

Ensemble Reasoning for Dust storm predictions

Tariq Saad Al Murayziq, Stelios Kapetanakis, Miltos Petridis

School of Computing, Engineering and Mathematics, University of Brighton, Moulsecoomb
Campus, Lewes road, Brighton BN2 4GJ, UK (t.s.a.murayziq, s.kapetanakis,
m.petridis}@brighton.ac.uk

Abstract. In meteorology, a dust storm is caused when sand and other similar particles are blown by the wind, typically across arid areas of the earth. Thus, soil and sand are transferred across varying distances by means of suspension and saltation. In an attempt to improve efforts to forecast dust storms, this paper proposes combining a Bayesian network with case-based reasoning, using historical experience to test the effectiveness of this method. As such, a hybrid method is devised that combines Bayesian classifiers with case-based reasoning to draw comparison between historical dust storms and the current experience. The results confirm the effectiveness of case-based reasoning for predicting dust storms. Furthermore, the possibility of using this technique to mitigate the effects of severe weather conditions are contemplated.

Keywords: case-based reasoning, Bayesian network, dust storm prediction, weather forecast prediction, artificial intelligence.

1 Introduction

Empirical research has indicated that the primary cause of dust storms is poor husbandry of arid land. For example, if a fallow system is not adhered to then dust storms are significantly more likely to occur and this can have consequences not only for agricultural output but also for the regional climate [4, 18]. Additional concerns relating to dust storms include medical conditions and the impact on members of the population if they are forced to remain indoors for prolonged periods of time. An inability to accurately forecast dust storms means that the population risks being ill-prepared and can give rise to numerous unwelcomed developments. For instance, a dust storm in Saudi Arabia during 2015 caused flights to be cancelled, schools to close and was believed to have caused several accidents across the country [14].

Dust storms have been known to spread disease across vast distances owing to the ability to transport virus spores that are swept up from the surface of earth and subsequently deposited elsewhere [20]. Dust storms can interrupt transport networks and prevent people from conducting their day-to-day activities. Therefore, the ability to reliably forecast dust storms would help communities to prepare by avoiding unnecessary travel, expensive precaution measures, securing their property, remaining indoors if they suffer from known medical conditions and taking steps to protect assets such as machinery and livestock that may be adversely affected by the storm. This is especially useful for agricultural areas if farmers need to harvest their crops early, move cattle indoors or protect expensive equipment from the elements.

This paper proposes a novel method for forecasting dust storms that combines Case-based Reasoning with Bayesian networks. The performance of this method is compared to that of lazy learning for classifying previous dust storms. In the new approach, a Bayesian network is initially used to classify historical dust storms and then case-based reasoning forecasts future dust storms by means of weighted Euclidean distance.

1.1 Motivation

Dust storms afflict numerous areas of the world, causing health problems, property damage and inconvenience. Saudi Arabia is especially prone to the adverse effects of dust storms and extreme weather conditions. Saudi Arabia is therefore selected as a testbed in this paper to confirm the effectiveness of combining the artificial intelligence techniques of case-based reasoning and a Bayesian network to classify and forecast dust storms in a timely manner so that populations can take evasive action to protect their health and their property.

1.2 Rationale for using a Bayesian network with case-based reasoning

Case-based reasoning is a natural choice for forecasting dust storms due to the features of the application area. Case-based reasoning assumes that different problems that share similar features will have similar solutions [7]. Consequently, case-based reasoning is well-suited for dust storms forecasting. Crucially, case-based reasoning not only forecasts events but also provides an indication of the probable success of a given recommendation. Again, this insight into the probable success of a venture is based on historical experience [17]. For instance, when a solution is proposed, case-based reasoning is able to offer an indication of how successful such approach could be [13]. In practical terms, this is achieved by gauging the scale of the current problem and setting this against historical experience gleaned from previous episodes while evaluating how effective the solution proved to be in the past.

One advantage associated with case-based reasoning is the flexibility that it offers residing to the ability to update the system with additional historical cases. These can then be used to enhance the decision-making process [19]. Moreover, a further advantage of case-based reasoning is its ability to accommodate incomplete data values and provide useful insight when faced with a large number of features. This is because case-based reasoning utilises informed guesses based on neighbouring values to fill in the gaps. Meanwhile, the key advantage of case-based reasoning is that it is able to reason when faced with uncertainty even when the model is not particularly well understood [16].

