Partial imputation to improve predictive modelling in insurance risk classification using a hybrid positive selection algorithm and correlation-based feature selection

Mlungisi Duma[1,*], Bhekisipho Twala[1], Fulufhelo V. Nelwamondo[2] and Tshilidzi Marwala[1]

[1]Department of Electrical Engineering and the Built Environment, University of Johannesburg, Auckland Park, Johannesburg, South Africa
[2]Modelling and Digital Science, Council for Scientific and Industrial Research, P O Box 395, Pretoria 0001, South Africa

We propose a hybrid missing data imputation technique using positive selection and correlation-based feature selection for insurance data. The hybrid is used to help supervised learning methods improve their classification accuracy and resilience in the presence of increasing missing data. The positive selection algorithm searches for potential candidates for imputation and the correlation-based feature selection method searches for attributes have a significant effect on the target outcome. The imputation is performed only on those attributes that have an impact on the target outcome. The results show that the classification accuracy and resilience of supervised learning methods improve significantly when applied with the imputation strategy under these assumptions.

Keywords: Insurance risk classification, missing data, positive selection, supervised learning.

DATASETS used in insurance underwriting or risk classification have a large number of variables and are susceptible to increasing missing data. Francis[1] outlines a number of reasons for this phenomenon, which include failure to disclose information and error or faulty handling of data by processing systems.

Here we propose a missing data partial imputation strategy (MDPIS) using a hybrid positive selection algorithm and a correlation-based feature selection (CFS) method. The positive selection algorithm performs a selection of potential candidates that can be used for imputation. The CFS method identifies the attributes that have great impact on the target outcome. The imputation is partial to permit reduced computational costs while improving the classification performance of supervised learning models in the presence of increasing missing data.

The supervised learning models chosen for the study are the ripper, naïve Bayes (NB), *k*-nearest neighbour (*k*-NN), logistic discriminant analysis (LgDA) and the support vector machines (SVMs). These models are chosen because they are now being adopted as predictive modelling methods in classification in credit and insurance risk analysis domains. The ripper has been employed in financial risk analysis to aid financial institutes select the appropriate policy for credit products, increase revenues and reduce losses[2]. NB has been applied in risk analysis of life insurance for clients, fraud claim analysis and determining if a client is a good or bad creditor[3–5]. *k*-NN and LgDA have been applied in credit risk analysis to segment loan applicants as good or bad creditors. SVMs have been employed to aid managers identify and manage credit risk as well as predicting solvency[6,7].

Missing data problems are not novel and there have been some significant attempts or strategies developed in the past few years to address the issue. These strategies either handle missing data by deletion[3], feature extraction[3,8], imputation or infer missing data using observable entries[4,9–11].

Zhang *et al.*[12] created a naïve Bayes and expected maximization model with an embedded strategy to handle or tolerate missing data and missing data imputation. The results showed improved performance and accuracy using the proposed strategies when compared with neural networks. The model works under the assumption that data are missing at random.

Lakshminarayan *et al.*[13] used a combination of AutoClass, an unsupervised algorithm designed to automatically discover clusters in data, and C4.5 for missing data imputation and prediction of a large-scale database. AutoClass is used as a hot-deck imputation for missing data imputation and C4.5 is used for learning and predicting values of the target variable. The proposed model is evaluated empirically with only the categorical variables with missing data. The results indicate accuracies of up to 80% imputation accuracies.

Gruenwald *et al.*[14] illustrated an algorithm that employs association rule mining and data clustering for missing data imputation for a multi-hop sensor network. (The issue with the sensor networks is that when missing data occur because of a malfunctioning sensor, the sensors either have to resend the entire message or ignore missing data. The former is an expensive solution and the latter is not viable.) The clustering algorithm is based on the distance between two sensors and is used for simultaneous missing data and phenomenon change in the surrounding environment. The experiments were conducted on synthetic and real-world datasets and the results show exceptional estimation accuracy, with error rates as low as 0.78% and 1.7% compared to other algorithms (such as SPIRIT and TinyDB). Furthermore, the algorithm illustrates the ability to preserve energy of the sensor networks.

Ramoni and Sebastian[15] presented a robust Bayesian estimator model, designed to learn conditional probability distributions from datasets with missing data. The model

*For correspondence. (e-mail: mlungisiduma@gmail.com)

does not make any assumptions about the missing data mechanism. The main strategy of the robust Bayesian estimator is to determine robust probability estimates in terms of different types of missing data. Robustness is achieved by supplying the probability intervals with estimates that can be adapted or learned from every observable datasets. The results of the experiment showed that when the model is trained by assigning missing entries, it performs better when training the model by ignoring the missing entries.

Nanni *et al.*[16] proposed an ensemble multiple imputation strategy based on a random subspace. A missing value is determined using a fuzzy clustering approach. Their experiments show that the proposed ensemble outperforms other state-of-the-art approaches to missing data imputation. They used the Wilcoxon single-rank test to illustrate that the ensemble is outperformed by the model trained using data with only >20% missing data. The ensemble achieves the best performance (or outperforms other classifiers) when there is a large number of missing data (about >30% missing data).

Polikar *et al*.[17] presented Learn^{++} MF, an ensemble of classifiers that uses random subspace selection strategy for handling missing values. The model does not perform missing data imputation to replace missing values. It builds or trains an ensemble of classifiers and each individual classifier is trained on a random subset of available variables. Records with missing data are classified using the majority voting of other classifiers whose training set does not contain missing values. The study shows that the ensemble can handle large amounts of missing data, with a decline in performance as the amount of missing data increases.

