. The results in Table 9 shows that lowest correlated classifiers, which constitute the reduct set generated by applying our method. This proves that the set of minimal classifiers satisfy the diversity required.

Table 9

Correlation between different classifiers for ozon day dataset

	RFTree	J48	DT	MLP	RD	REPTree	PART	RBFN	SMO	BNet	NV	IBk	LWL	lazyKStar
RFTree	1	0.8	0.88	0.74	0.79	0.75	0.81	0.87	0.68	0.56	0.56	0.58	0.76	0.54
J48	0.8	1	0.86	0.78	0.76	0.87	0.86	0.85	0.69	0.58	0.51	0.58	0.75	0.65
DT	0.88	0.86	1	0.81	0.82	0.52	0.79	0.79	0.78	0.64	0.61	0.64	0.66	0.69
MLP	0.74	0.78	0.81	1	0.9	0.89	0.85	0.86	0.84	0.74	0.74	0.65	0.85	0.66
RD	0.79	0.76	0.82	0.9	1	0.91	0.78	0.91	0.86	0.79	0.79	0.79	0.81	0.68
REPTree	0.75	0.87	0.52	0.89	0.91	1	0.91	0.94	0.51	0.36	0.26	0.76	0.45	0.69

PART	0.81	0.86	0.79	0.85	0.78	0.91	1	0.91	0.84	0.68	0.68	0.68	0.86	0.66
RBFN	0.87	0.85	0.79	0.86	0.91	0.94	0.91	1	0.86	0.69	0.69	0.69	0.86	0.68
SMO	0.68	0.69	0.78	0.84	0.86	0.51	0.84	0.86	1	0.33	0.22	0.82	0.39	0.84
BNet	0.56	0.58	0.64	0.74	0.79	0.36	0.68	0.69	0.33	1	0.38	0.95	0.42	0.75
NV	0.56	0.51	0.61	0.74	0.79	0.26	0.68	0.69	0.22	0.38	1	0.96	0.46	0.79
IBk	0.58	0.58	0.64	0.65	0.79	0.76	0.68	0.69	0.82	0.95	0.96	1	0.82	0.75
LWL	0.76	0.75	0.66	0.85	0.81	0.45	0.86	0.86	0.39	0.42	0.46	0.82	1	0.76
lazy.KStar	0.54	0.65	0.69	0.66	0.68	0.69	0.66	0.68	0.84	0.75	0.79	0.75	0.76	1

8.2.3 Q-Statistic

Between the different measures proposed to evaluate the diversity, the simplest pair Q-Statistic is widely used. In [28] authors mathematically and empirically prove that there is strong correlation between the Q-Statistic and the imbalance performance measurements we choose (Recall, Precision, F-measure).

Given two classifiers Di and, Dj the pair-wise Q-statistics measure are defined as the following:

$$Q_{i,j} = \frac{(TP \cdot FN - TN \cdot FP)}{(TP \cdot FN + TN \cdot FP)} \tag{14}$$

where

TP: the number of instances that are correctly classified by Di and Dj.

TN: the number of instances that are correctly classified by Di but incorrectly classified by Dj.

FP: the number of instances that are correctly classified by Dj but incorrectly classified by Di.

FN: the number of instances that are incorrectly classified by Di and Dj.

In this experiment the pair-wise Q-statistics between the 14 classifiers were calculated. The high value of pair-wise Q-statistics measure indicates that the diversity between them is low and vice versa. It means that ensemble learning using these two classifiers will not lead to good performance measures.

From Table 10, the low Q-statistics combination between classifiers (denoted as bold in the tables) matches the same result that the reduct set generated. For example the Q-statistics between NV and DT is high also between RD and PART. It indicates that the diversity between them is also high as well. On the other hand we found that the Q-statistics between NV and SMO is low (0.45) which indicates the diversity between them is very low; this emphasizes the output of the reduct that NV and SMO are in the reduct set.