The modelling framework applied in Bayesian networks is well-suited for making decisions when faced with uncertainty. Each Bayesian network comprises a series of nodes. These nodes signify a random variable in the domain with links directed between pairs of variables. The combination of nodes and arcs give a certain acyclic graph structure. Details of the conditional independence statements in the domain are reflected in the links or the lack of links. These links can be regarded as giving details about the causal mechanism [11].

Bayesian networks are increasingly being employed in order to explain causal inferences [8]. These Bayesian networks are well-suited to drawing out causal inferences and providing insight into diagnostic inference.

Within the confines of a monitored system, Bayesian reasoning is able to anticipate how likely certain events are to be realised. Bayesian networks not only utilise Bayesian reasoning but also illustrate the relationships between variables using graphical depictions. There are numerous examples of Bayesian networks in the empirical literature being used for purposes such as medical diagnosis and more generic modelling. For instance, Intellipath [2, 10] is a system deployed in the healthcare sector that exploits Bayesian networks for diagnosing conditions in pathology departments.

2 Literature Review

Case-based reasoning is able to develop a system for forecasting by logging successful efforts from previous experiences. For instance, case-based reasoning can be used to anticipate successful cyber attacks. Previous researchers have observed positive results when employing case-based reasoning [5]. By exploiting historical experience of intrusions [5, 10], case-based reasoning can be deployed for detection purposes. Similarly, the empirical literature confirms that the effectiveness of case-based reasoning can be enhanced by combining it with other methodologies [5, 10]. These additional methods could enhance case-based reasoning in various ways such as

speedier problem solving, offering support systems or by applying case-based reasoning to support other systems. For instance, case-based design systems have been employed for reasoning and support maintenance systems [15].

A four-stage process for case-based reasoning was advocated by Aamodt and Plaza [1]. Step one is retrieval and involves identifying similar cases based on a weighted sum of properties. Cases that fit within the parameters identified are selected [12]. Step two is re-use and involves using the derivational and transformational methods to alter the retrieved solutions. The derivational method applies algorithms to form the initial solution, thereby making it possible to devise a new solution that better suits the current case. Meanwhile, the transformational method employs transformational operators to adapt previous solutions so that they are better able to address the current problem.

The adapted solution is appraised in the revision step. If this proves to be unsuitable, it is repaired by exploiting domain-specific knowledge. In the final step, the solution is retained by storing details of the case in the system's memory so that this can be utilised when similar cases arise in future [9, 12].

There is a paucity of empirical studies that have combined a Bayesian network with case-based reasoning. However, the following two studies are relevant to this area of research.

A distributed case-based reasoning system was employed by Tran and Schönwälder [19] in an attempt to resolve issues with a communication system fault domain. In this instance, there are a number of symptoms (S) and a fault hypothesis (H) within the problem solution. The authors arrived at a two-stage reasoning process which involved ranking and then selection. Ranking involves finding the cases that are most similar to that being considered employing Bayesian network relations. Meanwhile, selection involves using the Bayesian network relations derived from the cases to form a Bayesian network. Subsequently, a suitable hypothesis is selected based on the information available at that stage.

Meanwhile, a software-based engineering application that draws upon case-based reasoning, Bayesian networks and WordNet was devised by Gomes [3] so as to assist software engineers wishing to re-use old designs. The problem description component of the cases include several WordNet synonym sets. The cases and synonym sets are nodes in the Bayesian network.

The parents from WordNet's hypernym relation are identified using the synonym sets from the problem description. These parents are then incorporated into the Bayesian network. The number of parents for the node determines the formulas that are used to create the conditionality probability tables. There are three stages to the RETRIEVE phase: synonym sets in the Bayesian Net are activated by the query case description; calculations of the Bayesian network nodes are taken and the most suitable cases are identified; the cases are ranked according to their probabilities.

Greater knowledge of the underlying causal structure of a domain can be gleaned by studying the structure of the Bayesian network model for a given domain. What is more, this approach can be employed to forecast quantities that would be difficult to measure on the grounds of expense or ethics, for example [11].

3 Research Methodology

This paper proposes a combination of Case-based reasoning and Bayesian networks in order to improve forecast accuracy and enable members of the public to take steps to mitigate the adverse effects of these dust storms. The dust storms will be categorised and classified by means of the Bayesian network. In the event that patterns are identified in the data, these will be added to the case-based reasoning database. Future dust storms can be forecast by using case-based reasoning to match these storms with historical knowledge of such storms. In order to improve on the current solutions, spatial data are required. Case-based reasoning is particularly well-suited for such demand since it is able to explain case domain knowledge to experts in this field.