Wagner *et al*.[18] presented a study aimed at constructing a multimodal, ensemble of classifiers for emotion recognition with missing values in one or multiple modalities (e.g. voice, face or gesture). The results show that classification accuracies of single modalities range between 42% and 51% while recognizing and dealing with missing values in observed channels. The ensemble on the other hand, achieved classification accuracies of 55%, which includes certain generic fusion schemes and emotion adapted strategies like arousal, valence and cross-axis.

There are four kinds of missing data mechanisms found in the literature, namely missing at random (MAR), missing not at random (MNAR), missing completely at random (MCAR) and missing by natural design (MBND)[3]. MAR refers to a case where the missing data are not related to or independent of the attributes themselves, but rather are related to values of other attributes in the dataset. MNAR refers to the case where the missing data are directly related to the attributes themselves and not any other value from the other attributes. MCAR refers to the case where missing data are independent of the attributes themselves and any other attribute in the dataset. MBND is the case where missing data occur because they are naturally deemed unmeasurable, even though they are required for analysis. In this case, the missing values are modelled using mathematical techniques[3].

MCAR is the approach used here for the problem under discussion. It is chosen so that single and multiple imputations return unbiased outcomes.

Here four datasets are used to conduct the experiment. The first dataset is a Texas insurance used to draw up an insurance report. The report provides an overview of various claims involving bodily injuries that were either settled in court or disposed off. This dataset is used to determine if the plaintiff has legal cover or not. It consists of just over 1,800 instances. One thousand instances are used for training and 800 instances are used for testing. There are 227 features (mixed with numerical and categorical values) which were trimmed down manually to 182 by removing features that are clearly unimportant or redundant for the experiment, like unique identities, dates and categorical attributes with a single value. The target attribute consists of two classes only.

The second dataset was obtained from the Medical Expenditure Panel (MEP) survey conducted in 1996 by Harvey Rosen (Princeton University, USA). The dataset consists of over 8000 instances, all completely observable. In this study, a total of 1,000 instances were used to conduct the experiment. The dataset also consists of a total of 11 attributes, pre-processed to 9 categorical attributes. The target attribute consists of two classes.

The third dataset is from a South African Insurance (SAI) company. The dataset consists of over 30,000 instances and over 150 attributes. There were only 5,000 used for this experiment and the number of attributes was trimmed to 16 by removing those attributes that were easy to identify as irrelevant, as with the Texas dataset. The 16 attributes are made up of 10 categorical and 6 continuous attributes. The target attribute contains two classes.

The fourth dataset is the COIL dataset from the UCI machine learning repository. The dataset is used to predict which customers are likely to have an interest in buying a caravan insurance policy. Here, we are interested in finding out customers who are likely to have a car insurance policy. The training dataset consists of over 5,400 instances, of which 2,000 were used for the experiment. The testing dataset consists of only 4,000 instances and 1,000 were used in this study. Each set has a total of 86 attributes with completely observable data, 5 of which are categorical attributes and 80 are continuous attributes. The target attribute consists of only two classes.

The implementation for the proposed MDPIS is illustrated in Figure 1. The strategy is performed using a completely observable training dataset and a testing dataset containing increasing missing values. We have already shown that the supervised learning methods used in this study do not perform well under this assumption[19].

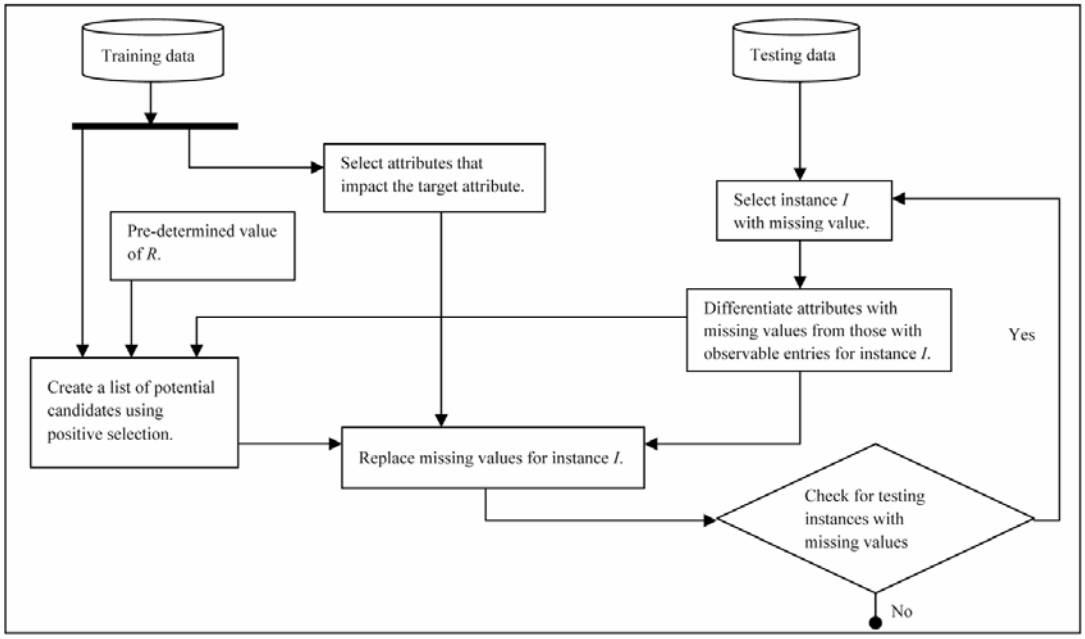

Figure 1. Missing data imputation process using positive selection.