Table 10
Pair-wise Q-statistics measure between 14 classifiers

	RFTree	J48	DT	MLP	RD	REPTree	PART	RBFN	SMO	BNet	NV	IBk	LWL	lazy.KStar
RFTree	1	0.9	0.86	0.5	0.65	0.68	0.55	0.68	0.85	0.91	0.94	0.89	0.65	0.67
J48	0.9	1	0.79	0.58	0.59	0.72	0.61	0.67	0.86	0.88	0.79	0.84	0.75	0.73
DT	0.86	0.79	1	0.54	0.61	0.65	0.54	0.64	0.88	0.79	0.91	0.86	0.84	0.68
MLP	0.58	0.58	0.54	1	0.89	0.96	0.94	0.94	0.72	0.69	0.78	0.74	0.63	0.63
RD	0.65	0.59	0.61	0.89	1	0.95	0.93	0.93	0.84	0.68	0.75	0.79	0.88	0.67
REPTree	0.68	0.72	0.65	0.96	0.95	1	0.86	0.94	0.37	0.44	0.29	0.81	0.44	0.8
PART	0.55	0.61	0.54	0.94	0.93	0.86	1	0.96	0.91	0.88	0.87	0.82	0.65	0.66
RBFN	0.68	0.67	0.64	0.94	0.93	0.94	0.96	1	0.68	0.69	0.69	0.72	0.84	0.78
SMO	0.85	0.86	0.88	0.72	0.84	0.37	0.91	0.68	1	0.31	0.45	0.55	0.41	0.49
BNet	0.91	0.88	0.79	0.69	0.68	0.44	0.88	0.69	0.31	1	0.35	0.89	0.51	0.58
NV	0.94	0.79	0.91	0.78	0.75	0.29	0.87	0.69	0.45	0.35	1	0.98	0.48	0.65
IBk	0.89	0.84	0.86	0.74	0.79	0.81	0.82	0.72	0.55	0.89	0.98	1	0.54	0.65
LWL	0.65	0.75	0.84	0.63	0.88	0.39	0.65	0.84	0.41	0.51	0.48	0.54	1	0.61
lazy.KStar	0.67	0.73	0.68	0.63	0.67	0.8	0.66	0.78	0.49	0.58	0.65	0.65	0.61	1

Conclusion

A variety ofMeta learning techniques have emerged recently. The ozone day prediction is an important issue due to its harmful effect onall creatures. The imbalance nature of the Ozone data set as well as the large number of features makes the prediction of the ozone day a challenging problem. In this paper we introduce a rough set as a Meta classifier technique with new entropy-based method to compute the weight of each classifier to improve the performance of ozone day prediction. The experiments show it to perform better over the well-known meta classifier techniques.

References

[1] Shannon, C.: The Mathematical Theory of Communication. Bell System Technical Journal, Vol. 27, 379-423, 1948

[2] S. A. Dudani: The Distance-weighted k-nearest Neighbor Rule. IEEE Trans. on Systems, Man and Cybernetics, Vol. 6:325-327, 1976

[3] Pawlak, Z.: Rough Sets, Journal of Computer and Information Science, Vol. 11, 341-356 (1982)

[4] B. V. Dasarathy: Nearest Neighbor Norms. NN Pattern Classification Techniques, IEEE ComputerSociety Press, Los Alamos, CA, 1991

[5] L. Hansen, P. Salamon: Neural Networks Ensembles. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 12(10), 993-1001, 1990

[6] Pawlak, Z.: Rough Sets. Theoretical Aspect of Reasoning about Data. Springer Verlag, 1991

[7] J. Benediktsson, P. Swain.: Consensus Theoretic Classification Methods. IEEE Transactions on System and Man Cybernetic, Vol. 22 (4), 688-704, 1992

[8] B. Dasarathy: Decision Fusion. IEEE Computer Society Press, Silver Spring, MD, 1994

[9] T. Ho, J. Hull, S. Srihari: Decision Combination In Multiple Classifier Systems. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 16 (1), 66-75, 1994

[10] M. Jordon, R. Jacobs: Hierarchical Mixtures of Expert and the EM Algorithm. Neural Computing, 181-214, 1994