The chosen methodology will enable meteorologists to recognise significant trends and appreciate the potential for the available solutions to detect dust storms in future. Any form of reliable early warning system would prove invaluable when helping to mitigate the effects of dust storms. For instance, prior knowledge of a storm would help to avoid financial loss in the farming community by taking appropriate evasive

action. Analysing the available data will help the public authorities to devise systematic solutions in a timely manner. Forecasts of dust storms are typically based on a combination of variables including temperature, humidity, soil moisture, air pressure and sun radiation [18].

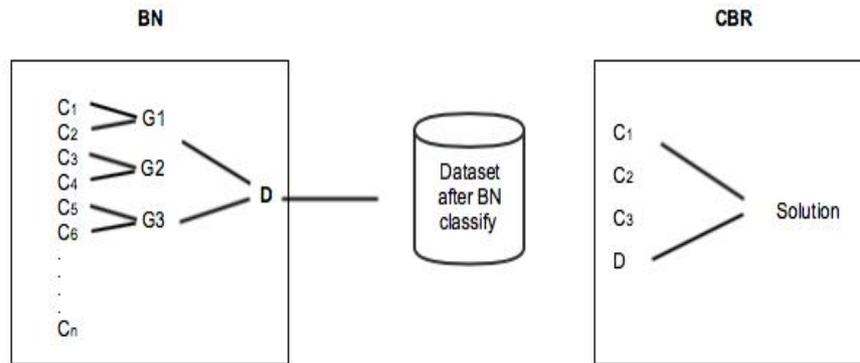


Fig. 1. Research Methodology

4 Results and Discussion

4.1 Dust Storm Attributes

It is only by identifying the main features of a dust storm that it is possible to establish the inputs. Based on the available evidence, the following variables were identified as being the most significant features of dust storms:

- Month* = month of the recorded dust case
- Rainfall* = how many mm its rain in that day
- W.S.M* = wind speed max
- P.M* = pressure max
- P.MIN* = pressure min
- P.S.M* = pressure sea level max
- P.S.MIN* = pressure sea level min
- H.M* = humidity max
- H.MIN* = humidity min
- A.T.M* = air temperature max
- A.T.MIN* = air temperature min
- S.A* = soil ability

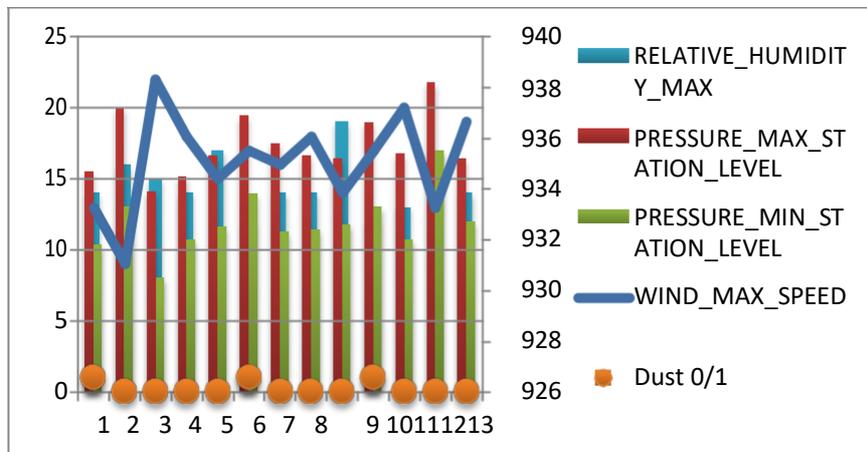


Fig. 2. Dust Storm Attributes

Figure 2 illustrates that a dust storm is likely to occur when a combination of factors coincides: high wind speed, low pressure and low humidity. Furthermore, a high soil temperature also increases the likelihood of a dust storm occurring.