The initial step of the process is choosing the attributes that have the greatest impact on the target attribute. To achieve this, the correlation-based feature selection (CFS) strategy is used[20]. CFS assesses the correlation of a group or a subset of attributes by calculating the predictive ability of each attribute per instance and the degree of dependency between the attributes. The selected attributes are the ones with a strong correlation with the target attribute and low correlation with other attributes.

CFS is a filtering method that orders a set of attributes using a correlation-based heuristic evaluation function. The function finds subsets of attributes that are strongly correlated with the target attribute and weakly correlated amongst each other. The function is expressed as follows[20]

$$D_X = \frac{n\overline{m_{ij}}}{\sqrt{n(1+(n-1)\overline{m_{ij}})}}. \quad (1)$$

D_X is the heuristic value of a set or a subset of attributes X and consists of n attributes. $\overline{m_{ij}}$ is the average attribute-target correlation, $j \in X$ and $\overline{m_{ij}}$ is the average attribute–attribute inter-correlation. The expression $n\overline{m_{ij}}$ is the predictive strength of the target value for a given subset of attributes. The denominator of eq. (1) indicates the redundancy between attributes. CFS is practical as it permits the imputation of missing values for attributes that impact the target variable. The effect is reduction in the computational cost for data imputation. Furthermore, as we do not infer missing values from other attributes, the weak correlation between attributes is insignificant.

As the attributes that impact the target variable are selected (from Figure 1), a list of potential candidates for imputation is determined. Positive selection algorithm is employed to achieve this. It takes the training dataset, a pre-determined R value and a testing instance I as inputs. The algorithm is expressed as follows:

1. Select an instance A from the training dataset.
2. Calculate the affinity between A and I.
3. If the affinity from step 2 is greater that R, then add A to a list of potential candidates C.
4. If there are instances in the training to be evaluated, go to step 1, otherwise go to step 5.
5. Select the strongest candidate B that best represents I.

Positive selection algorithm attempts to find the best candidate by comparing the affinity of each instance in the dataset with R. Calculating the affinity is done by comparing the similarities between a training instance and I in the order of succession between the attributes (attributes with missing values are ignored) and maintaining a count for each attribute that is same. For example, let S = {'ABBACEDDPUK', 'KBBAFLGDPUK'} represent the training data and P = {'HBBACL??PUK'} represent the testing set. If we compare '**ABBA**CEDDPUK' and 'H**BBAC**L??PUK', then the total is four (highlighted in bold), as there are four attributes in succession that have the same values and we cannot get a value greater than that. If we compare patterns 'KBBAF**L**GD**PUK**' and 'HBBAC**L**??**PUK**', the total is also four, as we ignore attributes with missing values while doing the comparison between the two instances. Hence, in the previous

