ICIC Express Letters
Volume **14**, Number **7**, July **2020**

 ISSN 1881-803X
pp. **653–659**

PREDICTING SEMINAL QUALITY WITH THE DOMINANCE-BASED ROUGH SETS APPROACH

Nassim Dehouche

Business Administration Division
Mahidol University International College
999 Phutthamonthon 4 Rd., Salaya, Phutthamonthon, Nakhon Pathom 73170, Thailand
nassim.deh@mahidol.edu

Received November 2019; accepted February 2020

Abstract. *The paper relies on the clinical data of a previously published study. We identify two very questionable assumptions of said work, namely confusing evidence of absence and absence of evidence, and neglecting the ordinal nature of attributes domains. We then show that using an adequate ordinal methodology such as the Dominance-based Rough Sets Approach (DRSA) can significantly improve the predictive accuracy of the expert system, resulting in almost complete accuracy for a dataset of 100 instances. Beyond the performance of DRSA in solving the diagnosis problem at hand, these results suggest the inadequacy and triviality of the underlying dataset. We provide links to open data from the UCI machine learning repository to allow for an easy verification/refutation of the claims made in this paper.*
Keywords: Decision support systems, Expert systems, Dominance-based rough sets approach, Diagnosis, Seminal quality

1. **Introduction.** Reports of a global decline in male fertility and declining sperm counts [1] have engendered significant academic interest in investigating the causes of this public health challenge. Among numerous approaches to undertake this challenge, artificial intelligence techniques offer promising decision support to clinicians in the detection of male fertility issues. In this regard, the classification of human sperm morphometry based on set standards has been a very successful line of research. Indeed, the visual appearance of sperm has been shown to correlate to male fertility potential [2] and automatic image processing techniques [3] can detect abnormal sperm shapes. An exhaustive study in [4] compared four supervised learning methods (1-Nearest Neighbor, Naive Bayes, decision trees and Support Vector Machine (SVM)) and three shape-based descriptors (Hu moments, Zernike moments and Fourier descriptors) for this task, finding that the best classification performance was achieved by the Fourier descriptor and SVM. More recently, deep learning techniques [5, 6] have been successfully applied to the same problem.

Another fruitful line of research aims at the early detection of male fertility issues based on lifestyle factor, which can indeed increase the success rate of treatment. However, despite their promises and predictive power, the performance of this approach is highly dependent on the quality and representativeness of the collected data. Thus, the present paper intends to highlight some existing limitations in the measurement of this aspect. We rely on the clinical data of [7], referred to as the "Assisted Reproduction" dataset, which was made publicly available on the UCI Machine Learning repository [8], by the first author of that publication, and used as a reference dataset by many studies [9, 10, 11].

DOI: 10.24507/icicel.14.07.653

2. **Data and Previous Results.** The dataset records 9 attributes pertaining to the lifestyle habits, socio-demographic and environmental factors and health status of 100 volunteers aged 18 to 36 years, who provided a semen sample analyzed according to the WHO 2010 criteria [12]. Based on this analysis, volunteers were classified into two classes *normal* (N), or *altered* (O), based on the sperm concentration they present. Table 1 lists and describes the attributes characterizing each volunteer. Figure 1 and Figure 2 respectively present the attribute values presented by *normal* and *altered* cases, in parallel coordinates form.

TABLE 1. List of attributes with their initial domains and normalized values

Attribute	Description	Domain	Normalization
Season	Season in which the analysis was performed	{Winter, Spring, Summer, Fall}	$\{-1, -0.33, 0.33, 1\}$
Age	Age of the volunteer at the time of analysis	$[18, 36]$	$[0, 1]$
Disease	Childish diseases (i.e., chicken pox, measles, mumps or polio)	{Yes, No}	$\{0, 1\}$
Trauma	Accidents or serious trauma	{Yes, No}	$\{0, 1\}$
Surgery	Surgical interventions	{Yes, No}	$\{0, 1\}$
Fever	High fevers in the last year	{Less than three months ago, More than three months ago, No}	$\{-1, 0, 1\}$
Alcohol	Frequency of alcohol consumption	{Several times a day, Every day, Several times a week, Once a week, Hardly ever or never}	$\{0.2, 0.4, 0.6, 0.8, 1\}$
Smoking	Smoking habit	{Never, Occasionally, Daily}	$\{-1, 0, 1\}$
Sitting	Number of hours spent sitting per day	$[0, 16]$	$[0, 1]$
Output	Diagnosis	{Normal, Altered}	$\{N, O\}$

This binary classification problem consists in predicting the *Output*, given the values of attributes *Season*, *Age*, *Disease*, *Trauma*, *Surgery*, *Fever*, *Alcohol*, *Smoking* and *Sitting*.