[11] Pawlak, Z., Grzymala-Busse J., Slowinski R., Ziarko, W.: Rough Sets. Communications of the ACM, Vol. 38, No. 11, 89-95, 1995

[12] T. G. Dietterich and G. Bakiri: Solving Multiclass Learning Problems via Error-Correcting Output Codes, Vol. 2, 263-286, 1995

[13] O. Matan: On Voting Ensembles of Classifiers. Proceeding of the 13th International Conference on Artificial Intelligence, 84-88, 1996

[14] Y. Freund and R. E. Schapire: A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. Journal of Computer And System Sciences, Vol. 55, 119-139, 1997

[15] Kittler, J., Hatef, M., Duin, R. P. W., Matas, J: On Combining Classifiers. Pattern Analysis and Machine Intelligence, IEEE Transactions, Vol. 20, 226-239, 1998

[16] L. Kuncheva: Switching Between Selection and Fusion in Combining Classifiers: An Experiment. IEEE Transactions on System and Man Cybernetic Part B, Vol. 32 (2), 146-156, 2002

[17] Rong Ji, Huan Liu: A Novel Approach to Model Generation for Heterogeneous Data Classification. Proceedings of the 19th international joint conference on Artificial intelligence, IJCAI'05, Edinburgh, Scotland, UK.30 July-5 August, 746-751, 2005

[18] Kun Zhang, Wei Fan, XiaoJing Yuan, Ian Davidson, Xiangshang Li: Forecasting Skewed Biased Stochastic Ozone Days: Analyses and Solutions. Proceedings of the Sixth International Conference on Data Mining, Hong Kong, China 18-22 December,753-764, 2006

[19] Costa, E., Lorena, A., Carvalho, A., Freitas: A., A Review of Performance Evaluation Measures for Hierarchical Classifiers. Association for the Advancement of Artificial Intelligence AAAI, 1-6, 2007

[20] Hu Q, Yu D, Xie Z, Li X: EROS: Ensemble Rough Subspaces. Pattern Recognition, Vol. 40(12), 3728-3739, 2007

[21] Suman SahA C. A. Murthy, Sankar K. Pal: Rough Set Based Ensemble Classifier. Lecture Notes in Computer Science, Vol. 6743, 27-33, 2011

[22] Hala S. Own, Ajith Abraham: A New Weighted Rough Set Framework Based Classification for Egyptian Neonatal Jaundice. Applied Soft Computing. Elsevier Vol. 12(3), 999-1005, 2012

[23] Deniz Akdemir: Ensemble Clustering with Logic Rules. eprint arXiv:1207.396, http://arxiv.org/abs/1207.3961, 2012

[24] Aydın Ulaş, Olcay Taner Yıldız, Ethem Alpaydın: Eigenclassifiers For Combining Correlated Classifiers. Vol. 187, 109-120, 2012

[25] Joseph Pun Yuri Lawryshyn: Improving Credit Card Fraud Detection using a Meta-Classification Strategy. International Journal of Computer Applications, Vol. 56(10), 41-46, 2012

[26] Mikel G., Alberto F., Edurne B., Humberto B., Francisco H.: A Review on Ensembles for the Class Imbalance Problem: Bagging-, Boosting-, and Hybrid-Based Approaches. IEEE Transactions on Systems, Man, and Cybernetics, Vol. 42, 463-484, 2012

[27] Mohamed B., Taklit A. A: Imbalanced Data Learning Approaches Review. International Journal of Data Mining & Knowledge Management Process (IJDKP) Vol. 3, 15-33, 2013

[28] Shuo Wang, Xin Yao: Relationships between Diversity of Classification Ensembles and Single-Class Performance Measures. IEEE Transactions on Knowledge and Data Engineering, Vol. 25, 206-219, 2013

[29] Rushi L., Snehlata S. D., Latesh M.: Class Imbalance Problem in Data Mining: Review. International Journal of Computer Science and Network (IJCSN), Vol. 2, 226-230, 2013

[30] UC Irvine Machine Learning Repository. http://archive.ics.uci.edu/ml/ last acesses 2014