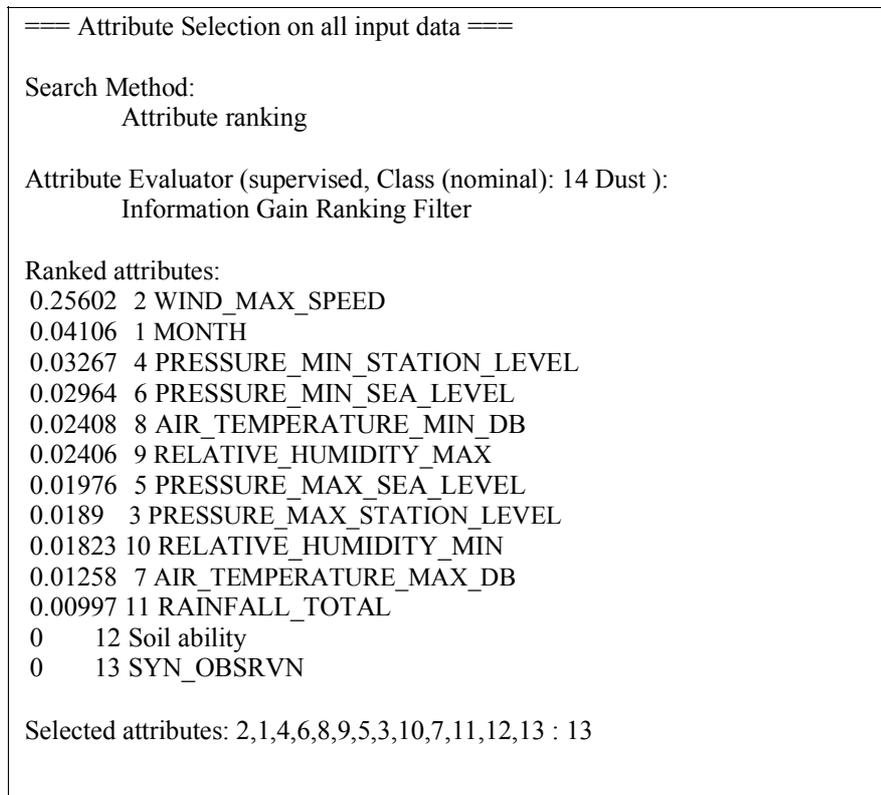


Fig. 3. Dust Storm Attribute Ranking

The box above shows how the different features of dust storms were ranked using the attribute ranker in the Weka tool. An information gain algorithm was applied which calculates information gain relative to the stated class. The results are in accordance with the ranking of the meteorological professionals.

$$InfoGain(Class, Attribute) = H(Class) - H(Class | Attribute).$$

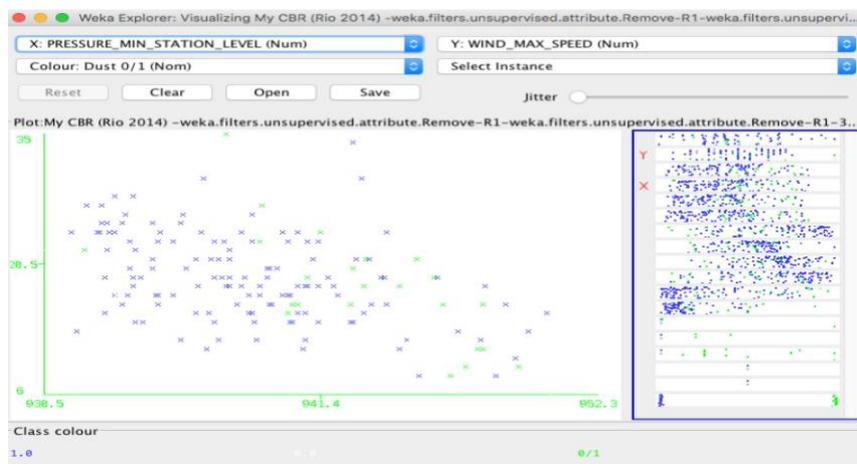


Fig. 4. The correlation between pressure and wind

Figure 4 illustrates how air pressure and wind speed are correlated. Dust storms are significantly more likely to arise when there is low air pressure and the wind speed is high.