To address this problem, the authors compare the performance of three Artificial Intelligence methods, Decision Trees (DT), MultiLayer Perception (MLP) and Support Vector Machines (SVM), using the following classical performance indicators based on the numbers of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) in the classification of the available 100 cases.

- Classification accuracy (%) $= \frac{TP+TN}{TP+FP+FN+TN} \times 100$
- Sensitivity (%) $= \frac{TP}{TP+FN} \times 100$
- Specificity (%) $= \frac{TN}{FP+TN} \times 100$
- Positive predictive value (%) $= \frac{TP}{TP+FP} \times 100$
- Negative predictive value (%) $= \frac{TN}{FN+TN} \times 100$

Results for the three methods, as they appear in [7] are presented in Table 3.

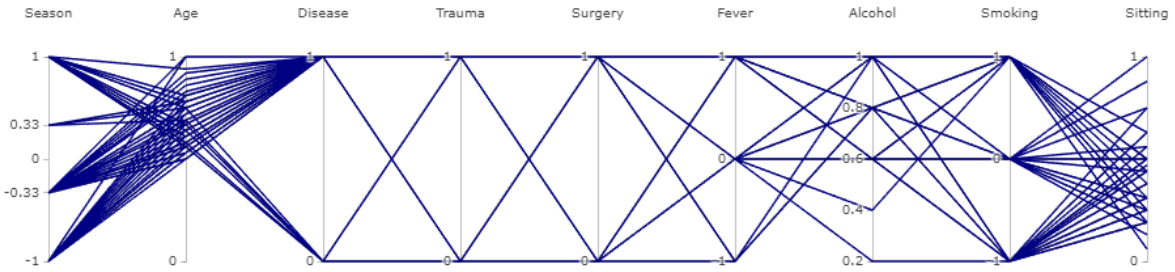

FIGURE 1. Parallel coordinates of the attribute values presented by *Normal* cases

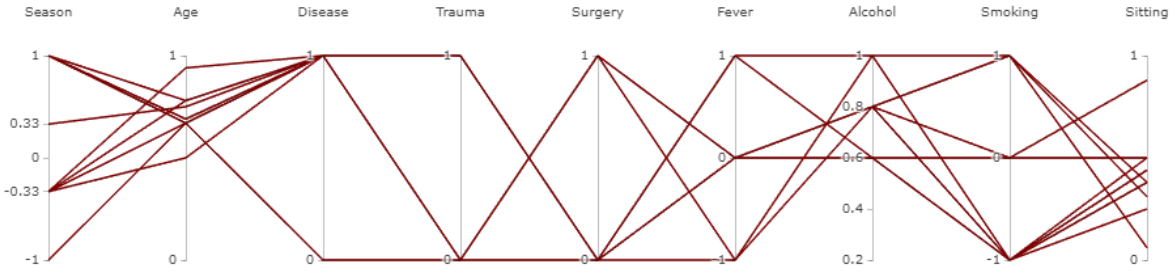

FIGURE 2. Parallel coordinates of the attribute values presented by *Altered* cases

3. **Literature Review.** We discuss two inconsistent assumptions in [7], namely confusing evidence of absence and absence of evidence (the former does not appear in environmental and lifestyle factors) and neglecting ordinal properties of attributes' domains. Thus the first questionable assumption of [7] lies in learning symmetrically from "Normal" and "Altered" cases. Indeed, there is a difference in nature between these two classes, in that the presence and interactions of the considered environmental factors can cause an Altered output. However, the absence of these environmental factors does not "cause", nor explain a Normal output. Consequently, we propose that learning should focus on the "Altered" class, the "Normal" class being considered a default class. Secondly, the environmental, lifestyle and occupational factors modeled by the classification attributes considered in [7] are known to negatively affect male fertility, which is admittedly the reason why they were considered in the survey of volunteers in that study. From a decision theoretic perspective, this means that attributes and classes are of an ordinal nature [13] and thus classification should be monotonic [14]. For instance, all other factors being equal a patient who consumes alcohol more frequently cannot generate a "Normal" output, when a patient who consumes less alcohol generates an "Altered" output.