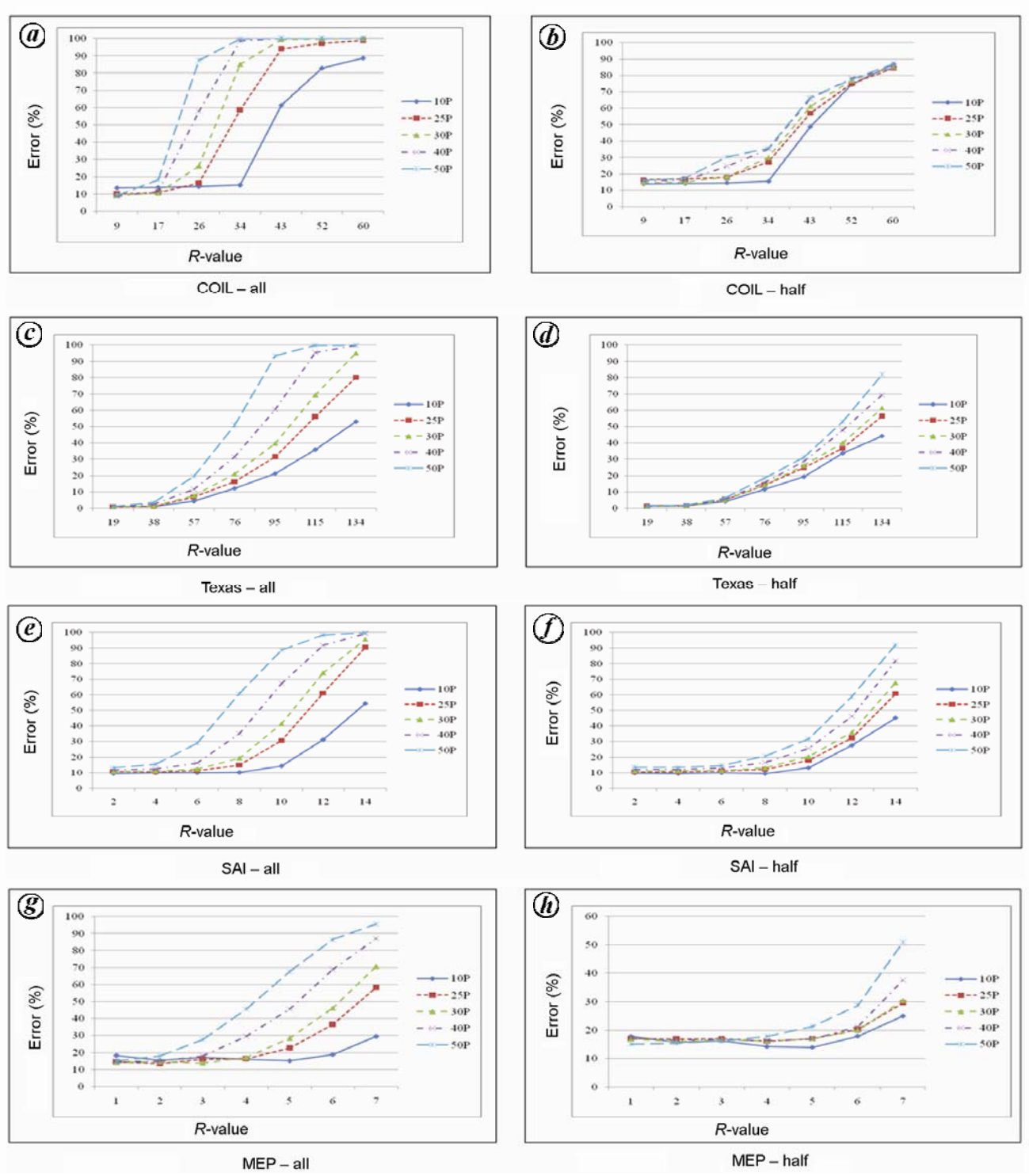

Figure 2. Datasets (***a***), (***c***), (***e***) and (***g***) having full set with missing values, and (***b***), (***d***), (***f***) and (***h***) having half the set with missing values.

example, we imply that in the pattern 'HBBACL??PUK', P succeeds L.

Once the list of candidates has been determined, the strongest candidate from the list is selected. The strongest candidate has the highest affinity. If no candidates are found, the value of *R* needs to be re-evaluated or adjusted.

The *R* value is used when performing a partial comparison between two instances to determine their similarities. This partial comparison has an advantage when dealing with datasets with a large number of variables. Even though we iterate through all the variables of an instance, we are only concerned if a certain number of

variables is the same or similar. Figure 2 illustrates the value of R for the datasets mentioned earlier. It illustrates how we derived the value of R for the experiments conducted here. It is clear from the figure that R is dataset-specific and hence the value is derived empirically.

Figure 2 illustrates the performance of the proposed MDPIS. There are five levels of proportions of missingness on the testing datasets that are generated (10%, 25%, 30%, 40% and 50%). At each level, the missingness is arbitrarily generated across the entire dataset, and then on half the attributes of the set. For the COIL dataset, if $R = 9$ and 10%, missing data are generated across the entire dataset, the data imputation strategy achieves high performance with an error of 13.57% (Figure 2 *a*). On an average, approximately 1,897 entries across the whole COIL dataset have missing data that affect the outcome. An error of 13.57% implies that only about 257 were replaced with incorrect data. If 25%, 30%, 40% or 50% is generated across the entire COIL dataset, the MDPIS still achieves high performance. This is significant, because if we consider the case where 40% missing data are generated across the entire dataset (which on average is approximately 8,127), we get approximately 708 entries that were incorrectly replaced.

If $R = 17$, the performance of MDPIS is similar to the case where $R = 9$. The exception is when 50% missing data are generated across the entire dataset. The error is double that compared to the case where $R = 9$. When $R = 26$, the performance of MDPIS decreases with increase in missingness of data. The gap in performance is quite significant and the performance is poor for the cases where there are 40% or 50% missing data across the entire dataset. When $R = 34$, MDPIS performs well only for the case where there are 10% missing data in the set. For other cases the performance is poor (or has dropped significantly) compared to when $R = 26$ or lower values. For R between 43 and 60, the performance of MDPIS is extremely poor. This behaviour is expected because the positive selection algorithm is performing a partial comparison of a large number of variables and in a situation where some or most variables have missing entries. The performance of MDPIS for Texas, SAI and MEP datasets is similar to that with COIL (Figure 2 *c*, *e* and *g*). The difference is that the performance deteriorates for different values of R per set.

Figure 2 *b* illustrates the case where missing data exist on half the attributes of the COIL dataset. For $R = 9$ and when 10% missing data are generated on half the attributes of the dataset, the average error is 14%. In this case, an average of 2006 entries on half the attributes for the COIL dataset have missing values, in which 281 were replaced with incorrect values. With half the attributes in the COIL dataset having missing values, the MDPIS shows no significant difference in performance for $R = 9$, 17, 26 or 34. This performance is similar to the case where 10% missingness is generated across the entire dataset. For $R = 9$ and with 40% missing data generated on half the dataset, there is an apparent difference in performance of the imputation strategy compared to when missing data are generated across the set. In this case the error is approximately 15.89% compared to 8.71% achieved when missing data are generated across the set. On an average, approximately 7,991 entries with missing data on half the attributes had missing values, in which an average of 1,270 were replaced incorrectly.

What is apparent from Figure 2 *a* and *b* is that at $R = 26$ or 34, MDPIS achieves high performance when half the attributes have missing values than when missing data are generated across the dataset. A similar pattern of behaviour can be observed with the Texas, SAI and MEP datasets in Figure 2. It can be observed that a small value of R is needed for the positive selection algorithm to achieve high performance in imputation. The effect of this, as we will observe later, is improved classification performance.

As an initial step of the experiment, we normalize per instance all the datasets that have numerical attributes. Thereafter, we distinguish between training and testing data. The SAI and MEP datasets have no testing data. Therefore, each dataset is partitioned into five folds (S_1, S_2, S_3, S_4, S_5) of approximately the same size as illustrated in Table 1. Table 1 resembles cross-validation with five folds with a small modification to the traditional approach.