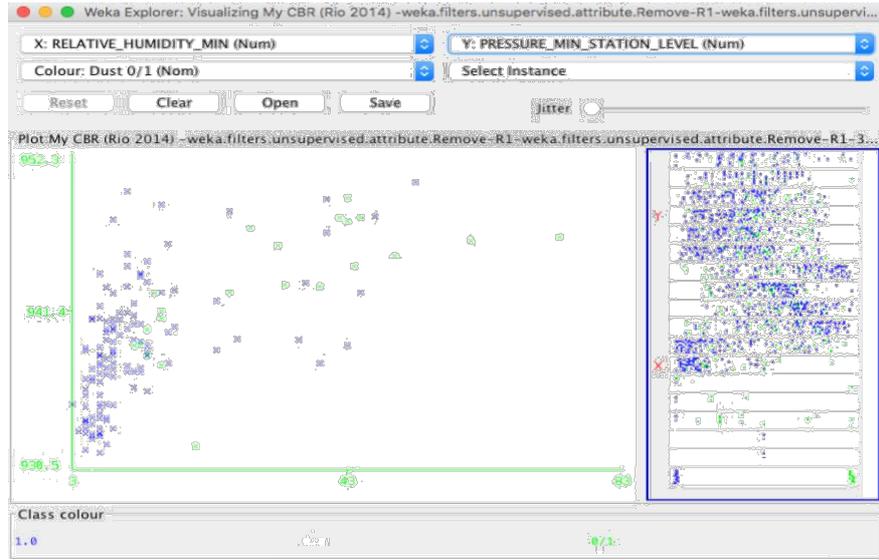


Fig.5. The correlation between Pressure and Humidity

Figure 5 illustrates how humidity and air pressure are correlated. Dust storms are significantly more likely to arise when there is low humidity and low air pressure.

4.2 Bayesian network

Bayesian probability is a logical method to select for reasoning purposes when it is known that there are uncertain statements. Importantly, Bayesian methods are applied when it is necessary to gain an understanding about a state of knowledge. A scale ranging from zero to one is used to express the degree of confidence associated with a given suggestion [6].

$$P(E|F) = \frac{P(E)P(F)}{P(F)}$$

Where two event E and F such that $P(E) \neq 0$ and $P(F) \neq 0$

In the current study, a Bayesian network is employed to classify historical dust storms and group these cases into various categories to depict their severity:

- **No** = No dust storm recorded.
- **Dusty** = Dust lifted from the ground in the immediate vicinity of the weather station but no evidence of a dust whirl and it did not develop into a full-blown dust storm.
- **Mid-Dust** = A moderate dust storm that has intensified within the previous sixty minutes.
- **Heavy dust storm** = A full dust storm that has started or become more intense within the previous sixty minutes.

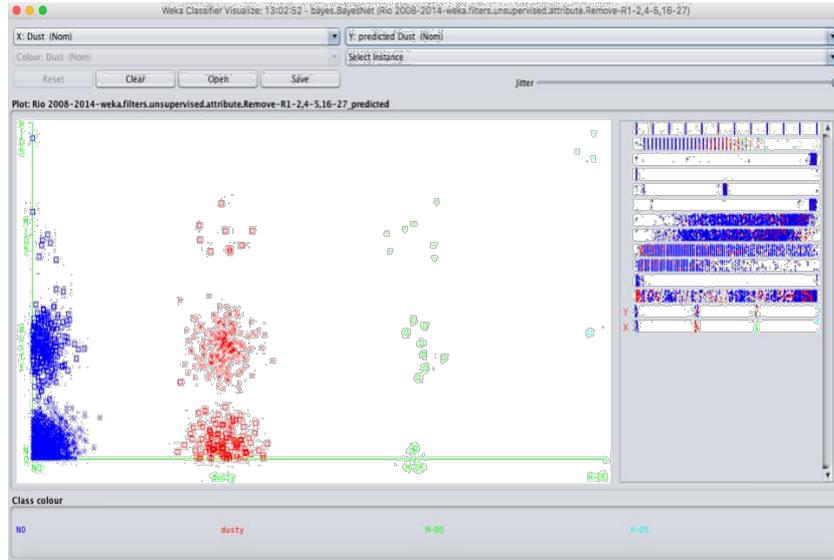


Fig.6. Using the Bayesian network to classify dust events

The Bayesian network has performed well; the majority of cases have been classified and the categories of uncertain events have been accurately predicted.

```

==== Stratified cross-validation ====
==== Summary ====

Correctly Classified Instances    2081    81.3208 %
Incorrectly Classified Instances  478     18.6792 %
Kappa statistic                   0.3849
Mean absolute error               0.1116
Root mean squared error          0.2708
Relative absolute error           88.7257 %
Root relative squared error       108.1556 %
Total Number of Instances        2559

==== Detailed Accuracy By Class ====

TP Rate FP Rate Precision Recall F-Measure MCC ROC Area PRC
Area Class
0.852 0.366 0.932 0.852 0.890 0.415 0.817 0.959 NO
0.595 0.148 0.391 0.595 0.472 0.379 0.797 0.424 dusty
0.353 0.006 0.300 0.353 0.324 0.321 0.933 0.259 M-DS
0.750 0.001 0.600 0.750 0.667 0.670 0.888 0.751 H-DS
W. Av 0.813 0.333 0.853 0.813 0.829 0.409 0.815 0.881

==== Confusion Matrix ====

a b c d <-- classified as
1863 317 7 0 | a = NO
134 209 7 1 | b = dusty
2 8 6 1 | c = M-DS
0 1 0 3 | d = H-DS

```

Fig.7. Bayesian network classification result

4.2.1 Bayesian network (BN) vs lazy learning

The dust storms were classified and categorised using lazy learning but it is apparent that these results were less accurate than those obtained using the Bayesian classifier. With lazy learning, computation does not take place until after classification. In the training time, there is no attempt to devise theories such as decision trees.