More formally, for data of an ordinal nature, the monotonicity requirement [15, 16] states that given two objects a and b to be classified, if a presents values that are no worse than those presented by b, on each attribute, then a classification system should assign a to a class that is ranked at least as high as the class a is assigned. For the classification problem at hand, this property means that for the same season of analysis, the output of formally consistent classification system, for an older subject b who would present more severe values for childish disease, trauma, surgery, fever, alcohol consumption, smoking and would sit longer than a subject b, cannot be "Normal", when the output for b is

"Altered". The three classification methods considered in [7] (DT, MLP and SVM) do not ensure that this common sense property is satisfied. In addition to ensuring the formal consistency of a classification system when the input and output are of an ordinal nature, ensuring that a classification system satisfies the monotonicity requirement facilitates the detection of inconsistencies among learning examples and substantially reduces the number of classification rules [15].

4. **Proposed Approach.** The Dominance-based Rough Sets Approach (DRSA) is an extension of the classical rough theory introduced by Pawlak [17] that explicitly gives consideration to attributes of an ordinal nature [18]. This approach is based on the dominance binary relation and computes two sets, known as the *upward* and *downward unions* associated for each class of the output.

The *upward union* associated with a class is composed of said class and all classes ranked higher, when the *downward union* the considered class and all classed ranked lower. Similarly, given a learning case a, the *dominating set* associated with it is defined as the set of all learning cases whose values on all attributes are at least as high as those of a. Finally, the *dominating set* associated with a is the set of all learning cases that do not present any value that is higher than that presented by a, on any attribute.

Logical rules induced through the DRSA aim at approximating the *upward* or *downward unions* of classes and have a classical "If (Conditions) Then (Output)" form, in which (Conditions) is a conjunction of elementary conditions in the form of lower or upper bounds on the attributes, and (Output) is an assignment to an *upward downward union* of classes. For the *upward union* $Cl^{\geq}$ (resp. the *downward union* $Cl^{\leq}$) of an output class $Cl^{\leq}$, the induced logical rules would suggest that an object satisfying their corresponding logical conditions should be assigned *at least* (resp. *at most*) to class Cl_t.

Finding a rule base G that would exhaustively cover all learning cases, with a minimum number of rules is known to be an NP-Hard problem [19]. The DOMLEM algorithm [18] aims at minimizing this number heuristically. Given a set of attributes F, let us denote by $F' \subseteq F$ a subset of attributes over which the elementary conditions of rules are stated, E denotes a conjunction of elementary conditions e, that is candidate to constituting the condition part of an elementary condition, while $[E]$ is the notation for a set of cases covered by E. In the DOMLEM algorithm E would be accepted as the condition part of a rule, if and only if $\cap_{e \in E}[e] \subseteq B$, where B is an *upward* or *downward union* of classes considered as input. The choice of elementary conditions e that would become part of conjunction E is based on the evaluation of $E \cup \{e\}$ by a function denoted $Evaluation()$. Several versions of this function may be used.

The version of the algorithm used here chooses the elementary rule providing the largest ratio $\frac{|[E \cup \{e\}] \cap G|}{|[E] \cup \{e\}|}$, in a strategy that consists in covering the maximum number of cases with the minimum number of elementary conditions. An alternative strategy would, for instance, aim at choosing the elementary rule e that minimizes the number of currently uncovered cases verifying it. To ensure minimality rules are checked iteratively redundant elementary conditions and rules are removed from the final rule base.