Each fold is made up of four parts used as a training set and the remaining part is used as a testing set. The training set has completely observable data and the testing set has simulated missing data. There are five levels of proportion of missingness simulated on the testing data (10%, 25%, 30%, 40% and 50%). At each level, the missingness is randomly generated across the entire dataset and then on half the variables of the set. This strategy ensures that we assess all possible scenarios of the missing data to best assess the performance of MDPIS and classifiers. The COIL and Texas datasets have testing sets with missing data simulated in the same way as the SAI and MEP datasets.

Once the partial imputation of missing data is completed, the testing data are supplied to each classifier for classification of unseen instances.

Table 2 provides a summary of the parameters used for each dataset to conduct the experiment. The positive selection algorithm is built using C# 3.5 programing language and the value of R is derived as discussed earlier.

Table 1. Splitting of datasets into training and testing sets

	Training set	Test set
Fold 1	B + C + D + E	A
Fold 2	A + C + D + E	B
Fold 3	A + B + D + E	C
Fold 4	A + B + C + E	D
Fold 5	A + B + C + D	E

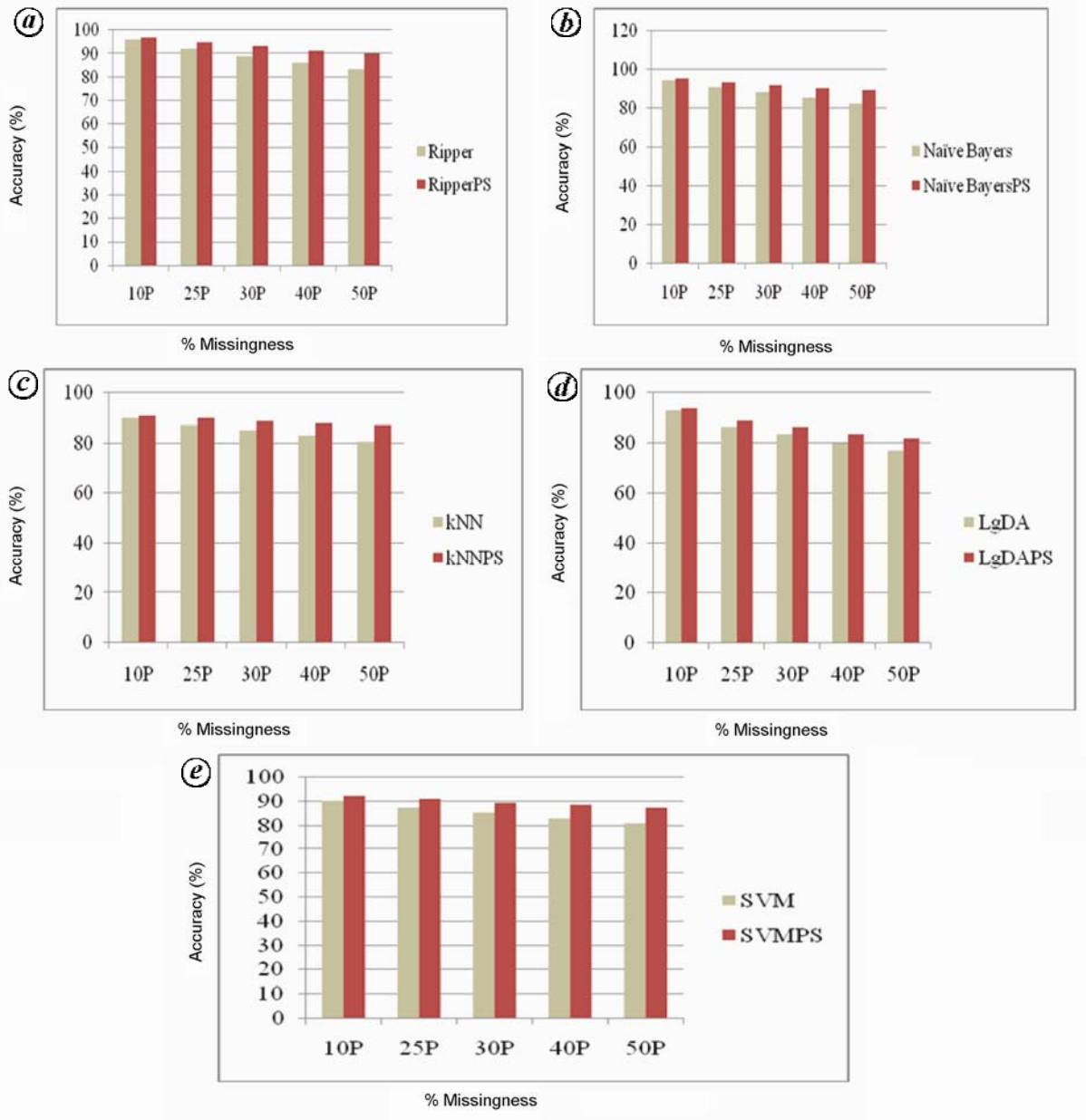

Figure 3. Overall performance of the classifiers from all the datasets.