Moreover, predictions are arrived at by lazy learning algorithms using training samples [21].

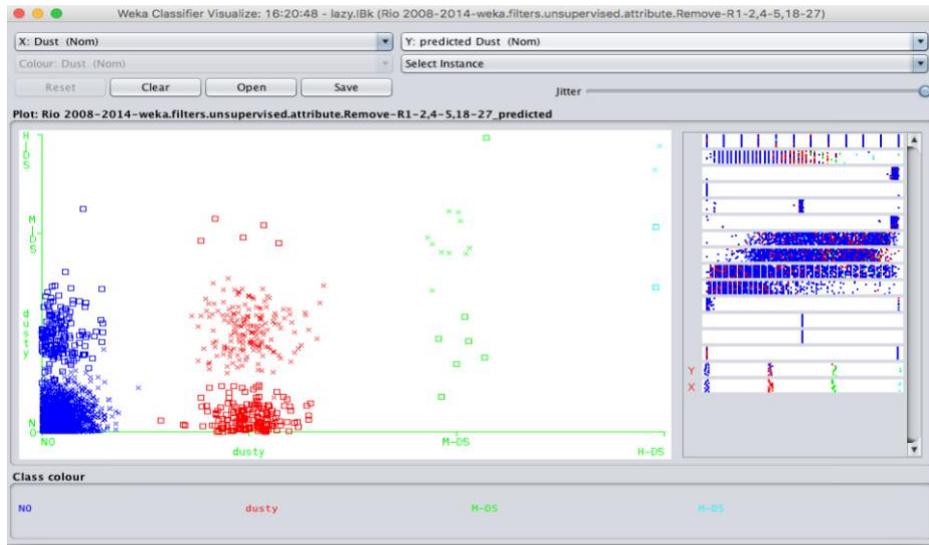


Fig.8. Applied lazy learning using Weka

```

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances    2207    86.2446 %
Incorrectly Classified Instances   352    13.7554 %
Kappa statistic                   0.452
Mean absolute error               0.0693
Root mean squared error           0.262
Relative absolute error           55.0904 %
Root relative squared error       104.6539 %
Total Number of Instances        2559

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  MCC      ROC Area  PRC
      0.922   0.454   0.923    0.922   0.922    0.467  0.734  0.918  NO
      0.507   0.079   0.506    0.507   0.506    0.428  0.714  0.325  dusty
      0.588   0.003   0.556    0.588   0.571    0.569  0.784  0.330  M-DS
      0.500   0.000   0.667    0.500   0.571    0.577  0.775  0.418  H-DS
Weighted Avg.   0.862  0.399  0.863    0.862   0.862    0.863  0.463  0.731  0.832

=== Confusion Matrix ===

 a  b  c  d <-- classified as
2017 169 1 0 | a = NO
 168 178 5  0 | b = dusty
   1  5 10  1 | c = M-DS
   0  0  2  2 | d = H-DS

```

Fig.9. Lazy learning classification result.

Figures 8 and 9 reveal that lazy learning is able to categorise ‘no dust’ and ‘dusty’ events relatively successfully but it underperforms when categorising ‘mid-dust’ and ‘heavy dust storms.’ It is these more intense events that this paper is primarily concerned with and, therefore, lazy learning does not appear to be suitable for these purposes.

4.3 Bayesian network with case-based reasoning

Old and new cases in the database are matched using case-based reasoning. By doing so, it will be possible to forecast dust storms. The selected methodology uses historical examples to set a benchmark so that dust storms can be more clearly understood.

Cases are classified using a Bayesian network and the following processes are adhered to:

1. A Euclidean distance equation is employed in order to recognise similarities between the cases in the database and the specific case being studied at present. The Euclidean distance equation is applied separately for each of the variables. It is important to initialise the different variable weights to 1.

$$\sqrt{\sum_{i=0}^n (x_i - y_i)^2} \quad (1)$$

Where $x=(x_1, x_2, \dots, x_n)$ and $y = (y_1, y_2, \dots, y_n)$ are two dust storm case vectors with i attributes and $n+1$ is the maximum length of the attributes. The next step is to aggregate the equation results for each of the cases, thereby giving the similarity. These values are then sorted according to their score of similarity in descending order. The equation above is used to calculate the degree of similarity between a new case and the database of old cases. It also gives a collection of cases that are similar to the specific case being studied at present.