As previously stated, for the problem at hand, we consider the "Normal" class to be a default class and focus learning on the *downward union* associated with cases from the "Altered" class (that is the class itself). Table 2 presents the resulting rule base of nine rules, to which we add a tenth rule assigning to the "Normal" class if none of the previous rules is satisfied.

5. **Results and Discussion.** Table 3 and Figure 3 compare the performance metrics of this classification system (DRSA) to those obtained in [7]. As can be observed, the dataset of 100 cases can be described almost exhaustively (98% accuracy) by the set of ten rules presented in Table 2. Further, the 2% inaccurate classification results from an

TABLE 2. Classification rules induced by the dominance-based rough sets approach

Rule	Logical expression	Support
Rule 1	If (Sitting = 0.06) & (Season = −0.33) Then (Output = O)	9.09%
Rule 2	If (Sitting = 0.25) & (Age ≤ 0.69) & (Surgery = 1) & (Disease = 1) Then (Output = O)	9.09%
Rule 3	If (Sitting = 0.31) & (Surgery = 1) & (Trauma = 0) Then (Output = O)	9.09%
Rule 4	If (Sitting = 0.38) & (Season = 1) & (Alcohol ≤ 0.80) Then (Output = O)	27.27%
Rule 5	If (Sitting = 0.44) & (Season = 0.33) Then (Output = O)	9.09%
Rule 6	If (Sitting = 0.44) & (Season = 1) & (Fever = −1) Then (Output = O)	9.09%
Rule 7	If (Sitting = 0.50) & (Disease = 0) Then (Output = O)	18.18%
Rule 8	If (Sitting = 0.50) & (Season = 0.33) & (Smoking = −1) & (Surgery = 0) Then (Output = O)	9.09%
Rule 9	If (Sitting = 0.88) & (Fever = −1) Then (Output = O or N)	100.00%
Rule 10	Else (Output = N)	100.00%

TABLE 3. Confusion matrix and performance indicators

	MLP [7]	**SVM** [7]	**DT** [7]	**DRSA**
TP	80	83	82	88
TN	6	3	2	12
FP	9	12	13	1
FN	5	2	3	1
Accuracy (%)	86.00	86.00	84.00	98.03
Sensitivity (%)	94.11	97.64	96.47	98.87
Specificity (%)	40.00	20.00	13.33	92.30
Positive Predictive Value (%)	89.88	87.36	86.31	98.87
Negative Predictive Value (%)	54.54	60.00	40.00	92.30

inconsistency in the original dataset, where cases number 67 and 71 present the exact same attribute values but are part of two different classes. The metrics achieved by DRSA in Table 3 and Figure 3 are thus the highest possible for this dataset and rather than indicating the performance of this approach, they clearly indicate the triviality of the original clinical dataset of [7], which was somehow hidden by the sub-optimal results obtained by methods MLP, SVM, DT in the original publication.

Despite the triviality of the reference dataset [8] and its small size of 100 instances, conclusions have been drawn by past studies concerning not only the relative technical merits of different machine learning methods [7], but also on medical aspects such as the importance of lifestyle factors on male fertility [9]. Thus, we insist on the importance of the representativeness of data in any machine learning endeavor and call for the development of objective statistical standards concerning the quality of datasets from which technical and medical conclusions can be drawn. As our results highlight, algorithmic accuracy indicators not only do not reflect the quality of datasets but can more worryingly hide the poor quality of some datasets with high but sub-optimal values.

6. **Conclusion.** In this research, the dominance-based rough sets approach was utilized on a widely studied reference dataset from the UCI machine learning repository. Due to the monotonic nature of the considered features, the proposed algorithm unsurprisingly outperformed previous machine learning approaches and highlighted serious issues with

(a) Accuracy

(b) Sensitivity

(c) Specificity

(d) Negative predictive value

(e) Positive predictive value

FIGURE 3. Performance indicators

the quality and representativeness of the reference dataset. Although, there exist elaborate statistical indicators for the performance of machine learning methods on particular datasets, our results suggest a gap in the literature concerning the nonexistence of objective standards for the measurement of data quality ex-ante, that is before a particular machine learning approach is considered. For instance, the explicit definition of conditions to be satisfied by the data (see for instance, conditions 1 to 4 of [20] in the context of natural language processing) seems to be a good practice that would warrant extension to medical diagnosis datasets.