Table 2. Parameters used for each dataset

	R-value	*k*-NN	SVM
SAI	4	k = 601, NNSA = Euclidian	KF = RBF, γ = 0.005
MEP	2	k = 5, NNSA = Euclidian	KF = RBF, γ = 0.005
COIL	17	k = 601, NNSA = Euclidian	KF = RBF, γ = 0.005
Texas	38	k = 101, NNSA = Euclidian	KF = RBF, γ = 0.005

CFS is constructed using Weka 3.6.2. The SVM model is built using libSVM 2.91, a library tool for SVMs designed by Chih-Chung Chang and Chin-Jen Lin. The model also used the radial basis function (RBF) as the kernel function (KF) and the gamma parameter was set to 0.005 (derived by trial and error) for the radial basis function. The *k*-NN model is constructed using IBk, a Weka 3.6.2 implementation of *k*-NN; the value of *k* is derived by trial and error. The Euclidean distance is employed as the nearest neighbour search algorithm (NNSA). Naïve Bayes in Weka 3.6.2 is used as a model for the NB algorithm. The MultiClassClassifier component using a logistic classifier is used to build the LgDA model, and JRip is used to construct the ripper model.

Figure 3 illustrates the overall performance of the classifiers using all the aforementioned datasets. It is apparent that there is significant improvement in classification performance when classifiers are used with proposed positive selection and imputation algorithm. All the

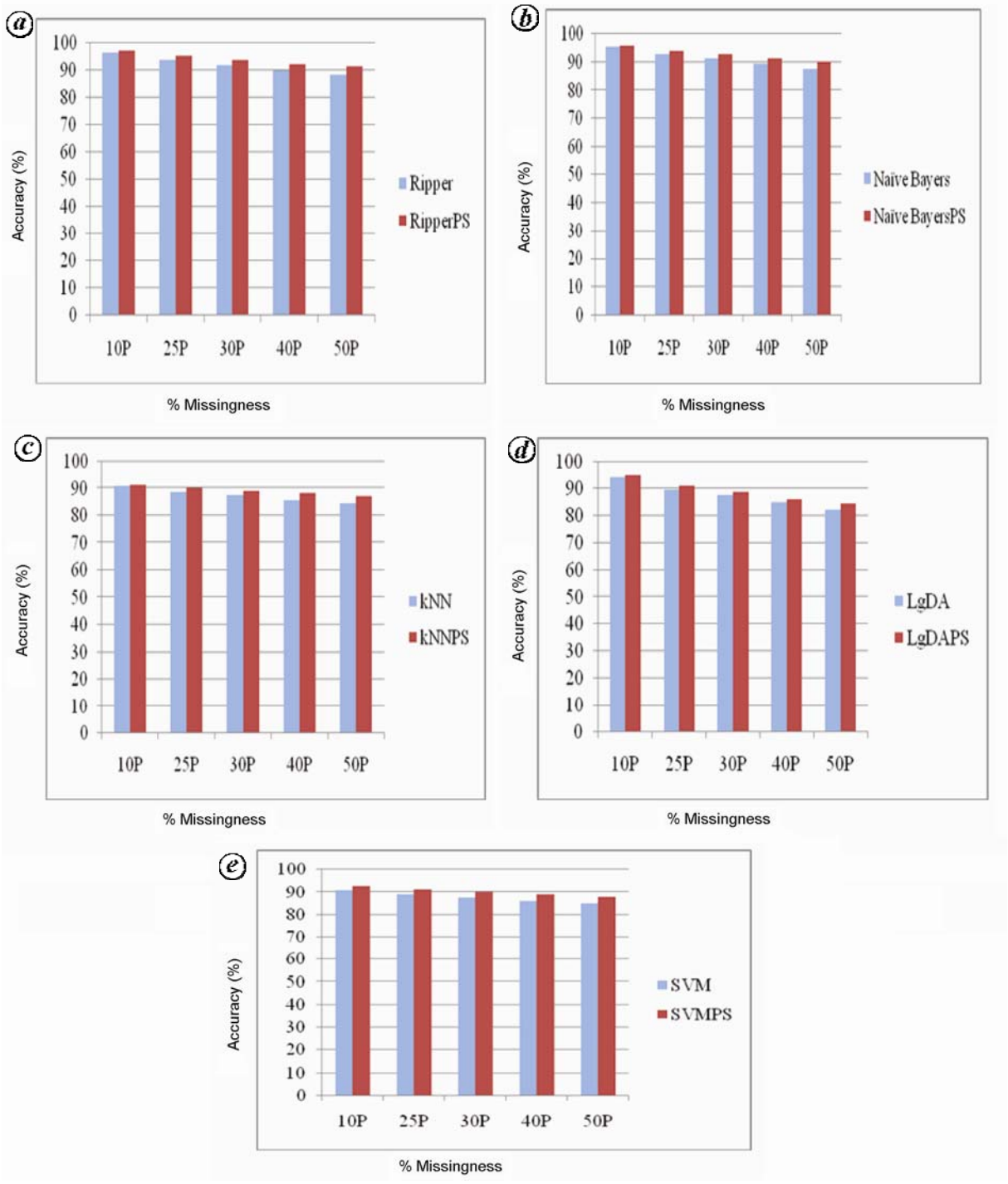

Figure 4. Overall performance of the classifiers from all the datasets with half the attributes with missing data.

classifiers show an improvement between 5% and 7% in classification accuracy for the cases where 40% or 50% of missing data are simulated across the dataset. While the ripper achieves the highest overall performance, the classifiers that achieved the most recognizable improvement when used with the proposed imputation strategy are the NB, *k*-NN and SVM. These models achieved over 6% improvement in accuracy and greater resilience to missing data compared to the ripper and LgDA.