2. Data are imported into the myCBR tool. By doing so it is possible to establish the similarity between cases by applying weighted Euclidean distance equation measurements (see Equation 2 below). It will be necessary to amend the weight values for some of the variables to reflect the main characteristics of the dust storm.

$$D(x, y) = \sqrt{\sum_{i=0}^n w_i (x_i - y_i)^2} \quad (2)$$

Case-based reasoning was employed in order to classify a randomly chosen subset from the database. These samples were typically 10%-12% of the database and the classification of these samples remained unknown. Instead, they were classified using case-based reasoning by means of similarity measures and 3NN classification. A total of ten experiments were carried out for each case and averages were taken of these results. Based on these results, the accuracy of case-based reasoning is found to be within the range of 60%-80%. However, the results are significantly more accurate when case-based reasoning is combined with a Bayesian network classifier. Indeed, combining these two yields accuracy results of between 80% and 90% when forecasting 'no dust', 'dusty' and 'mid-dust' storms. It is believed that this method would also perform well when forecasting 'heavy dust storms' but more samples would be required to test this reliably.

ID	Result of the case	Result of 1 st 3-KNN	Result of 2 nd 3-KNN	Result of 3 rd 3-KNN	Result of 4 th 3-KNN	Result of 5 th 3-KNN	Result of 6 th 3-KNN	Result of 7 th 3-KNN	Result of 8 th 3-KNN	Result of 9 th 3-KNN	Result of 10 th 3-KNN	Final result
5-2546	No	No	No	No	No	No	No	No	No	No	No	No
3-1910	No	No	No	No	No	No	No	No	No	No	No	No
3-1787	M-D	M-D	M-D	M-D	dusty	M-D	M-D	dusty	dusty	M-D	M-D	M-D
7-2555	dusty	dusty	dusty	dusty	dusty	No	dusty	dusty	No	dusty	dusty	dusty
4-1023	M-D	M-D	M-D	M-D	M-D	M-D	M-D	M-D	M-D	M-D	M-D	M-D
5-255	dusty	dusty	dusty	dusty	dusty	dusty	dusty	dusty	dusty	dusty	dusty	dusty
8-2557	M-D	H-D	H-D	H-D	dusty	dusty	dusty	No	dusty	No	No	Not Good results
4-1299	dusty	dusty	dusty	dusty	dusty	dusty	dusty	dusty	No	dusty	No	dusty
2-510	M-D	M-D	M-D	M-D	M-D	M-D	dusty	No	M-D	M-D	dusty	M-D
2-453	No	No	No	No	No	No	No	No	No	No	No	No

Fig.10. BN-CBR result.

5 Dataset Accessibility

This study has combined a Bayesian network with case-based reasoning in dust storms forecasting and it seems it can yield promising results. This method will be tested using real data from dust storms happened in Saudi Arabia spanning the period 2008-2014. This dataset will be provided by meteorologists in Saudi Arabia.

As previously stated, Saudi Arabia experiences large dust storms at regular intervals and these have adverse effects on industry, transport and the population in general. Furthermore, Saudi Arabia is particularly well-suited to studying forecasting ability in relation to dust storms because of the natural variation in its geography.

6 Conclusions

Dust storms have the potential to disrupt lives, cause inconvenience, spread disease, cause environmental damage and adversely affect economic output. This paper has demonstrated how case-based reasoning can be used to forecast the magnitude of dust storms before they occur. Case-based reasoning draws on historical experience to solve future problems that are expected to be similar in nature. Therefore, in such cases, the strategy proposed to mitigate future dust storms is likely to be similar to those that have been deployed in the past. It has been demonstrated that combining case-based reasoning with a Bayesian network can yield effective results; case-based reasoning is known to offer a viable method for solving current problems based on historical experience, while Bayesian networks have been shown to offer a suitable method for classifying historical episodes of dust storms. Future investigation to confirm the effectiveness of this combined case-based and Bayesian approach for forecasting heavy dust storms will require an improved dataset. Heavy dust storms could be simulated and additional algorithms could be applied in order to confirm this. In addition, a rule-based system should be developed in a bid to enhance safety and reduce the economic harm that is caused by dust storms.