REFERENCES

[1] P. Mishra and R. Singh, Seminal decline in semen quality in humans over the last 80 years, *Male Infertility: Understanding, Causes and Treatment*, pp.89-108, 2017.

[2] V. Bartak, Sperm count, morphology and motility after unilateral mumps orchitis, *Journal of Reproduction and Fertility*, vol.32, pp.491-494, 1973.

[3] J. Yaniz, S. Vicente-Fiel, S. Capistros, I. Palacn and P. Santolaria, Automatic evaluation of ram sperm morphometry, *Theriogenology*, vol.77, pp.1343-1350, 2012.

[4] V. Chang, A. Garcia, J. Hitschfeld and S. Hartel, Gold-standard for computer-assisted morphological sperm analysis, *Computers in Biology and Medicine*, vol.83, pp.143-150, 2017.

[5] S. Javadi and S. A. Mirroshandel, A novel deep learning method for automatic assessment of human sperm images, *Computers in Biology and Medicine*, vol.109, pp.182-194, 2019.
[6] J. Riordon, C. McCallum and D. Sinton, Deep learning for the classification of human sperm, *Computers in Biology and Medicine*, vol.111, 2019.
[7] D. Gil, J. Girela, M. Joaquin De Juan, J. Gomez-Torres and M. Johnson, Predicting seminal quality with artificial intelligence methods, *Expert Systems with Applications*, vol.39, no.16, pp.12564-12573, 2012.
[8] M. Lichman, *UCI Machine Learning Repository*, School of Information and Computer Science, University of California, Irvine, CA, 2013.
[9] H. Wang, Q. Xu and L. Zhou, Seminal quality prediction using clustering-based decision forests, *Algorithms*, vol.7, pp.405-417, 2014.
[10] A. Bidgoli, E. Komleh and S. Mousavirad, Seminal quality prediction using optimized artificial neural network with genetic algorithm, *Proc. of the 9th International Conference on Electrical and Electronics Engineering (ELECO)*, Bursa, pp.695-699, 2015.
[11] M. Sudha, Evolutionary and neural computing based decision support system for disease diagnosis from clinical data sets in medical practice, *Journal of Medical Systems*, vol.41, no.11, 2017.
[12] World Health Organization, *WHO Laboratory Manual for the Examination and Processing of Human Semen*, 5th Edition, 2010.
[13] J. C. Bioch and V. Popova, Rough sets and ordinal classification, *International Conference on Algorithmic Learning Theory ALT 2000: Algorithmic Learning Theory*, pp.291-305, 2000.
[14] J. R. Cano, P. Gutirrez, B. Krawczyk, M. Wozniak and S. Garca, Monotonic classification: An overview on algorithms, performance measures and data sets, *Neurocomputing*, vol.341, pp.162-182, 2019.
[15] A. Ben-David, L. Sterling and Y. Pao, Learning and classification of monotonic ordinal concepts, *Computational Intelligence*, vol.5, no.1, pp.45-49, 1999.
[16] J. R. Cano, J. Luengo and S. Garca, Label noise filtering techniques to improve monotonic classification, *Neurocomputing*, vol.353, pp.83-95, 2019.
[17] Z. Pawlak, *Rough Sets: Theoretical Aspects of Reasoning about Data*, Kluwer, Dordecht, 1991.
[18] S. Greco, B. Matarazzo and R. Slowinski, The use of rough sets and fuzzy sets in MCDM, in *Advances in Multiple Criteria Decision Making*, T. Gal, T. Hanne and T. Stewart (eds.), Boston, Kluwer Academic Publishers, 1999.
[19] T. L. Andersen and T. R. Martinez, NP-completness of minimum rule sets, *Proc. of the 10th International Symposium on Computer and Information Sciences*, pp.411-418, 1995.
[20] M. Murata, K. Orikane and R. Akae, Automatic selection and analysis of verb and adjective synonyms from Japanese sentences using machine learning, *International Journal of Innovative Computing, Information and Control*, vol.15, no.6, pp.2135-2147, 2019.