Figure 4 illustrates the overall performance of the classifiers when only half the attributes have missing data. The figure shows that the performance of the classifiers does not decrease significantly under these conditions. However, we can observe improved classification accuracies (when the imputation strategy is applied), with the ripper achieving the highest accuracy overall. NB, *k*-NN and SVM achieve better resilience to increasing missing data compared to LgDA.

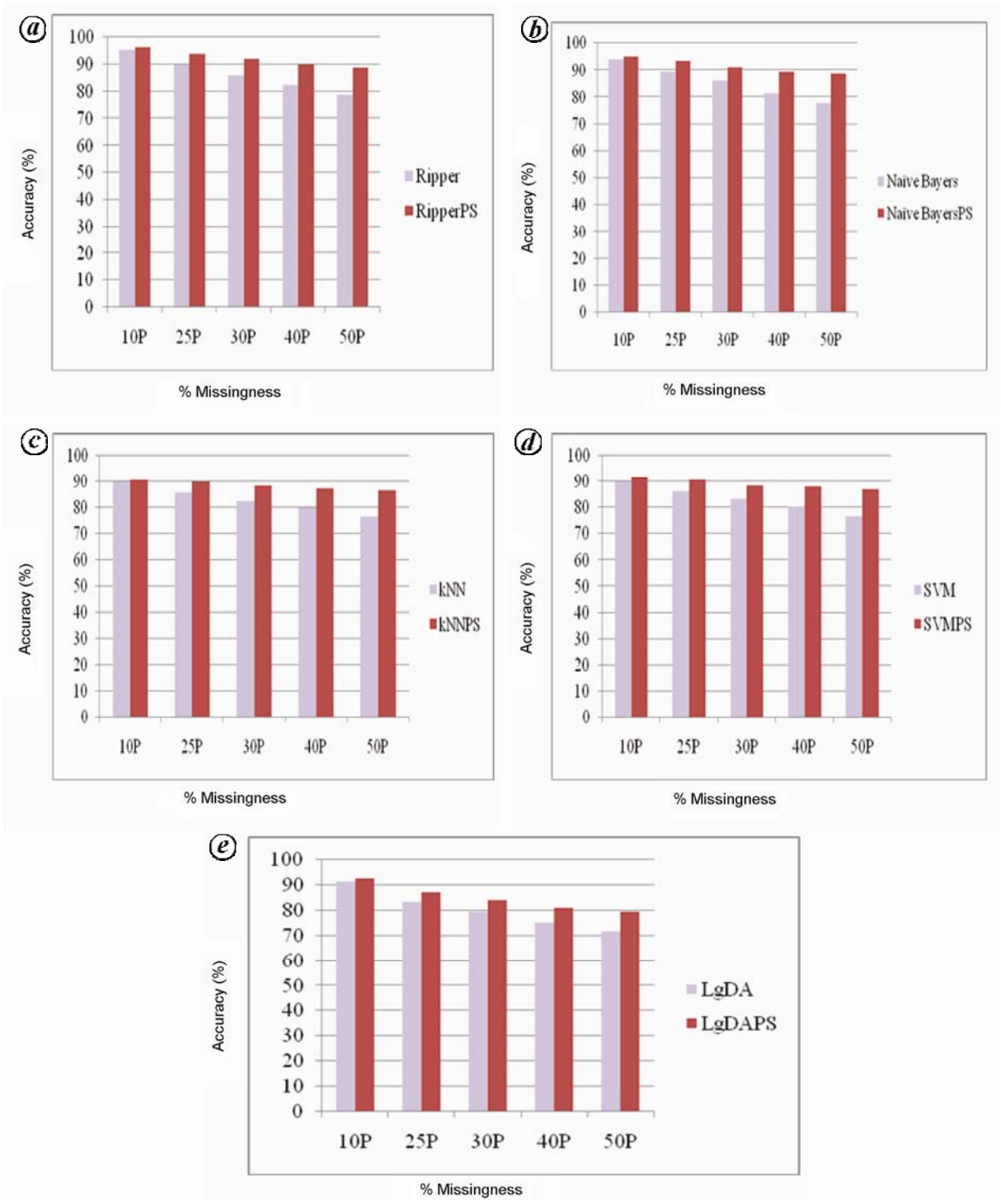

Figure 5. Overall performance of the classifiers from all the datasets with missing data generated across the entire dataset.

Figure 5 illustrates the performance of the classifiers with missing data generated across the entire datasets. We observe that under these conditions the performance of the classifiers decreases significantly. It can be seen that with the proposed MDPIS, the performance of the classifiers improves significantly. For example, NB, *k*-NN and SVM show improvements in accuracies ranging between 7% and 10%. Furthermore, all the classifiers show increase in resilience, similar to the results in Figure 4.

The proposed positive selection and data imputation strategy illustrates that by choosing the significant attribute to input missing data, the performance of the classifiers also improves. This approach sustains reduced computational cost as the number of attributes increase.

Furthermore, the resilience is increased for classifiers regardless of where missing data exist on half or across the attributes of a dataset.

In this communication we have illustrated a hybrid positive selection and correlation-based feature extraction method. We showed that the positive selection is dataset-specific and a small value of R is required for comparing attributes between two instances. This is ideal for cases where there is a large number of variables in a dataset. We also showed that using the CFS method to impute data only on those attributes that impact the outcome is significant enough to improve that classification accuracies of classifiers as well as increase their resilience to increasing missing data.