7 Acknowledgements

We would like to thank Dr Abdullah Almisnid from the School of Geography, Qassim University, KSA for his help and support on the dust storm domain and the dataset provision

8 References

1. Aamodt, A., Plaza, E. (1994). Case-Based Reasoning: Foundational Issues, Methodological Variations, and System Approaches. *AI Communications*. IOS Press, Vol. 7: 1, pp. 39-59.
2. Biocare Medical. (2009). *intelliPATH Automated Staining Instrument*. [Online]. Available at: <http://www.slidestainer.com/>.
3. Gomes, P. (2004). Software design retrieval using Bayesian Networks and WordNet. *Lecture Notes in Computer Science* 184–197
4. Griffin, D. W., & Kellogg, C. A. (2004). Dust storms and their impact on ocean and human health: dust in Earth's atmosphere. *EcoHealth*, 1(3), 284-295.
5. Kapetanakis, S., Filippopolitis, A, Loukas, G., Al Murayziq, T. S. (2014). Profiling cyber attackers using Case-based Reasoning. In proceeding of the 19th UK Workshop on Case- Based Reasoning (UKCBR 2014), 9th December 2014, Cambridge, UK, pp.39-48.
6. Kapetanakis, S. (2012). Intelligent monitoring of business processes using case-based reasoning. PhD thesis, University of Greenwich
7. Kolodner, J. L. (1992). An introduction to case-based reasoning. *Artificial Intelligence Review*, 6(1), 3-34.
8. Lacave, C., D'íez, F. (2003). A review of explanation methods for Bayesian networks. *The Knowledge Engineering Review* 17(02) 107–127
9. Lopez-Fernandez, H., Fdez-Riverola, F., Reboiro-Jato, M., Glez-Pena, D., & Mendez, J. R. (2011). Using CBR as Design Methodology for Developing Adaptable Decision Support Systems. *Efficient Decision Support Systems Practice and Challenges from Current to future*, Chiang J., (Ed.), USA: Intech Publishers, 123-126.
10. Main, J., Dilon, T. and Shiu, S. (2001). A Tutorial on Case-Based Reasoning, *Soft Computing in Case- Based Reasoning*, pp. 1-28.
11. Margaritis, D. (2003). *Learning Bayesian network model structure from data* (Doctoral dissertation, US Army).
12. Mascarenhas, S. (2010). *Case Based Reasoning & Fame IA*, Retrieved on 22 July 2014 from: <https://fenix.tecnico.ulisboa.pt/downloadFile/3779574683590/ASS-Case- Based%20Reasoning.pdf>
13. Pearl, J. (2000) *Causality: Models, Reasoning, and Inference*. Cambridge University Press
14. Polden, J. (2015). Engulfed by the sand storm: Man captures moment huge cloud hit streets in Saudi Arabia. Retrieved April 8, 2015 from <http://www.dailymail.co.uk/news/article-3029078/Sandstorm-wreaked-havoc- Saudi- Arabia-captured-camera-engulfing-road.html>
15. Roth-Berghofer, T., Goker, M.H., Guvernir, H.A. (2006). *Advances in Case-Based Reasoning: 8th European Conference, ECCBR 2006, Fethiye, Turkey, September 4-7,2006 Proceedings*. Berlin, Germany: Springer-Verlag Publishers.
16. Schiaffino, S. N., & Amandi, A. (2000). User profiling with Case-Based Reasoning and Bayesian Networks. In *IBERAMIA-SBIA 2000 open discussion track* (pp. 12-21).
17. Shiu, S. C. K. and Pal, S. K. (2004). Case-Based Reasoning: Concepts, Features and Soft Computing, *Applied Intelligence*, 21, pp . 233-238.
18. Sivakumar, M.V.K. (2005). Impacts of Sand/Dust Storms on Agriculture. M.V.K. Sivakumar, R.P. Motha, and H.P. Das (Eds.). *Natural Disasters and Extreme Events in Agriculture*. Springer-Verlag, Berlin, Germany. 367p
19. Tran, H., Schönwälder, J. (2008). Fault Resolution in Case-Based Reasoning. In: *Proceedings of the 10th Pacific Rim International Conference on Artificial Intelligence: Trends in Artificial Intelligence*, Springer 429
20. Vidal, J. (2009). Dust Storms Spread Deadly Diseases Worldwide. London:Guardian.co.uk. Retrieved April 23 2015.
21. Webb, G.I. (1996). A heuristic covering algorithm outperforms learning all rules. *Proceedings of the Conference, ISIS'96: Information, Statistics and Induction in Science* (pp. 20-30). Singapore: World Scientific.