1. Francis, L., Dancing with dirty data: methods for exploring and cleaning data. In Casualty Actuarial Society Forum, Casualty Actuarial Society, Virginia, USA, pp. 198–254.
2. Peng, Y. and Kou, G., A comparative study of classification methods in financial risk detection. In Proceedings of the 4th International Conference on Networked Computing and Advanced Information Management, Gyeongju, South Korea, 2008, pp. 9–12.
3. Marwala, T., *Computational Intelligence for Missing Data Imputation Estimation and Management Knowledge Optimization Techniques*, Information Science Reference, Hershey, NY, 2009.
4. Jurek, A. and Zakrzewska, D., Improving naïve Bayes models of insurance risk by unsupervised classification. In Proceedings of the International Multiconference on Computer Science and Information Technology, Wisla, Poland, 2008, pp. 137–144.
5. Viaene, S., Derrig, R. A. and Dedene, G., A case study of applying boosting naïve Bayes to claim fraud diagnosis. *J. IEEE Trans. Knowledge Data Eng.*, 2004, **16**, 612–620.
6. Chen, W. and Li, J., A model based on factor analysis and support vector machine for credit risk identification. In Proceedings of the 8th International Conference on Machine Learning and Cybernetics, Baoding, 2009, pp. 913–918.
7. Yang, C. and Duan, X., Credit risk assessment in commercial banks based on SVM using PCA. In Proceedings of the 7th International Conference on Machine Learning and Cybernetics, Kunming, 2008, pp. 1207–1211.
8. Han, J. and Kamber, M., *Data Mining, Concepts and Techniques*, Morgan Kaufmann, San Francisco, 2006, 2nd edn.
9. Nelwamondo, F. V., Goldinga, D. and Marwala, T., A dynamic programming approach to missing data estimation using neural networks. *J. ScienceDirect*, 2009 (in press).
10. Abdella, M. and Marwala, T., The use of genetic algorithms and neural networks to approximate missing data in database. *J. Comput. Artif. Intell.*, 2007, **24**, 577–589.
11. Quinlan, J. R., *C4.5: Programs for Machine Learning*, Morgan Kaufmann Publishers Inc., San Francisco, California, USA, 1993.
12. Zhang, W., Yang, Y. and Wang, Q., Handling missing data in software effort prediction with naive Bayes and EM algorithm. In Proceedings of the 7th International Conference on Predictive Models in Software Engineering, Association for Computing Machinery, New York, USA, 2011, vol. 4.
13. Lakshminarayan, K., Harp, S. A. and Samad, T., Imputation of missing data in industrial databases. *Appl. Intell.*, 1999, **11**, 259–275.
14. Gruenwald, L., Yang, H., Sadik, S. and Shukla, R., Using data mining to handle missing data in multi-hop sensor network applications. In Proceedings of the 9th ACM International Workshop on Data Engineering for Wireless and Mobile Access, Indiana, USA, 2010, pp. 9–16.
15. Ramoni, M. and Sebastian, P., Robust learning with missing data. *Mach. Learn*, 2001, **45**, 147–170.
16. Nanni, L., Lumini, A. and Brahnam, S., A classifier ensemble approach for the missing feature problem. *Artif. Intell. Med.*, 2012, **55**, 37–50.
17. Polikar, R., DePasquale, J., Mohammed, H. S., Brown, G. and Kuncheva, L. I., Learn^{++}. MF: A random subspace approach for the missing feature problem. *Pattern Recogn.*, 2010, **43**, 3817–3832.
18. Wagner, J., Lingenfelser, F., André, E. and Kim, J., Exploring fusion methods for multimodal emotion recognition with missing data. *IEEE Trans. Affect. Comp.*, 2011, **2**.
19. Duma, M., Twala, B., Marwala, T. and Nelwamondo, F. V., Classification performance measure using missing insurance data: a comparison between supervised learning models. In Proceedings of the International Conference on Computer and Computational Intell, Nanning, China, 2010, pp. 550–555.
20. Hall, M. A., *Correlation-based Feature Selection Machine Learning*, Ph D thesis, Department of Computer Science, University of Waikato, Hamilton, New Zealand, 1999.

Received 13 September 2011; revised accepted 8 August 2012

Development of village-wise flood risk index map using multi-temporal satellite data: a study of Nagaon district, Assam, India

S. V. Shiva Prasad Sharma*, G. Srinivasa Rao and V. Bhanumurthy

RS-Applications Area, National Remote Sensing Centre,
Indian Space Research Organisation, Balanagar,
Hyderabad 500 625, India

The Nagaon district in Assam is in a sub-humid region with a greater part of the district comprising alluvial soil ranging from pure sand on the banks of the Brahmaputra to stiff clay. The area is subjected to frequent flooding by rivers during a spell of 4 months in a year. In the present study, flood hazard layer is considered as the primary input and is integrated with land use/land cover, infrastructure and population data and weightages are assigned to each class. Based on this, village flood risk index map for Nagaon district has been generated. The results of analyses indicate that about 267 villages are in the moderate–high risk index zone. About 35,354 ha of the district is in high flood hazard zone and about 25,281 ha of crop area is affected annually. We conclude that use of multi-temporal satellite datasets, coupled with GIS tools, are useful in identifying vulnerability of infrastructure, population and land use in the event of flood disaster and in calculating the flood risk index.

Keywords: Flood flood hazard layer, risk, multi-temporal satellite data, vulnerability index.

*For correspondence. (e-mail: sharma2in@yahoo.co.in)